

Stochastik für Informatiker und Lehramtstudierende

PROF. DR. MARTIN KOLB

Studentische Mitschrift, überarbeitet von SHK
Stand: 18. September 2017

Inhaltsverzeichnis

1	Diskrete Wahrscheinlichkeitsräume	3
1.1	Grundbegriffe der Stochastik	3
1.2	Bedingte Wahrscheinlichkeiten	10
1.3	Stochastische Unabhängigkeit	16
1.4	Zufallsvariablen	20
1.5	Wichtige diskrete Verteilungen	35
1.6	Abschätzen von Wahrscheinlichkeiten	41
2	Kontinuierliche Wahrscheinlichkeitsräume	48
2.1	Grundlagen	48
2.2	Wichtige kontinuierliche Verteilungen	53
2.3	Stochastische Unabhängigkeit	56
2.4	Gemeinsame Verteilung	57
2.5	Wichtige Rechenregeln	58
2.6	Abschätzen von Wahrscheinlichkeiten	60
2.7	Der zentrale Grenzwertsatz	61
2.8	Anwendungen	65
2.8.1	Die Monte-Carlo-Methode	65
2.8.2	Rejection Sampling	66
2.8.3	Erzeugung reeller Zufallsvariablen mit vorgegebener Verteilung	68
3	Induktive Statistik	70
3.1	Schätztheorie	70
3.1.1	Ausblick: Bayes Schätzer	83
3.2	Testtheorie	87
4	Markov-Ketten	94
4.1	Grundlagen	95
4.2	Ergodisches Verhalten	101
4.3	Anwendungen	108
4.3.1	Markov-Chain-Monte Carlo	108
4.3.2	Stochastische Optimierung	109
4.3.3	Randomisierte Algorithmen	110
4.3.4	Modellierung in den Naturwissenschaften	111
5	Regression	112
5.1	Einfache lineare Regression	112
5.2	Lineare Regression	117

Anhang	II
A Tabelle zur Standardnormalverteilung	II
B Tabelle zur Student- t -Verteilung	III

Vorwort

Die vorliegende Vorlesung richtet sich an Studierende der Informatik und des gymnasialen Lehramts und basiert zu großen Teilen auf den sehr empfehlenswerten Lehrbüchern:

- Thomas Schickinger und Angelika Steger: *Diskrete Strukturen, Band 2, Wahrscheinlichkeitstheorie und Statistik*, Springer
- Wolfgang Tschirk: *Statistik: Klassisch oder Bayes: Zwei Wege im Vergleich*, Springer
- Ulrich Krengel: *Einführung in die Wahrscheinlichkeitstheorie und Statistik*, Vieweg
- Hans-Otto Georgii: *Stochastik: Einführung in die Wahrscheinlichkeitstheorie und Statistik*, de Gruyter
- Lutz Dümbgen: *Stochastik für Informatiker*, Springer
- Norbert Henze: *Stochastik für Einsteiger : Eine Einführung in die faszinierende Welt des Zufalls*, Springer
- Andreas Eichler und Markus Vogel: *Leitidee Daten und Zufall: Von konkreten Beispielen zur Didaktik der Stochastik*, Springer
- Michael Mitzenmacher und Eli Upfal: *Probability and Computing*, Cambridge University Press

sowie einem Skript von Prof. Schmalfuss. Die existierende Literatur zur Stochastik richtet sich typischerweise entweder an Informatik- oder an Lehramtsstudierende. Ziel dieses Skriptes besteht in der Zusammenstellung einer konsistenten Darstellung, die beiden Hörerkreisen soweit möglich gerecht wird. Insbesondere wurde versucht, alle wichtigen Ideen und Resultate anhand von Beispielen nochmals zu illustrieren.

Die Inhalte dieser Vorlesung finden sich in den genannten Büchern, lediglich die Grundkonzeption des vorliegenden Skripts unterscheidet sich leicht. Insbesondere wurden viele Beispiele aus den oben genannten Büchern übernommen. Die ersten zwei Kapitel folgen vor allem dem Buch von Schickinger und Steger, wobei wir einige eher theoretische Aspekte aus dem Buch von Krengel übernommen haben.

In der Vorlesung wird mehr Wert auf Verständnis der Aussagen gelegt als auf formale Korrektheit. Sehr saubere und detaillierte Beweise finden sich im Buch von Hans-Otto Georgii, das sich aber vor allem an Studierende der Mathematik richtet. Didaktische Aspekte stehen in dem Buch von Eichler und Vogel im Zentrum. Interessante weiterführende Aspekte und Anwendungen in der Informatik findet man im Buch von Mitzenmacher und Upfal.

Studierenden wird empfohlen, sich begleitend zum Besuch der Vorlesung mit diesen Büchern auseinanderzusetzen. Das vorliegende Skript beruht auf einer studentischen Mitschrift und ist bisher nicht vollständig überarbeitet worden. Aktuell dient das Skript vor allem dem Dozenten zur Strukturierung der Vorlesung und der Übungen. Auch wenn die vorliegende Version einen guten Überblick über relevante Themengebiete gibt, zählt für die Prüfungen ausschließlich der in der Vorlesung behandelte Stoff.

Abschließend sei also ausdrücklich darauf hingewiesen, dass es sich um eine studentische Mitschrift handelt und somit weder Korrektheit noch Vollständigkeit garantiert werden können. Diese Mitschrift ersetzt **WEDER** den Besuch der Vorelsungen bzw. Übungen **NOCH** die intensive Auseinandersetzung mit Übungsaufgaben. Verschiedene Skizzen und Illustrationen aus der Vorlesung finden sich nicht in dem vorliegenden Skript.

Diskrete Wahrscheinlichkeitsräume

1.1 Grundbegriffe der Stochastik

In diesem ersten fundamentalen Kapitel beschäftigen wir uns in einigem Detail mit den wichtigsten Konzepten der Stochastik in einem diskreten Kontext. Viele dieser Ideen, Methoden und Resultate lassen sich geeignet auf allgemeinere Wahrscheinlichkeitsräume übertragen, allerdings erfordert dies Kenntnisse aus der Maß- und Integrationstheorie. Aus diesem Grund werden die meisten Resultate nur im diskreten Kontext bewiesen.

Definition 1.1 (Diskreter Wahrscheinlichkeitsraum).

Ein Tripel $(\Omega, \mathcal{A}, \mathbb{P})$ bestehend aus

- i) einer abzählbaren oder endlichen Menge $\Omega = \{\omega_1, \omega_2, \omega_3, \dots\}$
- ii) der Potenzmenge $\mathcal{A} = \mathcal{P}(\Omega)$ von Ω . Elemente in \mathcal{A} heißen Ereignisse. Ein Ereignis E ist also eine Teilmenge von Ω , d. h. $E \in \mathcal{P}(\Omega) \iff E \subseteq \Omega$. Für $\omega \in \Omega$ heißt $\{\omega\}$ ein Elementarereignis.
- iii) einem Wahrscheinlichkeitsmaß $\mathbb{P}: \mathcal{A} \rightarrow [0, 1]$, mit

$$\begin{aligned} \mathbb{P}(\{\omega\}) &=: P(\omega), \quad \sum_{\omega \in \Omega} \mathbb{P}(\{\omega\}) = 1, \\ \mathbb{P}(E) &= \sum_{\omega \in E} \mathbb{P}(\{\omega\}) \quad \text{für } E \subset \Omega. \end{aligned}$$

heißt **diskreter Wahrscheinlichkeitsraum**.

Das Wahrscheinlichkeitsmaß induziert also eine Abbildung $P: \Omega \rightarrow [0, 1]$, $\omega \mapsto P(\omega) := \mathbb{P}(\{\omega\})$.

Durch Festlegen der Wahrscheinlichkeiten $\mathbb{P}(\{\omega\})$ beziehungsweise $P(\omega)$ auf allen Elementarereignissen $\{\omega\}$ ist also im diskreten Kontext die Wahrscheinlichkeit von beliebigen Ereignissen $E \subseteq \Omega$ bestimmt. Wir betrachten, das folgende einfache Beispiel.

Beispiel 1.2. Für das Zufallsexperiment „Ein Wurf mit sechsseitigem fairem Würfel“ hat man

$$\Omega = \{1, 2, 3, 4, 5, 6\}, \quad \mathcal{A} = \mathcal{P}(\Omega), \quad P(\omega) = \mathbb{P}(\{\omega\}) = \frac{1}{6} \quad \forall \omega \in \Omega.$$

Beispiel 1.3. Betrachte ein Rechnernetz, in dem eine Übertragung nur mit einer festen Wahrscheinlichkeit $p \in [0, 1]$ gelingt und sei $q := 1 - p$. Wir interessieren uns für die Frage mit welcher Wahrscheinlichkeit k -Versuche bis zur ersten erfolgreichen Übertragung notwendig sind.

i) $\Omega = \{\omega_1, \omega_2, \dots\}$, $\omega_i = i$ Versuche bis zur ersten erfolgreichen Übertragung

ii) $\mathcal{A} = \mathcal{P}(\Omega)$

iii) Es gilt $\mathbb{P}(\{\omega_i\}) = pq^{i-1}$, $\sum_{i=1}^{\infty} \mathbb{P}(\{\omega_i\}) = \sum_{i=1}^{\infty} pq^{i-1} = p \cdot \sum_{i=1}^{\infty} q^{i-1} \stackrel{\text{geom. Reihe}}{=} p \cdot \frac{1}{1-q} = 1$.

Von fundamentaler Bedeutung für die moderne Wahrscheinlichkeitstheorie ist die folgende auf Kolmogorov zurückgehende Axiomatik, auf die im Rahmen dieser Vorlesung nicht weiter eingegangen werden kann. Wir verweisen auf das Buch von Georgii.

Definition 1.4 (Allgemeiner Wahrscheinlichkeitsraum).

Ein Tripel $(\Omega, \mathcal{A}, \mathbb{P})$ heißt **Wahrscheinlichkeitsraum**, wenn

i) $\Omega \neq \emptyset$ eine beliebige Menge ist

ii) $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ aus Teilmengen von Ω besteht, so dass

1) $\Omega \in \mathcal{A}$

2) $A \in \mathcal{A} \implies A^c = \Omega \setminus A \in \mathcal{A}$

3) $A_1, A_2, \dots \in \mathcal{A} \implies \bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$

\mathcal{A} heißt dann σ -Algebra über Ω .

iii) $\mathbb{P}: \mathcal{A} \rightarrow [0, 1]$ eine Abbildung ist mit

1) $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(\Omega) = 1$

2) $A_1, A_2, \dots \in \mathcal{A}$, $A_i \cap A_j = \emptyset$ ($\forall i \neq j$) $\implies \mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$

\mathbb{P} heißt dann ein Wahrscheinlichkeitsmaß.

Bemerkung 1.5. Die Potenzmenge $\mathcal{P}(\Omega)$ ist offensichtlich eine σ -Algebra über Ω und damit sieht man leicht, dass jeder diskrete Wahrscheinlichkeitsraum aus Definition 1.1 auch die Bedingungen für einen Wahrscheinlichkeitsraum erfüllt.

Folgerung 1.6 (Elementare Eigenschaften).

a) $A, B \in \mathcal{A} \implies A \cap B \in \mathcal{A}$, weil $A \cap B = (A^c \cup B^c)^c$.

b) $A, B \in \mathcal{A}$, $A \cap B = \emptyset \implies \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$, denn dies lässt sich mit $A_1 = A$, $A_2 = B$ und $A_i = \emptyset$ für $i \geq 2$ aus den Eigenschaften eines Wahrscheinlichkeitsmaßes ableiten.

c) $A \in \mathcal{A}$, $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$, denn $1 = \mathbb{P}(\Omega) = \mathbb{P}(A \cup A^c) = \mathbb{P}(A) + \mathbb{P}(A^c)$

d) $A, B \in \mathcal{A}$, $A \subset B \implies \mathbb{P}(A) \leq \mathbb{P}(B)$, denn $\mathbb{P}(B) = \mathbb{P}(A \cup (B \setminus A)) = \mathbb{P}(A) + \underbrace{\mathbb{P}(B \setminus A)}_{\geq 0}$.

Die folgende Ungleichung ist besonders relevant, um die Wahrscheinlichkeit von Ereignissen abschätzen zu können.

Satz 1.7 (Boolesche Ungleichung / Subadditivität des Wahrscheinlichkeitsmaßes).

Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein W-Raum und seien $A_1, A_2, A_3, \dots \in \mathcal{A}$, dann gilt:

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

Besteht man statt einer Ungleichung für die Wahrscheinlichkeit einer Vereinigung von Ereignissen auf eine Gleichheit, dann kann man im Fall endlicher Vereinigungen auf das folgende Resultat zurückgreifen.

Satz 1.8 (Siebformel).

Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein W-Raum. Für Ereignisse $A_1, A_2, A_3, \dots, A_n \in \mathcal{A}$ gilt

$$\begin{aligned} \mathbb{P}\left(\bigcup_{i=1}^n A_i\right) &= \sum_{i=1}^n \mathbb{P}(A_i) - \sum_{1 \leq i_1 < i_2 \leq n} \mathbb{P}(A_{i_1} \cap A_{i_2}) \pm \dots \\ &\quad + (-1)^{l-1} \cdot \sum_{1 \leq i_1 < i_2 < \dots < i_l \leq n} \mathbb{P}(A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_l}) \\ &\quad \pm (-1)^{n-1} \cdot \mathbb{P}(A_1 \cap \dots \cap A_n). \end{aligned}$$

Für $n = 2$ gilt also

$$\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 \cap A_2)$$

und für $n = 3$:

$$\begin{aligned} \mathbb{P}(A_1 \cup A_2 \cup A_3) &= \mathbb{P}(A_1) + \mathbb{P}(A_2) + \mathbb{P}(A_3) - \mathbb{P}(A_1 \cap A_2) - \mathbb{P}(A_1 \cap A_3) - \mathbb{P}(A_2 \cap A_3) \\ &\quad + \mathbb{P}(A_1 \cap A_2 \cap A_3). \end{aligned}$$

BEWEIS. Wir zeigen die Aussage an dieser Stelle nur für den Fall $n = 2$. Seien A und B zwei Ereignisse und definiert man $C := A \setminus B = A \cap B^c$, dann sind C und $A \cap B$ disjunkt und somit erhalten wir wegen der Additivität von Wahrscheinlichkeitsmaßen

$$\mathbb{P}(A) = \mathbb{P}(C \cup (A \cap B)) = \mathbb{P}(C) + \mathbb{P}(A \cap B).$$

Wegen $A \cup B = C \cup B$ folgt hieraus

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

□

Als Übung veranschauliche und beweise der Leser wie in der Vorlesung anhand einer Skizze die Siebformel für den Fall $n = 3$. Wir werden später in Bemerkung 1.54 noch einen recht direkten Beweis der Siebformel sehen. An dieser Stelle genügt es, sich die Spezialfälle $n = 2$ und $n = 3$ geometrisch zu veranschaulichen.

Definition 1.9 (Prinzip von Laplace).

Wenn nichts dagegen spricht, dann gehen wir davon aus, dass alle Elementarereignisse gleichwahrscheinlich sind. Hierbei sei Ω eine endliche Menge.

$$\Omega = \{\omega_1, \dots, \omega_n\} \implies \mathbb{P}(\omega_i) = \frac{1}{|\Omega|}, \quad \mathbb{P}(E) = \sum_{\omega \in E} \frac{1}{|\Omega|} = \frac{|E|}{|\Omega|}.$$

Beispiel 1.10. A und B vereinbaren ein Spiel. A bekommt einen Punkt für jedes vorbeifahrende Auto in der Farbe *rot* oder *blau*, B bekommt einen Punkt für *gelbe* oder *weiße* Autos. Nachdem 3 Autos dieser Farben vorbeigefahren sind, wird abgerechnet.

- $\Omega = \{b, r, g, w\}^3 \ni \omega = (\omega_1, \omega_2, \omega_3)$
- $E = \{B \text{ sammelt keine Punkte}\} = \{rrr, rrb, rbr, brr, bbr, rbb, brb, bbb\}$.
- $\mathbb{P}(E) = \frac{|E|}{|\Omega|} = \frac{8}{4^3} = \frac{1}{8}$.

Die Annahme der Laplaceverteilung setzt also insbesondere voraus, dass die Autos der unterschiedlichen Farben etwa gleich häufig produziert werden und dass sich unsere Spieler zum Beispiel nicht gerade in der Nähe einer größeren Postfiliale befinden.

Einfache Urnenmodelle

Wir haben gesehen, dass verschiedene Fragestellungen aus dem Bereich der endlichen Wahrscheinlichkeitsräume auf die Analyse kombinatorischer Probleme reduzierbar sind. In diesem Exkurs werden wir einige wichtige kombinatorische Modelle einführen. Die Darstellung der Urnenmodelle ist aus Krenegels Buch entnommen.

In einer Urne seien N Kugeln, die wir uns mit $1, 2, \dots, N$ durchnummeriert denken. Sukzessive werden n Kugeln zufällig gezogen. Wir sprechen von Ziehung mit Zurücklegen, wenn Kugeln mehrfach gezogen werden können, ansonsten ohne Zurücklegen. Weiterhin können wir Wert auf die Reihenfolge der gezogenen Kugeln legen oder diese nicht beachten. Sei nun $\mathbb{A} := \{1, 2, \dots, N\}$.

- I) **Stichproben in Reihenfolge mit Zurücklegen:** Als Ereignisraum kann man

$$\Omega_I := \{\omega = (\omega_1, \dots, \omega_n) : \forall i = 1, \dots, n : \omega_i \in \mathbb{A}\}$$

Dann gilt

$$|\Omega_I| = N^n.$$

- II) **Stichproben in Reihenfolge ohne Zurücklegen:** Als Ereignisraum kann man

$$\Omega_{II} := \{\omega = (\omega_1, \dots, \omega_n) : \forall i = 1, \dots, n : \omega_i \in \mathbb{A}, \omega_i \neq \omega_j (i \neq j)\}$$

Dann gilt

$$|\Omega_{II}| = N \cdot (N - 1) \cdot \dots \cdot (N - n + 1).$$

Diese Formel ergibt sich durch die Überlegung, dass bei Zug der ersten Kugel noch N Möglichkeiten bestehen, bei Zug der zweiten nur noch $(N - 1)$ usw.

- III) **Stichproben ohne Reihenfolge ohne Zurücklegen:** Als Ereignisraum kann man

$$\Omega_{III} := \{\{\omega_1, \dots, \omega_n\} : \forall i = 1, \dots, n : \omega_i \in \mathbb{A}, \omega_i \neq \omega_j (i \neq j)\}$$

Wir können Ω_{III} auch durch Einführen einer Äquivalenzrelation beschreiben:

$$(\omega_1, \dots, \omega_n) \sim (\omega'_1, \dots, \omega'_n),$$

genau dann, wenn es eine Permutation π von $\{1, \dots, n\}$ gibt, mit $\omega'_i = \omega_{\pi(i)}$ für $i = 1, \dots, n$. Die Elemente von Ω_{III} sind dann die zu dieser Relation gehörigen Äquivalenzklassen.

Jede Klasse kann durch den Repräsentanten $(\omega_1, \omega_2, \dots, \omega_n)$ mit $\omega_1 < \omega_2 < \dots < \omega_n$ beschrieben werden. Jede Äquivalenzklasse besteht aus $n!$ Elementen. Also gilt

$$|\Omega_{II}| = n! \cdot |\Omega_{III}|$$

und somit

$$|\Omega_{III}| = \frac{N!}{n! \cdot (N-n)!} = \binom{N}{N-n} = \binom{N}{n} \quad (1 \leq n \leq N).$$

- **IV) Stichproben ohne Reihenfolge mit Zurücklegen:** Als Stichprobenraum Ω_{IV} können wir die Menge der Äquivalenzklassen unter der obigen Äquivalenzrelation in Ω_I nehmen. Indem wir aus jeder Äquivalenzklasse den Repräsentanten mit $\omega_1 \leq \omega_2 \leq \dots \leq \omega_n$ auswählen, sehen wir, dass sich Ω_{IV} auch als die Menge

$$\Omega_{IV} := \{(\omega_1, \dots, \omega_n) \in \mathbf{A}^n \mid \omega_1 \leq \omega_2 \leq \dots \leq \omega_n\}$$

beschreiben lässt. Es gilt

$$|\Omega_{IV}| = \binom{N+n-1}{n}.$$

Um dies einzusehen ordnet man den Elementen $\Omega \ni \omega = (\omega_1, \omega_2, \dots, \omega_n)$ die Folgen $(\omega'_1, \omega'_2, \dots, \omega'_n)$ mit $\omega'_i = \omega_i + i - 1$ zu. Durch diese Zuordnung wird Ω_{IV} bijektiv auf die Menge

$$\Omega'_{III} = \{(\omega'_1, \dots, \omega'_n) \in \mathbb{B}^n \mid \omega'_1 < \omega'_2 < \dots < \omega'_n\}$$

mit $\mathbb{B} = \{1, 2, \dots, N+n-1\}$ abgebildet. Und damit erhalten wir, dass Ω_{IV} und Ω'_{III} die gleiche Mächtigkeit besitzen.

Wir schließen diesen Exkurs mit einer alternativen Interpretationsmöglichkeit. Wir fragen nach der Anzahl der Möglichkeiten n Murmeln auf N Zellen genannte Plätze zu verteilen. Sind die Murmeln unterscheidbar, so ist eine Verteilung dadurch beschrieben, dass man für jedes i mit $1 \leq i \leq n$ die Nummer ω_i der Zelle angibt, in die man die i -te Murmel platziert hat. Eine Verteilung ist also beschrieben durch ein n -Tupel $\omega = (\omega_1, \dots, \omega_n)$ mit $1 \leq \omega_i \leq N$. Die Menge der Ergebnisse ist also wieder Ω_I aber jetzt mit der Uminterpretation

$$\begin{aligned} \text{Nummer der Ziehung} &\leftrightarrow \text{Nummer der Murmel} \\ \text{Nummer der Kugel} &\leftrightarrow \text{Nummer der Zelle.} \end{aligned}$$

Sind die Murmeln ununterscheidbar, so können wir zwischen Permutationen von $(\omega_1, \omega_2, \dots, \omega_n)$ nicht mehr unterscheiden. Diese werden also identifiziert. Ziehen ohne Rücklegen bedeutet, dass die Nummer einer Kugel in der Stichprobe nicht mehrfach auftreten darf. Dies bedeutet jetzt, dass jede Zelle nur einmal auftreten darf, dass man also in jede Zelle maximal eine Murmel legen darf. Man spricht von Verteilungen ohne bzw. mit Mehrfachbesetzung.

Beispiel 1.11. Wie groß ist die Wahrscheinlichkeit, dass mindestens zwei von den $n = 25$ Schülern einer Klasse am gleichen Tag Geburtstag haben?

Als Stichprobenraum wählen wir

$$\Omega_I = \{\omega = (\omega_1, \dots, \omega_n) : \forall i = 1, \dots, n : \omega_i \in \mathbb{A}\}$$

mit $n = 25$ und $N = 365$. Wir vernachlässigen Schaltjahre. Das Ereignis $(\omega_1, \dots, \omega_{25}) \in \Omega_I$ bedeutet dann, dass Schüler Nummer 1 am Tag ω_1 Geburtstag hat und dass Schüler Nummer 2 am Tag ω_2 Geburtstag hat usw. Das für uns interessante Ereignis ist das Komplement von

$$\Omega_{II} = \{ \{ \omega_1, \dots, \omega_n \} : \forall i = 1, \dots, n : \omega_i \in \mathbb{A}, \omega_i \neq \omega_j (i \neq j) \}.$$

Unter Annahme der Gleichverteilung (Laplace-Annahme) auf Ω_I gilt dann also

$$\mathbb{P}(\Omega_{II}^c) = 1 - \mathbb{P}(\Omega_{II})$$

und weiter

$$\mathbb{P}(\Omega_{II}) = \frac{|\Omega_{II}|}{|\Omega_I|} = 1 \cdot \left(1 - \frac{1}{N}\right) \cdot \left(1 - \frac{2}{N}\right) \cdot \dots \cdot \left(1 - \frac{n-1}{N}\right) \approx 0.44.$$

Wir erhalten also $\mathbb{P}(\Omega_{II}^c) = 1 - \mathbb{P}(\Omega_{II}) \geq 0,5$.

Geburtstagskollisionen sind also wohl häufiger als man vermuten könnte. Empirisch lässt sich diese Rechnung gut am Beispiel von Sportmannschaften überprüfen. Werfen wir z.B. einen Blick in die Kader der vier deutschen Nationalmannschaften, die bisher Weltmeister wurden.

- Kader des Teams bei der **WM 1954** in der Schweiz:
 - 1) Heinz Kwiatkowsky, Geburtsdatum 16.07.1926
 - 2) Alfred Pfaff, Geburtsdatum 16.07.1926
- Kader des Teams bei der **WM 1974** in Deutschland:
Keine Geburtstagskollisionen.
- Kader des Teams bei der **WM 1990** in Italien:
 - 1) Klaus Augenthaler, Geburtsdatum 26.09.1957
 - 2) Uwe Bein, Geburtsdatum 26.09.1960
- Kader des Teams bei der **WM 2014** in Brasilien:
Keine Geburtstagskollisionen.

Dieser Befund bestätigt die durch unsere theoretische Analyse gefundene Einsicht, dass Geburtstagskollisionen nicht allzu selten sind.

Beispiel 1.11 ist als Geburtstagsproblem bekannt. Die zugrundeliegende mathematische Fragestellung tritt aber in vielen unterschiedlichen Kontexten auf. Anwendung findet das Geburtstagsproblem beispielsweise bei hash tables. Dabei soll eine Menge M von Datensätzen möglichst gut auf eine Menge N von Speicherplätzen verteilt werden, wobei man vermeiden möchte, dass zwei Datensätze auf den selben Speicherplatz abgelegt werden. Verteilt man die Daten "zufällig", dann wissen wir schon, dass mit $|M| = 25$ und $|N| = 365$ die Wahrscheinlichkeit, dass zwei Datensätze auf den selben Speicherplatz abgelegt werden, bereits größer als 0,5 ist. Das Geburtstagsproblem taucht auch dem folgenden Kontext auf. Nehmen wir an, die Wahrscheinlichkeit, dass die DNA einer zufälligen Person mit einem gegebenen DNA-Profil übereinstimmt liegt bei 1 zu 13 Milliarden. Eine Forschungsgruppe fand jedoch bei einer Untersuchung einer Datenbank mit 65000 DNA-Profilen 10 Paare mit übereinstimmenden Profilen. Diese beiden Ergebnisse sind nicht widersprüchlich¹ und die Erklärung hierfür ist analog zum Geburtstagsproblem.

¹Für weitere Details siehe <http://freakonomics.com/2008/08/19/are-the-fbis-probabilities-about-dna-matches-crazy/>

Beispiel 1.12. In einer Urne befinden sich gut durchmischt n gleichartige Kugeln in den Farben Schwarz und Weiß, etwa s schwarze und w weiße ($s + w = n$). Man zieht ohne Zurücklegen willkürlich $m \leq n$ Kugeln und fragt nach der Wahrscheinlichkeit, daß darunter genau $k \leq s$ schwarze Kugeln sind. Zunächst besteht bei dem zugrundeliegenden Ereignisraum Ω_{II} unser gesuchtes Ereignis E aus allen Folgen $(\omega_1, \omega_2, \dots, \omega_m)$, bei denen genau k Zahlen $\omega_1, \omega_2, \dots, \omega_m$ auf schwarzen Kugeln stehen. Wir müssen also k Plätze aus m zur Verfügung stehenden auswählen und haben dafür $\binom{m}{k}$ Möglichkeiten. k schwarze Kugeln können auf genau $s(s-1) \dots (s-k+1)$ verschiedene Arten der Reihe nach gezogen werden und die verbleibenden $w(w-1) \dots (w-m+k+1)$ Arten mit Zahlen auf weiße Kugeln besetzt werden. Die gesuchte Wahrscheinlichkeit ist also

$$\binom{m}{k} \cdot \frac{s(s-1) \dots (s-k+1) \cdot w(w-1) \dots (w-m+k+1)}{n(n-1) \dots (n-m+1)} = \frac{\binom{s}{k} \binom{w}{m-k}}{\binom{n}{m}}.$$

Das folgende Beispiel aus dem Buch von Dümbgen illustriert das klassische Vorgehen.

Beispiel 1.13. Beim modernen Fünfkampf muss jeder Teilnehmer unter anderem einen Geländeritt absolvieren. Jeder der N Teilnehmer bringt ein Pferd zum Wettbewerb mit, die Pferde werden jedoch verlost. Man ermittelt also rein zufällig, welcher Reiter mit welchem Pferd an den Start geht. Wir wollen nun die Wahrscheinlichkeit dafür berechnen, dass Teilnehmer j sein eigenes Pferd reitet. Als Ereignisraum Ω wählen wir

$$\Omega_{II} = \Omega = \{(\omega_1, \omega_2, \dots, \omega_N) \mid \forall i = 1, \dots, N : \omega_i \in \{1, \dots, N\}, \omega_i \neq \omega_j (i \neq j)\}.$$

Das Element $(\omega_1, \omega_2, \dots, \omega_N) \in \Omega$ beschreibt uns, dass Teilnehmer 1 das Pferd von Teilnehmer ω_1 , Teilnehmer 2 das Pferd von Teilnehmer ω_2 usw. reitet. Wir legen die Laplace-Annahme zugrunde und erhalten also für das Ereignis

$$E_j = \{\text{Teilnehmer } j \text{ reitet sein eigenes Pferd}\} = \{\omega \in \Omega \mid \omega_j = j\}$$

die Wahrscheinlichkeit

$$\mathbb{P}(E_j) = \frac{|E_j|}{|\Omega|} = \frac{(N-1)!}{N!} = \frac{1}{N}.$$

Das folgende Beispiel illustriert insbesondere, wie man durch geschicktes Vorgehen Abzählprobleme löst.

Beispiel 1.14 (Skat).

Beim Skatspiel werden 32 Karten ausgeteilt. Es gibt 4 Buben. Jeder Spieler bekommt 10 Karten und 2 Karten gehen in den Skat.

Wie groß ist die Wahrscheinlichkeit für

$$A_1 = \{\text{Spieler 1 erhält alle Buben}\}$$

$$A_2 = \{\text{Jeder Spieler erhält genau einen Buben}\}$$

$$\omega = (\underbrace{\omega_1, \omega_2, \dots, \omega_{10}}_{\text{Spieler 1}}, \underbrace{\omega_{11}, \dots, \omega_{20}}_2, \underbrace{\omega_{21}, \dots, \omega_{30}}_3, \underbrace{\omega_{31}, \omega_{32}}_{\text{Skat}})$$

$$\begin{aligned} |A_1| &= 10 && (\text{Wo ist Kreuz-Bube?}) \\ &\cdot 9 && (\text{Wo ist Pik-Bube?}) \\ &\cdot 8 && (\text{Wo ist Herz-Bube?}) \\ &\cdot 7 && (\text{Wo ist Karo-Bube?}) \\ &\cdot 28! && (\text{Wo sind die anderen Karten?}) \end{aligned} \quad \Rightarrow \quad \mathbb{P}(A_1) = \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 28!}{32!} \approx 0,0058$$

$$\begin{aligned}
|A_2| &= 10^3 && \text{(Jeder Spieler bekommt einen Buben)} \\
&\cdot 2 && \text{(Ein Bube ist im Skat)} \\
&\cdot 4! && \text{(Reihenfolge der Buben egal)} \\
&\cdot 28! && \text{(Wo sind die anderen Karten?)}
\end{aligned}
\implies \mathbb{P}(A_2) \approx 0,0556$$

Die Anwendung der Siebformel wird anhand des Beispiels aus dem Fünfkampf illustriert, dieses ist wiederum dem Buch von Dübngn entnommen.

Beispiel 1.15. Wir untersuchen nochmals die Situation im Modernen Fünfkampf aus Beispiel 1.13. Diesmal interessieren wir uns für die Wahrscheinlichkeit des Ereignisses E , daß mindestens ein Teilnehmer sein eigenes Pferd reitet. Setzt man

$$A_i = \{\text{Teilnehmer } i \text{ reitet sein eigenes Pferd}\} = \{\omega \in \Omega_{II} \mid \omega_i = i\},$$

dann ist $\mathbb{P}(E) = \mathbb{P}(\bigcup_{i=1}^N A_i)$. Für eine beliebige k -elementige Teilmenge $J \subset \{1, \dots, N\}$ ist

$$|\bigcap_{i \in J} A_i| = (N - k)!$$

und somit

$$\mathbb{P}(\bigcap_{i \in J} A_i) = \frac{(N - k)!}{N!}.$$

Mit Hilfe der Siebformel berechnet man also

$$\mathbb{P}(\bigcup_{i=1}^N A_i) = \sum_{i=1}^N (-1)^{k-1} \binom{N}{k} \frac{(N - k)!}{N!} = \sum_{i=1}^N (-1)^{k-1} \frac{1}{k!}.$$

1.2 Bedingte Wahrscheinlichkeiten

Durch Bekanntwerden zusätzlicher Information *verändern* sich Wahrscheinlichkeiten. Das folgende Beispiel illustriert und erläutert diese Aussage.

Beispiel 1.16 (Five-Card-Draw Poker).

A und B spielen Poker. A hat 4 Asse und eine Herz 2. Somit kann B nur gewinnen, wenn B fünf Karten derselben Farbe in aufsteigender Reihenfolge hat.

$$\mathbb{P}(B \text{ gewinnt}) = ?$$

Angenommen die Karten sind markiert. A sieht also, dass B nur Karten der Farbe Kreuz auf der Hand hat. Wie hoch ist, aus der Sicht von A , die Wahrscheinlichkeit dafür, dass B gewinnt.

$$\mathbb{P}(B \text{ gewinnt} \mid B \text{ hat nur Kreuz}) = ?$$

Die Information, dass B nur Kreuz auf der Hand hat, wird uns Einschätzung der Gewinnwahrscheinlichkeit von B beeinflussen.

Definition 1.17. Es sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein W-Raum, für Ereignisse $A, B \in \mathcal{A}$ mit $\mathbb{P}(B) > 0$, definiere die **bedingte Wahrscheinlichkeit** von A unter der Bedingung B durch

$$\mathbb{P}(A \mid B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Bemerkung 1.18. Ist $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ ein diskreter W-Raum und $B \subset \Omega$ mit $\mathbb{P}(B) > 0$, so ist

$$\omega \rightarrow \mathbb{P}(\{\omega\} \mid B)$$

ein Wahrscheinlichkeitsmaß, denn es gilt

$$\sum_{\omega \in \Omega} \mathbb{P}(\{\omega\} \mid B) = \sum_{\omega \in \Omega} \frac{\mathbb{P}(\{\omega\} \cap B)}{\mathbb{P}(B)} = \frac{1}{\mathbb{P}(B)} \cdot \mathbb{P}(\Omega \cap B) = \frac{\mathbb{P}(B)}{\mathbb{P}(B)} = 1.$$

Das folgende Beispiel beleuchtet nochmals die Bedeutung der mathematischen Modellierung eines Sachverhalts.

Beispiel 1.19. Eine Familie hat zwei Kinder, eines davon kennen Sie und Sie wissen, dass es ein Mädchen ist.

Gesucht: Wahrscheinlichkeit dafür, dass die Familie zwei Mädchen hat?

$$\Omega = \{(w, w), (w, m), (m, w), (m, m)\}, \quad \mathbb{P}(\omega) = \frac{1}{4}, \quad \omega \in \Omega$$

$$A = \{\text{Beide Kinder sind Mädchen}\}, \quad B = \{\text{mind. ein Mädchen}\}$$

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(\{(w, w)\})}{\mathbb{P}(\{(w, w), (w, m), (m, w)\})} = \frac{1/4}{3/4} = \frac{1}{3}.$$

Es sei betont, dass sich die berechnete Wahrscheinlichkeit aus der gewählten mathematischen Formalisierung des Problems ergibt. In diese fließen notwendigerweise verschiedene Grundannahmen ein, die einer weiteren Analyse bedürfen. Das hier angegebene sogenannte zwei-Kinder-Paradox ist ein häufig diskutiertes Problem.

Beispiel 1.20. Die Wahrscheinlichkeit, die Stochastik Klausur im ersten Versuch zu bestehen beträgt 0,6. Die Wahrscheinlichkeit beim zweiten Versuch zu bestehen, wenn man beim ersten Versuch nicht bestanden hat, beträgt 0,9.

Mit welcher Wahrscheinlichkeit sind zum Bestehen mehr als 2 Versuche notwendig?

$$A = \text{„kein Erfolg bei Versuch 1“}, \quad \mathbb{P}(A) = 0,4$$

$$B = \text{„kein Erfolg bei Versuch 2“}, \quad \mathbb{P}(B \mid A) = 0,1$$

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \mathbb{P}(A) \cdot \mathbb{P}(B \mid A) = 0,04$$

Beispiel 1.21 (Zulassung zu einer amerikanischen Universität).

Das folgende Beispiel ist dem empfehlenswerten Buch von Krengel entnommen. Als Beauftragter für Geschlechtergleichheit wird Ihnen folgendes Dokument vorgelegt:

	Bewerb.	ang.	bed. Aufnahme-W-keit
m	2084	1036	$\mathbb{P}(A \mid M) \approx 0,49$
w	1067	349	$\mathbb{P}(A \mid W) \approx 0,33$

Nun stellt sich die Frage ob Frauen an dieser Universität tatsächlich diskriminiert wurden. Einen besseren Einblick gibt die folgende, ausführlichere Tabelle, in der nach verschiedenen Fachbereichen unterschieden wurde.

Fachb.	männl.			weibl.		
	Bewerb.	ang.	$\mathbb{P}_M(A B_i)$	Bewerb.	ang.	$\mathbb{P}_W(A B_i)$
B_1	826	551	0,67	108	89	0,82
B_2	560	353	0,63	25	17	0,68
B_3	325	110	0,34	593	219	0,37
B_4	373	22	0,06	341	24	0,07

Dies gibt Kenntnis darüber, dass

- a) Fachbereiche mit sehr hoher Zulassungsquote tendenziell mehr Bewerbungen von Männern erhalten, sowie
- b) Fachbereiche mit niedriger Zulassungsquote tendenziell mehr Bewerbungen von Frauen erhalten.

Dasselbe Phänomen tritt auch im nächsten aus dem Buch von Tschirk entnommenen Beispiel auf.

Beispiel 1.22. Eine Klinik A behandelt drei Krankheiten: K_1, K_2 und K_3 . Die Heilungswahrscheinlichkeiten sind: 0.4 bei K_1 , 0.9 bei K_2 und 0.8 bei K_3 . Von den Patienten leiden 60 % an K_1 und 30 % an K_2 und 10 % an K_3 und wir nehmen an, dass Patienten nur an einer der drei Krankheiten erkrankt sind. Die Wahrscheinlichkeit, dass ein zufälliger Patient geheilt wird ist dann

$$\begin{aligned}\mathbb{P}(\text{Heilung}) &= \mathbb{P}(\text{Heilung} | K_1)\mathbb{P}(K_1) + \mathbb{P}(\text{Heilung} | K_2)\mathbb{P}(K_2) \\ &\quad + \mathbb{P}(\text{Heilung} | K_3)\mathbb{P}(K_3) = 0.59.\end{aligned}$$

Eine Klinik B ist auf dieselben drei Krankheiten spezialisiert wie die Klinik A . Von den Patienten leiden 10 % an K_1 und 80 % an K_2 und 10 % an K_3 . Klinik B hat aber bei jeder dieser drei Krankheiten eine niedrigere Heilungswahrscheinlichkeit als A : nur 0.3 bei K_1 , 0.8 bei K_2 und 0.7 bei K_3 . Dennoch hat B eine höhere Heilungsrate als A :

$$\begin{aligned}\mathbb{P}(\text{Heilung}) &= \mathbb{P}(\text{Heilung} | K_1)\mathbb{P}(K_1) + \mathbb{P}(\text{Heilung} | K_2)\mathbb{P}(K_2) \\ &\quad + \mathbb{P}(\text{Heilung} | K_3)\mathbb{P}(K_3) = 0.74.\end{aligned}$$

Dies ist dadurch leicht zu erklären, dass Klinik B einen hohen Anteil an leichten Fällen hat, während in Klinik A viele schwere Fälle behandelt werden.

Effekte dieser Art kennt man unter dem Begriff *Simpson-Paradox*. Diese Beispiele zeigen, dass insbesondere bei Schlussfolgerungen aufgrund bedingter Wahrscheinlichkeiten sorgfältig nachgedacht werden muss.

Satz 1.23. Es sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein W-Raum und seien $A_1, A_2, \dots, A_n \in \mathcal{A}$ Ereignisse mit der Eigenschaft $\mathbb{P}(A_1 \cap \dots \cap A_n) > 0$. Dann gilt:

$$\mathbb{P}(A_1 \cap \dots \cap A_n) = \mathbb{P}(A_1) \cdot \mathbb{P}(A_2 | A_1) \cdot \mathbb{P}(A_3 | A_1 \cap A_2) \cdots \mathbb{P}(A_n | A_1 \cap \dots \cap A_{n-1})$$

BEWEIS. Beachte, dass

$$\begin{aligned} A_1 \supset A_1 \cap \dots \cap A_n &\implies \mathbb{P}(A_1) \geq \mathbb{P}(A_1 \cap \dots \cap A_n) > 0 \\ A_1 \cap A_2 \supset A_1 \cap \dots \cap A_n &\implies \mathbb{P}(A_1 \cap A_2) \geq \mathbb{P}(A_1 \cap \dots \cap A_n) > 0 \\ &\vdots \\ A_1 \cap \dots \cap A_{n-1} \supset A_1 \cap \dots \cap A_n &\implies \dots \end{aligned}$$

$\implies \mathbb{P}(A_2 | A_1), \dots, \mathbb{P}(A_n | A_1 \cap \dots \cap A_{n-1})$ sind wohldefiniert

Die rechte Seite der Gleichung lässt sich schreiben als

$$\frac{\cancel{\mathbb{P}(A_1)}}{1} \cdot \frac{\cancel{\mathbb{P}(A_1 \cap A_2)}}{\cancel{\mathbb{P}(A_1)}} \cdot \frac{\cancel{\mathbb{P}(A_1 \cap A_2 \cap A_3)}}{\cancel{\mathbb{P}(A_1 \cap A_2)}} \cdots \frac{\mathbb{P}(A_1 \cap \dots \cap A_n)}{\mathbb{P}(A_1 \cap \dots \cap A_{n-1})},$$

womit die Behauptung bewiesen ist. \square

Beispiel 1.24. Wie groß ist die Wahrscheinlichkeit, dass in einer m -köpfigen Gruppe mind. zwei Personen am selben Tag Geburtstag haben?

Alternativ: Werfe m Bälle in n Körbe. Wie groß ist die Wahrscheinlichkeit, dass jeder Ball alleine in seinem Korb landet?

O.E. $m \leq n$, denken uns Bälle nacheinander geworfen.

$A_i = \{\text{Ball } i \text{ landet in einem leeren Korb}\}$

$A = \{\text{Alle Bälle landen alleine in einem Korb}\}$

$$A = \bigcap_{i=1}^m A_i \implies \mathbb{P}\left(\bigcap_{i=1}^m A_i\right) = \mathbb{P}(A_1) \cdot \mathbb{P}(A_2 | A_1) \cdots \mathbb{P}\left(A_m \mid \bigcap_{i=1}^{m-1} A_i\right)$$

Untersuche für $1 \leq j \leq m$: $\mathbb{P}(A_j | \bigcap_{i=1}^{j-1} A_i)$

$$\mathbb{P}\left(A_j \mid \bigcap_{i=1}^{j-1} A_i\right) = \frac{n - (j-1)}{n} = 1 - \frac{j-1}{n}$$

$$\mathbb{P}(A) = \prod_{j=1}^m \left(1 - \frac{j-1}{n}\right) \stackrel{1-x \leq e^{-x}}{\leq} \prod_{j=2}^m e^{-\frac{j-1}{n}} = e^{-\frac{1}{n} \sum_{j=1}^{m-1} j} = e^{-\frac{m(m-1)}{2n}}$$

Für $m = 50$, $n = 365$ gilt $\mathbb{P}(A) \leq 0,05$.

Satz 1.25. Eine Folge $\{B_1, \dots, B_n, \dots\}$ von Mengen mit $B_i \in \mathcal{A}$ heißt **Zerlegung** von Ω , wenn die folgenden beiden Eigenschaften

- $B_i \cap B_j = \emptyset$ ($i \neq j$)
- $\Omega = \bigcup_{i=1}^{\infty} B_i$

erfüllt sind.

i) (**Formel von der totalen Wahrscheinlichkeit**)

Für jede Zerlegung $\{B_1, B_2, \dots\}$ und jedes Ereignis $A \in \mathcal{A}$ gilt:

$$\mathbb{P}(A) = \sum_k \mathbb{P}(B_k) \cdot \mathbb{P}(A | B_k).$$

(Ist $\mathbb{P}(B_k) = 0$, so interpretiere $\mathbb{P}(B_k) \cdot \mathbb{P}(A | B_k) = 0$)

ii) **(Formel von Bayes)**

Ist $\mathbb{P}(A) > 0$ und gelten die Voraussetzungen von (i), so gilt:

$$\mathbb{P}(B_i | A) = \frac{\mathbb{P}(B_i) \cdot \mathbb{P}(A | B_i)}{\sum_k \mathbb{P}(B_k) \cdot \mathbb{P}(A | B_k)}.$$

BEWEIS.

i) Es gilt, dass A eine disjunkte Vereinigung der $A \cap B_k$: $A = \bigcup_k (A \cap B_k)$ Es gilt zunächst

$$\sum_k \mathbb{P}(B_k) \cdot \mathbb{P}(A | B_k) = \sum_{k: \mathbb{P}(B_k) > 0} \mathbb{P}(B_k) \cdot \mathbb{P}(A | B_k) = \sum_{k: \mathbb{P}(B_k) > 0} \mathbb{P}(B_k) \cdot \frac{\mathbb{P}(A \cap B_k)}{\mathbb{P}(B_k)}$$

und weiter wegen $\mathbb{P}(B_k) = 0 \Rightarrow \mathbb{P}(B_k \cap A) = 0$ erhalten wir

$$\sum_{k: \mathbb{P}(B_k) > 0} \mathbb{P}(B_k) \cdot \underbrace{\mathbb{P}(A | B_k)}_{\frac{\mathbb{P}(A \cap B_k)}{\mathbb{P}(B_k)}} = \sum_{k: \mathbb{P}(B_k) > 0} \mathbb{P}(A \cap B_k) = \sum_k \mathbb{P}(A \cap B_k) = \mathbb{P}\left(\bigcup_k (A \cap B_k)\right) = \mathbb{P}(A)$$

ii) Formel von Bayes ergibt sich aus (i) durch $\mathbb{P}(B_i \cap A) = \mathbb{P}(B_i) \cdot \mathbb{P}(A | B_i)$:

$$\mathbb{P}(B_i | A) = \frac{\mathbb{P}(B_i \cap A)}{\mathbb{P}(A)} = \frac{\mathbb{P}(B_i) \cdot \mathbb{P}(A | B_i)}{\sum_k \mathbb{P}(B_k) \cdot \mathbb{P}(A | B_k)}.$$

Damit sind beide Aussagen des Satzes bewiesen. □

Beispiel 1.26 (Monty Hall Problem).

Eine Kandidatin einer TV-Show hat 3 Türen zur Auswahl. Hinter einer Tür befindet sich ein Auto, hinter den beiden anderen sind Ziegen. Die Kandidatin wählt eine Tür aus, nach der Wahl der Kandidatin öffnet der Moderator eine der Türen, hinter denen sich eine Ziege befindet und bietet der Kandidatin an, die Tür zu wechseln.

A = „Kandidatin hat bei Wahl 1 Tür mit Auto gewählt“

G = „Kandidatin gewinnt nach Wechseln der Tür“

Wir suchen die Wahrscheinlichkeit $\mathbb{P}(G)$.

Es gilt dabei, dass $\mathbb{P}(G | A) = 0$, $\mathbb{P}(G | A^c) = 1$. Also:

$$\mathbb{P}(G) = \mathbb{P}(A) \cdot \mathbb{P}(G | A) + \mathbb{P}(A^c) \cdot \mathbb{P}(G | A^c) = 0 \cdot \frac{1}{3} + 1 \cdot \frac{2}{3} = \frac{2}{3}$$

Bedingte Wahrscheinlichkeiten und eine Anwendung der Bayes Regel ist im medizinischen Kontext von Diagnoseverfahren immer wieder relevant (vgl. die Bücher von Tschirk und Eichler/Vogel).

Beispiel 1.27. Es sei Ω eine bestimmte Bevölkerungsgruppe, und sei K die Teilmenge aller Personen, welche an einer bestimmten Krankheit leiden. Für ein medizinisches Diagnoseverfahren, beispielsweise einen Bluttest, sei T die Menge der Personen, bei denen der Test positive ausfällt.

$$\begin{aligned} K &:= \{\text{Personen mit besagter Krankheit}\} \\ T &:= \{\text{Personen mit positivem Testergebnis}\}. \end{aligned}$$

Die Frage ist nun, inwieweit man vom Ausgang des Testergebnisses auf das Vorliegen oder Nichtvorliegen der Krankheit schliessen kann. Wir nennen die bedingte Wahrscheinlichkeiten

$$\mathbb{P}(T \mid K)$$

die *Sensitivität* des Tests bzw.

$$\mathbb{P}(T^c \mid K^c)$$

die *Spezifität* des Tests. In der Praxis interessieren wir uns natürlich besonders für die bedingten Wahrscheinlichkeiten

$$\mathbb{P}(K \mid T) \quad \text{und} \quad \mathbb{P}(K^c \mid T^c),$$

dem positiver bzw. dem negativen prädikativen Wert. Die Formel von Bayes liefert uns

$$\begin{aligned} \mathbb{P}(K \mid T) &= \frac{\mathbb{P}(K)\mathbb{P}(T \mid K)}{\mathbb{P}(K)\mathbb{P}(T \mid K) + \mathbb{P}(K^c)\mathbb{P}(T \mid K^c)} \\ &= \frac{\mathbb{P}(K)\mathbb{P}(T \mid K)}{\mathbb{P}(K)\mathbb{P}(T \mid K) + (1 - \mathbb{P}(K))(1 - \mathbb{P}(T^c \mid K^c))}, \end{aligned}$$

also

$$\mathbb{P}(K \mid T) = \frac{\mathbb{P}(K)\text{Sens}}{\mathbb{P}(K)\text{Sens} + (1 - \mathbb{P}(K))(1 - \text{Spez})}.$$

Analog zeigt man

$$\mathbb{P}(K^c \mid T^c) = \frac{(1 - \mathbb{P}(K))\text{Spez}}{(1 - \mathbb{P}(K))\text{Spez} + \mathbb{P}(K)(1 - \text{Sens})}.$$

Ist beispielsweise

$$\begin{aligned} \text{Sensitivität} &= \mathbb{P}(T \mid K) = 0,222 \\ \text{Spezifität} &= \mathbb{P}(T^c \mid K^c) = 0,993 \end{aligned}$$

und $\mathbb{P}(K) = 0.0264$. Dann ergibt sich

$$\mathbb{P}(K \mid T) \approx 0.462$$

und

$$\mathbb{P}(K \mid T) \approx 0.979,$$

also $\mathbb{P}(K^c \mid T^c) \approx 0.021$. In der Literatur² findet man als Basisrate für Brustkrebs den Wert $\mathbb{P}(K) = 0.01$ und für die Mammografie ergeben sich

$$\mathbb{P}(T \mid K) = 0,8 \quad \text{sowie} \quad \mathbb{P}(T^c \mid K^c) = 0,9.$$

In diesem speziellen Beispiel erhalten wir dann

$$\mathbb{P}(K \mid T) = \frac{8}{107} \approx 0,075.$$

Es kann hilfreich sein³, sich diesen Sachverhalt in absoluten Häufigkeiten darzustellen. Bei 1000 Patientinnen kann man in unserer Situation 10 erkrankte Patientinnen erwarten, bei diesen zehn Patientinnen wird man aufgrund der Sensitivität 8 positive Testergebnisse im Mittel erhalten.

Im Mittel sind 990 Patientinnen gesund, aufgrund der Spezifität des Tests werden wir mit 99 positiven Testergebnissen rechnen. Wir erhalten also

$$\mathbb{P}(K | T) = \frac{\mathbb{P}(K \cap T)}{\mathbb{P}(T)} = \frac{8}{8 + 99}.$$

Rechnungen dieser Art sind in unterschiedlichen Kontexten von großer Bedeutung.

1.3 Stochastische Unabhängigkeit

Die Begriffe der Unabhängigkeit bzw. Abhängigkeit sind zentrale Konzepte innerhalb der Stochastik. Für zwei Ereignisse lässt sich Unabhängigkeit wie folgt definieren.

Definition 1.28 (Unabhängigkeit von zwei Ereignissen).

Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein W-Raum und seien $A, B \in \mathcal{A}$ zwei Ereignisse. Die Ereignisse heißen (stochastisch) **unabhängig**, wenn

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B). \quad (1.1)$$

Das Konzept der Unabhängigkeit beinhaltet immer eine Art Produktformel in Analogie zur Gleichung (1.1). Die folgende Bemerkung stellt den Bezug zum Begriff der bedingten Wahrscheinlichkeit her.

Bemerkung 1.29. Sind A und B unabhängig, so gilt

$$\mathbb{P}(A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \mathbb{P}(A | B), \text{ wenn } \mathbb{P}(B) \neq 0.$$

Sind A und B unabhängig, so ändern sich durch Bedingen auf B die Wahrscheinlichkeit für das Ereignis B nicht.

Beispiel 1.30. Der Wurf von zwei unterscheidbaren Würfeln wird beschrieben durch $\Omega = \{1, \dots, 6\}^2$ mit der Gleichverteilung \mathbb{P} . Die Ereignisse

$$A = \{ \text{erster Würfel zeigt 6} \} = \{(6, l) \mid l \in \{1, \dots, 6\}\}$$

und

$$B = \{ \text{zweiter Würfel zeigt 6} \} = \{(k, 6) \mid k \in \{1, \dots, 6\}\}$$

sind wegen

$$\mathbb{P}(A \cap B) = \frac{1}{36} = \mathbb{P}(A) \cdot \mathbb{P}(B)$$

unabhängig. Sei weiter

$$C = \{ \text{Augensumme ist 7} \} = \{(k, l) \in \Omega \mid k + l = 7\},$$

dann sind A und C ebenfalls unabhängig. Unabhängigkeit ist also nicht dasselbe wie kausale Unabhängigkeit!

Sind weiter A, B, C paarweise unabhängig, so folgt im Allgemeinen **nicht**, dass auch A, B, C unabhängig sind. Dies erkennt man auch wieder am Würfelbeispiel.

Beispiel 1.31 (Zweimaliges Würfeln mit fairem Würfel).

A = Augenzahl im ersten Wurf gerade

B = Augenzahl im zweiten Wurf gerade

C = Summe der Augenzahlen beider Würfe ergibt 7

Ergebnismenge $\Omega = \{(i, j) : i, j \in \{1, 2, 3, \dots, 6\}\}$, $\mathbb{P}(\{(i, j)\}) = \frac{1}{36}$ für alle $(i, j) \in \Omega$. Es gelten

$$\begin{aligned} \mathbb{P}(A) &= \frac{1}{2} = \mathbb{P}(B), & \mathbb{P}(A \cap B) &= \mathbb{P}(A) \cdot \mathbb{P}(B), \\ \mathbb{P}(A \cap C) &= \frac{1}{12} = \mathbb{P}(A) \cdot \mathbb{P}(C), & \mathbb{P}(B \cap C) &= \mathbb{P}(B) \cdot \mathbb{P}(C) \end{aligned}$$

Allerdings ist

$$\mathbb{P}(A \cap B \cap C) = 0 \neq \mathbb{P}(A) \cdot \mathbb{P}(B) \cdot \mathbb{P}(C). \quad (1.2)$$

Aus der paarweisen Unabhängigkeit der Ereignisse A, B und C folgt als nicht die Gültigkeit der Produktformel (1.2) für die Mengen A, B und C .

Die Erweiterung des Begriffs der Unabhängigkeit auf Familien von Ereignissen ist also notwendigerweise etwas involvierter.

Definition 1.32. Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. Eine Familie $(A_i)_{i \in I}$, mit $A_i \in \mathcal{A}$ heißt **unabhängig**, wenn für jede *endliche* Teilmenge $J \subseteq I$ gilt

$$\mathbb{P}\left(\bigcap_{j \in J} A_j\right) = \prod_{j \in J} \mathbb{P}(A_j) \quad (1.3)$$

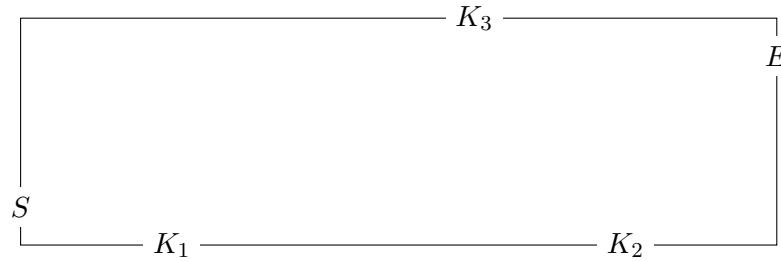
Die Definition der Unabhängigkeit einer Familie von Ereignissen beruht also auf der Gültigkeit der Produktformel (1.3) für jede *endliche* Teilfamilie von Ereignissen.

Beispiel 1.33. Die Mengen A_1, A_2, A_3, A_4 sind unabhängig, wenn

$$\begin{aligned} \mathbb{P}(A_1 \cap A_2) &= \mathbb{P}(A_1) \cdot \mathbb{P}(A_2), & \mathbb{P}(A_1 \cap A_3) &= \mathbb{P}(A_1) \cdot \mathbb{P}(A_3), & \mathbb{P}(A_1 \cap A_4) &= \mathbb{P}(A_1) \cdot \mathbb{P}(A_4), \\ \mathbb{P}(A_2 \cap A_3) &= \mathbb{P}(A_2) \cdot \mathbb{P}(A_3), & \mathbb{P}(A_2 \cap A_4) &= \mathbb{P}(A_2) \cdot \mathbb{P}(A_4), & \mathbb{P}(A_3 \cap A_4) &= \mathbb{P}(A_3) \cdot \mathbb{P}(A_4), \\ \mathbb{P}(A_1 \cap A_2 \cap A_3) &= \mathbb{P}(A_1) \cdot \mathbb{P}(A_2) \cdot \mathbb{P}(A_3), & \mathbb{P}(A_1 \cap A_2 \cap A_4) &= \mathbb{P}(A_1) \cdot \mathbb{P}(A_2) \cdot \mathbb{P}(A_4) \\ \mathbb{P}(A_1 \cap A_3 \cap A_4) &= \mathbb{P}(A_1) \cdot \mathbb{P}(A_3) \cdot \mathbb{P}(A_4), & \mathbb{P}(A_2 \cap A_3 \cap A_4) &= \mathbb{P}(A_2) \cdot \mathbb{P}(A_3) \cdot \mathbb{P}(A_4) \\ \mathbb{P}(A_1 \cap A_2 \cap A_3 \cap A_4) &= \mathbb{P}(A_1) \cdot \mathbb{P}(A_2) \cdot \mathbb{P}(A_3) \cdot \mathbb{P}(A_4) \end{aligned}$$

Die Forderung der unabhängigkeit der vier Mengen liefert uns somit bereits 11 Gleichungen. Dies demonstriert, dass die Forderung der Unabhängigkeit eine sehr starke Annahme ist mit entsprechend weitreichenden Konsequenzen.

Beispiel 1.34.



Zwischen Sender S und Empfänger E gibt es zwei verschiedene Routen: R_1 führt über die Knotenrechner K_1 und K_2 ; R_2 führt über K_3 .

Annahme: Jeder Knotenrechner funktioniert unabhängig voneinander. Mit Wahrscheinlichkeit p wird eine stabile Verbindung hergestellt.

$$K_i = \text{„}K_i \text{ intakt“}, \quad R_i = \text{„}R_i \text{ verfügbar“}$$

$$R_1 = K_1 \cap K_2, \quad R_2 = K_3$$

$$A := R_1 \cup R_2$$

$$\mathbb{P}(R_1) = \mathbb{P}(K_1 \cap K_2) = \mathbb{P}(K_1) \cdot \mathbb{P}(K_2) = p^2, \quad \mathbb{P}(R_2) = \mathbb{P}(K_3) = p$$

$$\begin{aligned} \mathbb{P}(A) &= \mathbb{P}(R_1 \cup R_2) = 1 - \mathbb{P}((R_1 \cup R_2)^c) = 1 - \mathbb{P}(R_1^c \cap R_2^c) = 1 - \mathbb{P}(R_1^c) \cdot \mathbb{P}(R_2^c) \\ &= 1 - (1 - p^2) \cdot (1 - p) = p + p^2 - p^3 \end{aligned}$$

Wir benutzen hierbei, dass die Unabhängigkeit von R_1 und R_2 auch die Unabhängigkeit von R_1^c und R_2^c impliziert. Dies ist natürlich noch zu beweisen.

Der folgende Satz zeigt zunächst, dass durch Einschränkung auf eine kleinere Klasse von Ereignissen die Unabhängigkeit nicht verloren geht. Desweiteren wird bewiesen, dass man zu einer unabhängigen Familie das sichere und das unmögliche Ereignis sowie Komplemente von Ereignissen hinzunehmen darf, ohne die Unabhängigkeit zu verlieren.

Satz 1.35.

- i) Jede Teilfamilie einer unabhängigen Familie von Ereignissen ist unabhängig.
- ii) Ist $(A_i)_{i \in I}$ eine Familie von unabhängigen Ereignissen, ist $k \notin I$ und $A_k \in \mathcal{A}$ mit $\mathbb{P}(A_k) = 1$ oder $\mathbb{P}(A_k) = 0$, so ist $(A_i)_{i \in I \cup \{k\}}$ unabhängig.
- iii) Ist $(A_i)_{i \in I}$ unabhängig und für jedes $i \in I$, sei B_i eines der Ereignisse $\emptyset, \Omega, A_i, A_i^c$, so ist $(B_i)_{i \in I}$ unabhängig.
- iv) Ist $I = \{1, 2, \dots, n\}$, so ist $(A_i)_{i \in I}$ unabhängig genau dann, wenn für jede Wahl $B_i \in \{A_i, A_i^c\}$ die Produktregel

$$\mathbb{P}(B_1 \cap \dots \cap B_n) = \mathbb{P}(B_1) \cdots \mathbb{P}(B_n)$$

gilt.

BEWEIS.

- i) Folgt aus der Definition.

ii) Es sei $J \subset I \cup \{k\}$ mit $|J| < \infty$ mit $k \in J$. Ist $\mathbb{P}(A_k) = 0$, so gilt:

$$\mathbb{P}\left(\overbrace{\bigcap_{j \in J} A_j}^{\subseteq A_k}\right) = 0, \quad \prod_{j \in J} \mathbb{P}(A_j) = 0$$

Ist dagegen $\mathbb{P}(A_k) = 1$, so gilt für jedes Ereignis A , dass $\mathbb{P}(A) = \mathbb{P}(A \cap A_k) + \mathbb{P}(A \cap A_k^c) = \mathbb{P}(A_k \cap A)$ und somit

$$\prod_{j \in J} \mathbb{P}(A_j) = \prod_{j \in J \setminus \{k\}} \mathbb{P}(A_j) = \mathbb{P}\left(\bigcap_{j \in J \setminus \{k\}} A_j\right) = \mathbb{P}\left(\bigcap_{j \in J} A_j\right)$$

iii) Wegen (ii) betrachte nur $B_i \in \{A_i, A_i^c\}$.

Durch Induktion über m beweisen wir: Ist $J \subset I$ endlich und ist $|\{j \in J : B_j = A_j^c\}| \leq m$, so gilt die Produktformel für $(B_j)_{j \in J}$.

Ist $m = 0$, so gilt die Produktformel für die $(B_j)_{j \in J}$; Sei nun die Induktionsannahme für m gültig und sei $J' \subseteq I$ eine endliche Teilmenge mit $|\{j \in J' : B_j = A_j^c\}| = m + 1$. O. E. sei $J' = \{1, \dots, N\}$, $N \geq m + 1$ und $B_1 = A_1^c$.

Wenden IV auf (A_1, B_2, \dots, B_N) und (B_2, \dots, B_N) an.

$$\begin{aligned} \mathbb{P}\left(\underbrace{\bigcap_{j=1}^N B_j}_{B_1 \cap \bigcap_{j=2}^N B_j = A_1^c \cap \bigcap_{j=2}^N B_j}\right) &= \mathbb{P}\left(\bigcap_{j=2}^N B_j\right) - \mathbb{P}\left(A_1 \cap \bigcap_{j=2}^N B_j\right) \\ &= \prod_{j=2}^N \mathbb{P}(B_j) - \mathbb{P}(A_1) \cdot \prod_{j=2}^N \mathbb{P}(B_j) \\ &= \prod_{j=2}^N \mathbb{P}(B_j) \cdot (1 - \mathbb{P}(A_1)) = \prod_{j=1}^N \mathbb{P}(B_j) \end{aligned}$$

\implies Beh.

iv) Die Notwendigkeit ist schon Teil (iii). Zur Umkehrung addiere Produktformel für B_1, B_2, \dots, B_n und B_1^c, B_2, \dots, B_n , so folgt

$$\begin{aligned} \mathbb{P}(B_2 \cap \dots \cap B_n) &= \mathbb{P}(B_1 \cap B_2 \cap \dots \cap B_n) + \mathbb{P}(B_1^c \cap B_2 \cap \dots \cap B_n) \\ &= \mathbb{P}(B_1) \prod_{j=2}^n \mathbb{P}(B_j) + \mathbb{P}(B_1^c) \prod_{j=2}^n \mathbb{P}(B_j) \\ &= \prod_{j=2}^n \mathbb{P}(B_j) \end{aligned}$$

und damit

$$\mathbb{P}(B_2 \cap \dots \cap B_n) = \prod_{j=2}^n \mathbb{P}(B_j).$$

Wir erhalten also die Produktformel für Durchschnitte von $(n - 1)$ Mengen. Iterativ erhält man die Behauptung. \square

Produkt Räume und Unabhängigkeit

Es seien n diskrete Wahrscheinlichkeitsräume $(\Omega_1, \mathcal{P}(\Omega_1), \mathbb{P}_1), (\Omega_2, \mathcal{P}(\Omega_2), \mathbb{P}_2), \dots, (\Omega_n, \mathcal{P}(\Omega_n), \mathbb{P}_n)$, die jeweils ein Zufallsexperiment beschreiben. Man kann nun das **unabhängige** Ausführen dieser n Experimente auf demselben Wahrscheinlichkeitsraum $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ mathematisch wie folgt beschreiben: Man definiert

$$\Omega := \Omega_1 \times \Omega_2 \times \dots \times \Omega_n.$$

Die σ -Algebra \mathcal{A} ist die Potenzmenge $\mathcal{P}(\Omega)$ von Ω und das Maß \mathbb{P} lässt sich wie folgt definieren: für $\omega = (\omega_1, \omega_2, \dots, \omega_n) \in \Omega$

$$\mathbb{P}(\{\omega\}) = \mathbb{P}_1(\{\omega_1\}) \cdot \mathbb{P}_2(\{\omega_2\}) \cdot \dots \cdot \mathbb{P}_n(\{\omega_n\}).$$

Dann gilt für beliebige Mengen $B_i \subset \Omega_i$ ($i = 1, \dots, n$)

$$\begin{aligned} \mathbb{P}(B_1 \times B_2 \times \dots \times B_n) &= \sum_{\omega \in B_1 \times B_2 \times \dots \times B_n} \mathbb{P}(\{\omega\}) = \sum_{(\omega_1, \dots, \omega_n) \in B_1 \times B_2 \times \dots \times B_n} \mathbb{P}_1(\{\omega_1\}) \dots \mathbb{P}_n(\{\omega_n\}) \\ &= \sum_{\omega_1 \in B_1} \sum_{\omega_2 \in B_2} \dots \sum_{\omega_n \in B_n} \mathbb{P}_1(\{\omega_1\}) \mathbb{P}_2(\{\omega_2\}) \dots \mathbb{P}_n(\{\omega_n\}) \\ &= \prod_{i=1}^n \left(\sum_{\omega_i \in B_i} \mathbb{P}_i(\{\omega_i\}) \right) = \prod_{i=1}^n \mathbb{P}_i(B_i). \end{aligned}$$

Setzt man $B_i = \Omega_i$ für jede $i \in \{1, \dots, n\}$, dann erhalten wir insbesondere

$$\mathbb{P}(\Omega) = \mathbb{P}(\Omega_1 \times \dots \times \Omega_n) = \prod_{i=1}^n \mathbb{P}_i(\Omega_i) = 1.$$

Wir haben also ein Wahrscheinlichkeitsmaß auf Ω erhalten. Man schreibt auch $\mathbb{P} = \mathbb{P}_1 \otimes \dots \otimes \mathbb{P}_n$. Wie bereits erwähnt modelliert der Wahrscheinlichkeitsraum $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ das unabhängige Ausführen der n durch $(\Omega_1, \mathcal{P}(\Omega_1), \mathbb{P}_1), \dots, (\Omega_n, \mathcal{P}(\Omega_n), \mathbb{P}_n)$ beschriebenen Telexperimente.

1.4 Zufallsvariablen

Bisher haben wir uns mit Wahrscheinlichkeiten von Ereignissen auseinandergesetzt. Wir werden im Verlauf der Vorlesung sehen, dass es sinnvoll und praktisch relevant ist, Ereignisse durch Zufallsvariablen zu beschreiben. Man kann sogar noch weiter gehen und den Begriff der Zufallsvariable als zentrales Objekt der Stochastik ansehen.

Definition 1.36 (Zufallsvariable).

Es sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein diskreter W-Raum über der Ergebnismenge Ω . Eine Abbildung

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

heißt **(reellwertige) Zufallsvariable** (kurz: ZV).

Der Wertebereich $W_X = X(\Omega) = \{x \in \mathbb{R} \mid \exists \omega \in \Omega : X(\omega) = x\}$ von X ist endlich oder abzählbar unendlich, wenn der zugrundeliegende Wahrscheinlichkeitsraum diskret ist.

Bemerkung 1.37. Ist $(\Omega, \mathcal{A}, \mathbb{P})$ kein diskreter W-Raum, dann ist Definition (1.36) nicht vollständig. Betrachte die Funktion

$$X = \text{Körpergröße}, \quad \underbrace{\mathbb{P}(\{X > 180\})}_{\in \mathcal{A}}$$

Zusätzlich sollte gefordert werden, dass für jedes $x \in \mathbb{R}$ gilt

$$\{X \leq x\} = \{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{A}$$

Man sagt, X muss **messbar** sein.

Beispiel 1.38 (3-maliger Wurf einer fairen Münze).

$$\Omega = \{K, Z\}^3 = \{(K, K, K), (K, K, Z), (K, Z, K), \dots\}$$

Sei Y eine Zufallsvariable (ZV) mit Y = Gesamtzahl der Würfe mit Ergebnis „Kopf“.

$$Y: \Omega \rightarrow \mathbb{R}, \text{ z.B. gilt } Y((K, K, W)) = 2 \\ Y((W, W, W)) = 0$$

$$E = \{2 \text{ mal Kopf}\} = \{\omega \in \Omega : Y(\omega) = 2\} = \{Y = 2\}.$$

Das Konzept der Zufallsvariable taucht in theoretischen wie auch angewandten Untersuchungen häufig und auf natürliche Art und Weise auf. Dies soll anhand des folgenden Beispiels illustriert werden.

Beispiel 1.39. In einigen Anwendungen möchte man prüfen, ob eine Bitfolge $\omega \in \Omega = \{0, 1\}^n$ der Länge n zufällig zustande kam. Die Zufallsvariable

$$X: \Omega \rightarrow \mathbb{R} \\ X(\omega) = X((\omega_1, \omega_2, \dots, \omega_n)) = |\{i \in \{2, \dots, n\} \mid \omega_{i-1} \neq \omega_i\}|$$

beschreibt die Anzahl der 'Wechsel' zwischen Einsen und Nullen. Es gilt zum Beispiel für $n = 5$

$$X((1, 1, 1, 0, 0)) = 1 \quad \text{und} \quad X((1, 1, 0, 1, 0)) = 3.$$

Ist \mathbb{P} die Gleichverteilung auf Ω , dann gilt für jedes $l = 0, \dots, n - 1$

$$\mathbb{P}(\{\omega \in \Omega \mid X(\omega) = l\}) = \frac{\binom{n-1}{l}}{2^{n-1}}.$$

Die Zufallsvariable X zerlegt also unseren Ereignisraum Ω in die Elementareignisse mit l Bitwechseln.

Definition 1.40. Sei $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ ein diskreter Wahrscheinlichkeitsraum und $X: \Omega \rightarrow \mathbb{R}$ eine Zufallsvariable. Dann definiere

$$f_X: \mathbb{R} \supseteq W_X \ni x \mapsto \mathbb{P}(\{X = x\}) \in [0, 1].$$

f_X nennt man **diskrete Dichtefunktion** (Zähldichte) von X . Die Funktion

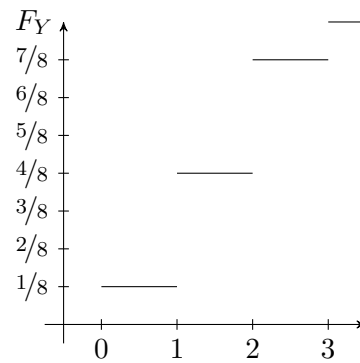
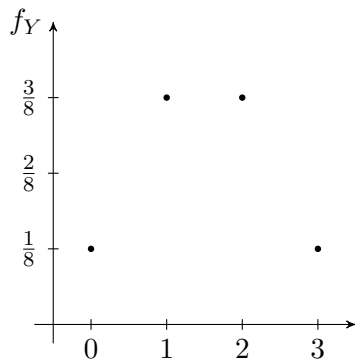
$$F_X: \mathbb{R} \rightarrow [0, 1], \quad x \mapsto F_X(x) = \mathbb{P}(\{X \leq x\})$$

heißt **Verteilungsfunktion** von X . Schreibe oft $\mathbb{P}(X \leq x)$ als Kurzform für $\mathbb{P}(\{X \leq x\})$.

Beispiel 1.41 (3-maliger Wurf einer fairen Münze).

$$Y = \# \text{Kopf}$$

$$\mathbb{P}(\{Y = 0\}) = \mathbb{P}(\{(Z, Z, Z)\}) = \frac{1}{8}, \quad \mathbb{P}(\{Y = 1\}) = \frac{3}{8}, \quad \mathbb{P}(\{Y = 2\}) = \frac{3}{8}, \quad \mathbb{P}(\{Y = 3\}) = \frac{1}{8}$$

**Definition 1.42 (Erwartungswert).**

Zu einer Zufallsvariable X auf dem diskreten W-Raum $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ definiere den Erwartungswert $\mathbb{E}[X]$ von X durch

$$\mathbb{E}[X] = \sum_{x \in W_X} x \cdot \mathbb{P}(\{X = x\}) = \sum_{x \in W_X} x \cdot f_X(x) \quad (1.4)$$

sofern

$$\sum_{x \in W_X} |x| \cdot f_X(x) < \infty \quad (1.5)$$

erfüllt ist

Die Bedingung (1.5) erlaubt es uns, die Sätze der Analysis zur Theorie der Reihen zu verwenden.

Beispiel 1.43 (3-maliger Wurf einer fairen Münze).

Betrachte die Situation aus Beispiel (1.38). $Y = \# \text{Kopf}$.

$$\begin{aligned} \mathbb{E}[Y] &= \sum_{i=0}^3 i \cdot \mathbb{P}(\{Y = i\}) = 0 \cdot \mathbb{P}(\{Y = 0\}) + 1 \cdot \mathbb{P}(\{Y = 1\}) + 2 \cdot \mathbb{P}(\{Y = 2\}) \\ &\quad + 3 \cdot \mathbb{P}(\{Y = 3\}) \\ &= 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = \frac{3}{2}. \end{aligned}$$

Beispiel 1.44. In einem Casino in Bahnhofsnähe wird das folgende Glücksspiel gezockt. Eine Münze wird solange geworfen, bis sie das erste Mal Kopf zeigt. Sei k die Anzahl der durchgeführten Würfe bis dahin. Wenn k ungerade ist, zahlt der Spieler an die Bank k Euro. Andernfalls muss die Bank k Euro an den Spieler zahlen.

Definiere

$G :=$ Gewinn der Bank

$\Omega = \{\omega_1, \omega_2, \dots\}$, $\omega_i = \{i \text{ Versuche bis Kopf}\}$

$G: \Omega \rightarrow \mathbb{R}$, $G(\omega_i) = \begin{cases} -i, & \text{falls } i \text{ gerade} \\ i, & \text{falls } i \text{ ungerade} \end{cases}$

$\mathbb{P}(\omega_i) = \left(\frac{1}{2}\right)^{i-1} \cdot \frac{1}{2} = \left(\frac{1}{2}\right)^i$

$\mathbb{E}[G] = \sum_{k=1}^{\infty} (-1)^{k-1} \cdot k \cdot \left(\frac{1}{2}\right)^k$, der EW existiert: $\sum_{k=1}^{\infty} k \cdot \left(\frac{1}{2}\right)^k < \infty$. Hierzu beachte, dass für $p \in (0, 1)$

$$\begin{aligned} \sum_{k=1}^{\infty} kp^k &= p \sum_{k=1}^{\infty} kp^{k-1} = p \sum_{k=1}^{\infty} \frac{d}{dp} p^k \\ &= p \frac{d}{dp} \sum_{k=1}^{\infty} p^k = p \frac{d}{dp} \sum_{k=0}^{\infty} p^k = p \frac{d}{dp} \frac{1}{1-p} \\ &= \frac{p}{(1-p)^2}. \end{aligned}$$

Wir erhalten also

$$\begin{aligned} \mathbb{E}[G] &= \sum_{j=1}^{\infty} \left[(2j-1) \left(\frac{1}{2}\right)^{2j-1} - 2j \left(\frac{1}{2}\right)^{2j} \right] = \sum_{j=1}^{\infty} \left[(2j-1) \left(\frac{1}{2}\right)^{2j-1} - j \left(\frac{1}{2}\right)^{2j-1} \right] \\ &= \sum_{j=1}^{\infty} \left(\frac{1}{2}\right)^{2j-1} \cdot ((2j-1) - j) = \sum_{j=1}^{\infty} \frac{1}{2} \left(\frac{1}{2}\right)^{2j-2} \cdot (j-1) \\ &= \frac{1}{2} \sum_{j=1}^{\infty} \left(\frac{1}{4}\right)^{j-1} \cdot (j-1) = \frac{1}{2} \cdot \sum_{j=1}^{\infty} (j-1) \cdot \left(\frac{1}{4}\right)^{j-1} = \frac{1}{2} \cdot \frac{1/4}{(1-1/4)^2} = \frac{2}{9}. \end{aligned}$$

Bemerkung 1.45 (Alternative Formel für den Erwartungswert). Sei $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ ein diskreter Wahrscheinlichkeitsraum und $X: \Omega \rightarrow \mathbb{R}$ eine Zufallsvariable mit existierendem Erwartungswert. Es gilt

$$\boxed{\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega) \cdot \mathbb{P}(\{\omega\}).} \quad (1.6)$$

Denn wegen

$$\mathbb{P}(X = x) = \mathbb{P}(\{\omega \mid X(\omega) = x\}) = \sum_{\omega: X(\omega)=x} \mathbb{P}(\{\omega\})$$

gilt

$$\mathbb{E}[X] = \sum_{x \in W_X} x \cdot \mathbb{P}(X = x) = \sum_{x \in W_X} \sum_{\substack{\omega \in \Omega \\ X(\omega)=x}} X(\omega) \mathbb{P}(\{\omega\}) = \sum_{\omega \in \Omega} X(\omega) \mathbb{P}(\{\omega\}).$$

Satz 1.46 (Monotonie des Erwartungswerts). Sei $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ ein diskreter Wahrscheinlichkeitsraum und seien $X, Y: \Omega \rightarrow \mathbb{R}$ Zufallsvariablen mit $X(\omega) \leq Y(\omega) \forall \omega \in \Omega$. Existieren $\mathbb{E}[X]$ und $\mathbb{E}[Y]$, so gilt:

$$\mathbb{E}[X] \leq \mathbb{E}[Y]$$

BEWEIS. Wir benutzen die alternative Darstellung des Erwartungswertes in (1.6) und erhalten

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega) \cdot \mathbb{P}(\{\omega\}) \leq \sum_{\omega \in \Omega} Y(\omega) \cdot \mathbb{P}(\{\omega\}) = \mathbb{E}[Y]. \quad \square$$

Beispiel 1.47.

Ist für alle $\omega \in \Omega$, $X(\omega) \in [a, b]$ Dann ist $\mathbb{E}[X] \in [a, b]$

Wende Satz (1.46) an mit $Y(\omega) = b \forall \omega \in \Omega$:

$$\mathbb{E}[X] \leq \mathbb{E}[Y] = \sum_{\omega \in \Omega} Y(\omega) \cdot \mathbb{P}(\{\omega\}) = b \cdot \underbrace{\sum_{\omega \in \Omega} \mathbb{P}(\{\omega\})}_{=1}.$$

Analog wende Satz (1.46) an mit $Z(\omega) = a \forall \omega \in \Omega$.

Bemerkung 1.48. Sei $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ ein diskreter W-Raum, seien $X, Y: \Omega \rightarrow \mathbb{R}$ Zufallsvariablen und $f: D \rightarrow \mathbb{R}$ eine Abbildung mit $W_X \subseteq D$. Dann ist $Y := f \circ X: \Omega \rightarrow \mathbb{R}$ ebenfalls eine Zufallsvariable. Es gilt für $y \in W_Y$

$$\mathbb{P}(Y = y) = \mathbb{P}(\{\omega \in \Omega : f(X(\omega)) = y\}) = \sum_{x: f(x)=y} \mathbb{P}(X = x).$$

Unter der Annahme, dass die Erwartungswerte existieren, gilt,

$$\mathbb{E}[\underbrace{f \circ X}_Y] = \sum_{y \in W_Y} y \cdot \mathbb{P}(\{Y = y\}) = \sum_{y \in W_Y} y \cdot \sum_{\substack{x \in W_X, \\ f(x)=y}} \mathbb{P}(\{X = x\}) = \sum_{x \in W_X} f(x) \cdot \underbrace{\mathbb{P}(\{X = x\})}_{f_X(x)}$$

Die Formel

$$\boxed{\mathbb{E}[f \circ X] = \mathbb{E}[f(X)] = \sum_{x \in W_X} f(x) \mathbb{P}(X = x)}$$

ist in konkreten Rechnungen sehr hilfreich und nützlich.

Satz 1.49. Für eine diskrete Zufallsvariable X und für $a, b \in \mathbb{R}$ gilt

$$\mathbb{E}[aX + b] = a \cdot \mathbb{E}[X] + b.$$

BEWEIS.

$$\begin{aligned} \mathbb{E}[aX + b] &= \sum_{\omega \in \Omega} (a \cdot X(\omega) + b) \cdot \mathbb{P}(\{\omega\}) = \sum_{\omega \in \Omega} a \cdot X(\omega) \cdot \mathbb{P}(\{\omega\}) + \sum_{\omega \in \Omega} b \cdot \mathbb{P}(\{\omega\}) \\ &= a \cdot \underbrace{\sum_{\omega \in \Omega} X(\omega) \cdot \mathbb{P}(\{\omega\})}_{\mathbb{E}[X]} + b \cdot \underbrace{\sum_{\omega \in \Omega} \mathbb{P}(\{\omega\})}_{=1} \end{aligned} \quad \square$$

Satz 1.50. Sei $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ ein diskreter W-Raum und sei $X: \Omega \rightarrow \mathbb{R}$ eine Zufallsvariable mit $W_X \subseteq \mathbb{N}_0$. Dann gilt

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} \mathbb{P}(X \geq i) = \sum_{i=1}^{\infty} (1 - F_X(i-1))$$

BEWEIS.

$$\begin{aligned} \mathbb{E}[X] &= \sum_{i=1}^{\infty} i \cdot \mathbb{P}(X = i) = \sum_{i=1}^{\infty} \sum_{j=1}^i \mathbb{P}(X = i) = \sum_{j=1}^{\infty} \sum_{i=j}^{\infty} \mathbb{P}(X = i) \\ &= \sum_{j=1}^{\infty} \mathbb{P}(X \geq j) \end{aligned} \quad \square$$

Der folgende Satz besagt, dass die Erwartungswertbildung eine lineare Größe ist. Der Erwartungswert einer Linearkombination von Zufallsvariablen ist also gleich der Linearkombination der Erwartungswerte. Die Bedeutung dieses Ergebnisses kann nicht überbewertet werden. Wir werden von diesem Ergebnis sehr häufig Gebrauch machen.

Satz 1.51 (Linearität des Erwartungswerts). Seien $X_1, \dots, X_n: \Omega \rightarrow \mathbb{R}$ diskrete Zufallsvariablen, $a_1, \dots, a_n \in \mathbb{R}$ und $\mathbb{E}[X_1], \dots, \mathbb{E}[X_n]$ existieren. Dann gilt

$$\mathbb{E}[a_1 X_1 + a_2 X_2 + \dots + a_n X_n] = a_1 \cdot \mathbb{E}[X_1] + a_2 \cdot \mathbb{E}[X_2] + \dots + a_n \cdot \mathbb{E}[X_n]$$

BEWEIS. Analog wie im Satz (1.49). □

Beispiel 1.52.

n betrunkene Seeleute torkeln nach Landgang in ihre Kojen. Jede Zuordnung Seemann \leftrightarrow Bett ist gleichwahrscheinlich, nur ein Seemann pro Bett. Wie viele Seeleute landen im eigenen Bett?

$$X_i = \begin{cases} 1, & \text{Seemann } i \text{ im eigenen Bett} \\ 0, & \text{sonst} \end{cases}$$

Definiere $X := X_1 + \dots + X_n$. Dies gibt uns die Anzahl der Seeleute im eigenen Bett.

$$\Omega = \{\pi : \pi \text{ Permutation von } \{1, \dots, n\}\}, |\Omega| = n!$$

$$\pi : \{1, \dots, n\} \xrightarrow{\text{bij.}} \{1, \dots, n\}, \pi(i) = j: \text{Seemann } i \text{ in Bett von } j$$

$$\mathbb{P}(X_i = 1) = \frac{|\{\pi \in \Omega : \pi(i) = i\}|}{|\Omega|} = \frac{(n-1)!}{n!} = \frac{1}{n}$$

$$\mathbb{E}[X_i] = 0 \cdot \mathbb{P}(X_i = 0) + 1 \cdot \mathbb{P}(X_i = 1) = \frac{1}{n}$$

$$\implies \mathbb{E}[X] = \mathbb{E}[X_1] + \dots + \mathbb{E}[X_n] = n \cdot \frac{1}{n} = 1.$$

Definition 1.53. Für ein Ereignis $A \in \mathcal{P}(\Omega)$ ist

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

die **Indikatorvariable** des Ereignisses A .

$$\mathbb{E}[\mathbb{1}_A] = 1 \cdot \underbrace{\mathbb{P}(\{\omega : \mathbb{1}_A(\omega) = 1\})}_{\mathbb{P}(A)} + 0 \cdot \underbrace{\mathbb{P}(\{\omega : \mathbb{1}_A(\omega) = 0\})}_{\mathbb{P}(A^c)} = \mathbb{P}(A)$$

$$\mathbb{E}[\mathbb{1}_{A_1} \cdots \mathbb{1}_{A_n}] = \mathbb{P}(A_1 \cap \dots \cap A_n)$$

Aussagen über Wahrscheinlichkeiten von Ereignissen lassen sich somit in Aussagen über Erwartungswerte von Indikatorvariablen übersetzen.

Bemerkung 1.54. Mit Hilfe des Erwartungswertes und von Indikatorvariablen kann man sehr elegant die Siebformel aus Satz 1.8 beweisen. Es seien dazu A_1, \dots, A_n Ereignisse und sei $B = A_1 \cup \dots \cup A_n$. Beachte dass das Produkt $\prod_{i=1}^n (1 - \mathbb{1}_{A_i})$ genau dann gleich Eins ist, wenn $\mathbb{1}_{A_1} = \dots = \mathbb{1}_{A_n} = 0$, d.h. wenn das Ereignis B nicht eintritt. Durch Ausmultiplizieren erhält man

$$\mathbb{1}_{B^c} = \prod_{i=1}^n (1 - \mathbb{1}_{A_i}) = 1 - \sum_{1 \leq i \leq n} \mathbb{1}_{A_i} + \sum_{1 \leq i_1 < i_2 \leq n} \mathbb{1}_{A_{i_1}} \mathbb{1}_{A_{i_2}} - \dots + (-1)^n \mathbb{1}_{A_1} \cdots \mathbb{1}_{A_n}.$$

Nimmt man auf beiden Seiten den Erwartungswert, dann erhält man die gewünschte Aussage.

Definition 1.55. Es sei $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ ein diskreter W-Raum und sei $A \in \mathcal{P}(\Omega)$ ein Ereignis mit $\mathbb{P}(A) > 0$. Sei $X : \Omega \rightarrow \mathbb{R}$ eine Zufallsvariable. Die bedingte Zufallsvariable $X | A$ besitzt die Dichte

$$f_{X|A}(x) = \mathbb{P}(\{X = x\} | A) = \frac{\mathbb{P}(\{X = x\} \cap A)}{\mathbb{P}(A)}.$$

Beachte, dass

$$\sum_{x \in W_X} f_{X|A}(x) = \sum_{x \in W_X} \frac{\mathbb{P}(\{X = x\} \cap A)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A)}{\mathbb{P}(A)} = 1$$

Der **bedingte Erwartungswert** für X unter der Bedingung A :

$$\mathbb{E}[X | A] = \sum_{x \in W_X} x \cdot f_{X|A}(x).$$

sofern $\sum_{x \in W_X} |x| \cdot f_{X|A}(x) < \infty$

Satz 1.56. Sei X eine diskrete Zufallsvariable mit Erwartungswert auf $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$, $(A_i)_{i \geq 1} \subseteq \mathcal{P}(\Omega)$ mit $\mathbb{P}(A_i) > 0$, so dass $(A_i)_{i \geq 1}$ eine disjunkte Zerlegung von Ω darstellt. Dann gilt

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} \mathbb{E}[X | A_i] \cdot \mathbb{P}(A_i).$$

BEWEIS. Erwartungswert von X existiert $\implies \sum_{x \in W_X} |x| \cdot \mathbb{P}(X = x) < \infty$.

Dann folgt

$$\sum_{x \in W_X} |x| \cdot f_{X|A_i}(x) = \sum_{x \in W_X} |x| \cdot \frac{\mathbb{P}(\{X = x\} \cap A_i)}{\mathbb{P}(A_i)} \leq \frac{1}{\mathbb{P}(A_i)} \sum_{x \in W_X} |x| \cdot \mathbb{P}(X = x) < \infty.$$

Damit sind die einzelnen Erwartungswerte wohldefiniert. Weiter gilt:

$$\begin{aligned} \mathbb{E}[X] &= \sum_{x \in W_X} x \cdot \mathbb{P}(X = x) = \sum_{x \in W_X} x \cdot \sum_{i=1}^{\infty} \mathbb{P}(X = x | A_i) \cdot \mathbb{P}(A_i) \\ &= \sum_{i=1}^{\infty} \mathbb{P}(A_i) \cdot \sum_{x \in W_X} x \cdot \underbrace{\mathbb{P}(X = x | A_i)}_{f_{X|A_i}} \\ &= \sum_{i=1}^{\infty} \mathbb{P}(A_i) \cdot \mathbb{E}[X | A_i]. \end{aligned}$$

Das Vertauschen der beiden Reihen wiederum ist wegen der Existenz des Erwartungswertes von X gestattet. \square

Beispiel 1.57. Werfe Münze solange bis zum ersten Mal Kopf erscheint. Kopf erscheint in jedem Wurf unabhängig mit W-keit p .

$$X = \# \text{ Würfe}$$

$$\mathbb{P}(X = k) = (1 - p)^{k-1} \cdot p$$

$$\begin{aligned} \mathbb{E}[X] &= \sum_{k=1}^{\infty} k \cdot \underbrace{(1-p)^{k-1}}_{:=q} \cdot p = p \cdot \frac{d}{dq} \left(\sum_{k=0}^{\infty} q^k - 1 \right) \\ &= p \cdot \frac{d}{dq} \left(\frac{1}{1-q} \right) = p \cdot \frac{1}{(1-q)^2} = \frac{1}{p}. \end{aligned}$$

Alternativ:

$$K_1 = \{\text{Im ersten Wurf fällt Kopf}\}$$

$$\mathbb{E}[X | K_1] = 1$$

$$\mathbb{E}[X] = \mathbb{E}[X \mid K_1] \cdot \mathbb{P}(K_1) + \mathbb{E}[X \mid K_1^c] \cdot \mathbb{P}(K_1^c) = 1 \cdot p + (1 + \mathbb{E}[X]) \cdot (1 - p).$$

$$\text{Löse nach } \mathbb{E}[X] \implies \mathbb{E}[X] = \frac{1}{p}$$

Das folgende Beispiel illustriert, dass neben dem Erwartungswert noch andere Kenngrößen zur Beschreibung von Zufallsvariablen notwendig sind.

Beispiel 1.58.

Richte Datenbanksystem ein für Mehrbenutzerbetrieb. Es sei für $i = 1, 2$:

$$G'_i = 100 + 50 \cdot G_i = \text{Anzahl der pro Minute bearbeiteten Anfragen unter Strategie } i$$

$$\mathbb{P}(G_1 = \pm 1) = \frac{1}{2}, \quad \mathbb{P}(G_2 = 35) = \frac{1}{36}, \quad \mathbb{P}(G_2 = -1) = \frac{35}{36}$$

$$\mathbb{E}[G_1] = 1 \cdot \frac{1}{2} + (-1) \cdot \frac{1}{2} = 0$$

$$\mathbb{E}[G_2] = 35 \cdot \frac{1}{36} + (-1) \cdot \frac{35}{36} = 0$$

$$\mathbb{E}[G'_1] = \mathbb{E}[100 + 50 \cdot G_1] = \mathbb{E}[100] + 50 \cdot \mathbb{E}[G_1] = 100$$

$$\mathbb{E}[G'_2] = 100$$

Der Erwartungswert ist also keine Kenngröße, die zwischen den Strategien 1 und 2 unterscheiden kann. Dennoch ist bereits anschaulich klar, dass beide Strategien sich stark unterscheiden. Bei Strategie 1 ist die Auslastung zeitlich relativ homogen, wohingegen bei Strategie 2 mit kleiner Wahrscheinlichkeit sehr hohe Auslastung auftritt.

Definition 1.59 (Varianz, Standardabweichung).

Für eine Zufallsvariable X auf dem diskreten W-Raum $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$, mit existierendem Erwartungswert $\mu := \mathbb{E}[X]$ definiere die **Varianz** von X durch

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \sum_{x \in W_X} (x - \mu)^2 \cdot \mathbb{P}(\{X = x\}).$$

Die Größe $\sigma(X) = \sqrt{\text{Var}(X)}$ heißt **Standardabweichung** von X .

Nota bene. Ist der W-Raum $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ nicht endlich, so ist $\text{Var}(X) = \infty$ möglich.

Satz 1.60. Es sei X eine diskrete Zufallsvariable auf $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ mit endlicher Varianz.

Dann gilt:

$$0 \leq \text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

BEWEIS.

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

□

Satz 1.61. Es sei X eine diskrete Zufallsvariable auf $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ mit existierendem Erwartungswert und endlicher Varianz und es seien $a, b \in \mathbb{R}$. Dann gilt:

$$\boxed{\text{Var}(a \cdot X + b) = a^2 \cdot \text{Var}(X).}$$

BEWEIS.

$$\begin{aligned} \text{Var}(a \cdot X + b) &= \mathbb{E}[(a \cdot X + b - \underbrace{\mathbb{E}[aX + b]}_{a\mathbb{E}[X] + b})^2] \\ &= \mathbb{E}[(\underbrace{aX - a\mathbb{E}[X]}_{a \cdot (X - \mathbb{E}[X])})^2] = \mathbb{E}[a^2 \cdot (X - \mathbb{E}[X])^2] \\ &= a^2 \cdot \mathbb{E}[(X - \mathbb{E}[X])^2] = a^2 \cdot \text{Var}(X). \end{aligned} \quad \square$$

Wir werden an den folgenden Beispielen die Bedeutung der Varianz weiter illustrieren.

Beispiel 1.62 (Varianz der Gleichverteilung auf $\{1, \dots, N\}$).

$\Omega = \{1, \dots, N\}$, Y sei eine Zufallsvariable auf Ω mit $\mathbb{P}(\{Y = i\}) = \frac{1}{N}$ für alle $i \in \Omega$. Dann gilt:

$$\begin{aligned} \mathbb{E}[Y] &= \sum_{i=1}^N i \cdot \mathbb{P}(Y = i) = \sum_{i=1}^N \left(i \cdot \frac{1}{N}\right) = \frac{1}{N} \cdot \sum_{i=1}^N i = \frac{1}{N} \cdot \frac{N(N+1)}{2} = \frac{N+1}{2} \\ \mathbb{E}[Y^2] &= \sum_{i=1}^N i^2 \cdot \mathbb{P}(Y = i) = \sum_{i=1}^N \left(i^2 \cdot \frac{1}{N}\right) = \frac{1}{N} \cdot \sum_{i=1}^N i^2 = \frac{1}{N} \cdot \frac{N(N+1)(2N+1)}{6} \\ &= \frac{(N+1)(2N+1)}{6} \end{aligned}$$

$$\implies \text{Var}(Y) = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 = \frac{N^2 - 1}{12}.$$

Beispiel 1.63. Betrachten wir statt einer gleichverteilten Zufallsvariable eine Zufallsvariable Z mit Werten in $\{0, \dots, N\}$ mit Dichte

$$\mathbb{P}(Z = k) = \binom{N}{k} \frac{1}{2^N}.$$

Wir werden noch sehen, dass dann $\mathbb{E}[Z] = N/2$ und $\text{Var}(Z) = N/4$ gilt.

Der Leser ist dazu angehalten, wie in der Vorlesung die Dichten der Zufallsvariablen Y und Z aus den gerade diskutierten Beispielen zu skizzieren und sich das unterschiedliche Verhalten der Varianz dadurch zu verdeutlichen. Für große Werte von N ist die Varianz der gleichverteilten Zufallsvariablen Y viel größer als die Varianz der Zufallsvariablen Z .

Definition 1.64. Es sei X eine diskrete Zufallsvariable und es sei $k \in \mathbb{N}$, so nennen wir $\mathbb{E}[X^k]$ das k -te Moment von X und $\mathbb{E}[(X - \mathbb{E}[X])^k]$ das k -te zentrale Moment, sofern die entsprechenden Erwartungswerte existieren.

Aus dem Verhalten der Momente lassen sich weitere Eigenschaften der zugrundeliegenden Verteilung ablesen. (Skizze in Vorlesung)

Definition 1.65 (Kovarianz).

Es sei $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ ein diskreter Wahrscheinlichkeitsraum. Seien X, Y zwei Zufallsvariablen auf Ω und es gelte $\mathbb{E}[X^2], \mathbb{E}[Y^2] < \infty$.

Dann heißt

$$\boxed{\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X]) \cdot (Y - \mathbb{E}[Y])]}$$

die **Kovarianz** von X und Y und

$$\boxed{\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y} \in [-1, 1]}$$

heißt **Korrelationskoeffizient**.

Wir starten mit der Herleitung einiger einfacher Eigenschaften der neu eingeführten Begriffe.

Satz 1.66. Es sei $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ ein diskreter Wahrscheinlichkeitsraum. Seien X, Y zwei Zufallsvariablen auf Ω und es gelte $\mathbb{E}[X^2], \mathbb{E}[Y^2] < \infty$. Dann gilt:

$$\boxed{\text{Cov}(X, Y) = \mathbb{E}[X \cdot Y] - \mathbb{E}[X] \cdot \mathbb{E}[Y].}$$

BEWEIS.

$$\begin{aligned} \text{Cov}(X, Y) &= \mathbb{E}[(X - \mathbb{E}[X]) \cdot (Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[XY - X \cdot \mathbb{E}[Y] - Y \mathbb{E}[X] + \mathbb{E}[X] \cdot \mathbb{E}[Y]] \\ &= \mathbb{E}[XY] - \mathbb{E}[X \cdot \mathbb{E}[Y]] - \mathbb{E}[Y \cdot \mathbb{E}[X]] + \mathbb{E}[\mathbb{E}[X] \cdot \mathbb{E}[Y]] \\ &= \mathbb{E}[X \cdot Y] - \mathbb{E}[X] \cdot \mathbb{E}[Y]. \end{aligned} \quad \square$$

Satz 1.67. Es sei $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ ein diskreter W-Raum und seien X_1, \dots, X_n Zufallsvariablen auf Ω mit $\mathbb{E}[X_1^2], \dots, \mathbb{E}[X_n^2] < \infty$.

$$\boxed{\text{Var}(X_1 + \dots + X_n) = \sum_{i=1}^n \text{Var}(X_i) + \sum_{i \neq j} \text{Cov}(X_i, X_j).}$$

BEWEIS. Betrachte o.B.d.A $\widetilde{X}_i = X_i - \mathbb{E}[X_i]$

$$\begin{aligned} \text{Var}(X_1 + \dots + X_n) &= \text{Var}(\widetilde{X}_1 + \dots + \widetilde{X}_n) \\ &= \mathbb{E}\left[(\widetilde{X}_1 + \dots + \widetilde{X}_n - \underbrace{\mathbb{E}[\widetilde{X}_1 + \dots + \widetilde{X}_n]}_{=0})^2\right] \\ &= \mathbb{E}\left[(\widetilde{X}_1 + \dots + \widetilde{X}_n)^2\right] \\ &= \mathbb{E}\left[\sum_{i=1}^n \widetilde{X}_i^2 + \sum_{i \neq j} \widetilde{X}_i \widetilde{X}_j\right] \\ &= \sum_{i=1}^n \underbrace{\mathbb{E}[\widetilde{X}_i^2]}_{\text{Var}(X_i)} + \sum_{i \neq j} \underbrace{\mathbb{E}[\widetilde{X}_i \widetilde{X}_j]}_{\text{Cov}(X_i, X_j)} \end{aligned} \quad \square$$

Satz 1.68 (Cauchy-Schwarz-Ungleichung). Seien X und Y zwei reellwertige Zufallsvariablen auf dem diskreten Wahrscheinlichkeitsraum Ω . Wenn $\mathbb{E}[X^2]$ und $\mathbb{E}[Y^2]$ endlich sind, dann gilt

$$|\mathbb{E}[X \cdot Y]|^2 \leq \mathbb{E}[X^2] \cdot \mathbb{E}[Y^2].$$

BEWEIS. Es seien $\alpha = \mathbb{E}[Y^2]$ und $\beta = -\mathbb{E}[XY]$. Man sieht leicht, dass aufgrund der Voraussetzungen an die Zufallsvariablen X und Y der Erwartungswert $\mathbb{E}[XY]$ wohldefiniert ist. Wir können ohne Einschränkung der Allgemeinheit annehmen, dass $\alpha > 0$ gilt. Damit erhalten wir

$$\begin{aligned} 0 &\leq \mathbb{E}[(\alpha X + \beta Y)^2] = \mathbb{E}[\alpha^2 X^2 + 2\alpha\beta XY + \beta^2 Y^2] \\ &= \alpha^2 \mathbb{E}[X^2] - 2\alpha\beta \mathbb{E}[XY] + \beta^2 \mathbb{E}[Y^2] \\ &= \alpha \mathbb{E}[X^2] \mathbb{E}[Y^2] - 2\mathbb{E}[Y^2] \mathbb{E}[XY]^2 + \mathbb{E}[Y^2] \mathbb{E}[XY]^2 \\ &= \alpha \mathbb{E}[X^2] \mathbb{E}[Y^2] - \mathbb{E}[Y^2] \mathbb{E}[XY]^2 \\ &= \alpha (\mathbb{E}[X^2] \mathbb{E}[Y^2] - \mathbb{E}[XY]^2), \end{aligned}$$

woraus die Behauptung folgt. \square

Interpretation: Für reellwertige X, Y bedeutet positive Kovarianz, dass eine Tendenz besteht, nach der $X(\omega)$ für diejenigen ω die größeren Werte annimmt, für die auch $Y(\omega)$ die größeren Werte annimmt. Dann wird nämlich häufig $X(\omega) - \mathbb{E}[X]$ das gleiche Vorzeichen haben wie $Y(\omega) - \mathbb{E}[Y]$ und damit

$$\text{Cov}(X, Y) = \sum_{\omega} (X(\omega) - \mathbb{E}[X])(Y(\omega) - \mathbb{E}[Y]) \mathbb{P}(\{\omega\})$$

positiv sein. Negative Kovarianz deutet auf die umgekehrte Tendenz hin. Eine analoge Analyse gilt für den Korrelationskoeffizienten. In den extremalen Fällen $\rho_{XY} = 1$ bzw. $\rho_{XY} = -1$ müssen X und Y einer Gleichung $Y = cX + d$ genügen. Grob gesprochen misst die Korrelation den Grad der linearen Abhängigkeit.

Definition 1.69. Sei $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ ein diskreter Wahrscheinlichkeitsraum, und seien X_1, \dots, X_n Zufallsvariablen auf Ω , so ist für $(x_1, \dots, x_n) \in W_{X_1} \times \dots \times W_{X_n}$

$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) := \mathbb{P}(\{X_1 = x_1\} \cap \dots \cap \{X_n = x_n\})$$

die **gemeinsame Dichte** der Zufallsvariablen (X_1, \dots, X_n) .

Beachte

$$\sum_{x_1 \in W_{X_1}, \dots, x_n \in W_{X_n}} \mathbb{P}(X_1 = x_1, \dots, X_n = x_n) = 1.$$

Beispiel 1.70. Aus einem Skatblatt mit 32 Karten ziehen wir zufällig eine Hand (10 Karten) und den Skat (2 Blatt)

X = Anzahl der Buben in den 10 Karten

Y = Anzahl der Buben im Skat

Dann ist

$$\mathbb{P}(X = x, Y = y) = \frac{\binom{4}{x} \cdot \binom{28}{10-x} \cdot \binom{4-y}{y} \cdot \binom{28-(10-x)}{2-y}}{\binom{32}{10} \cdot \binom{22}{2}}$$

die gemeinsame Dichte von (X, Y) .

Beispiel 1.71 (Zahlenbeispiel).

Sei $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ ein diskreter Wahrscheinlichkeitsraum. Außerdem seien $X, Y : \Omega \rightarrow \mathbb{R}$ zwei Zufallsvariablen. Die gemeinsame Verteilung sei gegeben durch folgende Tabelle

$X \setminus Y$	1	2	3	4	
0	0	0	0	$1/8$	$1/8 = \mathbb{P}(X = 0)$
1	$1/8$	$1/8$	$1/8$	0	$3/8 = \mathbb{P}(X = 1)$
2	$2/8$	$1/8$	0	0	$3/8 = \mathbb{P}(X = 2)$
3	$1/8$	0	0	0	$1/8 = \mathbb{P}(X = 3)$
	$\underbrace{1/2}_{\mathbb{P}(Y=1)}$	$\underbrace{1/4}_{\mathbb{P}(Y=2)}$	$\underbrace{1/8}_{\mathbb{P}(Y=3)}$	$\underbrace{1/8}_{\mathbb{P}(Y=4)}$	

Definition 1.72. Es seien X_1, \dots, X_n Zufallsvariablen auf einem diskreten Wahrscheinlichkeitsraum $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$. Sei weiter $\{i_1, \dots, i_k\}$ eine k -elementige Teilmenge von $\{1, \dots, n\}$. Dann heißt die gemeinsame Dichte der ZV $(X_{i_1}, \dots, X_{i_k})$ eine **k -dim. Marginal-** oder **Randverteilung**.

Im dem nächsten Beispiel werden wir insbesondere sehen, wie mit dem Konzept von gemeinsamen Verteilungen in konkreten Situationen gerechnet und umgegangen werden kann.

Beispiel 1.73 (Warum Freunde im Mittel immer mehr Freunde als man selbst haben).

Wir denken uns ein Soziales Netzwerk als Graph visualisiert: Ecken symbolisieren angemeldete Personen und Kanten stehen für eine 'Freundschaftrelation'. Weiter definieren wir:

n = Anzahl der Personen

$d(i)$ = Anzahl der Freunde von Person i

U = eine rein zufällig ausgewählte Person (nach diskreter Gleichverteilung)

F = eine rein zufällig ausgewählter Freund von U .

Wir setzen voraus, dass $d(i) \geq 1$ gilt für alle i .

Fakt: $\mathbb{E}[d(F)] \geq \mathbb{E}[d(U)]$

Beweis: Zunächst beachte, dass für alle $0 < x < \infty$

$$x + \frac{1}{x} \geq 2.$$

Schreibt man \mathcal{E} für die gerichteten Kanten (i, j) des Graphen so gelten

$$\mathbb{P}(U = i, F = j) = \frac{1}{n} \cdot \frac{1}{d(i)} \quad \text{für } (i, j) \in \mathcal{E}$$

$$\mathbb{P}(F = j) = \frac{1}{n} \sum_{i: (i, j) \in \mathcal{E}} \frac{1}{d(i)}.$$

und somit

$$\mathbb{E}[d(F)] = \sum_j \mathbb{P}(F = j) d(j) = \frac{1}{n} \sum_{(i, j) \in \mathcal{E}} \frac{1}{d(i)} d(j)$$

Da aber jedes ungeordnete Paar von Freunden $\{i, j\}$ zweimal in der Summe auftaucht, lässt sich dies zu

$$\mathbb{E}[d(F)] = \sum_j \mathbb{P}(F = j) d(j) = \frac{1}{n} \sum_{(i,j) \in \mathcal{E}} \frac{1}{2} \left(\frac{d(i)}{d(j)} + \frac{d(j)}{d(i)} \right)$$

umschreiben. Die zu Beginn des Beweises festgestellte Ungleichung liefert dann wegen

$$\mathbb{E}[d(U)] = \frac{1}{n} \sum_i d(i) = \frac{1}{n} \sum_{(i,j) \in \mathcal{E}} 1$$

die Behauptung $\mathbb{E}[d(F)] \geq \mathbb{E}[d(U)]$.

Beispiel 1.73 ist insoweit besonders interessant, da die Aussage auf sehr wenigen natürlichen Grundannahmen beruht. Der Leser möge die Aussage anhand der Anzahl seiner 'Freundes' 'Freunde' empirisch überprüfen.

Wir definieren nun das Konzept einer unabhängigen Familie von Zufallsvariablen, das auf dem entsprechenden Unabhängigkeitskonzept von Ereignissen beruht.

Definition 1.74. Es sei $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ ein diskreter W-Raum und sei $(X_i)_{i \in I}$ eine Familie von Zufallsvariablen, dann heißt die Familie $(X_i)_{i \in I}$ unabhängig, wenn für jede Wahl von Mengen $A_i \subseteq W_{X_i}$ die Familie von Ereignissen $(\{X_i \in A_i\})_{i \in I}$ unabhängig sind.

Satz 1.75. Es seien X_1, \dots, X_n Zufallsvariablen auf einem diskreten Wahrscheinlichkeitsraum $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$.

Dann sind folgenden Aussagen äquivalent:

- (i) X_1, \dots, X_n sind unabhängig.
- (ii) Für jede Wahl von $x_1 \in W_{X_1}, \dots, x_n \in W_{X_n}$ gilt:

$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^n \mathbb{P}(X_i = x_i).$$

- (iii) Für jedes $A_i \subset W_{X_i}$ ($i = 1, \dots, n$) gilt:

$$\mathbb{P}(\{X_1 \in A_1\} \cap \dots \cap \{X_n \in A_n\}) = \prod_{i=1}^n \mathbb{P}(X_i \in A_i).$$

Eine wichtige Folgerung aus der Unabhängigkeit gibt der folgende Satz.

Satz 1.76. Sei $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ ein diskreter Wahrscheinlichkeitsraum. Es seien außerdem X_1, \dots, X_n **unabhängige** Zufallsvariablen auf Ω mit existierendem Erwartungswert. Dann gilt:

$$\boxed{\mathbb{E}[X_1 X_2 \cdots X_n] = \mathbb{E}[X_1] \cdots \mathbb{E}[X_n].}$$

BEWEIS. Wir zeigen diese Aussage per Induktion: Betrachte zunächst $n = 2$.

$$\mathbb{E}[X \cdot Y] = \sum_{x \in W_X} \sum_{y \in W_Y} x \cdot y \cdot \underbrace{\mathbb{P}(X = x, Y = y)}_{\mathbb{P}(X=x) \cdot \mathbb{P}(Y=y)} = \underbrace{\sum_{x \in W_X} x \cdot \mathbb{P}(X = x)}_{\mathbb{E}[X]} \cdot \underbrace{\sum_{y \in W_Y} y \cdot \mathbb{P}(Y = y)}_{\mathbb{E}[Y]}.$$

Induktiv erhalten wir die Aussage indem wir $Y := X_1 \cdot \dots \cdot X_{n-1}$ setzen. Es sei abschließend bemerkt, dass dasselbe Argument auch die Existenz des Erwartungswertes des Produkts $X_1 \cdot \dots \cdot X_n$ liefert. \square

Der folgende Satz besagt i.w., dass Funktionen unabhängiger Zufallsvariablen unabhängig sind. Diese Invarianzeigenschaft ist von großer Bedeutung.

Satz 1.77. Es sei $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ ein diskreter W-Raum, X_1, \dots, X_n **unabhängige** diskrete Zufallsvariablen auf Ω und $f_i: D_i \rightarrow \mathbb{R}$ eine Abbildung mit $D_i \supseteq W_{X_i}$, so sind die Zufallsvariablen $f_1 \circ X_1, \dots, f_n \circ X_n$ unabhängig.

BEWEIS. Betrachte beliebige Werte z_1, \dots, z_n mit $z_i \in W_{f_i \circ X_i}$ für $i = 1, \dots, n$. Zu z_i definiere $S_i := \{x \in D_i : f_i(x) = z_i\}$. Dann gilt:

$$\begin{aligned} \mathbb{P}(f_1 \circ X_1 = z_1, \dots, f_n \circ X_n = z_n) &= \mathbb{P}(X_1 \in S_1, \dots, X_n \in S_n) = \mathbb{P}(X_1 \in S_1) \cdot \dots \cdot \mathbb{P}(X_n \in S_n) \\ &= \mathbb{P}(f_1 \circ X_1 = z_1) \cdot \dots \cdot \mathbb{P}(f_n \circ X_n = z_n). \end{aligned}$$

Und somit sind $f_1 \circ X_1, \dots, f_n \circ X_n$ unabhängig. \square

Bemerkung 1.78. Die Unabhängigkeits-Voraussetzung ist wesentlich. Ist z. B. X eine diskrete Zufallsvariable mit $\text{Var}(X) > 0$ und $Y = -X \implies$

$$\mathbb{E}[XY] = -\mathbb{E}[X^2] \neq -(\mathbb{E}[X])^2 = \mathbb{E}[X] \cdot \mathbb{E}[X],$$

denn $0 < \text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.

Beispiel 1.79 (Betrunkene Seeleute). Das folgende Beispiel demonstriert nochmals die bisher eingeführten Konzepte.

$$X_i = \begin{cases} 1, & \text{Seemann } i \text{ im eigenen Bett} \\ 0, & \text{sonst} \end{cases}$$

$$\mathbb{P}(X_i = 1) = \frac{1}{n}, \quad X = X_1 + \dots + X_n = \# \text{ Seeleute im eigenen Bett}$$

$$A_i = \{\text{Seemann } i \text{ im eigenen Bett}\}, \quad X_i = \mathbb{1}_{A_i}$$

$$\mathbb{E}[X_i \cdot X_j] = \mathbb{E}[\mathbb{1}_{A_i} \cdot \mathbb{1}_{A_j}] = \mathbb{P}(A_i \cap A_j) = \frac{1}{n(n-1)} \neq \mathbb{P}(A_i) \cdot \mathbb{P}(A_j) = \frac{1}{n^2}$$

$\implies X_i, X_j$ sind nicht unabhängig.

$$\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = 1$$

$$\begin{aligned} \mathbb{E}[X^2] &= \mathbb{E}[(X_1 + \dots + X_n)^2] = \mathbb{E}[X_1^2 + \dots + X_n^2 + \sum_{i \neq j} X_i X_j] \\ &= \sum_{i=1}^n \mathbb{E}[X_i^2] + \sum_{i \neq j} \underbrace{\mathbb{E}[X_i X_j]}_{= \frac{1}{n(n-1)}} = 2 \end{aligned}$$

$$\mathbb{E}[X_i^2] = 0 \cdot \mathbb{P}(X_i^2 = 0) + 1 \cdot \mathbb{P}(X_i^2 = 1) = 1 \cdot \mathbb{P}(X_i = 1) = \mathbb{P}(A_i) = \frac{1}{n}.$$

Bereits in Satz 1.67 haben wir uns mit der Varianz einer Summe von Zufallsvariablen beschäftigt. Sind die Zufallsvariablen *unabhängig*, so vereinfacht sich das Resultat und es zeigt sich, dass die Varianz der Summe *unabhängiger* Zufallsvariablen gleich der Summe der Varianzen ist.

Satz 1.80. Es sei $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ ein diskreter W-Raum und seien X_1, \dots, X_n **unabhängige** Zufallsvariablen. Dann gilt:

$$\boxed{\text{Var}(X_1 + \dots + X_n) = \text{Var}(X_1) + \dots + \text{Var}(X_n).}$$

BEWEIS.

$$\text{Var}(X_1 + \dots + X_n) = \sum_{i=1}^n \text{Var}(X_i) + \underbrace{\sum_{k \neq l} \text{Cov}(X_k, X_l)}_{=0}$$

Definiere $Z_k := (X_k - \mathbb{E}[X_k])$, $Z_l := (X_l - \mathbb{E}[X_l])$ unabhängig nach Satz 1.64. (*)

$$\begin{aligned} k \neq l: \quad \text{Cov}(X_k, X_l) &= \mathbb{E}[(X_k - \mathbb{E}[X_k]) \cdot (X_l - \mathbb{E}[X_l])] \\ &\stackrel{(*)}{=} \mathbb{E}[(X_k - \mathbb{E}[X_k])] \cdot \underbrace{\mathbb{E}[(X_l - \mathbb{E}[X_l])]}_{=\mathbb{E}[X_l] - \mathbb{E}[X_l] = 0} = 0 \quad \square \end{aligned}$$

Für zwei unabhängige Zufallsvariablen lässt sich die Verteilung der Summe leicht aus den Verteilungen der Zufallsvariablen berechnen. Ohne die Voraussetzung der Unabhängigkeit ist dies in der Regel nicht möglich.

Satz 1.81 (Faltungsformel). Seien $X_1, X_2: \Omega \rightarrow \mathbb{R}$ **unabhängige** Zufallsvariablen auf einem diskreten Wahrscheinlichkeitsraum $(\Omega, \mathcal{P}(\Omega), IP)$ mit Dichte f_{X_1} bzw. f_{X_2} . Dann gilt für $Z = X_1 + X_2$:

$$\boxed{f_Z(z) = \sum_{x \in W_{X_1}} f_{X_1}(x) \cdot f_{X_2}(z - x).}$$

f_Z heißt **Faltung** von f_{X_1} und f_{X_2} .

BEWEIS. Wir berechnen die Dichte von Z wie folgt:

$$\begin{aligned} f_Z(z) &= \mathbb{P}(Z = z) = \sum_{x \in W_{X_1}} \mathbb{P}(Z = z \mid X_1 = x) \cdot \mathbb{P}(X_1 = x) \\ &= \sum_{x \in W_{X_1}} \frac{\mathbb{P}(X_1 + X_2 = z, X_1 = x)}{\mathbb{P}(X_1 = x)} \cdot \mathbb{P}(X_1 = x) \\ &= \sum_{x \in W_{X_1}} \frac{\mathbb{P}(X_2 = z - x, X_1 = x)}{\mathbb{P}(X_1 = x)} \cdot \mathbb{P}(X_1 = x) \\ &\stackrel{\text{unabh.}}{=} \sum_{x \in W_{X_1}} \frac{\mathbb{P}(X_2 = z - x) \cdot \mathbb{P}(X_1 = x)}{\mathbb{P}(X_1 = x)} \cdot \mathbb{P}(X_1 = x) \\ &= \sum_{x \in W_{X_1}} \underbrace{\mathbb{P}(X_2 = z - x)}_{f_{X_2}(z-x)} \cdot \underbrace{\mathbb{P}(X_1 = x)}_{f_{X_1}(x)} \end{aligned}$$

Dies ist die Behauptung. □

Der Beweis zeigt deutlich, dass ohne die Voraussetzung der Unabhängigkeit die Formel nicht stimmt.

1.5 Wichtige diskrete Verteilungen

In diesem Abschnitt sammeln wir wichtige Klassen von diskreten Verteilungen und diskutieren anhand von Beispielen deren praktische Bedeutung.

Definition 1.82 (Bernoulli Verteilung).

Eine Zufallsvariable X mit $W_X = \{0, 1\}$ heißt **Bernoulli-verteilt** mit Parameter $p \in [0, 1]$, wenn

$$\mathbb{P}(X = 1) = p, \quad \mathbb{P}(X = 0) = 1 - p$$

Schreibe $X \sim \text{Ber}(p)$. Es gilt dann

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

und

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = (0 - p)^2 \cdot (1 - p) + (1 - p)^2 \cdot p = p^2 \cdot (1 - p) + (1 - p)^2 \cdot p = p \cdot (1 - p).$$

Definition 1.83 (Binomialverteilung).

Eine Zufallsvariable X mit Wertebereich $W_X = \{0, 1, \dots, n\}$ heißt **binomialverteilt** mit Parameter n und p , wenn für $k = 0, \dots, n$

$$\mathbb{P}(X = k) = b(k, n, p) = \binom{n}{k} \cdot p^k \cdot (1 - p)^{n-k}.$$

Schreibe auch hier $X \sim \text{Bin}(n, p)$.

Bemerkung 1.84. Es seien X_1, \dots, X_n **unabhängige** Bernoulli-verteilte Zufallsvariablen mit demselben Parameter p . Dann ist $X = \sum_{i=1}^n X_i$ binomialverteilt mit Parametern n, p . Es gilt

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

und

$$\text{Var}(X) = \text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \underbrace{\text{Var}(X_i)}_{p \cdot (1-p)} = n \cdot p \cdot (1 - p).$$

Einige wichtige Klassen von Verteilungen haben eine wichtige Invarianzeigenschaften. Addiert man z.B. zwei unabhängige binomialverteilten Zufallsvariablen mit demselben Erfolgsparameter p , dann ist die neue Zufallsvariable immer noch binomialverteilt.

Satz 1.85. Es sei X binomialverteilt mit Parameter n und p , und es sei Y eine von X unabhängige binomialverteilte Zufallsvariable mit Parameter \tilde{n} und p . Dann ist $Z = X + Y$ binomialverteilt mit Parameter $n + \tilde{n}$ und p .

BEWEIS.

$$X = \sum_{i=1}^n X_i, \quad X_i \text{ unabh. Bernoulli-verteilt, Parameter } p$$

$$Y = \sum_{i=1}^{\tilde{n}} Y_i, \quad Y_i \text{ unabh. (auch von } X_i) \text{ Bernoulli-verteilt, Parameter } p$$

$$Z = \sum_{i=1}^n X_i + \sum_{i=1}^{\tilde{n}} Y_i = \sum_{i=1}^{n+\tilde{n}} Z_i, \quad \text{wobei } Z_i \text{ unabh. Bernoulli-verteilt, Parameter } p$$

$$Z_i = \begin{cases} X_i, & i \in \{1, \dots, n\} \\ Y_i, & i \in \{n+1, \dots, n+\tilde{n}\} \end{cases}$$

Somit folgt die Behauptung. \square

Die nächste Klasse von Verteilungen besitzt keinen endlichen Wertebereich mehr und beschreibt in gewisser Weise Wartezeiten bzw. die Anzahl von Versuchen bis zum ersten Erfolg bei Bernoulliexperimenten.

Definition 1.86 (Geometrische Verteilung).

Eine Zufallsvariable X mit $W_X = \mathbb{N}$ heißt **geometrisch verteilt** mit Parameter p , wenn

$$\mathbb{P}(X = i) = p \cdot (1 - p)^{i-1}, \quad \text{für } i \in \mathbb{N}$$

Man schreibt hier auch $X \sim \text{Geo}(p)$. Wir haben

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} i \cdot \underbrace{(1-p)^{i-1}}_{:=q} \cdot p = p \cdot \frac{d}{dq} \left(\sum_{i=0}^{\infty} q^i - 1 \right) = p \cdot \frac{d}{dq} \left(\frac{1}{1-q} \right) = p \cdot \frac{1}{(1-q)^2} = \frac{1}{p}$$

und

$$\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \left(\frac{2}{p^2} - \frac{1}{p} \right) - \frac{1}{p^2} = \frac{1}{p^2} - \frac{1}{p} = \frac{1-p}{p^2}.$$

Beispiel 1.87. Eine Sonde, die am Rande eines Vulkankraters aufgestellt ist, soll einen bevorstehenden Ausbruch beobachten und dazu Messdaten aufnehmen. Sie sendet jede Sekunde ein Datenpaket an eine Empfängerstation.

Wir gehen davon aus, dass ab Beginn des Ausbruchs in jeder Sekunde eine Wahrscheinlichkeit $p_X = 0,05$ besteht, mit der die Sonde wegen zu großer Schäden ihren Betrieb einstellt.

X = Lebensdauer der Sonde in Sek.

$$\mathbb{P}(X = n) = \underbrace{(1 - p_X)^n}_{q_X} \cdot p_X \quad (X \text{ modifiziert geom. verteilt, } X+1 \text{ geom. verteilt})$$

Sei weiter Y die Anzahl der empfangenen Datenpakete, angenommen jedes Paket wird unabhängig mit W-keit $p_Y = 0,8$ empfangen.

Was ist $\mathbb{E}[Y]$?

$$\mathbb{P}(Y = t \mid X = n) = \binom{n}{t} \cdot p_Y^t \cdot (1 - p_Y)^{n-t}$$

$$\mathbb{E}[Y \mid X = n] = n \cdot p_Y$$

$$\begin{aligned} \mathbb{E}[Y] &= \sum_{i=0}^{\infty} \mathbb{E}[Y \mid X = i] \cdot \mathbb{P}(X = i) \\ &= \sum_{i=0}^{\infty} i \cdot p_Y \cdot (1 - p_X)^i \cdot p_X \\ &= 15,2 \end{aligned}$$

Beispiel 1.88 (Sammelbildproblem). Wenn es n verschiedene Paninibilder gibt, wie viele Bilder muss man im Mittel kaufen, bis man eine vollständige Sammlung hat? Wir nehmen an, dass bei jedem Kauf die Abziehbilder mit der gleichen Wahrscheinlichkeit auftreten.

X = Anzahl der Käufe bis Album voll

Phase i bezeichne die Schritte vom Erwerb des $(i-1)$ -ten Abziehbildes (ausschließlich) bis zum Erwerb des i -ten Bildes (einschließlich).

$$n = 4 : \quad \underbrace{2}_{\text{Phase 1}}, \underbrace{2, 1}_{\text{Phase 2}}, \underbrace{2, 2, 3}_{\text{Phase 3}}, \underbrace{1, 3, 2, 3, 1, 4}_{\text{Phase 4}}$$

X_i = #Käufe in Phase i ,

$$X = \sum_{i=1}^n X_i$$

Phase i wird beendet, wenn wir eines der $(n-i)+1$ Abziehbilder erhalten, die wir nicht besitzen.

X_i ist geometrisch verteilt mit Parameter $p = \frac{n-i+1}{n}$

$$\mathbb{E}[X_i] = \frac{n}{n-i+1}$$

$$\mathbb{E}[X] = \mathbb{E}[\sum_{i=1}^n X_i] = \sum_{i=1}^n \mathbb{E}[X_i] = n \cdot \sum_{i=1}^n \frac{1}{n-i+1} = n \cdot \sum_{i=1}^n \frac{1}{i} \approx n \cdot \ln(n)$$

Die erwartete Anzahl von notwendigen Käufen wächst also etwa wie $n \log n$ mit der Anzahl der verschiedenen Sammelbilder. Wir haben bereits gelernt, dass der Erwartungswert nicht immer eine gute Kenngröße für Zufallsvariablen ist. Es wäre also nun ein naheliegender nächster Schritt, mit der Varianz die Abweichung von X vom Erwartungswert zu quantifizieren.

Bemerkung 1.89. Netzwerk hat n User; wenn man durch Abhören des Netzwerkes alle Benutzer feststellen möchte und man annimmt, dass jedes Paket von einem zufälligen Nutzer stammt, so entspricht die Wartezeit zu dem Moment, wo alle Nutzer identifiziert wurden, dem Sammelbilderproblem.

Definition 1.90 (Poisson-Verteilung).

Eine Zufallsvariable X heißt **Poisson-verteilt** mit Parameter λ , wenn $W_X = \mathbb{N}_0$ und

$$\mathbb{P}(X = i) = e^{-\lambda} \cdot \frac{\lambda^i}{i!}, \quad i \in \mathbb{N}_0.$$

Schreibe auch $X \sim \text{Poi}(\lambda)$.

Beachte

$$\sum_{i=0}^{\infty} \mathbb{P}(X = i) = \sum_{i=0}^{\infty} e^{-\lambda} \cdot \frac{\lambda^i}{i!} = e^{-\lambda} \cdot \underbrace{\sum_{i=0}^{\infty} \frac{\lambda^i}{i!}}_{e^{\lambda}} = 1$$

Es gilt $\mathbb{E}[X] = \lambda$, sowie $\text{Var}(X) = \lambda$.

Satz 1.91. Sind X und Y **unabhängige** Zufallsvariablen mit $X \sim \text{Poi}(\lambda)$ und $Y \sim \text{Poi}(\mu)$, dann gilt

$$Z := X + Y \sim \text{Poi}(\lambda + \mu)$$

BEWEIS. f_Z ist die Faltung von f_X und f_Y . Es gilt dann

$$f_Z(z) = \sum_{x=0}^{\infty} f_X(x) \cdot f_Y(z-x) = \sum_{x=0}^z \frac{e^{-\lambda} \cdot \lambda^x}{x!} \cdot \frac{e^{-\mu} \cdot \mu^{z-x}}{(z-x)!} = e^{-(\lambda+\mu)} \cdot \frac{(\lambda+\mu)^z}{z!} \quad \square$$

Satz 1.92 (Poissonscher Grenzwertsatz).

Es seien X_1, \dots, X_n unabhängige Zufallsvariablen mit $\mathbb{P}(X_i = 1) = p_i$, $\mathbb{P}(X_i = 0) = 1 - p_i$, sei $S = X_1 + \dots + X_n$ und sei $\lambda = p_1 + \dots + p_n$.

Dann gilt:

$$\sum_{k=0}^n \left| \mathbb{P}(S = k) - \underbrace{e^{-\lambda} \cdot \frac{\lambda^k}{k!}}_{=\mathbb{P}(Y=k) \text{ für ZV } Y \sim \text{Poi}(\lambda)} \right| \leq 2 \cdot \sum_{i=1}^n p_i^2.$$

BEWEIS. Wir setzen $\Omega_i = \{-1, 0, 1, 2, \dots\}$, $P_i(0) = 1 - p_i$, $P_i(-1) = e^{-p_i} - (1 - p_i)$ und $P_i(k) = e^{-p_i} \frac{p_i^k}{k!}$ für $k \in \mathbb{N}$. Sei $\Omega = \Omega_1 \times \dots \times \Omega_n$ und definiere das Wahrscheinlichkeitsmaß \mathbb{P} auf Ω durch

$$\mathbb{P}(\{\omega\}) = P_1(\omega_1)P_2(\omega_2) \dots P_n(\omega_n)$$

für $\Omega \ni \omega = (\omega_1, \omega_2, \dots, \omega_n)$. Definiert man

$$X_i(\omega) = \begin{cases} 0 & \text{wenn } \omega_i = 0 \\ 1 & \text{sonst} \end{cases}$$

sowie

$$Y_i(\omega) = \begin{cases} k & \text{wenn } \omega_i \geq 1 \\ 0 & \text{sonst.} \end{cases}$$

Dann haben die X_i die geforderte Verteilung. Die Zufallsvariablen Y_i sind unabhängig und haben eine Poissonverteilung zum Parameter p_i . Es gilt weiter

$$\mathbb{P}(X_i = Y_i) = P_i(0) + P_i(1) = 1 - p_i + e^{-p_i} p_i$$

und damit ist

$$\mathbb{P}(X_i \neq Y_i) = p_i - e^{-p_i} p_i = p_i(1 - e^{-p_i}) \leq p_i^2.$$

Nach Satz 1.91 folgt die Zufallsvariable $T = Y_1 + \dots + Y_n$ einer Poissonverteilung mit Parameter λ . Wir erhalten somit

$$\begin{aligned} \sum_{k=0}^{\infty} |\mathbb{P}(S = k) - \mathbb{P}(T = k)| &\leq \sum_{k=0}^{\infty} (\mathbb{P}(S = k, T \neq k) + \mathbb{P}(S \neq k, T = k)) \\ &= 2\mathbb{P}(S \neq T) \leq 2 \sum_{i=1}^n \mathbb{P}(X_i \neq Y_i) \\ &\leq 2 \sum_{i=1}^n p_i^2. \end{aligned} \quad \square$$

Korollar 1.93.

Für $n \in \mathbb{N}$ sei $S_n \sim \text{Bin}(n, p_n)$, wobei $p_n = \frac{\lambda}{n}$. Dann gilt:

$$\mathbb{P}(S_n = k) = b(k, n, p_n) = \binom{n}{k} \cdot p_n^k \cdot (1 - p_n)^{n-k} \xrightarrow{n \rightarrow \infty} e^{-\lambda} \cdot \frac{\lambda^k}{k!}$$

BEWEIS.

1. Variante: Benutze Poissonschen-Grenzwertsatz mit $p_i = \frac{\lambda}{n}$ für alle $i \in \{1, \dots, n\}$, sowie X_1, X_2, \dots unabhängige Zufallsvariablen, mit $\mathbb{P}(X = 1) = p_n$, $\mathbb{P}(X_i = 0) = 1 - p_n$

Dann ist $S_n = X_1 + \dots + X_n \sim \text{Bin}(n, p_n)$

$$\begin{aligned} \Rightarrow \left| \mathbb{P}(S_n = k) - e^{-\lambda} \cdot \frac{\lambda^k}{k!} \right| &\leq 2 \cdot \sum_{i=1}^n p_i^2 = 2 \cdot \sum_{i=1}^n \frac{\lambda^2}{n^2} \\ &= 2 \cdot n \cdot \frac{\lambda^2}{n^2} \\ &= 2 \cdot \lambda^2 \cdot \frac{1}{n} \xrightarrow{n \rightarrow \infty} 0 \end{aligned}$$

2. Variante:

$$\begin{aligned} \mathbb{P}(S_n = k) &= b(k; n, p_n) = \binom{n}{k} \cdot p_n^k \cdot (1 - p_n)^{n-k} = \frac{n!}{k! \cdot (n-k)!} \cdot p_n^k \cdot (1 - p_n)^{n-k} \\ &= \frac{(n \cdot p_n)^k}{k!} \cdot \frac{n \cdot (n-1) \cdots (n-k+1)}{n^k} \cdot (1 - p_n)^{-k} \cdot (1 - p_n)^n \\ &= \frac{\lambda^k}{k!} \cdot \underbrace{\frac{n \cdot (n-1) \cdots (n-k+1)}{n^k}}_{\substack{1 \cdot (1 - \frac{1}{n}) \cdots (1 - \frac{k-1}{n}) \\ \rightarrow 1}} \cdot \underbrace{\left(1 - \frac{\lambda}{n}\right)^{-k}}_{\rightarrow 1, n \rightarrow \infty} \cdot \underbrace{\left(1 - \frac{\lambda}{n}\right)^n}_{\rightarrow e^{-\lambda}, n \rightarrow \infty} \quad \square \end{aligned}$$

Beispiel 1.94. Betrachte n Atome eines schwach radioaktiven Elements, wie C^{14} . Angenommen, in einem Zeitintervall fester Länge zerfällt jedes einzelne Atom unabhängig mit Wahrscheinlichkeit p . Dies entspricht einer Bernoulli-Folge.

$X_i = \{i\text{-tes Atom zerfällt in beobachtetem Zeitraum}\}$

Registriert man die Anzahl Z aller Zerfälle im Zeitintervall, so ist Z binomialverteilt mit Parameter n, p . Im Fall großer Anzahl n und kleiner Zerfallswahrscheinlichkeit p können wir Z durch eine $\text{Poi}(\lambda)$ -Zufallsvariable approximieren, wobei $\lambda = n \cdot p$

Beispiel 1.95. Ladislaus von Bortkewitsch berichtete in seinem Buch *Das Gesetz der kleinen Zahlen*, Teubner, 1898 verschiedene Datensätze, die gut zur Poissonverteilung passen.

Speziell in Abschnitt 12.4 ('Die durch Schlag eines Pferdes im preußischen Heere getöteten') werden für zwanzig Jahre (1875–1894) und zehn Armeekorps der preußischen Kavallerie, also insgesamt 200 Korpsjahre berichtet, in wievielen davon sich x Todesfälle durch Schlag eines Pferdes ereigneten.

Ergebnis x	Anzahl 'Korpsjahre'
0	109
1	65
2	22
3	3
4	1
≥ 5	0

Angenommen, die Anzahl durch Schlag eines Pferdes während eines Jahres in einem Korps getöteter Soldaten wäre $\text{Poi}(\lambda)$ mit $\lambda = 0.61$, so würden wir das Resultate x je

$200 \times \text{Poi}(\lambda)(\{x\})$ -mal erwarten.

Ergebnis x	Anzahl 'Korpsjahre'	$200 \times \text{Poi}(\lambda)(\{x\})$
0	109	108,67
1	65	66,29
2	22	20,22
3	3	4,11
4	1	0,63
≥ 5	0	0,08

Von Bortkewitsch schreibt hierzu: *Die Kongruenz der Theorie mit der Erfahrung lässt [...], wie man sieht, nichts zu wünschen übrig.*

Abschließend gehen wir auf eine weitere Verteilungsklasse ein. Die Klasse der hypergeometrischen Verteilungen treten bei der Entnahme von Stichproben ohne Zurücklegen aus einer Gesamtpopulation auf.

Definition 1.96 (hypergeometrischen Verteilungen).

Eine Zufallsvariable X heißt **hypergeometrisch** verteilt mit Parametern m , r und n wenn $W_X = \{0, \dots, n\}$ und für $k \in \{0, \dots, n\}$

$$\mathbb{P}(X = k) = \frac{\binom{r}{k} \binom{m-r}{n-k}}{\binom{m}{n}}$$

Schreibe auch $X \sim \text{Hyp}(m, r, n)$. Es gelten

$$\mathbb{E}[X] = n \frac{m}{r}, \quad \text{Var}(X) = n \frac{m}{r} \left(1 - \frac{m}{r}\right) \frac{n-r}{n-1}.$$

Beispiel 1.97. Wir betrachten eine Population S mit insgesamt m Objekten, z.B. die Kugeln in einer Urne oder die Wähler in einem Bundesland. Unter den m Objekten seien r , die eine gewisse Eigenschaft/Merkmal ausprägung besitzen (z.B. Wähler einer bestimmten Partei), und $m-r$, die diese Eigenschaft nicht besitzen. Wir wollen die Entnahme einer Zufallsstichprobe von n Objekten aus der Population beschreiben, wobei $n \leq \min(r, m-r)$ vorausgesetzt werde. Wir betrachten als Grundraum Ω

$$\Omega = \{I \subset S \mid |I| = n\},$$

oder im Urnenmodell Ω_{III} . Ω enthält dann $\binom{m}{n}$ Elemente. Gehen wir davon aus, dass alle Stichproben gleich wahrscheinlich sind, dann wählen wir als zugrundeliegende Wahrscheinlichkeitsverteilung in unserem Modell die Gleichverteilung. Für $\omega \in \Omega$ sei nun $X(\omega)$ die Anzahl der Elemente, die die Merkmalsausprägung aufweisen. Wir erhalten dann

$$\mathbb{P}(X = k) = \frac{|\{\omega \in \Omega \mid X(\omega) = k\}|}{|\Omega|} = \frac{\binom{r}{k} \binom{m-r}{n-k}}{\binom{m}{n}}.$$

1.6 Abschätzen von Wahrscheinlichkeiten

In vielen Situationen ist man damit zufrieden, die Wahrscheinlichkeiten für geeignete Ereignisse abschätzen zu können. Der folgende Satz ist ein fundamentales Hilfsmittel hierbei.

Satz 1.98 (Markov-Ungleichung).

Es sei $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ ein diskreter Wahrscheinlichkeitsraum und es sei X eine Zufallsvariable mit Werten in $[0, \infty[$.

Dann gilt für alle $t > 0$

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}.$$

BEWEIS.

$$\begin{aligned} \mathbb{E}[X] &= \sum_{x \in W_X} x \cdot \mathbb{P}(X = x) \geq \sum_{\substack{x \in W_X \\ x \geq t}} x \cdot \mathbb{P}(X = x) \\ &\geq \sum_{\substack{x \in W_X \\ x \geq t}} t \cdot \mathbb{P}(X = x) = t \cdot \sum_{\substack{x \in W_X \\ x \geq t}} \mathbb{P}(X = x) = t \cdot \mathbb{P}(X \geq t) \end{aligned} \quad \square$$

Satz 1.99 (Chebyshev-Ungleichung).

Es sei $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ ein diskreter Wahrscheinlichkeitsraum und es sei $t > 0$. Ist X eine Zufallsvariable mit endlicher Varianz, so gilt

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq \frac{\text{Var}(X)}{t^2}.$$

BEWEIS. Es gilt zunächst

$$\mathbb{P}(|X - \mathbb{E}[X]| > t) = \mathbb{P}((X - \mathbb{E}[X])^2 > t^2)$$

die Zufallsvariable $Y := (X - \mathbb{E}[X])^2$ hat nach Definition der Varianz den Erwartungswert

$$\mathbb{E}[Y] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \text{Var}(X)$$

Markov-Ungleichung liefert uns dann:

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq t) = \mathbb{P}((X - \mathbb{E}[X])^2 \geq t^2) = \mathbb{P}(Y \geq t^2) \leq \frac{\mathbb{E}[Y]}{t^2} = \frac{\text{Var}(X)}{t^2} \quad \square$$

Beispiel 1.100. Bei einer Telefon-Hotline gibt es im Durchschnitt 50 Anrufe pro Stunde, bei einer Streuung von 10. Mit welcher Wahrscheinlichkeit kommt es vor, dass es in einer Stunde mindestens 70 Anrufe gibt?

Da die Verteilung der Anzahl X der Anrufe pro Stunde nicht bekannt ist, kann

$$\mathbb{P}(X \geq 70)$$

nicht direkt berechnet werden. Allerdings kann man mit Hilfe der Markov bzw. der Chebyshev Ungleichung die gesuchte Wahrscheinlichkeit abschätzen

$$\mathbb{P}(X \geq 70) \leq \frac{50}{70} = 5/7$$

bzw.

$$\mathbb{P}(X \geq 70) = \mathbb{P}(|X - 50| \geq 20) \leq \frac{100}{20^2} = 1/4.$$

Hier liefert Chebyshev die bessere Abschätzung.

Der folgende Satz ist eine fundamentale Folgerung aus der Chebyshev-Ungleichung.

Satz 1.101 (Schwaches Gesetz der großen Zahlen).

Gegeben sei eine diskrete Zufallsvariable X mit endlicher Varianz $\text{Var}(X) < \infty$; sind X_1, X_2, \dots unabhängige Zufallsvariablen mit derselben Dichte wie X und ist $S_n := \frac{1}{n}(X_1 + \dots + X_n)$, so gilt für jedes $\delta > 0$

$$\mathbb{P}(|S_n - \underbrace{\mathbb{E}[X]}_{=\mathbb{E}[S_n]}| \geq \delta) \leq \frac{\text{Var}(X)}{\delta^2 \cdot n}.$$

Insbesondere gilt $\lim_{n \rightarrow \infty} \mathbb{P}(|S_n - \mathbb{E}[X]|) = 0$.

BEWEIS. Da die Zufallsvariablen X_i dieselbe Dichte haben wie X gilt zunächst

$$\mathbb{E}[X_i] = \mathbb{E}[X] \quad \text{sowie} \quad \mathbb{E}[S_n] = \mathbb{E}[X].$$

Weiter gilt

$$\begin{aligned} \mathbb{P}(|S_n - \mathbb{E}[S_n]| \geq \delta) &= \mathbb{P}\left(\left|\frac{1}{n} \cdot \sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right| \geq \delta\right) \\ &\stackrel{\text{Chebyshev}}{\leq} \text{Var}\left(\frac{1}{n} \cdot \sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right) \cdot \frac{1}{\delta^2} \\ &= \frac{1}{\delta^2 \cdot n^2} \cdot \text{Var}\left(\sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right) \\ &= \frac{\sum_{i=1}^n \text{Var}(X_i)}{\delta^2 \cdot n^2} = \frac{n \cdot \text{Var}(X)}{n^2 \cdot \delta^2} \\ &= \frac{\text{Var}(X)}{\delta^2 \cdot n}. \quad \square \end{aligned}$$

Die im schwachen Gesetz der großen Zahlen auftretende Art der Konvergenz ist die sogenannte stochastische Konvergenz, auch Konvergenz in Wahrscheinlichkeit genannt.

Definition 1.102. Sei Y_n eine Folge von Zufallsvariablen auf $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ und sei Y eine weitere Zufallsvariable. Man sagt, dass Y_n **stochastisch** oder auch **in Wahrscheinlichkeit** gegen Y konvergiert, wenn für jedes $\varepsilon > 0$

$$\lim_{n \rightarrow \infty} \mathbb{P}(|Y_n - Y| > \varepsilon) = 0.$$

Man schreibt dann auch $Y_n \xrightarrow{\mathbb{P}} Y$.

Bemerkung 1.103. Sind X_1, X_2, \dots unabhängige Zufallsvariablen mit gleicher Dichte und endlicher Varianz, dann konvergiert also $S_n = \frac{X_1 + \dots + X_n}{n}$ stochastisch gegen $\mathbb{E}[X_1]$. Beachte wieder, dass

$$\mathbb{E}[X_1] = \sum_{x \in W_{X_1}} x \cdot f_{X_1}(x) = \sum_{x \in W_{X_2}} x \cdot f_{X_2}(x) = \mathbb{E}[X_2] \dots$$

Bemerkung 1.104. Es sei A ein Ereignis und $X = \mathbb{1}_A$,

$$\mathbb{1}_A(w) = \begin{cases} 1, & \omega \in A \\ 0, & \omega \notin A. \end{cases}, \quad \mathbb{P}(A) = p$$

Dann ist X Bernoulli-verteilt mit Parameter p . Es ist also $\mathbb{E}[X] = p$. Betrachte für n unabhängige Versionen von X genannt X_1, \dots, X_n die folgende Zufallsvariable.

$$S_n = \frac{X_1 + \dots + X_n}{n} = \frac{\text{Anzahl der Versuche, bei denen } A \text{ eintritt}}{n = \text{Anzahl der Versuche}}$$

Dann folgt, dass

$$\mathbb{P}(|S_n - p| > \delta) \leq \frac{\text{Var}(X)}{n \cdot \delta^2} = \frac{p \cdot (1-p)}{\delta^2 \cdot n}.$$

Die Wahrscheinlichkeit, dass die **relative** Erfolgshäufigkeit sich von mehr als ein beliebig fest vorgegebener Wert von p unterscheidet, konvergiert gegen 0. Die Wahrscheinlichkeit verhält sich heuristisch gesprochen also approximativ wie der Grenzwert der **relativen** Häufigkeiten.

Beispiel 1.105.

- 1000-maliger Münzwurf einer fairen Münze, $X = \#$ von Kopf, $X \sim \text{Bin}(1000, \frac{1}{2})$

$$\mathbb{E}[X] = n \cdot p = 500, \quad \text{Var}(X) = \frac{n}{4} = 250$$

(10 % Abweichung von Erwartungswert)

$$\mathbb{P}(X \geq 500 + 50) \leq \mathbb{P}(|X - 500| \geq 50) \stackrel{\text{Cheby.}}{\leq} \frac{\text{Var}(X)}{50^2} = \frac{250}{50^2} = 0,1$$

- 10 000-maliger Münzwurf

$$\mathbb{E}[X] = n \cdot p = 5\,000, \quad \text{Var}(X) = 2\,500$$

(10 % Abweichung von Erwartungswert)

$$\mathbb{P}(X \geq 5\,500) \leq \frac{2\,500}{500^2} = 0,01$$

Die Chebyshev-Ungleichung nutzt lediglich die Endlichkeit des zweiten Moments. Stellt man weitere Forderungen an die Zufallsvariablen, so lassen sich die Abschätzungen noch weiter verbessern.

Definition 1.106. Als (wahrscheinlichkeits)erzeugende Funktion einer Zufallsvariablen X mit Werten in \mathbb{N}_0 bezeichnen wir die Potenzreihe

$$g_X(t) := \sum_{n=0}^{\infty} \mathbb{P}(X = n) t^n.$$

Da die Koeffizienten der Potenzreihe nicht-negativ sind und ihre Summe 1 ergibt, konvergiert die Reihe mindestens für alle t mit $|t| \leq 1$. Mit Methoden der Analysis lässt sich folgender Satz zeigen, hinter dem i.w. die Erkenntnis liegt, dass

$$\mathbb{P}(X = k) = \frac{1}{k!} \frac{d^k}{ds^k} \Big|_{s=0} \mathbb{E}[s^X].$$

Zu rechtfertigen ist natürlich das Vertauschen von Differentiation und der unendlichen Reihe.

Satz 1.107. Die Dichte und die Verteilung einer Zufallsvariablen X mit Werten in \mathbb{N}_0 sind durch ihre wahrscheinlichkeitserzeugende Funktion g_X eindeutig bestimmt.

Weiter gilt

Satz 1.108. Sei X eine \mathbb{N}_0 -wertige Zufallsvariable.

- i) Für $0 \leq t \leq 1$ ist g_X stetig, monoton wachsend und konvex, und es ist $g(0) = \mathbb{P}(X = 0)$ und $g(1) = 1$.
- ii) Sei $k \geq 1$ und sei $g_X^{(k)}(1-)$ der linkseitige Grenzwert der k -ten Ableitung von g_X

$$g^{(k)}(1-) = \lim_{t \rightarrow 1-} g^{(k)}(t).$$

Dann gilt

$$\mathbb{E}[X(X-1)\dots(X-k+1)] = g^{(k)}(1-).$$

Definition 1.109. Zu einer Zufallsvariablen X definiere die momentenerzeugende Funktion durch

$$m_X(t) := \mathbb{E}[e^{tX}].$$

Wir bemerken hier, dass wir $m_X(t) = \infty$ setzen für diejenige t , für die der Erwartungswert $\mathbb{E}[e^{tX}]$ nicht nach unserer Konvention definiert ist. Die momentenerzeugende Funktion ist sehr nützlich, wenn m_X in einer Umgebung der Null wohldefiniert und endlich ist. Wir werden für den Fall von unabhängigen Bernoulli-verteilten Zufallsvariablen sog. Tailabschätzungen herleiten, die weit bessere Aussagen erlauben als die mit Hilfe der Chebyshev-Ungleichung erreichbaren. Diese Abschätzung finden häufig insbesondere in der komplexitätstheoretischen Analyse von Algorithmen Verwendung.

Satz 1.110 (Chernoff-Ungleichung).

Seien X_1, \dots, X_n unabhängige Bernoulli-verteilte Zufallsvariablen mit $\mathbb{P}(X_i = 1) = p_i$ und $\mathbb{P}(X_i = 0) = 1 - p_i$. Dann gilt für $S_n = \sum_{i=1}^n X_i$, $\mu_n = \mathbb{E}[S_n] = \sum_{i=1}^n p_i$ und jedes $\delta > 0$ gilt

$$\mathbb{P}(S_n \geq (1 + \delta)\mu_n) \leq \left(\frac{e^\delta}{(1 + \delta)^{1+\delta}} \right)^{\mu_n}.$$

Beispiel 1.111. Wie in Beispiel 1.89.

10 000-maliger Münzwurf:

$$\mathbb{P}(X \geq 5500) = \mathbb{P}(X \geq (1 + 0,1) \cdot 5000) \leq \left(\frac{e^{0,1}}{(1 + 0,1)^{1+0,1}} \right)^{5000} \leq 0,308 \cdot 10^{-10}$$

Im Vergleich hierzu ergab die Chebyshev-Abschätzung den Wert 0,01

BEWEIS (CHERNOFF UNGLEICHUNG).

Wir skizzieren relativ ausführlich die grundlegenden Ideen. Für $t > 0$ gilt

$$\mathbb{P}(S_n \geq (1 + \delta)\mu_n) = \mathbb{P}(e^{t \cdot S_n} \geq e^{t \cdot (1+\delta)\mu_n})$$

und mit Hilfe der Markov-Ungleichung folgt

$$\mathbb{P}(S_n \geq (1 + \delta)\mu_n) \leq \frac{\mathbb{E}[e^{tS_n}]}{e^{t(1+\delta)\mu_n}}$$

Aufgrund der Unabhängigkeit der Zufallsvariablen X_1, \dots, X_n gilt

$$\mathbb{E}[e^{tS_n}] = \mathbb{E}[e^{\sum_{i=1}^n tX_i}] = \mathbb{E}\left[\prod_{i=1}^n e^{tX_i}\right] \stackrel{\text{unabh.}}{=} \prod_{i=1}^n \mathbb{E}[e^{tX_i}].$$

Weiter ist für $i = 1, \dots, n$:

$$\mathbb{E}[e^{tX_i}] = e^{t \cdot 1} \cdot p_i + e^{t \cdot 0} \cdot (1 - p_i) = e^t \cdot p_i + (1 - p_i) = 1 + p_i \cdot (e^t - 1)$$

und damit

$$\begin{aligned} \mathbb{P}(S_n \geq (1 + \delta)\mu_n) &\leq \frac{\prod_{i=1}^n (1 + p_i \cdot (e^t - 1))}{e^{t(1+\delta)\mu_n}} \leq \frac{\prod_{i=1}^n (\exp(p_i(e^t - 1)))}{e^{t(1+\delta)\mu_n}} \\ &= \frac{\exp(\sum_{i=1}^n p_i(e^t - 1))}{e^{t(1+\delta)\mu_n}} = \frac{\exp(\mu_n(e^t - 1))}{e^{t(1+\delta)\mu_n}} := f(t). \end{aligned}$$

Wir suchen nun ein t , so dass $f(t)$ minimal ist, nämlich $t = \ln(1 + \delta)$. Dann gilt

$$f(t) = \frac{e^{\delta\mu_n}}{(1 + \delta)^{(1+\delta)\mu_n}}.$$

und wir erhalten die Behauptung. □

Analog zum eben geführten Beweis ergibt sich

Satz 1.112 (Chernoff-Ungleichung).

Seien X_1, \dots, X_n unabhängige Bernoulli-verteilte Zufallsvariablen mit $\mathbb{P}(X_i = 1) = p_i$ und $\mathbb{P}(X_i = 0) = 1 - p_i$. Dann gilt für $S_n = \sum_{i=1}^n X_i$, $\mu_n = \mathbb{E}[S_n] = \sum_{i=1}^n p_i$ und jedes $\delta > 0$ gilt

$$\boxed{\mathbb{P}(S_n \geq (1 + \delta)\mu_n) \leq \left(\frac{e^\delta}{(1 + \delta)^{1+\delta}} \right)^{\mu_n}}.$$

Zur Herleitung einer etwas bequemen Arbeitsversion dieser Ungleichungen zeigen wir eine analytische Hilfsaussage.

Lemma 1.113. Für $0 \leq \delta < 1$ gilt

$$(1 - \delta)^{1-\delta} \geq e^{-\delta+\delta^2/2} \quad \text{und} \quad (1 + \delta)^{1+\delta} \geq e^{\delta+\delta^2/3}$$

BEWEIS. Wir betrachten

$$f(x) = (1 - x) \ln(1 - x) \quad \text{und} \quad g(x) = -x + \frac{1}{2}x^2.$$

Für $0 \leq x < 1$ gilt

$$g'(x) = x - 1 \leq -\ln(1 - x) - 1 = f'(x)$$

und ebenso

$$f(0) = 0 = g(0),$$

woraus sich die erste Behauptung ergibt. Die Herleitung der zweiten Ungleichung erfolgt analog. □

Korollar 1.114 (Arbeitsversion der Chernoff-Ungleichungen).

Seien X_1, \dots, X_n unabh. Zufallsvariablen mit $\mathbb{P}(X_i = 1) = p_i$, $\mathbb{P}(X_i = 0) = 1 - p_i$ (Bernoulli-verteilt). Dann gilt für $S_n = \sum_{i=1}^n X_i$, $\mu_n = \mathbb{E}[S_n] = \sum_{i=1}^n p_i$ und $0 < \delta \leq 1$

$$1) \mathbb{P}(S_n \geq (1 + \delta)\mu_n) \leq e^{-\mu_n \cdot \frac{\delta^2}{3}} \quad (\forall \delta \in (0, 1]).$$

$$2) \mathbb{P}(S_n \leq (1 - \delta)\mu_n) \leq e^{-\mu_n \cdot \frac{\delta^2}{2}}.$$

$$3) \mathbb{P}(|S_n - \mu_n| \geq \delta\mu_n) \leq 2 \cdot e^{-\mu_n \cdot \frac{\delta^2}{3}}.$$

$$4) \mathbb{P}(S_n \geq (1 + \delta)\mu_n) \leq \left(\frac{e}{1 + \delta}\right)^{(1 + \delta)\mu_n}.$$

$$5) \mathbb{P}(S_n \geq t) \leq 2^{-t} \text{ für } t \geq 2e\mu_n.$$

Konstruktion von Wahrscheinlichkeitsräumen aus bedingten Wahrscheinlichkeiten

Ein aus n Telexperimenten bestehendes Experiment lässt sich oft wie folgt mathematisch beschreiben. Es sei Ω_i der Ergebnisraum des i -ten Telexperiments. Die Wahrscheinlichkeit für das Eintreten von ω_1 im ersten Experiment sei mit $\mathbb{P}_1(\{\omega_1\})$ bezeichnet. Für $2 \leq i \leq n$ legen wir in Abhängigkeit der Ergebnisse der Telexperimente $1, \dots, i - 1$ die Wahrscheinlichkeit für das Eintreten von ω_i fest, indem wir die bedingte Wahrscheinlichkeit

$$\mathbb{P}_i(\{\omega_i\} \mid \omega_1, \dots, \omega_{i-1})$$

für ω_i spezifizieren, wobei $\omega_1, \dots, \omega_{i-1}$ die Ergebnisse der Stufen $1, \dots, i - 1$ sind. Dann wird auf $\Omega = \Omega_1 \times \dots \times \Omega_n$ durch

$$\mathbb{P}(\{\omega\}) = \mathbb{P}_1(\{\omega_1\})\mathbb{P}_2(\{\omega_2\} \mid \omega_1) \cdots \mathbb{P}_n(\{\omega_n\} \mid \omega_1, \dots, \omega_{n-1}) \quad (1.7)$$

für $\omega = (\omega_1, \dots, \omega_n) \in \Omega$ eine Wahrscheinlichkeitsverteilung definiert. Wir können nun die Zufallsvariablen X auf Ω durch $X_i((\omega_1, \dots, \omega_n)) := \omega_i$ definieren. Im folgenden Satz werden Eigenschaften der obigen Konstruktion formuliert und bewiesen.

Satz 1.115. i) Durch (1.7) wird auf Ω eine Wahrscheinlichkeitsverteilung definiert.

ii) Für alle $\eta_i \in \Omega_i$ ($i = 1, \dots, n$) ist

$$\mathbb{P}(\{X_1 = \eta_1\}) = \mathbb{P}_1(\eta_1)$$

und

$$\mathbb{P}(\{X_i = \eta_i \mid \{X_1 = \eta_1\} \cap \dots \cap \{X_{i-1} = \eta_{i-1}\}\}) = \mathbb{P}_i(\{\eta_i\} \mid \eta_1, \dots, \eta_{i-1})$$

iii) Eigenschaften i) und ii) charakterisieren (1.7) eindeutig.

BEWEIS. Zum Beweis von ii) berechne

$$\begin{aligned} \mathbb{P}(\{X_1 = \eta_1\} \cap \dots \cap \{X_i = \eta_i\}) &= \sum_{(\omega_{i+1}, \dots, \omega_n)} \mathbb{P}_1(\eta_1) \mathbb{P}_2(\{\eta_2\} \mid \eta_1, \dots) \mathbb{P}_i(\{\eta_i\} \mid \eta_1, \dots, \eta_{i-1}) \\ &\quad \cdot \mathbb{P}_{i+1}(\{\omega_{i+1}\} \mid \eta_1, \dots, \eta_i) \\ &\quad \cdot \mathbb{P}_{i+2}(\{\omega_{i+2}\} \mid \eta_1, \dots, \eta_i, \omega_{i+1}) \cdots \\ &\quad \cdot \mathbb{P}_n(\{\omega_n\} \mid \eta_1, \dots, \eta_i, \omega_{i+1}, \dots, \omega_{n-1}) \end{aligned}$$

Bei der Summation über $\omega_n \in \Omega_n$ und festem $\eta_1, \dots, \eta_i, \dots, \omega_{n-1}$ ist die Summe der $\mathbb{P}_n(\{\omega_n\} \mid \dots) = 1$, so dass wir den letzten Faktor der Summe streichen können. Dann fällt der vorletzte Faktor durch Summation über ω_{n-1} weg usw. Wir erhalten dann

$$= \mathbb{P}_1(\{\eta_1\}) \cdot \dots \cdot \mathbb{P}_i(\{\eta_i\} \mid \eta_1, \dots, \eta_{i-1}).$$

Im Fall $i = 1$ steht nur der erste Faktor da und die erste Aussage von ii) ist gezeigt. im Fall $i > 1$ erhalten wir die zweite Aussage von ii) durch Einsetzen in die definierende Gleichung. Weiter beachte, dass $\mathbb{P}(\{\omega\}) \geq 0$ und dass durch Summation

$$\sum_{\omega \in \Omega} \mathbb{P}(\{\omega\}) = 1$$

folgt. Die Aussage iii) ist leicht zu sehen. □

Kontinuierliche Wahrscheinlichkeitsräume

2.1 Grundlagen

Wir erinnern zunächst an einige bereits im ersten Kapitel erwähnte allgemeine Konzepte. Ziel dieses Kapitels ist ein Verständnis grundlegender Konzepte aus der Theorie der kontinuierlichen Wahrscheinlichkeitsräume, das allerdings auf eine theoretische Fundierung leider verzichten muß. An dieser Stelle sei dem Leser das Lehrbuch *Stochastik* von Hans-Otto Georgii für eine ausführlichere Darstellung empfohlen.

Erinnerung 2.1. Sei $\Omega \neq \emptyset$ und $\mathcal{A} \subset \mathcal{P}(\Omega)$, dann heißt \mathcal{A} σ -Algebra, wenn folgendes gilt

$$\text{E1) } \Omega \in \mathcal{A}$$

$$\text{E2) } E \in \mathcal{A} \implies E^c \in \mathcal{A}$$

$$\text{E3) } A_1, A_2, \dots, A_n \in \mathcal{A} \implies \bigcup_{i=1}^n A_i \in \mathcal{A}$$

Eine Abbildung $\mathbb{P}: \mathcal{A} \rightarrow [0, 1]$ heißt Wahrscheinlichkeitsmaß, wenn folgendes gilt

$$\text{W1) } \mathbb{P}(\emptyset) = 0, \mathbb{P}(\Omega) = 1$$

$$\text{W2) für paarweise disjunkte } A_1, A_2, A_3, \dots \in \mathcal{A} \text{ gilt } \mathbb{P}\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} \mathbb{P}(A_k)$$

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

Eigenschaften 2.2.

$$i) A \in \mathcal{A} \implies \mathbb{P}(A^c) = 1 - \mathbb{P}(A)$$

$$ii) A, B \in \mathcal{A}, A \subseteq B \implies \mathbb{P}(A) \leq \mathbb{P}(B)$$

$$iii) A_i \in \mathcal{A} \quad \forall i \in \mathbb{N}, A_1 \subseteq A_2 \subseteq A_3 \subseteq \dots \implies \lim_{n \rightarrow \infty} \mathbb{P}(A_n) = \mathbb{P}\left(\bigcup_{k=1}^{\infty} A_k\right)$$

$$iv) A_i \in \mathcal{A} \quad \forall i \in \mathbb{N}, A_1 \supseteq A_2 \supseteq A_3 \supseteq \dots \implies \lim_{n \rightarrow \infty} \mathbb{P}(A_n) = \mathbb{P}\left(\bigcap_{k=1}^{\infty} A_k\right)$$

Die Gültigkeit dieser Eigenschaften wurde in der Vorlesung aus den Axiomen hergeleitet.

Definition 2.3. Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum, dann heißt die Abbildung

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

eine **reellwertige Zufallsvariable**, wenn für jedes $x \in \mathbb{R} : \{X \leq x\} \in \mathcal{A}$.

Bemerkung 2.4. Im Rahmen dieser Veranstaltung nehmen wir alle Abbildungen $X: \Omega \rightarrow \mathbb{R}$ als Zufallsvariablen an.

Definition 2.5. Sei X eine reellwertige Zufallsvariable auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$. Dann nennen wir die Abbildung

$$\begin{aligned} F_X: \mathbb{R} &\rightarrow [0, 1], \\ F_X(x) &= \mathbb{P}(X \leq x) \end{aligned}$$

Verteilungsfunktion der Zufallsvariable X .

Wir sagen, dass die Zufallsvariable X die **Dichte** f_X hat, wenn

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^x f_X(t) dt.$$

Eigenschaften von F_X und f_X :

- i) $\forall t \in \mathbb{R} : f_X(t) \geq 0$.
- ii) $\int_{-\infty}^{\infty} f_X(t) dt = \mathbb{P}(X \in \mathbb{R}) = 1$.
- iii) $F_X(x) \in [0, 1]$.
- iv) F_X ist nicht fallend, d. h. $x_1 \geq x_2 \implies F_X(x_1) \geq F_X(x_2)$.
- v) $\lim_{x \rightarrow -\infty} F_X(x) = 0, \lim_{x \rightarrow \infty} F_X(x) = 1$.
- vi) F_X ist rechtsseitig stetig und hat linksseitige Grenzwerte.

Diese Eigenschaften lassen sich leicht direkt nachprüfen, die entsprechenden Argumente werden in der Vorlesung bzw. den Übungen gegeben.

Bemerkung 2.6 (Heuristik). Es sei X eine Zufallsvariable mit stetiger Dichte f_X , es gilt also

$$\mathbb{P}(X \in [c, d]) = \int_c^d f_X(t) dt$$

und damit für kleine $\delta > 0$

$$\mathbb{P}(X \in [x, x + \delta]) = \int_x^{x+\delta} f_X(t) dt \approx \delta \cdot f_X(x).$$

Wir folgern also

$$f_X(x) = \lim_{\delta \rightarrow 0+} \frac{1}{\delta} \mathbb{P}(X \in [x, x + \delta]).$$

Insbesondere gilt für jedes feste $x \in \mathbb{R}$

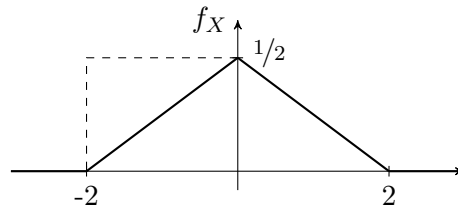
$$\mathbb{P}(X = x) = \lim_{\delta \rightarrow \infty} \mathbb{P}(X \in [x, x + \delta]) = \int_x^x f_X(t) dt = 0.$$

Es ist für stetige Zufallsvariablen mit Dichte f_X also im Gegensatz zum diskreten Fall nicht sehr sinnvoll nach der Wahrscheinlichkeit von $\{X = x\}$ zu fragen.

Beispiel 2.7.

Es sei

$$f_X(x) = \begin{cases} \frac{1}{2} + \frac{x}{4}, & -2 \leq x \leq 0 \\ \frac{1}{2} - \frac{x}{4}, & 0 < x \leq 2 \\ 0, & \text{sonst} \end{cases}$$



Dann ist f_X eine Wahrscheinlichkeitsdichte einer Zufallsvariable X . Es genügt zu wissen dass,

a) $\forall x \in \mathbb{R} : f_X(x) \geq 0$

b) $\int_{-\infty}^{\infty} f_X(t) dt = 1$

Zeige, dass $\int_{-\infty}^{\infty} f_X(t) dt = 1$ und bestimme $\mathbb{P}(X \geq \frac{1}{2})$.

$$\begin{aligned} \int_{-\infty}^{\infty} f_X(t) dt &= \underbrace{\int_{-\infty}^{-2} f_X(t) dt}_{=0} + \int_{-2}^0 f_X(t) dt + \int_0^2 f_X(t) dt + \underbrace{\int_2^{\infty} f_X(t) dt}_{=0} \\ &= \int_{-2}^0 f_X(t) dt + \int_0^2 f_X(t) dt \\ &= \int_{-2}^0 \left(\frac{1}{2} + \frac{t}{4} \right) dt + \int_0^2 \left(\frac{1}{2} - \frac{t}{4} \right) dt \\ &= \left[\frac{1}{2}t + \frac{1}{8}t^2 \right]_{-2}^0 + \left[\frac{1}{2}t - \frac{1}{8}t^2 \right]_0^2 \\ &= 0 - \left(-1 + \frac{1}{2} \right) + \left(1 - \frac{1}{2} \right) - 0 = 2 - 1 = 1 \end{aligned}$$

$$\begin{aligned} \mathbb{P}\left(X \geq \frac{1}{2}\right) &= \int_{\frac{1}{2}}^{\infty} f_X(t) dt = \int_{\frac{1}{2}}^2 f_X(t) dt = \int_{\frac{1}{2}}^2 \left(\frac{1}{2} - \frac{t}{4} \right) dt \\ &= \left[\frac{1}{2}t - \frac{1}{8}t^2 \right]_{\frac{1}{2}}^2 = \left(1 - \frac{1}{2} \right) - \left(\frac{1}{4} - \frac{1}{32} \right) = \frac{7}{32} \end{aligned}$$

Beispiel 2.8. Es sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein W-Raum und sei X eine Zufallsvariable auf Ω , dann gilt für (stetige) Funktionen $g: \mathbb{R} \rightarrow \mathbb{R}$, dass $Y = g \circ X = g(X)$ ebenfalls eine Zufallsvariable ist. Es sei beispielsweise X eine nicht-negative Zufallsvariable und $g: \mathbb{R} \rightarrow \mathbb{R}$ mit:

$$g(x) = \begin{cases} x^2, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Außerdem habe die Zufallsvariable X die Dichte f_X . Dann lässt sich mit Hilfe der Integrationsregeln aus der Analysis die Dichte von Y berechnen:

Für $y < 0$ gilt:

$$\mathbb{P}(Y \leq y) = 0$$

Für $y \geq 0$ gilt:

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y, X \geq 0) = \mathbb{P}(X \leq \sqrt{y}) = F_X(\sqrt{y}) = \int_{-\infty}^{\sqrt{y}} f_X(x) dx \\ &= \int_0^{\sqrt{y}} f_X(x) dx \stackrel{x=\sqrt{t}}{\stackrel{dx=\frac{1}{2\sqrt{t}} dt}{=}} \int_0^y \underbrace{f_X(\sqrt{t}) \cdot \frac{1}{2\sqrt{t}}}_{f_Y(t)} dt \end{aligned}$$

Hieraus lässt sich aus dem Zusammenhang von Dichte und Verteilungsfunktion herauslesen, dass

$$f_Y(t) = \begin{cases} f_X(\sqrt{t}) \cdot \frac{1}{2\sqrt{t}}, & x > 0 \\ 0, & x < 0 \end{cases}$$

die Dichte der Zufallsvariablen Y darstellt. $\implies F_Y(y) = \mathbb{P}(Y \leq y) = \int_{-\infty}^y f_Y(t) dt$

Wir erinnern an die Substitutionsregel aus der Analysis

Satz 2.9. Seien I ein reelles Intervall, $f : I \rightarrow \mathbb{R}$ eine stetige Funktion und $\varphi : [a, b] \rightarrow I$ eine stetig differenzierbare Funktion. Dann gilt

$$\int_a^b f(\varphi(t)) \varphi'(t) dt = \int_{\varphi(a)}^{\varphi(b)} f(t) dt.$$

Hieraus ergibt sich sofort

Satz 2.10. Sei X eine Zufallsvariable mit Werten in einem offenen möglicherweise unbeschränkten Intervall I und stetiger Dichte f_X , sei weiter $J \subset I$ und sei $\varphi : I \rightarrow J$ bijektiv, stetig differenzierbar mit nirgends verschwindender Ableitung φ' . Dann hat die Zufallsvariable $Y := \varphi(X)$ die Dichte

$$f_Y(y) = \begin{cases} \frac{f_X(\varphi^{-1}(y))}{|\varphi'(\varphi^{-1}(y))|} & y \in J \\ 0 & \text{sonst} \end{cases}$$

BEWEIS. Beachte, dass wegen der Voraussetzung, dass die Ableitung von φ keine Nullstelle hat, die Funktion φ streng monoton ist. Wir betrachten den strikt wachsenden Fall. Für $z < \varphi(a)$ gilt natürlich $\mathbb{P}(Y \leq z) = 0$ und für $z > \varphi(b)$ dann $\mathbb{P}(Y \geq z) = 1$. Für $z \in [\varphi(a), \varphi(b)]$ folgern wir

$$\begin{aligned} \mathbb{P}(Y \leq z) &= \mathbb{P}(\varphi(X) \leq z) = \mathbb{P}(X \leq \varphi^{-1}(z)) \\ &= \int_{-\infty}^{\varphi^{-1}(z)} f_X(x) dx = \int_{-\infty}^z f_X(\varphi^{-1}(y)) \frac{1}{|\varphi'(\varphi^{-1}(y))|} dy. \end{aligned}$$

Hieraus ergibt sich die Behauptung. □

Man erhält aus der Substitutionsregel der Analysis folgenden Satz als Spezialfall

Satz 2.11. Ist $Y = X + a$ und hat die Zufallsvariable X die Dichte f , so hat Y die Dichte

$$g(y) = f(y - a).$$

Ist $Y = cX$ mit $c \neq 0$ und hat die Zufallsvariable X die Dichte f , so hat Y die Dichte

$$g(y) = \frac{1}{|c|} f\left(\frac{y}{c}\right).$$

BEWEIS. *Übung*

□

Bemerkung 2.12 (Einführende Analysis des Begriffs des Erwartungswerts). Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein W-Raum und sei X eine Zufallsvariable mit abzählbarem Wertebereich mit möglichen Werten $\alpha_1, \alpha_2, \dots$. Dann ist der Erwartungswert von X definiert durch

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} \alpha_i \cdot \mathbb{P}(X = i).$$

Wiederrum sagen wir, dass der Erwartungswert existiert, wenn diese Reihe absolut konvergiert. Ist nun \mathcal{D} die Klasse der Zufallsvariablen mit nur abzählbar vielen Werten, deren Erwartungswert existiert. Dann gelten in \mathcal{D} die alten Rechenregeln.

Für allgemeine reellwertige Zufallsvariablen X definieren wir den Erwartungswert mittels einer Approximation. Für $k \in \mathbb{Z}$ und $n \in \mathbb{N}$ sei

$$A_{n,k} = \{k/n \leq X < (k+1)/n\}$$

und

$$X_n = \sum_{k=-\infty}^{\infty} \frac{k}{n} \mathbf{1}_{A_{n,k}}.$$

Dann ist $X_n \leq X < X_n + 1/n$ und $|X_n - X_m| \leq 1/n + 1/m$. Existiert daher $\mathbb{E}[X_n]$ für ein n , so existiert $\mathbb{E}[X_n]$ für alle n , und es gilt $|\mathbb{E}[X_n] - \mathbb{E}[X_m]| \leq 1/n + 1/m$. Wir sagen dass $\mathbb{E}[X]$ existiert, wenn $\mathbb{E}[X_1]$ existiert, und setzen

$$\mathbb{E}[X] = \lim_{n \rightarrow \infty} \mathbb{E}[X_n].$$

Im Fall von Zufallsvariablen X mit einer stetigen Dichte ergibt sich mit Hilfe der Bemerkung 2.12 die folgende Definition.

Definition 2.13. Es sei X eine Zufallsvariable mit Dichte f_X . Dann ist der Erwartungswert von X definiert durch

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} t \cdot f_X(t) dt,$$

sofern das Integral $\int_{-\infty}^{\infty} |t| \cdot f_X(t) dt < \infty$ absolut konvergiert.

Sei X eine Zufallsvariable mit Dichte f_X , so dass $\mathbb{E}[X]$ existiert, dann heißt

$$\text{Var}(X) = \int_{-\infty}^{\infty} (t - \mathbb{E}[X])^2 \cdot f_X(t) dt$$

Varianz der Zufallsvariable X .

Satz 2.14. Es sei X eine reellwertige Zufallsvariable mit einer stückweise stetigen Dichte f_X . Sei $g: \mathbb{R} \rightarrow \mathbb{R}$ eine [stetige] Abbildung. Dann ist die Komposition $g \circ X = g(X)$ eine Zufallsvariable und es gilt

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(t) \cdot f_X(t) dt,$$

sofern das Integral $\int_{-\infty}^{\infty} |g(t)| \cdot f_X(t) dt < \infty$.

Bemerkung 2.15. Die Varianz ist definiert durch $\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$. Es sei $g: \mathbb{R} \rightarrow \mathbb{R}$ eine Funktion mit $g(t) = (t - \mathbb{E}[X])^2$.

Dann gilt

$$\text{Var}(X) = \mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(t) \cdot f_X(t) dt = \int_{-\infty}^{\infty} (t - \mathbb{E}[X])^2 \cdot f_X(t) dt.$$

2.2 Wichtige kontinuierliche Verteilungen

Definition 2.16 (Gleichverteilung).

Eine Zufallsvariable X mit Werten in \mathbb{R} heißt **gleichverteilt** auf $[a, b]$, $a < b$, wenn X die Dichte

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

besitzt. Es ist zu beachten, dass

$$f_X(t) \geq 0 \quad \forall t \in \mathbb{R}$$

sowie

$$\int_{-\infty}^{\infty} f_X(t) dt = \int_a^b f_X(t) dt = \int_a^b \frac{1}{b-a} dt = \frac{b-a}{b-a} = 1.$$

Die Zufallsvariable X sei gleichverteilt auf $[0, 1]$. Angenommen, durch Zusatzinformation ist bekannt, dass X einen Wert kleiner als $1/2$ realisiert hat, dann ist die auf diese Information bedingte Verteilung gleichverteilt auf $[1/2]$. Wir formalisieren dies in folgenden Theorem.

Satz 2.17. Es sei X eine auf $[a, b]$ ($a < b$) gleichverteilte Zufallsvariable und sei $a \leq c \leq d \leq b$. Dann gilt

$$\mathbb{P}(X \leq c \mid X \leq d) = \frac{c-a}{d-a}.$$

BEWEIS.

$$\begin{aligned} \mathbb{P}(X \leq c \mid X \leq d) &= \frac{\mathbb{P}(\{X \leq c\} \cap \{X \leq d\})}{\mathbb{P}(\{X \leq d\})} \\ &= \frac{\mathbb{P}(\{X \leq c\})}{\mathbb{P}(\{X \leq d\})} \\ &= \frac{c-a}{d-a}. \end{aligned}$$

Folglich ist die auf $X \leq d$ bedingte Verteilung von X gleich verteilt auf $[a, d]$. □

Die Klasse der exponentialverteilten Zufallsvariablen ist das stetige Analogon zur Klasse der geometrisch verteilten Zufallsvariablen. Insbesondere werden Exponentialverteilungen häufig zur Beschreibung von Wartezeiten benutzt.

Definition 2.18 (Exponentialverteilung).

Eine Zufallsvariable X heißt **exponentialverteilt** mit Parameter λ , wenn X nicht-negativ ist und die Dichte

$$f_X(x) = \begin{cases} \lambda \cdot e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

besitzt. Für die zugehörige Verteilungsfunktion gilt, falls $x \geq 0$

$$F_X(x) = \int_{-\infty}^x f_X(t) dt = \int_0^x f_X(t) dt = \int_0^x \lambda \cdot e^{-\lambda t} dt = \left[-e^{-\lambda t} \right]_0^x = 1 - e^{-\lambda x}$$

(für alle $x < 0$ ist $F_X(x) = 0$).

Satz 2.19. Es sei X eine exponentialverteilte Zufallsvariable mit Parameter $\lambda > 0$. Dann gilt

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

BEWEIS. Es gilt:

$$\begin{aligned} \mathbb{E}[X] &= \int_0^\infty t \cdot f_X(t) dt = \int_0^\infty t \cdot \lambda \cdot e^{-\lambda t} dt = - \int_0^\infty t \cdot \frac{d}{dt} e^{-\lambda t} dt \\ &\stackrel{\text{part.}}{\stackrel{\text{Int.}}{=}} \left[-(e^{-\lambda t}) \cdot t \right]_0^\infty + \int_0^\infty e^{-\lambda t} dt = \int_0^\infty e^{-\lambda t} dt = \frac{1}{\lambda}. \end{aligned}$$

Hier haben wir im vorletzten Schritt benutzt, dass

$$\lim_{t \rightarrow \infty} t \cdot e^{-t} = 0$$

gilt. Weiter gilt $\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$, also genügt es $\mathbb{E}[X^2]$ zu berechnen

$$\begin{aligned} \mathbb{E}[X^2] &= \int_0^\infty t^2 \cdot f_X(t) dt = \int_0^\infty t^2 \cdot \lambda \cdot e^{-\lambda t} dt = - \int_0^\infty t^2 \frac{d}{dt} e^{-\lambda t} dt \\ &= \left[t^2 \cdot (-e^{-\lambda t}) \right]_0^\infty + 2 \cdot \int_0^\infty t \cdot e^{-\lambda t} dt = \frac{2}{\lambda} \cdot \underbrace{\int_0^\infty t \cdot \lambda \cdot e^{-\lambda t} dt}_{\mathbb{E}[X]} = \frac{2}{\lambda^2} \end{aligned}$$

$$\implies \text{Var}(X) = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2} \quad \square$$

Satz 2.20. Es sei X eine exponentialverteilte Zufallsvariable mit Parameter $\lambda > 0$.

Dann gilt für $t, s > 0$

$$\boxed{\mathbb{P}(X > t + s \mid X > s) = \mathbb{P}(X > t).}$$

BEWEIS. Für $t, s > 0$ gilt wegen $\{X > t + s\} \subset \{X > s\}$ und der Definition der elementaren bedingten Wahrscheinlichkeit

$$\begin{aligned} \mathbb{P}(X > t + s \mid X > s) &= \frac{\mathbb{P}(X > t + s)}{\mathbb{P}(X > s)} = \frac{\int_{t+s}^\infty \lambda e^{-\lambda r} dr}{\int_s^\infty \lambda e^{-\lambda r} dr} \\ &= \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} = e^{-\lambda t} = \mathbb{P}(X > t). \end{aligned} \quad \square$$

Bemerkung 2.21. Diese Eigenschaft – **Gedächtnislosigkeit** genannt – charakterisiert die Exponentialverteilung. In der Praxis wird die Exponentialverteilung deshalb bei der Beschreibung von Wartezeiten auf Ereignisse benutzt, wenn kein 'Alterungsprozess' des zugrundeliegenden Systems zu erwarten ist.

Beispiel 2.22. Radioaktive Materialien altern nicht; Ausfall von Glühlampen lassen sich approximativ durch Exponentialverteilungen modellieren ...

Die in der folgenden Definition eingeführte Klasse von Verteilungen spielt aufgrund des noch zu diskutierenden Zentralen Grenzwertsatzes eine besonders wichtige Rolle.

Definition 2.23 (Normalverteilung).

Eine Zufallsvariable X mit Werten in \mathbb{R} heißt **normalverteilt** mit Parameter $\mu \in \mathbb{R}$ und $\sigma \in \mathbb{R}_+$, wenn X die Dichte

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{(x-\mu)^2}{2\sigma^2}} =: \varphi(x; \mu, \sigma), \quad \varphi(x) := \varphi(x; 0, 1)$$

besitzt. Man sagt

$X \sim \mathcal{N}(\mu, \sigma^2) \iff X$ ist normalverteilt mit Parameter μ und σ .

$X \sim \mathcal{N}(0, 1) \iff X$ ist standardnormalverteilt.

Die Verteilungsfunktion

$$F_X(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt =: \Phi(x; \mu, \sigma), \quad \Phi(x) := \Phi(x; 0, 1)$$

einer normalverteilten Zufallsvariable ist nicht explizit berechenbar, d.h. die Werte von F_X müssen numerisch bestimmt werden oder in Tabellen (\rightarrow A) nachgeschlagen werden.

Um die Normierung der Standardnormalverteilung einzusehen, beachte man zunächst, dass

$$I^2 := \left(\int_{\mathbb{R}} e^{-\frac{x^2}{2}} dx \right)^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{x^2+y^2}{2}} dx dy.$$

Übergang zu Polarkoordinaten liefert dann

$$I^2 = \int_0^{2\pi} \int_0^{\infty} e^{-r^2/2} r dr d\varphi = 2\pi.$$

Die korrekte Normierung der Dichte der allgemeinen Normalverteilung lässt sich mit Hilfe eines noch zu beweisenden Satzes hieraus ableiten. Zur Bedeutung der Parameter zeigen wir den folgenden Satz.

Satz 2.24. Sei X eine standardnormalverteilte Zufallsvariable, d.h. $X \sim \mathcal{N}(0, 1)$.

Dann gilt

$$\mathbb{E}[X] = 0, \quad \text{Var}(X) = 1.$$

BEWEIS. Zunächst weisen wir nach, dass der Erwartungswert existiert:

$$\begin{aligned} \int_{-\infty}^{\infty} |t| \cdot f_X(t) dt &= \int_{-\infty}^{\infty} |t| \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{t^2}{2}} dt \\ &= 2 \cdot \int_0^{\infty} t \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{t^2}{2}} dt = -2 \cdot \frac{1}{\sqrt{2\pi}} \cdot \int_0^{\infty} \frac{d}{dt} e^{-\frac{t^2}{2}} dt \\ &= -2 \cdot \frac{1}{\sqrt{2\pi}} \cdot \left[e^{-\frac{t^2}{2}} \right]_0^{\infty} = 2 \cdot \frac{1}{\sqrt{2\pi}} < \infty. \end{aligned}$$

Es gilt:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} t \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{t^2}{2}} dt = \int_{-\infty}^0 t \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{t^2}{2}} dt + \int_0^{\infty} t \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{t^2}{2}} dt = 0.$$

Weiter gilt:

$$\begin{aligned}
 1 &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}} dt, \quad f_X \text{ ist W-Dichte} \\
 \sqrt{2\pi} &= \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}} dt = \int_{-\infty}^{\infty} \left(\frac{d}{dt} t \right) \cdot e^{-\frac{t^2}{2}} dt \stackrel{\text{part. Int.}}{=} \underbrace{\left[t \cdot e^{-\frac{t^2}{2}} \right]_{-\infty}^{\infty}}_{=0} + \int_{-\infty}^{\infty} t^2 \cdot e^{-\frac{t^2}{2}} dt \\
 1 &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t^2 \cdot e^{-\frac{t^2}{2}} dt = \mathbb{E}[X^2]
 \end{aligned}$$

$$\implies \text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = 1. \quad \square$$

Satz 2.25. Sei X normalverteilt mit $X \sim \mathcal{N}(\mu, \sigma^2)$. Dann gilt für beliebige $a \in \mathbb{R} \setminus \{0\}$ und $b \in \mathbb{R}$, dass

$$Y := a \cdot X + b$$

normalverteilt ist mit $Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.

BEWEIS. Untersuche den Fall $a > 0$. Es gilt

$$\begin{aligned}
 \mathbb{P}(Y \leq y) &= \mathbb{P}(aX + b \leq y) = \mathbb{P}\left(X \leq \frac{y-b}{a}\right) \\
 &= \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \int_{-\infty}^{(y-b)/a} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt.
 \end{aligned}$$

Substituiere wie folgt: $t = \frac{r-b}{a}$, $dt = \frac{1}{a} dr$. Damit erhalten wir

$$\mathbb{P}(Y \leq y) = \frac{1}{\sqrt{2\pi a^2 \sigma^2}} \cdot \int_{-\infty}^y e^{-\frac{(r-a\mu-b)^2}{2a^2\sigma^2}} dr.$$

Also gilt wie behauptet $Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$. Der Fall $a < 0$ verläuft analog. \square

Bemerkung 2.26. Der obige Satz (2.20) ermöglicht es, beliebige $\mathcal{N}(\mu, \sigma^2)$ -verteilte Zufallsvariablen X durch die folgende Transformation zu standardisieren

$$X \sim \mathcal{N}(\mu, \sigma^2) \implies Y := \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1).$$

2.3 Stochastische Unabhängigkeit

Wir gehen in direkter Analogie zum diskreten Fall vor und halten die Darstellung deshalb und aufgrund zeitlicher Einschränkungen auch sehr kurz und knapp.

Definition 2.27. Es sei $(X_n)_{n \in \mathbb{N}}$ eine Folge von Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ mit Dichte f_{X_1}, f_{X_2}, \dots . Dann heißen $(X_k)_{k \in \mathbb{N}}$ **unabhängig**, wenn für jede endliche Teilmenge $J \subseteq \mathbb{N}$ und jede Wahl von $x_j \in \mathbb{R}$ mit $j \in J$ gilt

$$\mathbb{P}\left(\bigcap_{j \in J} \{X_j \leq x_j\}\right) = \prod_{j \in J} \mathbb{P}(X_j \leq x_j).$$

Satz 2.28 (Additivität der Normalverteilung).

Die Zufallsvariablen X_1, \dots, X_n seien **unabhängig** und normalverteilt mit Parametern μ_i und σ_i für $1 \leq i \leq n$. Es gilt: Die Zufallsvariable

$$Z := a_1 X_1 + \dots + a_n X_n$$

ist normalverteilt mit Erwartungswert $\mu = a_1\mu_1 + \dots + a_n\mu_n$ und Varianz $\sigma^2 = a_1^2\sigma_1^2 + \dots + a_n^2\sigma_n^2$.

Satz 2.29. Sind X_1, \dots, X_n **unabhängige** Zufallsvariablen mit Dichten f_{X_1}, \dots, f_{X_n} , so gilt

$$F_{(X_1, \dots, X_n)}(x_1, \dots, x_n) = \mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n) = \int_{-\infty}^{x_1} f_{X_1}(t_1) dt_1 \cdots \int_{-\infty}^{x_n} f_{X_n}(t_n) dt_n \quad (*)$$

und für jede (stetige) Funktion $g: \mathbb{R} \rightarrow \mathbb{R}$ gilt

$$\mathbb{E}[g(X_1, \dots, X_n)] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(t_1, \dots, t_n) \cdot f_{X_1}(t_1) \cdots f_{X_n}(t_n) dt_1 \dots dt_n$$

sofern

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} |g(t_1, \dots, t_n)| \cdot f_{X_1}(t_1) \cdots f_{X_n}(t_n) dt_1 \dots dt_n < \infty.$$

Sind andererseits X_1, \dots, X_n Zufallsvariablen mit Dichten f_{X_1}, \dots, f_{X_n} und gilt (*) für alle $x_1, \dots, x_n \in \mathbb{R}$, so sind X_1, \dots, X_n unabhängig.

2.4 Gemeinsame Verteilung

Wir greifen eine Formel aus der Definition der Unabhängigkeit von n Zufallsvariablen auf.

Definition 2.30.

Es seien X_1, \dots, X_n stetige Zufallsvariablen, dann heißt die Funktion

$$\mathbb{R}^n \ni (x_1, \dots, x_n) \mapsto F_{(X_1, \dots, X_n)}(x_1, \dots, x_n) = \mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n)$$

gemeinsame Verteilungsfunktion von X_1, \dots, X_n .

Existiert ein $f_{(X_1, \dots, X_n)}: \mathbb{R} \rightarrow [0, \infty)$, so dass

$$F_{(X_1, \dots, X_n)}(x_1, \dots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_{(X_1, \dots, X_n)}(t_1, \dots, t_n) dt_n \dots dt_1$$

gilt für alle $x_1, \dots, x_n \in \mathbb{R}$, dann heißt $f_{(X_1, \dots, X_n)}$ die **gemeinsame Dichte** von (X_1, \dots, X_n) .

Bemerkung 2.31. Die Zufallsvariablen X_1, \dots, X_n auf $(\Omega, \mathcal{A}, \mathbb{P})$ mit Dichten f_{X_1}, \dots, f_{X_n} sind genau dann unabhängig, wenn die gemeinsame Dichte

$$f_{(X_1, \dots, X_n)}(x_1, \dots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n)$$

erfüllt. Die Zufallsvariablen sind genau dann unabhängig, wenn also

$$F_{(X_1, \dots, X_n)}(x_1, \dots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_{X_1}(t_1) \cdots f_{X_n}(t_n) dt_n \dots dt_1.$$

Bei **unabhängigen** Zufallsvariablen ist die gemeinsame Verteilung also das Produkt der einzelnen Verteilungen.

Definition 2.32. Sind X_1, \dots, X_n Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{S}, \mathbb{P})$ und ist $\{i_1, \dots, i_k\} \subset \{1, \dots, n\}$ eine k -elementige Teilmenge, dann heißt die Funktion

$$\mathbb{R}^k \ni (x_{i_1}, \dots, x_{i_k}) \mapsto \mathbb{P}(X_{i_1} \leq x_{i_1}, \dots, X_{i_k} \leq x_{i_k}) = F_{(X_{i_1}, \dots, X_{i_k})}(x_{i_1}, \dots, x_{i_k})$$

eine k -dimensionale Randverteilung.

Beispiel 2.33. (X_1, X_2) haben die gemeinsame Dichte $f_{(X_1, X_2)}$, dann sind die 1-dimensionalen Randverteilungen

$$\mathbb{P}(X_1 \leq x_1) = \mathbb{P}(X_1 \leq x_1, X_2 < \infty) = \int_{-\infty}^{x_1} \int_{-\infty}^{\infty} f_{(X_1, X_2)}(t_1, t_2) dt_2 dt_1$$

$$\mathbb{P}(X_2 \leq x_2) = \mathbb{P}(X_1 < \infty, X_2 \leq x_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{x_2} f_{(X_1, X_2)}(t_1, t_2) dt_2 dt_1$$

$$\Rightarrow \mathbb{P}(X_1 \leq x_1) = \int_{-\infty}^{x_1} g_{X_1}(t_1) dt_1, \text{ wobei } g_{X_1}(t_1) = \int_{-\infty}^{\infty} f_{(X_1, X_2)}(t_1, t_2) dt_2$$

$\Rightarrow g_{X_1}$ ist Dichte von X_1 .

Die folgende Aussage demonstriert nochmals gut, auf welche Art und Weise die Unabhängigkeitsvoraussetzung im kontinuierlichen Kontext konkret ausgenutzt werden kann.

Satz 2.34. Seien X_1, X_2, \dots, X_n unabhängige exponentialverteilte Zufallsvariablen mit Parametern $\lambda_1, \lambda_2, \dots, \lambda_n$. Dann ist die Zufallsvariable $Z := \min(X_1, X_2, \dots, X_n)$ exponentialverteilt mit Parameter $\sum_{i=1}^n \lambda_i$ und es gilt für jedes $1 \leq i \leq n$

$$\mathbb{P}(\min(X_1, X_2, \dots, X_n) = X_i) = \frac{\lambda_i}{\sum_{j=1}^n \lambda_j}.$$

BEWEIS. Die erste Behauptung des Satzes überlassen wir als Übungsaufgabe. Zum Beweis der zweiten können wir ohne Einschränkung annehmen, dass $n = 2$ ist. Die gemeinsame Verteilung der Zufallsvariablen X_1 und X_2 ist wegen der vorausgesetzten Unabhängigkeit gegeben durch die Dichte

$$f(x_1, x_2) = \lambda_1 e^{-\lambda_1 x_1} \lambda_2 e^{-\lambda_2 x_2}.$$

Es gilt also

$$\begin{aligned} \mathbb{P}(X_1 < X_2) &= \int_0^{\infty} \int_0^{x_2} f(x_1, x_2) dx_1 dx_2 \\ &= \int_0^{\infty} \lambda_2 e^{-\lambda_2 x_2} \int_0^{x_2} \lambda_1 e^{-\lambda_1 x_1} dx_1 dx_2 \\ &= \int_0^{\infty} \lambda_2 e^{-\lambda_2 x_2} (1 - e^{-\lambda_1 x_2}) dx_2 \\ &= \int_0^{\infty} (\lambda_2 e^{-\lambda_2 x_2} - \lambda_2 e^{-(\lambda_1 + \lambda_2)x_2}) dx_2 \\ &= 1 - \frac{\lambda_2}{\lambda_1 + \lambda_2} = \frac{\lambda_1}{\lambda_1 + \lambda_2}. \end{aligned}$$

□

2.5 Wichtige Rechenregeln

Man wird zurecht erwarten, dass mit Erwartungswerten analog zum diskreten Fall gerechnet werden kann.

Satz 2.35 (Rechenregeln für Erwartungswerte).

- i) Es seien X_1, \dots, X_n Zufallsvariablen auf einem W-Raum $(\Omega, \mathcal{A}, \mathbb{P})$ mit Dichten f_{X_1}, \dots, f_{X_n} und seien $\lambda_1, \dots, \lambda_n \in \mathbb{R}$. Angenommen alle Erwartungswerte $\mathbb{E}[X_1], \dots, \mathbb{E}[X_n]$ existieren, dann existiert auch der Erwartungswert der Zufallsvariable $\lambda_1 X_1 + \dots + \lambda_n X_n$ und es gilt

$$\mathbb{E}[\lambda_1 X_1 + \dots + \lambda_n X_n] = \lambda_1 \mathbb{E}[X_1] + \dots + \lambda_n \mathbb{E}[X_n].$$

- ii) Es seien X_1, \dots, X_n zusätzlich **unabhängig** und seien $g_1, \dots, g_n: \mathbb{R} \rightarrow \mathbb{R}$ stetige Abbildungen mit der Eigenschaft, dass für alle $i = 1, \dots, n$: $\int_{-\infty}^{\infty} |g_i(t)| f_{X_i}(t) dt < \infty$, dann gilt

$$\mathbb{E}[g(X_1) \cdots g(X_n)] = \mathbb{E}[g(X_1)] \cdots \mathbb{E}[g(X_n)].$$

Es wird nicht überraschen, dass auch die Varianz im kontinuierlichen Kontext die zum diskreten Kontext analogen Rechenregeln erfüllt.

Satz 2.36 (Rechenregeln für Varianz).

- i) Ist X eine Zufallsvariable mit Dichte f_X und endlicher Varianz, so gilt für $a, b \in \mathbb{R}$

$$\text{Var}(a \cdot X + b) = a^2 \cdot \text{Var}(X).$$

- ii) Es gilt weiter:

$$\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

- iii) Es seien X_1, \dots, X_n **unabhängige** Zufallsvariablen mit Dichten f_{X_1}, \dots, f_{X_n} , dann gilt

$$\text{Var}(X_1 + \dots + X_n) = \text{Var}(X_1) + \dots + \text{Var}(X_n).$$

BEWEIS.

$$\begin{aligned} i) \quad \text{Var}(a \cdot X + b) &= \mathbb{E}[(a \cdot X + b - \underbrace{\mathbb{E}[a \cdot X + b]}_{a\mathbb{E}[X]+b})^2] = \mathbb{E}[(aX - a\mathbb{E}[X])^2] \\ &= a^2 \cdot \mathbb{E}[(X - \mathbb{E}[X])^2] = a^2 \cdot \text{Var}(X_1) \end{aligned}$$

$$\begin{aligned} ii) \quad \text{Var}(X) &= \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2 - 2 \cdot X \cdot \mathbb{E}[X] + \mathbb{E}[X]^2] \\ &= \mathbb{E}[X^2] - \mathbb{E}[2 \cdot X \cdot \mathbb{E}[X]] + \mathbb{E}[X]^2 = \mathbb{E}[X^2] - 2 \cdot \mathbb{E}[X] \cdot \mathbb{E}[X] + \mathbb{E}[X]^2 \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \end{aligned}$$

- iii) Für $n = 2$ gilt:

$$\begin{aligned} \text{Var}(X_1 + X_2) &= \mathbb{E}[(X_1 + X_2 - \mathbb{E}[X_1 + X_2])^2] = \mathbb{E}[(X_1 - \mathbb{E}[X_1] + X_2 - \mathbb{E}[X_2])^2] \\ &= \mathbb{E}[(X_1 - \mathbb{E}[X_1])^2 + 2 \cdot (X_1 - \mathbb{E}[X_1]) \cdot (X_2 - \mathbb{E}[X_2]) + (X_2 - \mathbb{E}[X_2])^2] \\ &= \underbrace{\mathbb{E}[(X_1 - \mathbb{E}[X_1])^2]}_{\text{Var}(X_1)} + \underbrace{\mathbb{E}[(X_2 - \mathbb{E}[X_2])^2]}_{\text{Var}(X_2)} + \underbrace{2 \cdot \mathbb{E}[(X_1 - \mathbb{E}[X_1]) \cdot (X_2 - \mathbb{E}[X_2])]}_I \end{aligned}$$

Aufgrund der Unabhängigkeit der Zufallsvariablen X_1 und X_2 schließt man

$$\begin{aligned} I &= \mathbb{E}[(X_1 - \mathbb{E}[X_1]) \cdot (X_2 - \mathbb{E}[X_2])] = \mathbb{E}[X_1] - \mathbb{E}[X_1] - \mathbb{E}[X_2] - \mathbb{E}[X_2] \\ &= 0. \end{aligned}$$

Per Induktion folgt die Behauptung. \square

2.6 Abschätzen von Wahrscheinlichkeiten

Wir zeigen nun, dass wichtige Ungleichungen ebenfalls im stetigen Kontext richtig bleiben.

Satz 2.37 (Markov-Ungleichung).

Es sei X eine nicht-negative Zufallsvariable mit Dichte f_X . Dann gilt für $t > 0$

$$t \cdot \mathbb{P}(X \geq t) \leq \mathbb{E}[X],$$

sofern $\mathbb{E}[X]$ existiert, d. h. $\int_{-\infty}^{\infty} |t| \cdot f_X(t) dt < \infty$.

BEWEIS.

$$\begin{aligned} t \cdot \mathbb{P}(X \geq t) &= t \cdot \int_t^{\infty} f_X(r) dr = \int_t^{\infty} t \cdot f_X(r) dr \\ &\leq \int_t^{\infty} r \cdot f_X(r) dr \leq \int_0^{\infty} r \cdot f_X(r) dr = \mathbb{E}[X]. \end{aligned} \quad \square$$

Wie für diskrete Wahrscheinlichkeitsräume erhält man mit Hilfe der Markov-Ungleichung das folgende Resultat.

Satz 2.38 (Chebyshev-Ungleichung).

Es sei X eine Zufallsvariable mit Dichte f_X , so dass der Erwartungswert und Varianz existieren. Dann gilt für $\delta > 0$

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \delta) \leq \frac{\text{Var}(X)}{\delta^2}.$$

BEWEIS. Die Ungleichung ist eine Folgerung aus der Markov-Ungleichung:

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \delta) = \mathbb{P}((X - \mathbb{E}[X])^2 \geq \delta^2) \leq \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{\delta^2} = \frac{\text{Var}(X)}{\delta^2}. \quad \square$$

Hieraus folgern wir wieder das schwache Gesetz der großen Zahlen, das einen Zusammenhang zwischen empirischen Größen und dem Erwartungswert herstellt.

Satz 2.39 (Gesetz der großen Zahlen).

Es sei X eine Zufallsvariable mit Dichte f_X und endlicher Varianz; seien weiter X_1, \dots, X_n unabhängige Zufallsvariable mit derselben Dichte wie X .

Dann gilt für jedes $\varepsilon > 0$

$$\mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E}[X]\right| \geq \varepsilon\right) \leq \frac{\text{Var}(X)}{n \cdot \varepsilon^2}.$$

Insbesondere gilt für jedes $\varepsilon > 0$

$$\lim_{n \rightarrow \infty} \mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E}[X]\right| \geq \varepsilon\right) = 0.$$

BEWEIS. Sei $\varepsilon > 0$. Es gilt

$$\begin{aligned} \mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E}[X]\right| \geq \varepsilon\right) &= \mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right| \geq \varepsilon\right) \\ &\leq \frac{\text{Var}\left(\frac{1}{n} \sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right)}{\varepsilon^2} = \frac{\text{Var}\left(\sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right)}{\varepsilon^2 \cdot n^2} \\ &= \frac{1}{\varepsilon^2 \cdot n^2} \cdot \sum_{i=1}^n \text{Var}(X) = \frac{\text{Var}(X)}{\varepsilon^2 \cdot n}. \end{aligned} \quad \square$$

2.7 Der zentrale Grenzwertsatz

Wir formulieren zunächst das entsprechende Analogon zur im diskreten Kontext bereits gezeigten Faltungsformel.

Satz 2.40 (Faltungsformel). Sind X_1 und X_2 **unabhängige** Zufallsvariablen mit Dichten f_{X_1} und f_{X_2} , so hat $Z = X_1 + X_2$ die Dichte

$$f_Z(z) = \int_{-\infty}^{\infty} f_{X_1}(z-v) \cdot f_{X_2}(v) dv =: (f_{X_1} * f_{X_2})(z).$$

Die Dichte $f_{X_1} * f_{X_2}$ heißt Faltung der Dichten f_{X_1} und f_{X_2} .

BEWEIS.

$$\mathbb{P}(Z \leq z) = \mathbb{P}(X_1 + X_2 \leq z) = \mathbb{P}((X_1, X_2) \in B) \text{ mit } B = \{(t_1, t_2) : t_1 + t_2 \leq z\}$$

X_1, X_2 unabhängig \implies die gemeinsame Dichte von (X_1, X_2) ist $f_{X_1}(\cdot) \cdot f_{X_2}(\cdot)$.

$$\begin{aligned} \mathbb{P}((X_1, X_2) \in B) &= \iint_B f_{X_1}(t_1) \cdot f_{X_2}(t_2) dt_1 dt_2 \\ &\stackrel{u=t_1+t_2}{\stackrel{v=t_2}{=}} \int_{-\infty}^z \int_{-\infty}^{\infty} f_{X_1}(u-v) \cdot f_{X_2}(v) dv du \end{aligned}$$

Damit gilt für die Verteilungsfunktion von Z :

$$F_Z(z) = \int_{-\infty}^z \underbrace{\int_{-\infty}^{\infty} f_{X_1}(u-v) \cdot f_{X_2}(v) dv}_{\text{Dichte von } Z} du$$

Dadurch folgt die Behauptung. \square

Hiermit lässt sich die wichtige Invarianz der Normalverteilung zeigen.

Satz 2.41. Seien $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ und (X_1, X_2) **unabhängig**, dann gilt

$$X_1 + X_2 \sim \mathcal{N}(\mu, \sigma^2), \quad \text{wobei } \mu = \mu_1 + \mu_2 \text{ und } \sigma^2 = \sigma_1^2 + \sigma_2^2$$

BEWEIS. Sei $Z = X_1 + X_2$, dann gilt für die Dichte f_Z von Z

$$f_Z(z) = \int_{-\infty}^{\infty} f_{X_1}(z-y) \cdot f_{X_2}(y) dy = \frac{1}{2\pi\sigma_1\sigma_2} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\frac{(z-y-\mu_1)^2}{\sigma_1^2}} \cdot e^{-\frac{1}{2}\frac{(y-\mu_2)^2}{\sigma_2^2}} dy$$

Das Berechnen des Integrals wird dem interessierten Leser als Übungsaufgabe empfohlen. \square

Der folgende Satz ist Grundlage vieler statistischer Verfahren und rechtfertigt die zentrale Bedeutung der Normalverteilung innerhalb der Wahrscheinlichkeitstheorie und Statistik.

Satz 2.42 (Zentraler Grenzwertsatz).

Seien X_1, \dots, X_n, \dots unabhängige Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \mathbb{P})$ mit derselben Verteilungsfunktion und existierendem Erwartungswert $\mu \in \mathbb{R}$ und Varianz $\sigma^2 \in (0, \infty)$. Weiter sei Y_n definiert durch $Y_n = X_1 + \dots + X_n$, $n \geq 1$. Dann ist die Zufallsvariable

$$Z_n := \frac{Y_n - n \cdot \mu}{\sigma \cdot \sqrt{n}}$$

asymptotisch (standard)normalverteilt, d. h. $\forall x \in \mathbb{R}$

$$\lim_{n \rightarrow \infty} F_{Z_n}(x) = \lim_{n \rightarrow \infty} \mathbb{P}(Z_n \leq x) = \Phi(x).$$

BEWEIS (SKIZZE). Wir begründen, dass für jede zweimal stetig differenzierbare Funktion $f : \mathbb{R} \rightarrow \mathbb{R}$ mit beschränkten und gleichmäßig stetigen Ableitungen f' und f''

$$\lim_{n \rightarrow \infty} \mathbb{E}[f(Z_n)] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} f(t) dt$$

gilt. In der Tat lässt sich hieraus mit etwas zusätzlicher Arbeit der Zentrale Grenzwertsatz herleiten. Ohne Einschränkung sei $\mu = 0$ und $\sigma = 1$. Sei weiter $(N_i)_{i \geq 1}$ eine Folge von unabhängigen, standardnormalverteilten Zufallsvariablen, welche ebenfalls von $(X_i)_{i \geq 1}$ unabhängig ist. Wir definieren $\zeta_n := \sum_{i=1}^n \frac{N_i}{\sqrt{n}}$. Weil die Zufallsvariablen ζ_n standardnormalverteilt sind, genügt der Nachweis von

$$\lim_{n \rightarrow \infty} \mathbb{E}[f(Z_n) - f(\zeta_n)] = 0.$$

Wir stellen die Differenz

$$f(Z_n) - f(\zeta_n) \tag{2.1}$$

als Teleskopsumme dar. Mit den Abkürzungen $X_{i,n} = X_i/\sqrt{n}$ und $N_{i,n} = N_i/\sqrt{n}$ und $W_{i,n} = \sum_{j=1}^{i-1} N_{j,n} + \sum_{j=i+1}^n X_{j,n}$ gilt offenbar für jedes $1 < i \leq n$

$$W_{i,n} + X_{i,n} = W_{i-1,n} + N_{i-1,n}$$

und somit

$$f(Z_n) - f(\zeta_n) = \sum_{i=1}^n [f(W_{i,n} + X_{i,n}) - f(W_{i,n} + N_{i,n})]. \tag{2.2}$$

Eine Anwendung der Taylorformel ergibt

$$f(W_{i,n} + X_{i,n}) = f(W_{i,n}) + f'(W_{i,n})X_{i,n} + \frac{1}{2}f''(W_{i,n})X_{i,n}^2 + R_{X,i,n}$$

wobei der Fehlerterm

$$R_{X,i,n} = \frac{1}{2} X_{i,n}^2 [f''(W_{i,n} + \vartheta X_{i,n}) - f''(W_{i,n})]$$

mit einem geeigneten $\vartheta \in [0, 1]$ die Abschätzung $|R_{X,i,n}| \leq X_{i,n}^2 \sup_x |f''(x)|$ erfüllt. Weil f'' gleichmäßig stetig ist erhalten wir dann für vorgegebenes $\varepsilon > 0$ und geeignetes $\delta > 0$

$$|R_{X,i,n}| \leq X_{i,n}^2 [\varepsilon \mathbf{1}_{\{|X_{i,n}| \leq \delta\}} + \sup_x |f''(x)| \mathbf{1}_{\{|X_{i,n}| \geq \delta\}}].$$

Eine vollkommen analoge Abschätzung erhält man für $f(W_{i,n} + N_{i,n})$.

Setzt man diese Taylor-Approximation in (2.2) ein und bildet den Erwartungswert, so verschwinden alle Terme bis auf die Restglieder, da

$$\mathbb{E}[X_{i,n}] = \mathbb{E}[N_{i,n}] = 0,$$

$$\mathbb{E}[X_{i,n}^2] = \frac{1}{n} = \mathbb{E}[N_{i,n}^2]$$

und andererseits wegen der Unabhängigkeitsvoraussetzungen z.B.

$$\mathbb{E}[f''(W_{i,n})[X_{i,n}^2 - N_{i,n}^2]] = \mathbb{E}[f''(W_{i,n})] \mathbb{E}[X_{i,n}^2 - N_{i,n}^2] = 0.$$

Folglich gilt

$$\begin{aligned} \mathbb{E}[f(Z_n) - f(\zeta_n)] &\leq \sum_{i=1}^n \mathbb{E}[|R_{X,i,n}| + |R_{N,i,n}|] \\ &\leq \sum_{i=1}^n \left(\varepsilon \mathbb{E}[X_{i,n}^2 + N_{i,n}^2] + \sup_x |f''(x)| \mathbb{E}[X_{i,n}^2 \mathbf{1}_{\{|X_{i,n}| \geq \delta\sqrt{n}\}} + N_{i,n}^2 \mathbf{1}_{\{|N_{i,n}| \geq \delta\sqrt{n}\}}] \right) \\ &= 2\varepsilon + \sup_x |f''(x)| [\mathbb{E}[X_{i,n}^2 \mathbf{1}_{\{|X_{i,n}| \geq \delta\sqrt{n}\}}] + \mathbb{E}[N_{i,n}^2 \mathbf{1}_{\{|N_{i,n}| \geq \delta\sqrt{n}\}}]]. \end{aligned}$$

Es ergibt sich also

$$\limsup_{n \rightarrow \infty} \mathbb{E}[f(Z_n) - f(\zeta_n)] \leq 2\varepsilon$$

und da $\varepsilon > 0$ beliebig war, erhalten wir die Behauptung.

Um den Beweis des Satzes abzuschließen zeigt man dass für $a < b$ und beliebiges $\delta > 0$ zwei Funktionen f_1, f_2 existieren, so dass die die Eigenschaften der obigen Funktion f gelten und

- Für alle $x \in \mathbb{R}$ gilt $f_2(x) \leq \mathbf{1}_{[a,b]}(x) \leq f_1(x) \leq 1$
- Für alle $x < a$, $x \in [a + \delta, b - \delta]$ und alle $x > b + \delta$ gilt $f_2(x) = \mathbf{1}_{[a,b]}(x)$
- Für alle $x < a - \delta$, $x \in [a, b]$ und $x > b + \delta$ gilt $f_1(x) = \mathbf{1}_{[a,b]}(x)$.

Dann erhalten wir aus der obigen Analyse

$$\mathbb{E}[f_2(N)] \leq \liminf_{n \rightarrow \infty} \mathbb{E}[\mathbf{1}_{[a,b]}(Z_n)] \leq \limsup_{n \rightarrow \infty} \mathbb{E}[\mathbf{1}_{[a,b]}(Z_n)] \leq \mathbb{E}[f_2(N)],$$

wobei N standardnormal ist. Weil

$$\mathbb{E}[f_2(N)] \geq \mathbb{P}(a + \delta \leq N \leq b - \delta)$$

und

$$\mathbb{E}[f_1(N)] \leq \mathbb{P}(a - \delta \leq N \leq b + \delta)$$

erhält man die Behauptung. \square

Bemerkung 2.43. Der zentrale Grenzwertsatz lässt auch in einem gewissen Sinne als Antwort auf die Frage nach der Konvergenzgeschwindigkeit im Gesetz der großen Zahlen interpretieren. Das Gesetz der großen Zahl besagt, dass

$$\frac{1}{n} \sum_{i=1}^n (X_i - \mu) \rightarrow 0$$

für $n \rightarrow \infty$ im Sinne der stochastischen Konvergenz. Der zentrale Grenzwertsatz beantwortet nun die Frage, mit welchem Faktor die linke Seite aufgeblasen werden muss, damit im Sinne der Verteilungskonvergenz ein nicht-trivialer Grenzwert entsteht. Als Antwort gibt der zentrale Grenzwertsatz i.w. den Faktor \sqrt{n} .

Dieser Faktor bewirkt, dass die Varianz konstant bleibt:

$$\begin{aligned} \text{Var}(Z_n) &= \text{Var}\left(\frac{Y_n - n \cdot \mu}{\sigma \cdot \sqrt{n}}\right) = \frac{1}{\sigma^2 \cdot n} \cdot \text{Var}(Y_n - n \cdot \mu) \\ &= \frac{1}{\sigma^2 \cdot n} \text{Var}\left(\sum_{i=1}^n X_i\right) = \frac{1}{\sigma^2 \cdot n} \sum_{i=1}^n \text{Var}(X_i) \\ &= 1. \end{aligned}$$

Die Zufallsvariablen Z_n haben also Erwartungswert 0 und Varianz 1. Dies erklärt zumindest für die im Limes auftauchende Normalverteilung ebenfalls dieselben Parameter zu erwarten sind.

Beispiel 2.44. Wir wollen nun die Situation eines Roulette Spielers mit der Strategie „In jedem Durchgang setze einen Euro auf rot“ beschreiben. Besonders interessieren wir uns für die Wahrscheinlichkeit, nach n Spielen nicht als Verlierer vom Feld zu gehen. Hierzu führen wir zunächst unabhängige Zufallsvariablen X_1, \dots, X_n ein, die beschreiben, ob wir das i -te Spiel gewonnen oder verloren haben. Beachte, dass 18 der 37 Felder rot sind. Wir haben also

$$\mathbb{P}(X_i = 1) = \frac{18}{37}, \quad \mathbb{P}(X_i = -1) = \frac{19}{37}.$$

Die Zufallsvariable

$$S_n = \sum_{i=1}^n X_i$$

beschreibt unseren Gewinn nach n Spielen. Es gilt natürlich $\mathbb{E}[X_i] = -\frac{1}{37}$, $\mathbb{E}[S_n] = -n \cdot \frac{1}{37}$, $\text{Var}(X_i) \approx 0,99993$. Zur Vereinfachung: $\sigma = 1$, $n = 37^2$.

$$\begin{aligned} \mathbb{P}(S_n \geq 0) &= \mathbb{P}\left(\frac{S_n - \mathbb{E}[S_n]}{\sigma \sqrt{n}} \geq \frac{-n\mathbb{E}[X_i]}{\sigma \sqrt{n}}\right) \\ &= \mathbb{P}\left(\frac{S_n - n\mathbb{E}[X_i]}{\sigma \sqrt{n}} \geq 1\right) \\ &\approx \mathbb{P}(\chi \geq 1) = 1 - \Phi(1) \approx 0,1587, \quad \chi \sim \mathcal{N}(0, 1). \end{aligned}$$

Mit Wahrscheinlichkeit von approximativ 0,16 verlässt man als Spieler nach 37^2 Spielen das Casino ohne Verlust.

Bemerkung 2.45 (Heuristische Beschreibung des Zentralen Grenzwertsatzes). Es sei X eine Zufallsvariable mit Dichte f_X und Erwartungswert μ und endlicher Varianz $\sigma^2 \in (0, \infty)$.

Seien weiter X_1, X_2, \dots, X_n unabhängige Zufallsvariablen mit derselben Dichte wie X . Dann besagt der zentrale Grenzwertsatz, dass für reelle Zahlen $a < b$

$$\lim_{n \rightarrow \infty} \mathbb{P}\left(a \leq \frac{\sum_{i=1}^n (X_i - \mu)}{\sigma \sqrt{n}} \leq b\right) = \int_a^b \varphi(x) dx$$

gilt. Spezialisiert man auf ganze Zahlen $a = -l$ und $b = l$ so erhält man

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbb{P}\left(a \leq \frac{\sum_{i=1}^n (X_i - \mu)}{\sigma \sqrt{n}} \leq b\right) &= \lim_{n \rightarrow \infty} \mathbb{P}\left(n \cdot \mu - l \cdot \sqrt{n} \sigma \leq \sum_{i=1}^n X_i \leq n \mu + l \cdot \sqrt{n} \sigma\right) \\ &= \int_{-l}^l \varphi(x) dx = 2\Phi(l) - 1. \end{aligned}$$

In einem Tabellenwerk findet man für $l = 1, 2, 3$ die approximativen Werte

$$2\Phi(1) - 1 \approx 0,682, \quad 2\Phi(2) - 1 \approx 0,954, \quad 2\Phi(3) - 1 \approx 0,997.$$

Approximativ erhält man also, dass die Summe $\sum_{i=1}^n X_i$ von n unabhängigen Zufallsvariablen mit Mittel μ und Varianz σ^2 mit der Wahrscheinlichkeit

- 0,682 in dem Bereich $n \cdot \mu \pm 1 \cdot \sqrt{n} \cdot \sigma$
- 0,954 in dem Bereich $n \cdot \mu \pm 2 \cdot \sqrt{n} \cdot \sigma$
- 0,997 in dem Bereich $n \cdot \mu \pm 3 \cdot \sqrt{n} \cdot \sigma$

zu finden ist. Betrachtet man zum Beispiel unser klassisches Würfelbeispiel, dann lässt sich folgern, dass die Augensumme bei 100 Würfeln approximativ mit Wahrscheinlichkeit von 0.954 in dem Bereich

$$n \cdot \mu \pm 2 \cdot \sqrt{n} \sigma = 100 \cdot 3.5 \pm 2 \cdot \sqrt{100} \cdot \frac{35}{12},$$

also zwischen 316 und 384, liegt. Die gerade beschriebene Überlegung bezeichnet man manchmal auch als \sqrt{n} -Regel.

2.8 Anwendungen

2.8.1 Die Monte-Carlo-Methode

Angenommen, wir wollen den Wert der Kreiszahl π numerisch approximieren. Mit Hilfe der eingeführten Konzepte und Resultate können wir wie folgt vorgehen. Seien X und Y unabhängige Zufallsvariablen, die beide gleichverteilt auf $[-1, 1]$ sind. Dann hat (X, Y) Werte in $[-1, 1] \times [-1, 1]$. Die Kreisscheibe mit Radius 1 und Zentrum im Punkt $(0, 0)$ ist eine Teilmenge von $[-1, 1] \times [-1, 1]$ und ihr Flächeninhalt beträgt π . Führt man die Zufallsvariable

$$Z = \begin{cases} 1 & \text{wenn } \sqrt{X^2 + Y^2} \leq 1 \\ 0 & \text{sonst} \end{cases}$$

ein, dann ergibt eine einfache Rechnung

$$\mathbb{P}(Z = 1) = \frac{\pi}{4}.$$

Wir betrachten nun $2m$ unabhängige jeweils auf $[-1, 1]$ gleichverteilte Zufallsvariablen

$$X_1, Y_1, X_2, Y_2, \dots, X_m, Y_m$$

und definieren wie oben

$$Z_i = \begin{cases} 1 & \text{wenn } \sqrt{X_i^2 + Y_i^2} \leq 1 \\ 0 & \text{sonst.} \end{cases}$$

Dann sind die Zufallsvariablen Z_1, Z_2, \dots, Z_m unabhängig und identisch verteilt. Das Gesetz der großen Zahlen impliziert, dass für $m \rightarrow \infty$

$$\frac{1}{m} \sum_{i=1}^m Z_i \rightarrow \mathbb{E}[Z_1] = \frac{\pi}{4}$$

im Sinne der stochastischen Konvergenz gilt. Definiert man $W := \frac{4}{m} \sum_{i=1}^m Z_i$ so ergibt eine Anwendung der Chernoff Schranken

$$\mathbb{P}(|W - \pi| \geq \varepsilon \pi) \leq 2e^{-m\pi\varepsilon^2/12}.$$

Dieses Vorgehen ist ein Beispiel für die sogenannte Monte-Carlo-Methode. Das (starke) Gesetz der großen Zahlen garantiert uns die Konvergenz des Verfahrens und z.B. die Chernoff-Ungleichung kann zum Abschätzen der Wahrscheinlichkeit zu großer Abweichungen benutzt werden.

2.8.2 Rejection Sampling

Die Idee des Rejection-Samplings klang in der obigen Anwendung des Monte-Carlo Verfahrens bereits an. Wir geben nur noch ein paar zusätzliche zentrale Ideen, gehen aber aus Zeitgründen nicht weiter in die Thematik ein.

$(\Omega, \mathcal{F}, \mathbb{P})$ sei ein Wahrscheinlichkeitsraum. und sei B ein Ereignis mit $\mathbb{P}(B) > 0$

Accept-Reject-Algorithmus:

- Generate $x \sim \mathbb{P}$
- If $x \in B$ then accept $x \rightarrow y$
else reject: return to Step 1.

Von einem eher mathematischen Standpunkt her gesehen, bedeutet das, dass die Zufallsvariablen

$$X_1, X_2, \dots \sim \mathbb{P}$$

unabhängig sind,

$$Y := X_T,$$

wobei

$$T := \min\{n \in \mathbb{N} \mid X_n \in B\}.$$

Es gilt dann

Satz 2.46. Unter den obigen Annahmen gilt

- $T \sim \text{Geom}(p)$, $p = \mathbb{P}(B)$, und somit $\mathbb{E}[T] = \frac{1}{p}$
- $Y \sim \mathbb{P}(\cdot \mid B)$

BEWEIS. Die erste Behauptung ist offensichtlich, die zweite Behauptung ergibt sich aus

$$\begin{aligned}
 \mathbb{P}(Y \in A) &= \sum_{n=1}^{\infty} \mathbb{P}(Y \in A, T = n) \\
 &= \sum_{n=1}^{\infty} \mathbb{P}(X_n \in A, X_1 \notin B, \dots, X_{n-1} \notin B) \\
 &= \sum_{n=1}^{\infty} \mathbb{P}(A) \mathbb{P}(B^c)^{n-1} = \mathbb{P}(A \mid B) \quad \square
 \end{aligned}$$

Wir werden an dieser Stelle ausnahmsweise nochmals im rein diskreten Kontext arbeiten, auch wenn die Aussagen allgemeiner formulierbar sind. Es sei S eine endliche oder abzählbare Menge, μ eine Wahrscheinlichkeitsverteilung auf S mit Dichte p und ν eine Wahrscheinlichkeitsverteilung mit Dichte q , i.e.

$$\mu(A) = \sum_{x \in A} p(x) \quad \text{und} \quad \nu(A) = \sum_{x \in A} q(x).$$

Angenommen, wir können gemäß der Verteilung ν unabhängige Stichproben generieren. Wie können wir daraus Stichproben von μ erhalten? Hierfür treffen wir folgende Annahme:

$$\text{Es gibt ein } c \in [1, \infty) : \quad p(x) \leq cq(x) \quad \text{für alle } x \in S.$$

Aus der Annahme folgt dann, dass für alle $x \in S$

$$\frac{p(x)}{cq(x)} \leq 1$$

gilt. Wir werden diesen Quotienten als Akzeptanzwahrscheinlichkeit wählen.

Algorithmus: (Acceptance–Rejection–Verfahren)

INPUT: Gewicht $p(y), q(y), c$ ($y \in S$),

OUTPUT: Stichprobe x von μ .

repeat

erzeuge Stichprobe $x \sim \nu$

erzeuge Stichprobe $u \sim \text{Unif}[0, 1]$

until $\frac{p(x)}{cq(x)} \geq u$ {akzeptiere mit Wahrscheinlichkeit $\frac{p(x)}{cq(x)}$ }

return x

ANALYSE DES ALGORITHMUS:

Für die verwendeten Zufallsvariablen gilt:

$$X_1, X_2, \dots \sim \nu,$$

$$U_1, U_2, \dots \sim \text{Unif}[0, 1].$$

Die Zufallsvariablen sind unabhängig. Seien

$$T = \min\{n \in \mathbb{N} \mid \frac{p(X_n)}{cq(X_n)} \geq U_n\}$$

und

$$X_T(\omega) := X_{T(\omega)}(\omega).$$

Wir erhalten dann

Satz 2.47. Folgende Aussagen gelten

- T ist geometrisch verteilt mit Parameter $1/c$.
- $X_T \sim \mu$.

BEWEIS. Zur Aussage i): Sei

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

Aufgrund der Unabhängigkeit der Zufallsvariablen $X_1, U_1, X_2, U_2, \dots$ folgt, dass die Ereignisse A_1, A_2, \dots unabhängig sind. Unter Ausnutzung der Unabhängigkeit von X_n und U_n ergibt sich damit

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

Folglich ist T geometrisch mit Parameter $1/c$.

Zur Aussage ii): Wir rechnen

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

□

2.8.3 Erzeugung reeller Zufallsvariablen mit vorgegebener Verteilung

Definition 2.48. Es sei F_X die Verteilungsfunktion einer reellwertigen Zufallsvariablen X . Die (verallgemeinerte) inverse Funktion oder auch Quantilsfunktion von F_X ist für $t \in [0, 1]$ definiert durch

$$F_X^{-1}(t) := \inf\{x \in \mathbb{R} \mid F_X(x) \geq t\},$$

wobei wir mit der Konvention $\inf \emptyset = \infty$ arbeiten.

Lemma 2.49. Es sei F_X^{-1} die verallgemeinerte Inverse einer Verteilungsfunktion F_X . Dann gilt

$$F_X^{-1}(y) \leq x \quad \text{genau, wenn} \quad y \leq F_X(x).$$

BEWEIS. Es gelte zunächst $F_X^{-1}(y) \leq x$. Nach Definition haben wir

$$F_X^{-1}(y) = \inf\{u \mid F_X(u) \geq y\} =: u_{\inf} \leq x.$$

Weil F_X als Verteilungsfunktion rechtsstetig ist, gilt $F_X(u_{\inf}) \geq y$ und wegen der Monotonie dann $F_X(x) \geq y$.

Es dagegen $F_X(x) \geq y$ für ein x , dann folgt wieder unter Benutzung der Monotonie

$$F_X^{-1}(y) = \inf\{u \mid F_X(u) \geq y\} \leq x. \quad \square$$

Satz 2.50 (Erzeugung reeller ZV mit vorgegebener Verteilung). Sei F_X eine Verteilungsfunktion einer reellwertigen Zufallsvariablen X mit verallgemeinertem Inversen F_X^{-1} und sei U eine gleichverteilt auf $[0, 1]$. Dann besitzt die Zufallsvariable

$$Y := F_X^{-1}(U)$$

dieselbe Verteilung wie X

BEWEIS. Es gilt $F_X^{-1}(t) \leq x$ genau, wenn $F_X(x) \geq t$ und somit ist für $x \in \mathbb{R}$

$$\begin{aligned} \mathbb{P}(Y \leq x) &= \mathbb{P}(F_X^{-1}(U) \leq x) = \mathbb{P}(U \leq F_X(x)) \\ &= \mathbb{P}(0 \leq U \leq F_X(x)) = F_X(x). \end{aligned} \quad \square$$

Beispiel 2.51. Es sei

$$f_X(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-\lambda x} & x \geq 0 \end{cases}$$

die Verteilungsfunktion einer exponentialverteilten Zufallsvariable X mit Parameter λ . Dann gilt für $y \in (0, 1)$

$$f_X^{-1}(y) = -\frac{1}{\lambda} \ln(1 - y).$$

Ist dann U eine gleichverteilte Zufallsvariable, dann ist

$$-\frac{1}{\lambda} \ln(1 - U)$$

eine exponentialverteilte Zufallsvariable mit Parameter λ .

Bemerkung 2.52 (Zufallszahlen). Realisierung von auf $[0, 1]$ gleichverteilten Zufallsvariablen nennt man auch *Zufallszahlen*. Zufallszahlen findet man in Tabellenwerken oder im Internet. Zufallszahlen werden manchmal durch 'echten' Zufall erzeugt¹, in den meisten Fällen werden diese jedoch deterministisch generiert. Eine übliche Methode zur Erzeugung von Zufallszahlen ist die sog. lineare Kongruenz Methode.

Wähle sorgfältig einen 'Modulus' m (e.g. $m = 2^{32}$) sowie einen Faktor $a \in \mathbb{N}$ und ein Inkrement $b \in \mathbb{N}$. Nun wähle $k_0 \in \{0, \dots, m-1\}$ und definiere die Folge $k_{i+1} = ak_i + b \bmod m$. Die Pseudo-Zufallszahlen bestehen dann aus der Folge $(k_i/m)_{i \geq 1}$. Für eine geeignete Wahl von a und b hat die Folge (k_i) Periode m und besteht die üblichen Tests auf Unabhängigkeit. Man sollte jedoch stets genau prüfen, ob die benutzten Zufallszahlen für die die benutzten Zwecke angemessen sind. Der IBM Zufallsgenerator randu aus den 60-er Jahren hatte die ungünstige Eigenschaft, dass die Tripel (u_i, u_{i+1}, u_{i+2}) in lediglich 15 verschiedenen parallelen Ebenen des \mathbb{R}^3 liegen. Für details konsultiere man die Standardreferenz Donald Knuth: *The art of computer programming*, Addison Wesley, 1997.

¹vgl. www.rand.org/pub/monograph/reports/MR1418/index.html

Induktive Statistik (Schätz- und Testtheorie)

3.1 Schätztheorie

In der Wahrscheinlichkeitstheorie ist das stochastische Modell typischerweise gegeben und die Aufgabe besteht in der Berechnung von Wahrscheinlichkeiten für das Auftreten gewisser Ereignisse. In gewissem Sinne befasst sich die Statistik mit einem dazu inversen Problem. Die zugrundeliegende Wahrscheinlichkeitsverteilung ist in der Statistik nicht explizit bekannt. Um Rückschlüsse auf die zugrundeliegende Verteilung zu ziehen, erhebt man dafür z.B. durch Experimente Daten. Die Frage, auf welche Weise man aufgrund von Daten Charakteristika der zugrundeliegenden Verteilung ermitteln kann, ist ein zentraler Gegenstand der Statistik.

Beispiel 3.1 (Motivation).

$X \triangleq$ Anzahl von Lesezugriffen auf eine Festplatte bis zum ersten Lesefehler. Wir spezifizieren eine geeignete Klasse von Wahrscheinlichkeitsverteilungen, die die gegebene Problemstellung geeignet modellieren. Hierbei ergibt sich zum Beispiel der folgende erste sinnvolle Ansatz:

$$\mathbb{P}_p(X = i) = (1 - p)^{i-1} \cdot p, p \in [0, 1].$$

Man nimmt hierbei an, dass bei jedem Zugriff unabhängig und mit derselben Wahrscheinlichkeit p ein Lesefehler auftreten kann.

Unsere Aufgabenstellung lässt sich nun wie folgt präziser formulieren: Man schätze auf eine sinnvolle Art und Weise den Wert p aus vorhandenen Daten.

Das Gesetz der großen Zahlen liefert einen Lösungsansatz. Wir wissen, dass $\mathbb{E}_p[X] = \frac{1}{p}$, d. h. wir können auch ebenso $\mathbb{E}_p[X]$ ermitteln. Betrachte n baugleiche Platten und die zugehörigen unabhängigen Zufallsvariablen X_1, \dots, X_n , mit gleicher Verteilung wie X . Um $\mathbb{E}_p[X]$ empirisch zu ermitteln, bilde

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

Das arithmetische Mittel schätzt den Erwartungswert und liefert damit eine Schätzung für $\frac{1}{p}$.

Das typische Setting der induktiven Statistik ist also ein Erbenissraum Ω , eine σ -Algebra \mathcal{A} über Ω , die Ereignisse enthält sowie eine Familie $(\mathbb{P}_\vartheta)_{\vartheta \in \Theta}$ von Verteilungen auf \mathcal{A} , die wir als Kandidaten betrachten. Man versucht dann den 'richtigen' Parameter ϑ zu schätzen. Mit dem Begriff des statistischen Modell meinen wir genau eine solche Spezifikation von Ω , \mathcal{A} und $(\mathbb{P}_\vartheta)_{\vartheta \in \Theta}$.

Definition 3.2. Gegeben sei eine Zufallsvariable X (Grundgesamtheit). Als mathematische Stichprobe vom Umfang n dieser Grundgesamtheit bezeichnet man eine Menge an Zufallsvariablen X_1, \dots, X_n mit folgenden Eigenschaften:

- Jedes X_i besitze dieselbe Verteilung wie X
- Die Zufallsvariablen X_1, \dots, X_n sind unabhängig

Ermittelt man n konkrete Werte x_1, \dots, x_n ($x_1 \in W_{X_1}, \dots, x_n \in W_{X_n}$), so nennen wir (x_1, \dots, x_n) eine konkrete Stichprobe.

Definition 3.3. Gegeben sei eine mathematische Stichprobe X_1, \dots, X_n und eine (messbare) Funktion $T: W_{X_1} \times \dots \times W_{X_n} \rightarrow \mathbb{R}$. Dann nennt man die Zufallsvariable

$$T(X_1, \dots, X_n): \Omega \rightarrow \mathbb{R}$$

eine mathematische Schätzfunktion bzw. Schätzer.

Hierbei haben wir im Hinterkopf, dass der Schätzer T aus der Stichprobe einen Parameter der zugrundeliegenden Verteilung schätzt. Man beachte, dass aufgrund der Allgemeinheit der Definition fast jede Funktion ein zulässiger Schätzer ist. Wir müssen also noch Kriterien erarbeiten, die die Güte von Schätzern beschreiben.

Beispiel 3.4. Sei $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$. Die folgenden Funktionen sind Schätzer.

- i) $T(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n X_i$
- ii) $\tilde{S}^2 = T(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$
- iii) $S^2 = T(X_1, \dots, X_n) = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$

Beispiel 3.5. Wir nehmen an, dass X_1^*, \dots, X_n^* aus X_1, \dots, X_n gebildet werden, indem wir die Werte der Größe nach ordnen (d. h. $X_1^* = \min(X_1, \dots, X_n), \dots, X_n^* = \max(X_1, \dots, X_n)$).

Dann ist

$$X_{0,5} = T(X_1, \dots, X_n) = \begin{cases} X_{\frac{n+1}{2}}^* & n \text{ ungerade} \\ \frac{1}{2} \left(X_{\frac{n}{2}}^* + X_{\frac{n+1}{2}}^* \right) & n \text{ gerade} \end{cases}$$

eine mathematische Schätzfunktion und heißt Stichprobenmedian.

Bei der Bestimmung von Parametern einer Grundgesamtheit mittels einer Stichprobe können natürlich Abweichungen des Schätzwerts vom Parameter auftreten. Es wird also gute und schlechte Schätzer für einen Parameter geben und wir werden uns mit einigen Gütekriterien befassen.

Definition 3.6. Es sei eine mathematische Stichprobe X_1, \dots, X_n für die Grundgesamtheit X auf Ω gegeben. Ein Schätzer $T(X_1, \dots, X_n)$ heißt **erwartungstreu** für Parameter p , wenn

$$\mathbb{E}[T(X_1, \dots, X_n)] = p.$$

$T(X_1, \dots, X_n)$ heißt **asymptotisch erwartungstreu**, wenn

$$\lim_{n \rightarrow \infty} \mathbb{E}[T(X_1, \dots, X_n)] = p.$$

Beispiel 3.7. Sei X die Grundgesamtheit und X_1, \dots, X_n eine mathematische Stichprobe für X . Dann ist $T(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n X_i$ ein erwartungstreuer Schätzer für $\mathbb{E}[X]$, denn es gilt

$$\mathbb{E}[T(X_1, \dots, X_n)] = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_i] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X] = \mathbb{E}[X].$$

Beispiel 3.8.

Betrachte aus Beispiel (3.4) den Schätzer $\tilde{S}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$. Dann ist \tilde{S}^2 nicht erwartungstreu für $p = \text{Var}(X)$.

Wir untersuchen $\mathbb{E}[\tilde{S}^2]$.

Zunächst gilt:

$$(X_i - \bar{X}) = (X_i - \mathbb{E}[X]) - (\bar{X} - \mathbb{E}[X]) = (X_i - \mathbb{E}[X]) - \left(\frac{1}{n} \sum_{i=1}^n (X_i - \mathbb{E}[X])\right)$$

Also nach bin. Formel:

$$\mathbb{E}[(X_i - \bar{X})^2] = \mathbb{E}[(X_i - \mathbb{E}[X])^2] + \mathbb{E}[(\bar{X} - \mathbb{E}[X])^2] - \frac{2}{n} \cdot \mathbb{E}\left[\sum_{j=1}^n (X_j - \mathbb{E}[X])(X_i - \mathbb{E}[X])\right]$$

$$\mathbb{E}\left[\sum_{j=1}^n (X_j - \mathbb{E}[X])(X_i - \mathbb{E}[X])\right] = \sum_{j=1}^n \mathbb{E}[(X_j - \mathbb{E}[X])(X_i - \mathbb{E}[X])]$$

Für $j \neq i$:

$$\mathbb{E}[(X_j - \mathbb{E}[X])(X_i - \mathbb{E}[X])] = \underbrace{\mathbb{E}[X_j - \mathbb{E}[X]]}_{\mathbb{E}[X] - \mathbb{E}[X] = 0} \cdot \mathbb{E}[X_i - \mathbb{E}[X]] = 0$$

Damit:

$$\mathbb{E}\left[\sum_{j=1}^n (X_j - \mathbb{E}[X])(X_i - \mathbb{E}[X])\right] = \mathbb{E}[(X_i - \mathbb{E}[X])^2] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \text{Var}(X)$$

Folglich:

$$\begin{aligned} \mathbb{E}[(X_i - \bar{X})^2] &= \text{Var}(X) + \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n (X_i - \mathbb{E}[X])^2\right] - \frac{2}{n} \text{Var}(X) \\ &= \text{Var}(X) + \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X) - \frac{2}{n} \text{Var}(X) = \frac{n-1}{n} \text{Var}(X) \end{aligned}$$

Die im eben behandelten Beispiel durchgeführte Rechnung liefert uns direkt den folgenden Satz.

Satz 3.9. Die mathematische Schätzfunktion

$$S^2 = T(X_1, \dots, X_n) = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

ist erwartungstreu für die Varianz der Grundgesamtheit X .

Die nächste Definition präzisiert folgende wünschenswerte Eigenschaft mathematischer Schätzer: Angenommen wir geben uns eine Fehlergrenze $\varepsilon > 0$ vor, dann ist es wünschenswert, durch

eventuelles Vergrößern des Stichprobenumfangs mit beliebig hoher Wahrscheinlichkeit einen Schätzwert in einer ε -Umgebung von p zu erhalten.

Definition 3.10. Für jedes $n \in \mathbb{N}$ sei $T(X_1, \dots, X_n)$ eine mathematische Schätzfunktion für die mathematische Stichprobe (X_1, \dots, X_n) der Grundgesamtheit X . Die Folge $(T(X_1, \dots, X_n))_{n \in \mathbb{N}}$ von Schätzern heißt **konsistent**, wenn für Parameter p und jedes $\varepsilon > 0$ gilt

$$\lim_{n \rightarrow \infty} \mathbb{P}(|T(X_1, \dots, X_n) - p| > \varepsilon) = 0.$$

Beispiel 3.11. Sei (X_1, \dots, X_n) eine mathematische Stichprobe für die Grundgesamtheit X . Weiter besitze X einen Erwartungswert und es gelte $\text{Var}(X) < \infty$. Dann ist $T(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n X_i$ ein konsistenter Schätzer für $\mathbb{E}[X]$, denn nach dem Gesetz der großen Zahlen gilt für jedes $\varepsilon > 0$

$$\mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E}[X]\right| > \varepsilon\right) \xrightarrow{n \rightarrow \infty} 0.$$

Ein weitere Möglichkeit, die Güte zweier Schätzer für einen Parameter miteinander zu vergleichen, gibt die folgende Definition.

Definition 3.12. Es sei (X_1, \dots, X_n) eine mathematische Stichprobe für die Grundgesamtheit X und seien $T_1(X_1, \dots, X_n)$ und $T_2(X_1, \dots, X_n)$ zwei erwartungstreue Schätzer für einen Parameter p der Grundgesamtheit X . Der Schätzer $T_1(X_1, \dots, X_n)$ heißt effizienter als $T_2(X_1, \dots, X_n)$, wenn

$$\text{Var}\left(T_1(X_1, \dots, X_n)\right) < \text{Var}\left(T_2(X_1, \dots, X_n)\right).$$

Die Frage, welche wünschenswerten Eigenschaften wie z.B. Erwartungstreue ein Schätzer erfüllen soll, hängt von der Situation ab. In verschiedenen Kontexten ist es durchaus denkbar, dass man bewusst nicht erwartungstreue Schätzer wählt. Man denke hier zum Beispiel an Messungen des Blutzuckerwertes. Niedrige Blutzuckerwerte sind kurzfristig gefährlicher als höhere.

Das folgende Beispiel illustriert einen verwandten Aspekt

Beispiel 3.13 (Ein guter verzerrter Schätzer). Wir betrachten das Schätzproblem der Erfolgswahrscheinlichkeit bei einer Bernoullikette, i.e. $\mathcal{X} = \{0, \dots, n\}$, $\Theta = [0, 1]$ und

$$\mathbb{P}_\vartheta(\{x\}) = \binom{n}{k} \vartheta^k (1 - \vartheta)^{n-k}.$$

Der ML-Schätzer für den Parameter ϑ ist gegeben durch $T(x) = \frac{x}{n}$. Dieser Schätzer ist sogar derjenige erwartungstreue Schätzer mit minimaler Varianz. Der **Mittlere Quadratische Fehler** des Schätzers T ist gegeben durch

$$\text{MQF}(T) = \mathbb{E}_\vartheta[|T - \vartheta|^2] = \frac{\vartheta(1 - \vartheta)}{n}.$$

Der Schätzer

$$S(x) = \frac{x + 1}{n + 2}$$

für den Parameter ϑ hat einen kleineren Mittleren Quadratischen Fehler. Zunächst sieht man, dass $S(x) \geq T(x)$ genau wenn $T(x) \leq 1/2$. Folglich ist in diesen Fällen S näher am Zentrum von $[0, 1]$. Der **Bias** von S ist gegeben durch

$$\mathbb{E}_\vartheta[S] - \vartheta = \frac{n\vartheta + 1}{n + 2} - \vartheta = \frac{1 - 2\vartheta}{n + 2},$$

Der Mittlere Quadratische Fehler ist also gegeben durch

$$\text{MQF}(S) = \text{Var}_{\vartheta}(S) + \text{Bias}_{\vartheta}(S) = \frac{n\vartheta(1-\vartheta) + (1-2\vartheta)}{(n+2)^2}.$$

Für ϑ mit $|\vartheta - 1/2| \leq 1/\sqrt{8}$ ist der MQF von S kleiner als der von T .

Nun befassen wir uns mit einem sehr weitverbreiteten Verfahren zur Konstruktion von Schätzern, der sogenannten Maximum-Likelihood-Methode. Maximum-Likelihood-Schätzer besitzen nicht immer alle wünschenswerten Eigenschaften, sind aber in einem asymptotischen Sinne meistens gut und in algorithmischer Weise auf große Modellklassen anwendbar.

Definition 3.14. Es sei X eine Grundgesamtheit und seien X_1, \dots, X_n eine mathematische Stichprobe vom Umfang n . Sei weiter $f(x_i, p)$ die Dichte der Zufallsvariable X_i .

Dann ist die **Likelihood-Funktion** $L(x, p)$ für $x = (x_1, \dots, x_n)$ wie folgt definiert

$$L(x, p) = L_x(p) = \prod_{k=1}^n f(x_k, p).$$

Man beachte, dass die Likelihood-Funktion eine Funktion in den Parametern ist.

Definition 3.15. Der Schätzer $T(X_1, \dots, X_n)$ heißt Maximum-Likelihood-Schätzer für einen Parameter p und eine konkrete Stichprobe x_1, \dots, x_n , falls für alle möglichen Werte des Parameters gilt

$$L(x_1, \dots, x_n, p) \leq L(x_1, \dots, x_n, T(x_1, \dots, x_n)).$$

Wir maximieren bei gegebenen Daten die Likelihood-Funktion im Parameter.

Zusammenfassend sei also nochmals explizit angemerkt, dass man eine **Maximum-Likelihood-Schätzung** in zwei Schritten vollzieht:

- Aufstellen einer zum statistischen Modell gehörenden Likelihood-Funktion.
- Maximierung dieser Funktion.

Der Maximum-Likelihood-Schätzer ist also derjenige Parameterwert, der zur Verteilung gehört, bzgl. der die gegebenen Daten so wahrscheinlich wie möglich sind.

Um mit den bisher eingeführten Konzepten vertrauter zu werden, untersuchen wir zunächst das folgende Beispiel.

Beispiel 3.16. Wir untersuchen die Lebensdauer eines technischen Produkts, wobei zum Beispiel wegen der experimentell untersuchten Gedächtnislosigkeit angenommen wird, dass die Lebensdauer exponentialverteilt ist mit unbekanntem Parameter $\lambda > 0$.

Das Experiment liefert Ihnen eine konkrete Stichprobe vom Umfang n : $(x_1, \dots, x_n) \in (0, \infty)^n$.

Wir bestimmen einen Schätzer über die Maximum-Likelihood-Methode: Für $x = (x_1, \dots, x_n)$ ist die Likelihood-Funktion gegeben durch

$$L(x, \lambda) = \mathbb{1}_{[0, \infty)}(x_1) \cdots \mathbb{1}_{[0, \infty)}(x_n) \cdot \lambda^n e^{-\lambda x_1} \cdots e^{-\lambda x_n}$$

Die Maximum-Likelihood-Methode besagt, dass wir die Likelihood-Funktion im Parameter λ maximieren. Betrachte $x_1, \dots, x_n > 0$ stattdessen die Log-Likelihood-Funktion

$$\log L(x, \lambda) = n \cdot \log(\lambda) - \lambda \cdot \sum_{i=1}^n x_i.$$

Wir bestimmen einen Kandidaten durch

$$\frac{d}{d\lambda} \log(L(x, \lambda)) = n \cdot \frac{1}{\lambda} - \sum_{i=1}^n x_i \implies \frac{d}{d\lambda} \log(L(x, \lambda)) = 0 \iff \lambda_{\text{ML}} = \frac{1}{\frac{1}{n} \sum_{i=1}^n x_i}$$

Leicht ist zu sehen, dass λ_{ML} ein Maximum ist und damit erhalten wir

$$T_{\text{ML}}(X_1, \dots, X_n) = \frac{1}{\frac{1}{n} \sum_{i=1}^n X_i}$$

Beispiel 3.17. Es soll die Wahrscheinlichkeit $p = \mathbb{P}(\{\text{Kopf}\})$ beim Münzwurf ermittelt werden. Dazu wird die Münze 6-mal geworfen.

Sei

$$X_i = \begin{cases} 1 & \text{, wenn im } i\text{-ten Wurf „Kopf“} \\ 0 & \text{, wenn im } i\text{-ten Wurf „Zahl“} \end{cases}$$

$$\begin{aligned} L(x_1, \dots, x_n, p) &= \mathbb{P}(\{X_1 = x_1\} \cap \dots \cap \{X_n = x_n\}) = \mathbb{P}(\{X_1 = x_1\}) \cdots \mathbb{P}(\{X_n = x_n\}) \\ &= p^{\sum_{i=1}^6 x_i} \cdot (1-p)^{6-\sum_{i=1}^6 x_i} \end{aligned}$$

Die Beobachtungen seien

$$1 \quad 1 \quad 0 \quad 1 \quad 0 \quad 1.$$

Die Likelihoodfunktion ist dann durch

$$L(1, 1, 0, 1, 0, 1, p) = p^4 \cdot (1-p)^2$$

gegeben. Wir erhalten zum Beispiel:

p	0,1	0,2	0,3	...	0,5	...	0,8	0,9
$L(1, 1, 0, 1, 0, 1, p)$	0,000081	0,001024	...		0,015625	...	0,016384	0,006561

Zur Bestimmung des ML-Schätzers müssen wir die Abbildung $p \mapsto L(x_1, \dots, x_n, p)$ auf Maxima untersuchen. Statt einer Analyse der Abbildung $p \mapsto L(x_1, \dots, x_n, p)$ ist es oft bequemer die Abbildung

$$p \mapsto \log \left(L(x_1, \dots, x_n, p) \right)$$

zu untersuchen:

$$\log L(1, 1, 0, 1, 0, 1, p) = 4 \cdot \log(p) + 2 \cdot \log(1-p).$$

Zur Bestimmung von Kandidaten für Extremalstellen bestimmen wir die Ableitung und setzen diese Null:

$$\frac{d}{dp} \log L(1, 1, 0, 1, 0, 1, p) = \frac{4}{p} - \frac{2}{1-p} \stackrel{!}{=} 0$$

Dann ergibt sich als Kandidat

$$\frac{4}{p} - \frac{2}{1-p} = 0 \iff 4(1-p) = 2p \iff p = \frac{2}{3} = \frac{4}{6}$$

Leicht überzeugt man sich davon, dass es sich um eine Maximalstelle handelt.

Beispiel 3.18. An die folgenden 10 Beobachtungen soll eine Poissonverteilung angepasst werden

15 14 19 20 23 25 24 11 15 18

Für die Poissonverteilung mit Parameter $\lambda > 0$ gilt: $\mathbb{P}(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$

Bestimmung eines Schätzers mit der ML-Methode: Gegeben sei eine Stichprobe $x = (x_1, \dots, x_n)$. Wir stellen zunächst die Likelihood-Funktion auf.

$$L(x, \lambda) = \prod_{i=1}^n \left(\frac{\lambda^{x_i} e^{-\lambda}}{(x_i)!} \right)$$

$$\log(L(x, \lambda)) = \left(\sum_{i=1}^n x_i \right) \cdot \log(\lambda) - \left(\sum_{i=1}^n \log((x_i)!) \right) - n\lambda$$

Bestimmung eines Kandidaten für Maximum:

$$\frac{d}{d\lambda} \log(L(x, \lambda)) = \frac{\sum_{i=1}^n x_i}{\lambda} - n \stackrel{!}{=} 0$$

$$\implies \lambda_{\text{ML}} = \frac{1}{n} \cdot \sum_{i=1}^n x_i.$$

Es ist wiederum leicht zu sehen, dass in der Tat ein Maximum gefunden wurde. Im konkreten Fall

$$\lambda_{\text{ML}}(15, 14, 19, 20, \dots, 18) = \frac{1}{10} \cdot (15 + 14 + \dots + 18) = 18,4.$$

Beispiel 3.19 (Taxiproblem).

In einer großen Stadt gibt es N Taxen, die die Nummern $\{1, \dots, N\}$ tragen. Ein Passant steht an einer vielbefahrenen Straße und beobachtet die Nummern x_1, \dots, x_n der vorbeifahrenden Taxen, wobei Wiederholungen ignoriert werden. Aufgrund dieser Daten soll die unbekannte Anzahl N der Taxen geschätzt werden. Unser Wahrscheinlichkeitsmodell sieht wie folgt aus:

$$\Omega = \left\{ \{x_1, \dots, x_n\} : x_i \in \{1, \dots, N\}, x_i \neq x_j \text{ für } i \neq j \right\}$$

$$\mathbb{P}(\{x_1, \dots, x_n\}) = \frac{1}{|\Omega|} = \binom{N}{n}^{-1}$$

Klar ist, dass $N \geq \max(x_1, \dots, x_n)$.

Beachte, dass $\binom{N}{n}^{-1}$ größer wird, wenn N kleiner wird \implies

$$T_{\text{ML}}(x_1, \dots, x_n) = \max(x_1, \dots, x_n)$$

Beispiel 3.20. Die Grundgesamtheit X sei normalverteilt mit unbekanntem Erwartungswert μ und bekannter Varianz $\sigma = 1$. Für eine konkrete Stichprobe (x_1, \dots, x_n) vom Umfang n gilt:

$$L(x_1, \dots, x_n, \mu) = \left(\frac{1}{\sqrt{2\pi}} \right)^n \prod_{i=1}^n e^{-\frac{(x_i - \mu)^2}{2}}$$

Zur Bestimmung des ML-Schätzers betrachten wir:

$$\begin{aligned} \log(L(x_1, \dots, x_n, \mu)) &= -n \log(\sqrt{2\pi}) + \sum_{i=1}^n \left(-\frac{(x_i - \mu)^2}{2} \right) \\ \frac{d}{d\mu} \log(L(x_1, \dots, x_n, \mu)) &= \sum_{i=1}^n (x_i - \mu) \stackrel{!}{=} 0 \iff \mu = \frac{1}{n} \cdot \sum_{i=1}^n x_i \end{aligned}$$

Weiter gilt

$$\begin{aligned} \frac{d^2}{d\mu^2} \log(L(x_1, \dots, x_n, \mu)) &= -n \\ \implies \mu_{\text{ML}} &= \frac{1}{n} \cdot \sum_{i=1}^n x_i \text{ ist Maximum.} \end{aligned}$$

Bemerkung 3.21. Die Differenz zwischen Schätzwert und dem wahren Parameter der Grundgesamtheit kann groß sein. Ist ein Schätzer konsistent, so weiß man zumindest, dass die Wahrscheinlichkeit für Abweichungen vom wahren Parameter für größer werdenden Umfang der Stichprobe gegen 0 konvergiert.

Definition 3.22. Es sei X_1, \dots, X_n eine mathematische Stichprobe vom Umfang n der Grundgesamtheit X . Sei $\alpha \in (0, 1)$ vorgegeben. Ein Intervall $(G_u^p(X_1, \dots, X_n), G_o^p(X_1, \dots, X_n))$ heißt **Konfidenzintervall** für den Parameter p mit Irrtumswahrscheinlichkeit α , wenn

$$\mathbb{P}(p \in (G_u^p(X_1, \dots, X_n), G_o^p(X_1, \dots, X_n))) \geq 1 - \alpha$$

gilt. Dabei sind $G_u^p(X_1, \dots, X_n)$ und $G_o^p(X_1, \dots, X_n)$ Schätzer für die linke bzw. rechte Intervallgrenze sind.

Einseitige Konfidenzintervalle

$$(G_u^p(X_1, \dots, X_n), \infty) \quad \text{bzw.} \quad (-\infty, G_o^p(X_1, \dots, X_n))$$

für Parameter p und Irrtumswahrscheinlichkeit α erfüllen

$$\mathbb{P}(p \in (G_u^p(X_1, \dots, X_n), \infty)), \quad \mathbb{P}(p \in (-\infty, G_o^p(X_1, \dots, X_n))) \geq 1 - \alpha.$$

Bemerkung 3.23. Typische Werte für α sind: $\alpha = 0,001; 0,005; 0,01; 0,05$

α klein	große Sicherheit, dass das Intervall den Parameter enthält	Intervall sehr groß, d. h. ungenaue Eingrenzung vom Parameter
α groß	kleine Sicherheit, dass das Intervall den Parameter enthält	Intervall klein, d. h. gute Eingrenzung des Parameters

Definition 3.24. Es sei X eine stetige Zufallsvariable mit Dichte f_X und Verteilungsfunktion F_X . Sei $\gamma \in (0, 1)$. Dann heißt eine Zahl x_γ mit

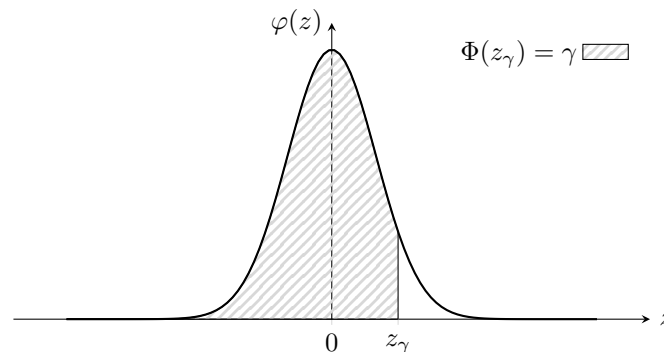
$$F_X(x_\gamma) = \gamma$$

γ -Quantil von X bzw. F_X .

Definition 3.25. Für eine Standardnormalverteilung bezeichnet z_γ das γ -Quantil, wenn

$$\Phi(z_\gamma) = \gamma$$

Anschaulich:



Satz 3.26. Es sei X normalverteilt mit unbekanntem Erwartungswert μ und bekannter Varianz σ^2 . Sei X_1, \dots, X_n eine mathematische Stichprobe vom Umfang n für die Grundgesamtheit X und sei $\alpha \in (0, 1)$ gegeben. Dann ist

$$\left(\bar{X} - \frac{z_{(1-\frac{\alpha}{2})} \cdot \sigma}{\sqrt{n}}, \bar{X} + \frac{z_{(1-\frac{\alpha}{2})} \cdot \sigma}{\sqrt{n}} \right)$$

ein α -Konfidenzintervall für den Parameter μ .

BEWEIS. Wir wissen bereits, dass $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ ein erwartungstreuer Schätzer für den Parameter μ ist. Außerdem folgt aus den Resultaten in Kapitel 2, dass $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$.

Es ist naheliegend ein Konfidenzintervall zu suchen, das symmetrisch um den erwartungstreuen Schätzer \bar{X} für den Parameter μ ist. Wir setzen

$$Z := \sqrt{n} \cdot \frac{\bar{X} - \mu}{\sigma} \sim \mathcal{N}(0, 1) \quad (*).$$

Für Z suche c , so dass

$$\mathbb{P}(-c \leq Z \leq c) \stackrel{!}{=} 1 - \alpha$$

Also

$$\mathbb{P}(-c \leq Z \leq c) = \mathbb{P}(Z \leq c) - \mathbb{P}(Z \leq -c) = F_Z(c) - F_Z(-c) = \Phi(c) - \Phi(-c) \stackrel{!}{=} 1 - \alpha$$

Die Symmetrie der Normalverteilung liefert

$$\Phi(-c) = 1 - \Phi(c)$$

Also erhalten wir $\Phi(c) - \Phi(-c) = 2\Phi(c) - 1 \stackrel{!}{=} 1 - \alpha \implies \Phi(c) = 1 - \frac{\alpha}{2}, c = z_{(1-\frac{\alpha}{2})}$.

Weiter gilt

$$\begin{aligned} 1 - \alpha &= \mathbb{P}(-c \leq -Z \leq c) = \mathbb{P}\left(-c \leq -\left(\sqrt{n} \cdot \frac{\bar{X} - \mu}{\sigma}\right) \leq c\right) \\ &\stackrel{(*)}{=} \mathbb{P}\left(\bar{X} - \frac{c\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{c\sigma}{\sqrt{n}}\right) \\ &= \mathbb{P}\left(\mu \in \left(\bar{X} - \frac{c\sigma}{\sqrt{n}}, \bar{X} + \frac{c\sigma}{\sqrt{n}}\right)\right). \end{aligned}$$

Damit ist die Behauptung gezeigt.

Beispiel 3.27. Es soll ein Intervall bestimmt werden, sodass die mittlere Reißfestigkeit von Bremsseilen mit Wahrscheinlichkeit 0,95 im Intervall liegt. Wir nehmen an, dass die Reißfestigkeit normalverteilt ist bei bekannter Standardabweichung 10 N/mm^2 . Es ergeben sich 20 Messwerte:

$$95,5 \quad 79,1 \quad 79,4 \quad \dots \quad 82,9 \quad \dots \quad 77,8$$

Es gilt

$$\bar{X} = \frac{1}{20} \cdot (95,5 + 79,1 + 79,4 + \dots + 82,9 + \dots + 77,8) = 80,59.$$

$$z_{(1-\frac{0,05}{2})} = z_{0,975} = 1,96, \sigma = 10$$

Damit ergibt sich das folgende Konfidenzintervall:

$$\left(80,59 - 1,96 \cdot \frac{10}{\sqrt{20}}, 80,59 + 1,96 \cdot \frac{10}{\sqrt{20}}\right) = (76,2, 84,97)$$

Beispiel 3.28. Vom Standpunkt eines Qualitätskontrolleurs stelle man sich vor, dass die Bremsseile in den Verkauf gehen, wenn 95 % der Seile eine Reißfestigkeit $\geq 75 \text{ N/mm}^2$ besitzen. Dann bietet sich der Einsatz eines einseitigen Konfidenzintervalls an. Analoges Vorgehen wie bei zweiseitigen Konfidenzintervall:

Das Intervall

$$\left(\bar{X} - z_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}, \infty\right)$$

ist ein einseitiges Konfidenzintervall mit Irrtumswahrscheinlichkeit α . Daten wie in Beispiel 3.26 und $\alpha = 0,05 \implies (75,39, \infty)$

Definition 3.29. Es sei X eine normalverteilte Grundgesamtheit und sei X_1, \dots, X_n eine mathematische Stichprobe vom Umfang n . Dann ist $S^2 = \frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X})^2$ ein erwartungstreuer Schätzer für die Varianz.

Die Verteilung der Zufallsvariable

$$T_{n-1} := \frac{\bar{X} - \mu}{S} \cdot \sqrt{n}$$

heißt **Studentverteilung** oder **t-Verteilung** mit $n - 1$ Freiheitsgraden. Die Dichte der t-Verteilung mit $n - 1$ Freiheitsgraden ist durch

$$f_{n-1}(x) := \frac{\Gamma(\frac{n}{2})}{\sqrt{n\pi} \Gamma(\frac{n-1}{2})} \left(1 + \frac{x^2}{n}\right)^{-\frac{n}{2}}$$

gegeben. Dabei ist

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt$$

die Gamma-Funktion.

Die Dichte der Studentverteilung ist symmetrisch um den Ursprung und approximiert für große Werte von n die Normalverteilung.

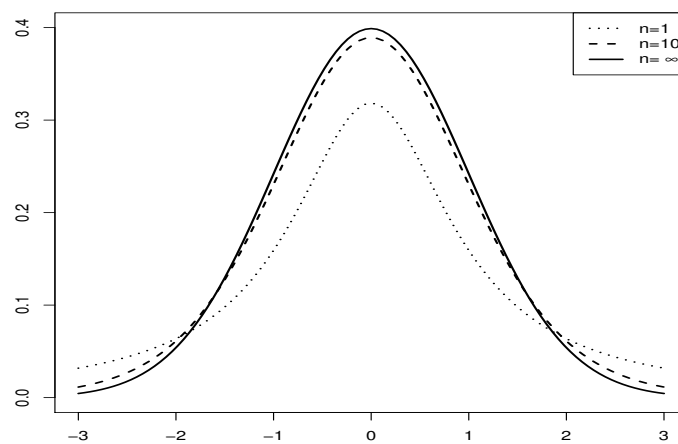


Abbildung 3.1: Dichte der t -Verteilung für $n = 1$, und $n = 10$. Für große Werte von n kann die Dichte der t -Verteilung gut durch die Normalverteilung approximiert werden

Satz 3.30. Es sei X eine normalverteilte Grundgesamtheit und sei X_1, \dots, X_n eine mathematische Stichprobe vom Umfang n . Bezeichnet $t_{1-\frac{\alpha}{2}, n-1}$ das $(1 - \frac{\alpha}{2})$ -Quantil einer studentverteilten Zufallsvariable mit $(n - 1)$ Freiheitsgraden, so ist

$$\left(\bar{X} - t_{1-\frac{\alpha}{2}, n-1} \cdot \frac{S}{\sqrt{n}}, \bar{X} + t_{1-\frac{\alpha}{2}, n-1} \cdot \frac{S}{\sqrt{n}} \right)$$

ein Konfidenzintervall mit Irrtumswahrscheinlichkeit α für den Parameter $\mathbb{E}[X] = \mu$.

Einseitige Konfidenzintervalle sind

$$\left(\bar{X} - t_{1-\alpha, n-1} \cdot \frac{S}{\sqrt{n}}, \infty \right) \quad \text{bzw.} \quad \left(-\infty, \bar{X} + t_{1-\alpha, n-1} \cdot \frac{S}{\sqrt{n}} \right)$$

Beispiel 3.31. Um die Qualität einer Abfüllanlage zu beschreiben, soll ein Intervall angegeben werden, das mit Wahrscheinlichkeit $1 - \alpha = 0,95$ den Erwartungswert der Grundgesamtheit enthält. Es soll angenommen werden, dass die Füllmenge X normalverteilt ist.

Es liegt folgende konkrete Stichprobe vor:

549 547 549 549 545 550 550 545 550 544 543 549 548

Berechne

$$\bar{X} = \frac{1}{13}(549 + 547 + \dots + 548) = 547,54$$

$$S^2 = \frac{1}{12}((549 - 547,54)^2 + \dots + (548 - 547,54)^2) \approx 2,179^2$$

\Rightarrow Tabellarisch erhalten wir das Konfidenzintervall $(546,05, 549,03)$.

Bemerkung 3.32. Bezeichnet man $\chi_{1-\frac{\alpha}{2}, n-1}^2$ bzw. $\chi_{\frac{\alpha}{2}, n-1}^2$ die $(1 - \frac{\alpha}{2})$ - bzw. $\frac{\alpha}{2}$ -Quantile der sog. χ^2 -Verteilung mit $n - 1$ Freiheitsgraden. Dann ist

$$\left(\frac{(n-1)S^2}{\chi_{1-\frac{\alpha}{2}, n-1}^2}, \frac{(n-1)S^2}{\chi_{\frac{\alpha}{2}, n-1}^2} \right)$$

ein Konfidenzintervall für die Varianz.

Beispiel 3.33 (Konstruktion eines Konfidenzintervall mit Chebyshev-Ungleichung).

Anhand einer konkreten Fragestellung werden wir nun ein weiteres Verfahren zur Konstruktion von Konfidenzintervallen im Binomialmodell vorstellen. Ein Reißnagel wird mehrmals hintereinander geworfen. Wir stellen uns die Frage, mit welcher Wahrscheinlichkeit $\vartheta \in (0, 1)$ er auf die Spitze fällt? Die Anzahl der Versuche, bei denen er auf die Spitze stellt, wird als binomialverteilt angenommen. Es sei X Bernoulli-verteilt mit Parameter ϑ . Weiter sei X_1, \dots, X_n eine mathematische Stichprobe vom Umfang n zur Grundgesamtheit X . Ein Schätzer für ϑ ist

$$T(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n X_i.$$

Für die Konstruktion eines Konfidenzintervalls machen wir folgenden natürlichen Ansatz

$$\left[T(X_1, \dots, X_n) - \varepsilon, T(X_1, \dots, X_n) + \varepsilon \right]$$

Bei vorher fixierter Irrtumswahrscheinlichkeit α fordern wir nun, dass

$$\mathbb{P} \left(\left| \vartheta - \frac{\sum_{i=1}^n X_i}{n} \right| > \varepsilon \right) \leq \alpha.$$

Mit Hilfe der Chebyshev-Ungleichung erhalten wir

$$\mathbb{P} \left(\left| \vartheta - \frac{\sum_{i=1}^n X_i}{n} \right| > \varepsilon \right) \leq \frac{\text{Var}(\text{Bin}(\vartheta, n))}{n^2 \cdot \varepsilon^2} = \frac{\vartheta(1-\vartheta)}{n \cdot \varepsilon^2} \leq \frac{1}{4 \cdot n \cdot \varepsilon^2}$$

Da wir ϑ nicht kennen, sondern schätzen wollen, maximieren wir die rechte Seite in ϑ :

$$f(\vartheta) = \vartheta \cdot (1 - \vartheta), \quad f'(\vartheta) = 1 - 2\vartheta, \quad f'(\vartheta) \stackrel{!}{=} 0 \iff \vartheta = \frac{1}{2}$$

$$f''(\vartheta) = -2 \implies \vartheta = \frac{1}{2} \text{ Maximum}$$

Wählen wir nun ε mit der Eigenschaft, dass

$$4 \cdot n \cdot \varepsilon^2 \cdot \alpha \geq 1$$

dann können wir folgern, dass

$$]\bar{X} - \varepsilon, \bar{X} + \varepsilon[$$

ein Konfidenzintervall für ϑ mit Irrtumswahrscheinlichkeit α definiert., ist

$$\text{z. B. } n = 1000, \alpha = 0,025 \implies \varepsilon = \frac{1}{\sqrt{100}} = 0,1.$$

Beispiel 3.34. In einem Transaktionssystem für Flugbuchungen können angefangene Transaktionen ohne Durchführung einer Buchung abgebrochen werden. Wir gehen davon aus, dass vorzeitige Abbrüche voneinander unabhängig erfolgen. Wie groß ist die Wahrscheinlichkeit p einer abgebrochenen Transaktion?

Stichprobenvariablen X_1, \dots, X_n :

$$X_i = \begin{cases} 1 & \text{wenn } i\text{-te Transaktion abgebrochen} \\ 0 & \text{sonst} \end{cases}$$

$H_n = X_1 + \dots + X_n$ ist die absolute Anzahl der abgebrochenen Transaktionen.

$h_n = \frac{H_n}{n} = \frac{1}{n} \sum_{i=1}^n X_i$ ist ein Schätzer für p .

$Z = \frac{H_n - np}{\sqrt{n \cdot (1-p)p}} = \frac{h_n - p}{\sqrt{\frac{(1-p)p}{n}}}$ ist nach dem zentralen Grenzwertsatz für große n approximativ standardnormalverteilt.

$\tilde{Z} = \frac{h_n - p}{\sqrt{\frac{h_n(1-h_n)}{n}}}$ ist typischerweise für große n approximativ standardnormalverteilt.

Konfidenzintervall (approximativ):

$$\left[h_n - z_{1-\frac{\alpha}{2}} \sqrt{\frac{h_n(1-h_n)}{n}}, h_n + z_{1-\frac{\alpha}{2}} \sqrt{\frac{h_n(1-h_n)}{n}} \right]$$

3.1.1 Ausblick: Bayes Schätzer

Wir geben in diesen Abschnitt nur einen kurzen Einblick in die sogenannte Bayes Statistik. Dieser statistische Ansatz spielt heutzutage in vielen Bereichen wie zum Beispiel dem Maschinellen Lernen eine fundamentale Rolle. Im Rahmen dieser Vorlesung können wir nur einen kleinen Einblick in die zugrundeliegende Methodik geben und verweisen für weiterführende Betrachtungen auf die Bücher von Tschirk und Efron/Hastie. Das Buch *Computer Age Statistical Inference* der beiden weltweit führenden Statistiker Bradley Efron und Trevor Hastie ist eine besonders empfehlenswerte Lektüre für alle, die sich eingehender mit wichtigen (computer-gestützten) statistischen Verfahren bis hin zu modernen Entwicklungen vertraut machen wollen. Wir starten wieder mit einem Beispiel

Beispiel 3.35. Ein Kunde möchte sein Auto versichern. Sein persönlicher Fahrstil führt mit Wahrscheinlichkeit $\vartheta \in [0, 1]$ zu mindestens einer Schadensmeldung im Jahr führt. Die Wahrscheinlichkeit ϑ ist dem Versicherungsunternehmen natürlich nicht bekannt, jedoch existieren natürlich ausführliche Statistiken bzgl. der gesamten Population. Diese werden zu einer Verteilung \mathbb{P} auf dem Parameterraum $\Theta = [0, 1]$ und wir nehmen der Einfachheit halber an, dass die Verteilung \mathbb{P} eine Dichte $\rho(\vartheta)$ besitzt. Das Versicherungsunternehmen geht also davon aus, dass die Wahrscheinlichkeit dafür, dass der Kunde in k von n Jahren Ansprüche geltend macht durch

$$\int_0^1 d\vartheta \rho(\vartheta) \binom{n}{k} \vartheta^k (1 - \vartheta)^{n-k}$$

gegeben ist. Die angegebene Wahrscheinlichkeit wird in zwei Schritten erzeugt. Zunächst zieht man mit Verteilung \mathbb{P} den Parameter ϑ und anschließend gemäß einer Binomialverteilung mit Parameter n und ϑ die Anzahl der Schadensmeldungen. Es sei nun angenommen, dass der Kunde nach n Jahren einen neuen Vertrag aushandeln will. Der Versicherungsmakler weiß dann, dass es in $x \in \{0, \dots, n\}$ Jahren eine Schadensmeldung gab. Es ist dann naheliegend, die sogenannte **Prior**verteilung \mathbb{P} auf den Parametern durch die **Posterior**verteilung mit der Dichte

$$\pi_x(\vartheta) = \frac{\rho(\vartheta) \binom{n}{x} \vartheta^x (1 - \vartheta)^{n-x}}{\int_0^1 \rho(p) \binom{n}{x} \vartheta^p (1 - \vartheta)^{n-p} dp}$$

zu ersetzen. Die angegebene Dichte ist die Dichte der bedingten Verteilung von ϑ unter der Bedingung, dass es x Jahre mit Schadensmeldungen gab. Als Schätzwert kann der Versicherungsmakler nun den Erwartungswert

$$T(x) := \int_{\Theta} \vartheta \cdot \pi_x(\vartheta) d\vartheta.$$

nehmen. Dieser Bayes-Schätzer minimiert

$$\int_{\Theta} \rho(\vartheta) \sum_{x=0}^n \binom{n}{x} \vartheta^x (1 - \vartheta)^{n-x} (T(x) - \vartheta)^2 d\vartheta,$$

das sogenannte Bayes-Risiko.

Ähnlich wie im klassischen oben untersuchten frequentistischen Ansatz legen wir auch in der Bayes Sichtweise eine Familie

$$\mathcal{F} = \{f(\cdot, \vartheta) \mid \vartheta \in \Theta\}$$

zugrunde. Als Statistiker beobachtet man $x \in \mathcal{X}$, das gemäß der Dichte $f(\cdot, \vartheta)$ verteilt ist und zieht Rückschlüsse auf den Parameter ϑ . Betrachten wir als Beispiele die normale Familie

$$f(x, \vartheta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\vartheta)^2}{2}},$$

d.h. der unbekannte Parameter ist der Mittelwert. Als zweites Beispiel betrachte

$$f(x, \vartheta) = e^{-\vartheta} \frac{\vartheta^x}{x!},$$

wobei $x \in \mathcal{X}$ und $\vartheta \in \Theta = (0, \infty)$. Bayes Inferenz benötigt nun die Angabe einer **Prior**verteilung auf den Parametern, gegeben z.B. durch eine Dichte

$$\rho(\vartheta), \vartheta \in \Theta.$$

Dieser Prior stellt die Vorabinformation bzgl. des Parameters ϑ vor der Beobachtung von x dar. Mit Hilfe einer geeigneten Formulierung der Bayes Regel kombiniert man nun die durch ρ gegebene **Prior**information mit den Daten x und erhält die sogenannte **Posterior**verteilung

$$\rho(\vartheta | x) = \frac{\rho(\vartheta) f(x, \vartheta)}{\int_{\Theta} f(x, \vartheta) \rho(\vartheta) d\vartheta}.$$

Ist $\Theta \subset \mathbb{R}$ z.B. ein Intervall so minimiert der Schätzer

$$T(x) := \int_{\Theta} \rho(\vartheta | x) \vartheta d\vartheta$$

den quadratischen Fehler

$$\int_{\Theta} \rho(\vartheta) \int f(x, \vartheta) (T(x) - \vartheta)^2 dx d\vartheta.$$

Zum Abschluß des Ausblicks auf Bayes Verfahren noch ein sehr einfaches Beispiel, das eine grundlegende Schlussweise innerhalb der Bayes Statistik mit Hilfe der Bayes Regel auch hinsichtlich der Bayes Inferenz nochmals auf sehr elementare Weise veranschaulicht.

Beispiel 3.36. Eine Frau ist mit Zwillingen schwanger und stellt sich die Frage, mit welcher Wahrscheinlichkeit die Zwillinge eineiig sind. Medizinische Erfahrung lehrt, dass, ein Drittel der Zwillinge eineiig sind. Der unbekannte Parameter ϑ ist also ein- oder zweieiig. Die Priorverteilung ist $1/3$ und $2/3$. Bei einem Sonogramm zeigt sich, dass bei den Zwillingen dasselbe Geschlecht vorliegt. Bei eineiigen Zwillingen liegt immer dasselbe Geschlecht zugrunde bei zweieiigen nur mit Wahrscheinlichkeit $1/2$. Die Bayes Regel liefert dann

$$\frac{\mathbb{P}(\text{eineiig} | \text{gleiches Geschlecht})}{\mathbb{P}(\text{zweieiig} | \text{gleiches Geschlecht})} = \frac{1/3}{2/3} \cdot \frac{1}{1/2} = 1.$$

Die Posteriorwahrscheinlichkeiten für die beiden Fälle sind somit gleich $1/2$.

Zum Abschluß präsentieren wir weiteres sehr interessantes Beispiel, das wohl eine der ersten Anwendung der unter dem Schlagwort 'empirical Bayes' laufenden Methodik Anwendungen sein dürfte. Die Analyse dieses Beispiels ist anspruchsvoller als die bisherigen, dennoch sollten die grundlegenden Ideen prinzipiell nachvollziehbar sein.

Beispiel 3.37 (Missing-Species-Problem). Der Naturforscher Alexander Corbet sammelte zwei Jahre lang in Malaysia Schmetterling: Von den seltenen Arten (118 an der Zahl) hat Corbet nur ein Exemplar fangen können. Insgesamt 74 Arten fing er zweimal und 44

Arten konnte er dreimal fangen. Von den weniger seltenen Arten hatte Corbet natürlich oft über hundert Exemplare in seiner Sammlung. Interessant sind natürlich die selteneren Arten.

x	1	2	3	4	5	6	7	8	9	10	11	12
y	118	74	44	24	29	22	20	19	20	15	12	14

x	13	14	15	16	17	18	19	20	21	22	23	24
y	6	12	6	9	9	6	10	10	11	5	3	3

Corbet stellte die folgende Frage: Wenn er noch ein Jahr weiter Schmetterlinge fängt, wie viele neue Arten wird er schätzungsweise fangen? Zunächst könnte man davon ausgehen, dass auf diese Frage wohl kaum sinnvolle Antworten existieren. Der Statistiker R. A. Fischer (Begründer der ML-Methode) schlug folgende Lösung vor.

Es sei angenommen, es gibt insgesamt S Arten und x_k , die Anzahl der Arten k , die in einer Zeiteinheit (hier zwei Jahre) gefangen wird, folgt einer Poissonverteilung mit Parameter ϑ_k . Beachte, dass dann für die Tabelleneinträge

$$\{x_k = x\} \quad \text{für } x = 1, \dots, 24$$

gilt. Seit $t = 1/2$ und sei $x_k(t)$ die Anzahl, mit der Art k in der halben Zeiteinheit gefangen wird. Fischer trifft die Modellierungsvoraussetzung

$$x_k(t) \sim \text{Poi}(\vartheta_k, t)$$

unabhängig von x_k . Die Wahrscheinlichkeit, dass eine Art k nicht in der Anfangsperiode, aber in der neuen Periode gefangen wird, beträgt dann also

$$e^{-\vartheta_k}(1 - e^{-\vartheta_k t}).$$

Die erwartete Anzahl der in der neuen Periode neu gefangenen Arten ist dann also

$$E(t) := \sum_{k=1}^S e^{-\vartheta_k}(1 - e^{-\vartheta_k t}).$$

Für das folgende ist es bequem, statt der Summe ein Integral zu schreiben

$$E(t) = S \int_0^\infty e^{-\vartheta}(1 - e^{-\vartheta t})g(\vartheta) d\vartheta,$$

wobei die empirische Dichte g das Gewicht $1/S$ auf jeden der Werte ϑ_k setzt. Entwickelt man $1 - e^{-\vartheta t}$ in eine Potenzreihe, so erhält man

$$E(t) = S \int_0^\infty e^{-\vartheta}(\vartheta t - (\vartheta t)^2/2 + (\vartheta t)^3/3! + \dots)g(\vartheta) d\vartheta.$$

Definiert man $e_x = \mathbb{E}[y_x]$, d.h.

$$e_x = \mathbb{E}[y_x] = \sum_{k=1}^S e^{-\vartheta_k} \vartheta_k^x / x! = S \int_0^\infty [e^{-\vartheta} \vartheta^x / x!]g(\vartheta) d\vartheta.$$

Damit erhalten wir

$$E(t) = e_1 t - e_2 t^2 + e_3 t^3 - \dots$$

Die Werte e_1, e_2, \dots sind nicht bekannt, ersetzt man diese jedoch durch die empirisch ermittelten Werte y_1, y_2, \dots , so erhält man einen rational nachvollziehbare Schätzwert

$$\hat{E}(t) = y_1 t - q_w t^2 + y_3 t^3 - \dots,$$

d.h.

$$\hat{E}(1/2) = 44.2.$$

3.2 Testtheorie

Wir starten diesen Abschnitt mit einem Beispiel, das bereits wesentliche Ideen und Konzepte motivierend beleuchtet.

Beispiel 3.38 (Motivation). Eine Lady behauptet, dass sie – wenn sie Tee probiert, der einen Zusatz Milch enthält – unterscheiden könne, ob zuerst Milch oder zuerst Tee eingegossen worden ist. Der Lady soll n -mal die Aufgabe gestellt werden zwei Tassen, von denen eine vom Typ 1 und eine vom Typ 2 ist, korrekt zu klassifizieren. Die beiden Tassen werden ihr jeweils in einer zufälligen durch Münzwurf bestimmten Reihenfolge gegeben. Damit die Lady unabhängig von früheren Entscheidungen urteilen kann, wird jedes Telexperiment an einem anderen Tag ausgeführt. X sei die Zahl der Tage, an denen sie die beiden Tassen richtig klassifiziert.

Als Modell für diese Versuchsanordnung bietet es sich an, X als binomialverteilt mit Parametern n und p anzunehmen. Die *Nullhypothese* entspricht dem Fall $p = 1/2$ und die Alternative, dass die Lady tatsächlich bessere Erfolgschancen hat, als es dem reinen Zufall entspricht, kann man durch $p > 1/2$ beschreiben. Man nimmt also an, dass die Lady, wenn sie Recht hat, an jedem Tag unabhängig von den anderen Tagen mit der Wahrscheinlichkeit $p > 1/2$ einen Erfolg erzielt.

Das für die Versuchsanordnung gewählte Modell ist durch

$$\mathcal{X} = \{0, 1, \dots, n\}, \quad \Theta = [1/2, 1], \quad \vartheta = p$$

und

$$\mathbb{P}_p(X = x) = b_{n,p}(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

beschrieben. Die Hypothese $H_0 = \{1/2\}$ und die Alternative $H_1 = (1/2, 1]$. Wir sind bereit unsere Nullhypothese zu verwerfen, wenn die Anzahl der Erfolge unwahrscheinlich hoch ist. Unseren Verwerfungsbereich wählen wir also der Form

$$R := \{x \in \mathcal{X} \mid x \geq t\}.$$

Es sei nun angenommen, dass die Wahrscheinlichkeit dafür, die Nullhypothese fälschlicherweise zu verwerfen, höchstens $\alpha = 0.05$ betragen soll. Dann muss bei $n = 5$ notwendigerweise $t = 5$ gesetzt werden, denn es gilt

$$\mathbb{P}_{1/2}(X \geq 4) = \left(\frac{1}{2}\right)^5 + \binom{5}{1} \left(\frac{1}{2}\right)^5 > \alpha$$

und $\mathbb{P}_{1/2}(X \geq 5) \leq \alpha$. Die Hypothese wird also nur verworfen, wenn die Lady alle 5 Paare richtig klassifiziert.

Wurde $\alpha = 0.05$ gewählt und $n = 20$. Es gilt

$$\mathbb{P}_{1/2}(X \geq 14) \sim 0.0577$$

und

$$\mathbb{P}_{1/2}(X \geq 15) \sim 0.0207.$$

$R := \{15, 16, \dots, 20\}$ definiert also einen zulässigen Verwerfungsbereich.

Grundbegriffe der Testtheorie

Von einem Testproblem spricht man, wenn eine zufällige Größe X mit einer unbekannten Verteilung \mathbb{P}_ϑ beobachtet wird, und man aufgrund des beobachteten Wertes x der Zufallsvariablen entscheiden soll, ob \mathbb{P}_ϑ einer bestimmten Menge von Verteilungen angehört oder nicht.

Im Folgenden \mathcal{X} die Menge der möglichen Werte einer Zufallsvariablen X und $(\mathbb{P}_\vartheta)_{\vartheta \in \Theta}$ sei die Menge der in Betracht gezogenen Verteilungen von X . Durch Übergang zum Bildraum können wir immer annehmen, dass X die Identität auf \mathcal{X} ist. Sei Θ die disjunkte Vereinigung von Θ_0 und Θ_1 . Ein **Test** ist eine Entscheidungsregel, die für jeden möglichen Wert x von X festlegt, ob man sich für die **Nullhypothese** $H_0 : \vartheta \in \Theta_0$ oder für die **Alternative** $H_1 : \vartheta \in \Theta_1$ entscheidet. Die Entscheidung für H_0 nennt man Annahme der Hypothese, die Entscheidung für die Alternative nennt man Verwerfen der Hypothese. Ein Test ist also beschrieben durch die Angabe der Menge R derjenigen x , für die die Hypothese verworfen werden soll. R heißt auch **Verwerfungsbereich** oder **kritischer Bereich**.

Es sind zwei Arten von Fehlern möglich:

- **Fehler erster Art:** $\vartheta \in \Theta_0$ und die Hypothese wird verworfen.
- **Fehler zweiter Art:** $\vartheta \in \Theta_1$ und die Hypothese H_0 wird nicht verworfen.

In den Anwendungen gibt man R mit Hilfe einer 'Statistik' $T(x)$ an. Man wählt T oft so, dass besonders große Werte von T gegen die Nullhypothese sprechen. Also

$$R := \{x \mid T(x) > t\}.$$

Wir sagen, dass der Test **Niveau** $\alpha \in (0, 1)$ hat, wenn für alle $\vartheta \in \Theta_0$ gilt

$$\beta(\vartheta) := \mathbb{P}_\vartheta(X \in R) \leq \alpha,$$

die Wahrscheinlichkeit eines Fehlers erster Art ist dann beschränkt durch α . Für $\vartheta \in \Theta_1$ heißt $\beta(\vartheta)$ die **Macht** des Tests in ϑ . Ist die Macht $\beta(\vartheta)$ nahe 1, so ist die Wahrscheinlichkeit $1 - \beta(\vartheta)$ eines Fehlers zweiter Art klein.

Testen im Normalmodell

Wir nehmen an, dass die Grundgesamtheit X normalverteilt ist mit bekannter Varianz σ^2 und unbekanntem Mittelwert und wir untersuchen

$$H_0 : \mathbb{E}[X] = \mu_0 \quad \text{gegen} \quad \mathbb{E}[X] \neq \mu_0.$$

Für eine mathematische Stichprobe X_1, \dots, X_n vom Umfang n ist die Zufallsvariable

$$Z := \frac{\bar{X} - \mu_0}{\sigma} \sqrt{n}, \quad \text{wobei} \quad \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

unter der Annahme von H_0 standardnormalverteilt, d.h. $Z \sim \mathcal{N}(0, 1)$. \bar{X} ist ein erwartungstreuer Schätzer von $\mathbb{E}[X]$. Ist also $|\bar{X} - \mu_0|$ groß, so ist man eher geneigt, die Hypothese H_0 zu verwerfen. Ist das Niveau α spezifiziert, so können wir durch Wahl von

$$R : \{|Z| > z_{1-\alpha/2}\}$$

sicherstellen, dass unter der Annahme von H_0

$$\begin{aligned}\mathbb{P}(\{|Z| \leq z_{1-\alpha/2}\}) &= \Phi(z_{1-\alpha/2}) - \Phi(-z_{1-\alpha/2}) \\ &= 2\Phi(z_{1-\alpha/2}) - 1 = 1 - \alpha.\end{aligned}$$

Die Wahrscheinlichkeit für einen Fehler erster Art ist also

$$\mathbb{P}(\{|Z| > z_{1-\alpha/2}\}) = \alpha.$$

Damit erhalten wir das folgende Rezept.

Rezept 1: Gaußtest

Annahmen: Die Zufallsvariablen X_1, \dots, X_n seien unabhängig und identisch verteilt mit $X_i \sim \mathcal{N}(\mu, \sigma^2)$, wobei σ^2 bekannt sei. Es gelte also $\mathbb{E}[X_i] = \mu$ und $\text{Var}(X_i) = \sigma^2$

Hypothesen:

- a) $H_0 : \mu = \mu_0$ gegen $H_1 : \mu \neq \mu_0$
- b) $H_0 : \mu \geq \mu_0$ gegen $H_1 : \mu < \mu_0$
- c) $H_0 : \mu \leq \mu_0$ gegen $H_1 : \mu > \mu_0$

Teststatistik:

$$Z := \frac{\bar{X} - \mu_0}{\sigma} \sqrt{n},$$

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

Ablehnungskriterium für H_0 bei Niveau α :

- a) $|Z| > z_{1-\frac{\alpha}{2}}$
- b) $Z < z_\alpha$
- c) $Z > z_{1-\alpha}$

Beispiel 3.39. Ein Abfüllautomat füllt immer 500 g in eine Packung. Es wird angenommen, dass die Füllmenge normalverteilt ist, wobei die Varianz $\sigma^2 = 1,5^2$ bekannt ist.

$$H_0 : \mathbb{E}[X] = 500, \quad H_1 : \mathbb{E}[X] \neq 500$$

Es wurde die folgende Stichprobe ermittelt:

499 500 498 500 498 498 498 498 495 499 500 496 499 497 500

Sei $\alpha = 0,01$, $1 - \frac{\alpha}{2} = 0,995$, $z_{1-\frac{\alpha}{2}} = 2,58$.

Damit ist der Akzeptanzbereich

$$I = \left(500 - 2,58 \cdot \frac{1,5}{\sqrt{3,87}}, 500 + 2,58 \cdot \frac{1,5}{\sqrt{3,87}} \right)$$

Prüfe, ob $\bar{X} = \frac{1}{15}(499 + \dots + 500) \in I$.

$$\begin{aligned} Z \in (-z_{1-\frac{\alpha}{2}}, z_{1-\frac{\alpha}{2}}) &\iff \frac{\bar{X} - \mu_0}{\sigma} \cdot \sqrt{n} \in (-z_{1-\frac{\alpha}{2}}, z_{1-\frac{\alpha}{2}}) \\ &\iff \bar{X} \in \left(\mu_0 - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \mu_0 + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right) = I \end{aligned}$$

Ist die Varianz σ^2 unbekannt, so ist es wie im Fall von Konfidenzintervallen naheliegend, einen erwartungstreuen Schätzer der Varianz zu Hilfe zu ziehen. Diese Idee mündet in das folgende Vorgehen.

Rezept 2: t-Test

Annahmen: Die Zufallsvariablen X_1, \dots, X_n seien unabhängig und identisch verteilt mit $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Es gelte also $\mathbb{E}[X_i] = \mu$ und $\text{Var}(X_i) = \sigma^2$. Wir kennen die Varianz nicht.

Hypothesen:

- a) $H_0 : \mu = \mu_0$ gegen $H_1 : \mu \neq \mu_0$
- b) $H_0 : \mu \geq \mu_0$ gegen $H_1 : \mu < \mu_0$
- c) $H_0 : \mu \leq \mu_0$ gegen $H_1 : \mu > \mu_0$

Teststatistik:

$$T := \frac{\bar{X} - \mu_0}{S} \sqrt{n},$$

wobei $\bar{X} := \frac{1}{n}(X_1 + \dots + X_n)$ und $S^2 := \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$.

Ablehnungskriterium für H_0 bei Niveau α :

- a) $|T| > t_{n-1, 1-\frac{\alpha}{2}}$
- b) $T < t_{n-1, \alpha}$
- c) $T > t_{n-1, 1-\alpha}$

Beispiel 3.40. Die Wirkung zweier Schlafmittel A und B sollen verglichen werden. Dazu werden $n = 10$ Patienten in zwei aufeinander folgenden Nächten die Medikamente A und B verabreicht werden und die Differenz der jeweiligen Schlafdauer gemessen. Letztere wird als normalverteilt angenommen mit unbekannten Parametern μ und σ . Es ergibt sich für die Differenz der Schlafdauern:

Patient	1	2	3	4	5	6	7	8	9	10
Differenz	1.2	2.4	1.3	1.3	0.0	1.0	1.8	0.8	4.6	1.4

Wir wählen das Niveau $\alpha = 0.01$. Für den vorliegenden Datensatz ergeben sich

$$\bar{x} = 1.58 \quad S^2 = 1.513.$$

Wir testen die Nullhypothese $H_0 : \mu = 0$ gegen $H_1 : \mu \neq 0$. Für den obigen Datensatz ergibt sich $T(x) = 1.58 \sqrt{\frac{10}{1.513}} = 4.06$. Da $t_{9, 0.995} = 3.25$ gilt wird die Nullhypothese aufgrund der Daten verworfen.

Tests im Binomialmodell

Es sei X_1, \dots, X_n eine mathematische Stichprobe einer zum Parameter p Bernoulliverteilten Grundgesamtheit X vom Umfang n . Wir veranschaulichen mögliche Konstruktionserfahren statistischer Hypothesentests für den Parameter $p \in (0, 1)$

Erinnerung:

Die Anzahl der Erfolge

$$Y = \sum_{i=1}^n X_i$$

ist binomialverteilt mit Parametern n und p , d.h. insbesondere gilt für $k = 0, \dots, n$

$$\mathbb{P}(Y = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

sowie

$$\mathbb{E}[Y] = n \cdot p \quad \text{und} \quad \text{Var}(Y) = n \cdot p \cdot (1-p).$$

Vorgehen:

Das Niveau α sei nun fixiert und im folgenden interessieren wir uns für die Hypothesen

$$H_0 : p \leq p_0 \in (0, 1) \quad \text{gegen} \quad H_1 : p > p_0$$

Unseren Verwerfungsbereich wollen wir in der Form

$$R := \{Y \geq k\}$$

für geeignetes k wählen. Die Intuition dahinter ist natürlich, dass *zu viele* Erfolge nicht gut mit der Nullhypothese verträglich sind.

Wir müssen sicherstellen, dass die Wahrscheinlichkeit für einen Fehler erster Art nicht α überschreitet, d.h. dass

$$\forall p \leq p_0 : \beta(p) = \mathbb{P}_p(Y \geq k) \leq \alpha$$

gilt. Wir haben gesehen, dass die Abbildung

$$(0, 1) \ni p \mapsto \mathbb{P}_p(Y \geq k)$$

monoton wachsend ist. Somit genügt es sicherzustellen, dass

$$\mathbb{P}_{p_0}(Y \geq k) \leq \alpha$$

erfüllt ist. Zur Festlegung des Verwerfungsbereichs R definieren wir also

$$k = k(n, \alpha) = \min\{l = 0, 1, \dots, n \mid \sum_{j=l}^n \binom{n}{j} p_0^j (1-p_0)^{n-j} \leq \alpha\}.$$

Sollte

$$\{l = 0, 1, \dots, n \mid \sum_{j=l}^n \binom{n}{j} p_0^j (1-p_0)^{n-j} \leq \alpha\} = \emptyset$$

gelten, dann setze formal $k = \infty$. In diesem letzten Fall ist der Verwerfungsbereich also die leere Menge und wir verwerfen H_0 nie.

Bemerkung 3.41. By den Hypothesen

$$H_0 : p = p_0 \in (0, 1) \quad \text{gegen} H_1 : p \neq p_0$$

sprechen sowohl zu viele als auch zu wenige Erfolge gegen H_0 . Man wird also als Ablehnungsbereich R einen Ansatz

$$R = \{Y \leq k\} \cup \{Y \geq l\}$$

als sinnvoll erachten können. Die Grenzen l und k müssen dann in Analogie zu obigem Vorgehen ermittelt werden.

Rezept 3: Der Binomialtest

Annahmen: Die Zufallsvariablen X_1, \dots, X_n seien unabhängig und identisch verteilt mit $\mathbb{P}(X_i = 1) = p$ und $\mathbb{P}(X_i = 0) = 1 - p$, wobei p unbekannt sei. Folglich folgt $Y := \sum_{i=1}^n X_i$ einer Binomialverteilung mit Parametern n, p .

Hypothesen:

- a) $H_0 : p = p_0$ gegen $H_1 : p \neq p_0$
- b) $H_0 : 0 \leq p \leq p_0$ gegen $H_1 : 1 \geq p > p_0$
- c) $H_0 : 0 \geq p \geq p_0$ gegen $H_1 : 1 \leq p > p_0$

Teststatistik:

$$Y := \sum_{i=1}^n X_i,$$

wobei $Y := X_1 + \dots + X_n$ die Anzahl der Ereignisse mit $X_i = 1$ bezeichnet.

Ablehnungskriterium für H_0 bei Niveau α :

- a) $\{0, \dots, k, l, \dots, n\}$ mit $\mathbb{P}_{p_0}(X \leq k), \mathbb{P}_{p_0}(X \geq l) \leq \alpha/2$
- b) $\{l, \dots, n\}$ mit $\mathbb{P}_{p_0}(X \geq l) \leq \alpha$
- c) $\{0, \dots, k\}$ mit $\mathbb{P}_{p_0}(X \leq k), \mathbb{P}_{p_0}(X \geq l) \leq \alpha$.

Abschließend halten wir noch folgende Beobachtung fest. Unter den Annahmen dieses Abschnittes zur Tests im Binomialmodell gelten die folgenden Aussagen:

- Ist der Umfang n sehr groß und $p = p_0$, dann ist die Zufallsvariable

$$Z := \frac{Y - np_0}{\sqrt{np_0(1 - p_0)}}$$

als Folgerung aus dem Zentralen Grenzwertsatz approximativ standard normalverteilt.

- Ist der Umfang n sehr groß und $p \neq p_0$, dann ist die Zufallsvariable

$$Z := \frac{Y - np_0}{\sqrt{np_0(1 - p_0)}}$$

als Folgerung aus dem Zentralen Grenzwertsatz approximativ normalverteilt mit Erwartungswert $\sqrt{n} \frac{p-p_0}{\sqrt{p_0(1-p_0)}}$ und Varianz $\frac{p(1-p)}{p_0(1-p_0)}$

Ein Vorgehen wie im Gausstest liefert nun das folgende Rezept.

Rezept 4: Approximativer Binomialtest

Annahmen: Die Zufallsvariablen X_1, \dots, X_n seien unabhängig und identisch verteilt mit $\mathbb{P}(X_i = 1) = p$ und $\mathbb{P}(X_i = 0) = 1 - p$, wobei p unbekannt sei. Wir setzen voraus, dass n hinreichend groß ist, so dass eine Anwendung des Zentralen Grenzwertsatzes hinreichend gute Ergebnisse liefert.

Hypothesen:

- a) $H_0 : p = p_0$ gegen $H_1 : p \neq p_0$
- b) $H_0 : p \geq p_0$ gegen $H_1 : p < p_0$
- c) $H_0 : p \leq p_0$ gegen $H_1 : p > p_0$

Teststatistik:

$$Z := \frac{Y - np_0}{\sqrt{np_0(1-p_0)}},$$

wobei $Y := X_1 + \dots + X_n$ die Anzahl der Ereignisse mit $X_i = 1$ bezeichnet.

Ablehnungskriterium für H_0 bei Niveau α :

- a) $|Z| > z_{1-\frac{\alpha}{2}}$
- b) $Z < z_\alpha$
- c) $Z > z_{1-\alpha}$

Wir möchten abschließend darauf hinweisen, dass noch eine große Anzahl von Testverfahren existieren, aufgrund des einführenden Charakters dieser Vorlesung und der zeitlichen Beschränkungen haben wir uns auf wenige exemplarische Testverfahren reduziert und vor allem das allgemeine Vorgehen bei statistischen Testverfahren in den Mittelpunkt gesetzt. Wir verweisen hierzu exemplarisch auf das Buch *Introduction to Probability and Statistics for Engineers and Scientists* von Sheldon Ross.

Markov-Ketten

Ein stochastischer Prozess in diskreter Zeit ist eine Familie $(X_t)_{t \in T}$ von Zufallsvariablen, wobei X_t für jedes $t \in T$ auf demselben Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ definiert ist und Werte einer Menge E annimmt. Die Zufallsvariable X_t beschreibt anschaulich gesprochen den Zustand des Systems zur Zeit t , also zum Beispiel den Ort eines Teilchens zur Zeit t .

In diesem Kurs ist $T = \mathbb{N}_0$ und E ist höchstens abzählbar. Eine gut analysierbare Klasse von stochastischen Prozessen bilden die Markovketten. Hier hängen die Wahrscheinlichkeiten für den nächsten Zustand des Prozesses nur vom aktuellen Zustand ab, nicht aber von der übrigen Vergangenheit. Markovketten haben also ein kurzes Gedächtnis.

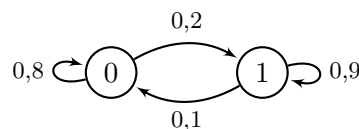
Beispiel 4.1. Wir nehmen an, dass vom Rechner A in festen Zeitschritten Datenpakete an Rechner B übertragen werden, um die Funktionsfähigkeit der Übertragung zu prüfen. Wir interessieren uns für die Zufallsvariable $(t \in \mathbb{N}_0)$

$$X_t = \begin{cases} 1 & , \text{ falls Übertragung in } t\text{-ten Schritt erfolgreich} \\ 0 & , \text{ sonst} \end{cases}.$$

Im Normalfall nehmen wir an, dass Rechner B mit Wahrscheinlichkeit 0,9 erreicht werden kann; wenn im vorherigen Schritt ein Übertragungsfehler vorlag, so setzen wir diese mit 0,2.

$$\begin{aligned} \mathbb{P}(X_{t+1} = 1 \mid X_t = 1) &= 0,9 & \mathbb{P}(X_{t+1} = 1 \mid X_t = 0) &= 0,2 \\ \mathbb{P}(X_{t+1} = 0 \mid X_t = 1) &= 0,1 & \mathbb{P}(X_{t+1} = 0 \mid X_t = 0) &= 0,8 \end{aligned}$$

Veranschaulichung:



Angenommen zur Zeit 0 starten wir in 1. Berechne die Wahrscheinlichkeit

$$\begin{aligned} \mathbb{P}(X_2 = 1 \mid X_0 = 1) &= \mathbb{P}(X_2 = 1, X_1 = 1 \mid X_0 = 1) + \mathbb{P}(X_2 = 1, X_1 = 0 \mid X_0 = 1) \\ &= 0,9 \cdot 0,9 + 0,1 \cdot 0,2 \end{aligned}$$

4.1 Grundlagen

Wir beginnen mit der mathematischen Beschreibung einiger zentraler Grundlagen. Es ist zu erwähnen, dass wir auf die Konstruktion der Markovkette zu einer gegebenen Übergangsmatrix in diesem Rahmen verzichten müssen.

Definition 4.2. Sei $E \neq \emptyset$ eine höchstens abzählbare Menge, eine Matrix $P = (p_{ij})_{i,j \in E}$, $p_{ij} \geq 0$, heißt **stochastische Matrix**, wenn für jedes $i \in E$ gilt

$$\sum_{j \in E} p_{ij} = 1$$

Stochastische Matrizen sind diejenigen Objekte, die die sogenannten Ein-Schritt-Wahrscheinlichkeiten von Markovketten kodieren. Dies wird in der folgenden Definition präzisiert.

Definition 4.3. Eine Folge von Zufallsvariablen $(X_n)_{n \geq 0}$ von Zufallsvariablen auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ mit Werten in einem abzählbaren Zustandsraum E heißt eine **Markovkette** mit Übergangsmatrix $P = (p_{ij})_{i,j \in E}$, wenn für alle $n \geq 0$ und Zustände $i_0, i_1, \dots, i_{n+1} \in E$

$$\begin{aligned} \mathbb{P}(X_{n+1} = i_{n+1} \mid X_0 = i_0, X_1 = i_1, \dots, X_n = i_n) \\ = \mathbb{P}(X_{n+1} = i_{n+1} \mid X_n = i_n) \\ = p_{i_n i_{n+1}}, \end{aligned}$$

sofern $\mathbb{P}(X_0 = i_0, X_1 = i_1, \dots, X_n = i_n) > 0$. Die Verteilung von X_0 heißt Startverteilung α , wobei für $i \in E$ gilt $\alpha(i) = \mathbb{P}(X_0 = i)$. Die obige Eigenschaft wird auch Markoveigenschaft genannt.

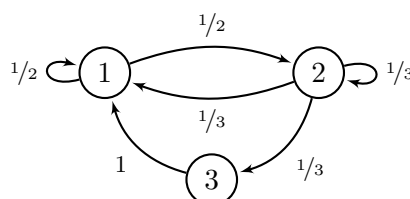
Beispiel 4.4. Betrachte erneut die Markovkette aus Beispiel 4.1. Wir erhalten als Übergangsmatrix zu dieser Markovkette

$$P = \begin{matrix} & \begin{matrix} 0 & 1 \end{matrix} \\ \begin{matrix} 0 \\ 1 \end{matrix} & \begin{pmatrix} 0,8 & 0,2 \\ 0,1 & 0,9 \end{pmatrix} \end{matrix}, \quad \text{Startverteilung } \alpha: \alpha(0) = 0, \alpha(1) = 1.$$

Bemerkung 4.5. Interpretiere X_n als Zustand eines Systems zur Zeit n . Der Prozess $(X_n)_{n \geq 0}$ hat die Markoveigenschaft, wenn die Wahrscheinlichkeit, zur Zeit $n+1$ in einen beliebigen Zustand zu gehen, nur vom Zustand zur Zeit n abhängt.

Eine sinnvolle Art, sich Markovketten zu veranschaulichen, besteht in der Darstellung seiner Übergangsdarstellung als gerichtete Graphen. Wir illustrieren dies im folgenden Beispiel.

Beispiel 4.6. Betrachte die folgende Markovkette. $E = \{1, 2, 3\}$, $P = \begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/3 & 1/3 & 1/3 \\ 1 & 0 & 0 \end{pmatrix}$



Das folgende Beispiel ist das klassische Gambler's ruin setting.

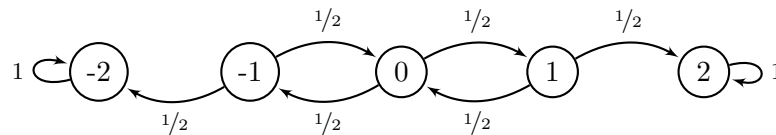
Beispiel 4.7. Zwei Spieler A und B sind im Besitz von a bzw. b Euro. Sie werfen wiederholt eine faire Münze. Je nach Ergebnis zahlt einer der beiden an den anderen 1€. Das Spiel ist beendet, wenn ein Spieler pleite ist.

Sei X_n der Gewinn von Spieler A nach Spielen $1, \dots, n$. Die Zufallsvariablen haben Werte in $E = \{-a, \dots, b\}$.

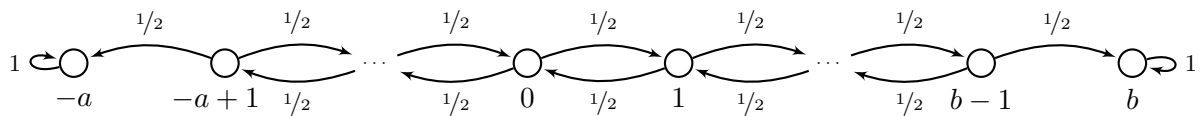
Bilde die Markov-Kette

$$p_{ij} = \begin{cases} 1/2 & \text{falls } -a < i < b, |j - i| = 1 \\ 1 & \text{falls } i = j \in \{-a, b\} \\ 0 & \text{sonst} \end{cases}$$

Für $a = b = 2$



Für a, b beliebig



Satz 4.8 (Markoveigenschaft). Sei $(X_k)_{k \geq 0}$ eine Markovkette mit Zustandsraum E , Übergangsmatrix $P = (p_{ij})_{i,j \in E}$ und Startverteilung α . Dann gilt für alle $0 < n < N$ sowie $A \subseteq E^n$, $B \subseteq E^{N-n}$ und $i_n \in E$

$$\begin{aligned} & \mathbb{P}\left((X_{n+1}, \dots, X_N) \in B \mid (X_0, \dots, X_{n-1}) \in A, X_n = i_n\right) \\ &= \mathbb{P}\left((X_{n+1}, \dots, X_N) \in B \mid X_n = i_n\right), \end{aligned}$$

sofern $\mathbb{P}((X_0, \dots, X_{n-1}) \in A, X_n = i_n) > 0$

Berechnen von Wahrscheinlichkeiten 4.9.

Es seien $i_0, \dots, i_n \in E$ und es sei (X_n) eine Markov-Kette mit Zustandsraum E , Übergangsmatrix $P = (p_{ij})_{i,j \in E}$ und Startverteilung α .

Berechne

$$\mathbb{P}(X_0 = i_0, X_1 = i_1, \dots, X_n = i_n)$$

Angenommen $\mathbb{P}(X_0 = i_0, X_1 = i_1, \dots, X_n = i_n) \neq 0$, dann gilt

$$\begin{aligned}
 & \mathbb{P}(X_0 = i_0, X_1 = i_1, \dots, X_n = i_n) \\
 &= \underbrace{\mathbb{P}(X_n = i_n \mid X_0 = i_0, X_1 = i_1, \dots, X_{n-1} = i_{n-1})}_{\frac{\mathbb{P}(X_n = i_n, X_{n-1} = i_{n-1}, \dots, X_0 = i_0)}{\mathbb{P}(X_{n-1} = i_{n-1}, \dots, X_0 = i_0)}} \\
 & \quad \cdot \underbrace{\mathbb{P}(X_{n-1} = i_{n-1} \mid X_0 = i_0, \dots, X_{n-2} = i_{n-2})}_{\frac{\mathbb{P}(X_{n-1} = i_{n-1}, X_{n-2} = i_{n-2}, \dots, X_0 = i_0)}{\mathbb{P}(X_{n-2} = i_{n-2}, \dots, X_0 = i_0)}} \\
 & \quad \cdots \mathbb{P}(X_1 = i_1 \mid X_0 = i_0) \cdot \mathbb{P}(X_0 = i_0) \\
 &= \alpha(i_0) \cdot p_{i_0 i_1} \cdot p_{i_1 i_2} \cdots p_{i_{n-1} i_n}
 \end{aligned} \tag{*}$$

Durch Summation in (*) über alle möglichen Zustände i_1, \dots, i_{n-1} , erhält man

$$\boxed{\mathbb{P}(X_n = i_n \mid X_0 = i_0) = (P^n)_{i_0 i_n}}$$

In der Theorie der Markovketten steht man häufig vor dem Problem der Berechnung der Wahrscheinlichkeit, ein Gebiet an einer gewissen Stelle zu verlassen. In diesem Abschnitt werde wir einige einfache Aussagen hierzu herleiten.

Definition 4.10. Ein Zustand $z \in E$ heißt absorbierend bzgl. der Übergangsmatrix $P = (p_{ij})_{i,j \in E}$, wenn $p_{zz} = 1$.

In diesem Fall heißt

$$\begin{aligned}
 h_z(x) &= \mathbb{P}(X_n = z \text{ schließlich} \mid X_0 = x) =: \mathbb{P}^x(X_n = z \text{ schließlich}) \\
 &= \mathbb{P}\left(\bigcup_{N \geq 0} \underbrace{\bigcap_{n \geq N} \{X_n = z\}}_{\forall n \geq N: X_n = z}\right) \\
 &= \underbrace{\quad}_{\exists N \forall n \geq N: X_n = z}
 \end{aligned}$$

die Absorptionswahrscheinlichkeit in Zustand z bei Start in Zustand x .

Definition 4.11. Für einen beliebigen Zustand $z \in E$ heißt

$$\tau_z = \inf\{n \geq 1 : X_n = z\}$$

die Eintrittszeit oder Trefferzeit in den Zustand z oder bei Start in z die Rückkehrzeit zu z .

Bemerkung 4.12. Für alle $n \geq 1$ gilt

$$\{\tau_z = n\} = \{(X_0, \dots, X_{n-1}) \in B, X_n = z\},$$

wobei $B = E \times (E \setminus \{z\})^{n-1}$, d. h. $\{\tau_z = n\}$ hängt nur von (X_0, \dots, X_n) ab, nicht aber von der Zukunft nach n . Solche Zeiten heißen **Stoppzeiten**.

Satz 4.13. Für einen absorbierenden Zustand $z \in E$ und alle Startwerte $x \in E$ gilt

$$h_z(x) = \mathbb{P}^x(\tau_z < \infty) = \lim_{n \rightarrow \infty} \mathbb{P}^x(X_n = z)$$

und h_z ist die kleinste nicht-negative Funktion $E \rightarrow [0, 1]$ mit

$$h_z(z) = 1 \quad \text{und} \quad \sum_{y \in E} p_{xy} \cdot h_z(y) = h_z(x) \quad \forall x \in E$$

Denkt man sich also h_z als Spaltenvektor, dann gilt also $P \cdot h_z = h_z$. Also beschreibt h_z einen Eigenvektor der stochastischen Matrix P zum Eigenwert 1.

BEWEIS. Die erste Aussage ist intuitiv klar. Offenbar gilt $h_z(z) = 1$, $h_z(x) \geq 0 \forall x \in E$.

$$\begin{aligned} \sum_{y \in E} p_{xy} \cdot h_z(y) &= \sum_{y \in E} \mathbb{P}^x(X_1 = y) \cdot \mathbb{P}^y(\{X_n = z \text{ schließlich}\}) \\ &\stackrel{\text{Markov-Eigenschaft}}{=} \sum_{y \in E} \mathbb{P}(X_1 = y) \cdot \mathbb{P}^x(X_{n+1} = z \text{ schließlich} \mid X_1 = y) \\ &\stackrel{\text{totale W-keit}}{=} h_z(x) \end{aligned}$$

Für jede weitere Funktion $h \geq 0$ mit $h(z) = 1$ und $\underbrace{\sum_{y \in E} p_{xy} \cdot h_z(y)}_{Ph=h, \dots, P^n h=h} = h_z(x)$

$$h(x) = (P^n h)(x) \geq (P^n)_{xz} = \mathbb{P}^x(X_n = z) \xrightarrow{n \rightarrow \infty} h_z(x) \implies h \geq h_z \quad \square$$

Beispiel 4.14 (Münzwurfbeispiel aus 4.10).

Untersuche „Ruinwahrscheinlichkeit“

$$r_A := h_{-a}(0) = \mathbb{P}^0(\tau_{-a} < \infty)$$

Für $-a < x < b$ gilt:

$$\begin{aligned} h_{-a}(x) &= (Ph_{-a})(x) = \frac{1}{2}h_{-a}(x-1) + \frac{1}{2}h_{-a}(x+1) \\ \iff \frac{1}{2}h_{-a}(x) + \frac{1}{2}h_{-a}(x) &= \frac{1}{2}h_{-a}(x-1) + \frac{1}{2}h_{-a}(x+1) \\ \iff \frac{1}{2}h_{-a}(a) - \frac{1}{2}h_{-a}(x-1) &= \frac{1}{2}h_{-a}(x+1) - \frac{1}{2}h_{-a}(x) \end{aligned}$$

\implies der Wert $c := h_{-a}(x+1) - h_{-a}(x)$ ist also unabhängig von x ab und wegen

$$h_{-a}(-a) = 1, h_{-a}(b) = 0$$

folgt

$$h_{-a}(x) - h_{-a}(-a) = (x+a) \cdot c \quad \wedge \quad c = \frac{1}{a+b}$$

Also gilt: $r_A = h_{-a}(0) = 1 - \frac{a}{a+b} = \frac{b}{a+b}$.

Bemerkung 4.15. Sei $(X_n)_{n \in \mathbb{N}}$ eine Markov-Kette mit Zustandsraum E und Übergangsmatrix $P = (p_{ij})_{i,j \in E}$ und sei $B \subseteq E$

$$\sigma_B = \min\{n \geq 0 : X_n \in B\}$$

Wir setzen $\sigma_B = \infty$, falls $\{n \geq 0 : X_n \in B\} = \emptyset$. Weiterhin sei für $x \in E$

$$e_B(x) = \mathbb{E}^x[\sigma_B] = \sum_{j=0}^{\infty} j \cdot \mathbb{P}^x(\sigma_B = j) + \infty \cdot \mathbb{P}(\sigma_B = \infty).$$

Proposition 4.16. Für $B \subseteq E$ ist e_B die kleinste nicht-negative Lösung von

$$\begin{aligned} \forall x \in E \setminus B: e_B(x) &= 1 + (Pe_B)(x) \\ \forall x \in B: e_B(x) &= 0 \end{aligned} \quad (*)$$

BEWEIS. Wir beachten, dass für $x \in B$ nach Definition von σ_B natürlich $e_B(x) = 0$ gilt. Weiter gilt mit Hilfe einer Ein-Schritt-Analyse

$$\begin{aligned} e_B(x) &= \mathbb{E}_x[\sigma_B] = \mathbb{E}_x[\sigma_B \mathbf{1}_B(X_1)] + \mathbb{E}_x[\sigma_B \mathbf{1}_{E \setminus B}(X_1)] \\ &= \mathbb{E}_x[\mathbf{1}_B(X_1)] + \sum_{k \in E \setminus B} \mathbb{E}_x[\sigma_B; X_1 = k] \\ &= \mathbb{E}_x[\mathbf{1}_B(X_1)] + \sum_{y \in E \setminus B} p_{xy}(1 + \mathbb{E}_y[\sigma_B]) \\ &= \mathbb{E}_x[\mathbf{1}_B(X_1)] + \mathbb{E}_x[\mathbf{1}_{E \setminus B}(X_1)] + \sum_{y \in E \setminus B} p_{xy} \mathbb{E}_y[\sigma_B] = 1 + \sum_{y \in E \setminus B} p_{xy} \mathbb{E}_y[\sigma_B]. \end{aligned}$$

Zeige nun die Minimalität von e_B .

Sei e eine beliebige nicht-negative Funktion mit

$$\begin{aligned} \forall x \in E \setminus B: e(x) &= 1 + (Pe)(x) \\ \forall x \in B: e(x) &= 0 \end{aligned}$$

Ziel: $e_B \leq e$

Zur Erinnerung: Für jede nicht-negative diskrete Zufallsvariable T gilt

$$\mathbb{E}[T] = \sum_{k=1}^{\infty} \mathbb{P}(T \geq k)$$

Denn:

$$\begin{aligned} \mathbb{E}[T] &= \sum_{j=0}^{\infty} j \cdot \mathbb{P}(T = j) + \infty \cdot \mathbb{P}(T = \infty) = \sum_{j=1}^{\infty} j \cdot \mathbb{P}(T = j) + \infty \cdot \mathbb{P}(T = \infty) \\ &= \sum_{j=1}^{\infty} \sum_{k=1}^j \mathbb{P}(T = j) + \infty \cdot \mathbb{P}(T = \infty) = \sum_{k=1}^{\infty} \sum_{j=k}^{\infty} \mathbb{P}(T = j) + \infty \cdot \mathbb{P}(T = \infty) \\ &= \sum_{k=1}^{\infty} \mathbb{P}(T \geq k) \end{aligned}$$

Wir zeigen, dass

$$e(x) \geq \sum_{k=1}^n \mathbb{P}^x(\sigma_B \geq k) \quad \forall n \in \mathbb{N}_0, x \in E \quad (**)$$

Die Ungleichung (**) gilt offensichtlich für $x \in B$ bzw. wenn $n = 0$.

Weiter gilt für $x \in E \setminus B$

$$\begin{aligned} \sum_{k=1}^{n+1} \mathbb{P}^x(\sigma_B \geq k) &= 1 + \sum_{k=2}^{n+1} \sum_{y \in E} p_{xy} \cdot \mathbb{P}^y(\sigma_B \geq k-1) = 1 + \sum_{y \in E} p_{xy} \cdot \sum_{k=2}^{n+1} \mathbb{P}^y(\sigma_B \geq k-1) \\ &= 1 + \sum_{y \in E} p_{xy} \cdot \underbrace{\sum_{k=1}^n \mathbb{P}^y(\sigma_B \geq k)}_{\substack{\text{IA} \\ \leq e(y)}} \leq 1 + \sum_{y \in E} p_{xy} \cdot e(y) = 1 + (Pe)(y). \quad \square \end{aligned}$$

Beispiel 4.17 (Münzwurfbeispiel aus 4.10). Definieren wir $B := \{-a, b\}$ und setzen $e_B(x) := \mathbb{E}_x[\sigma_B]$. Dann gelten $e_B(-a) = 0$ und $e_B(b) = 0$ und für $-a < x < b$

$$e_B(x) = 1 + \frac{1}{2}e_B(x-1) + \frac{1}{2}e_B(x+1).$$

Es zeigt sich, dass

$$e_B(x) = (x+a)(b-x)$$

die gesuchte Lösung ist.

Definition 4.18. Eine Markov-Kette mit Zustandsraum E und Übergangsmatrix P heißt **irreduzibel**, wenn es für alle Paare von Zuständen $i, j \in E$ ein $n \in \mathbb{N}$ gibt, so dass

$$(P^n)_{ij} := p_{ij}^{(n)} > 0$$

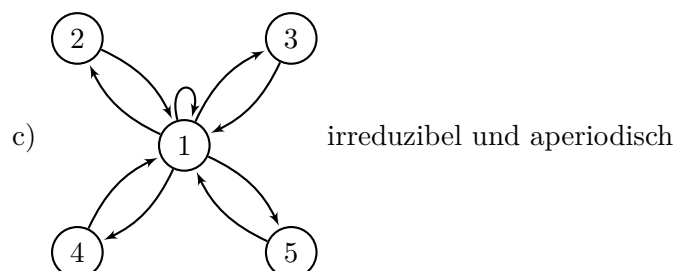
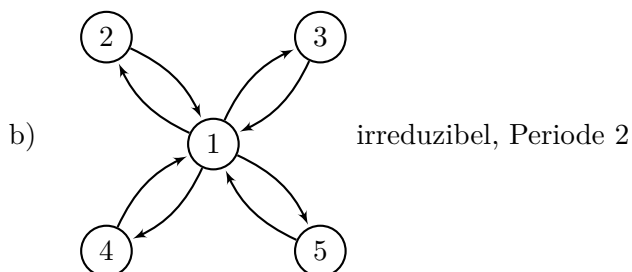
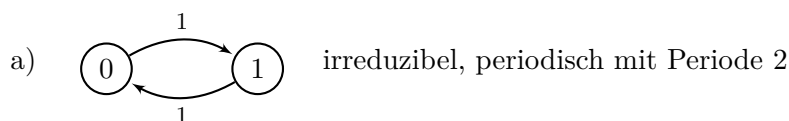
Bemerkung 4.19. Anschaulich besagt diese Definition, dass jeder Zustand j von jedem anderen Zustand i mit positiver Wahrscheinlichkeit in endlicher Zeit erreicht werden kann.

Definition 4.20. Die Periode d_j eines Zustandes j ist definiert durch

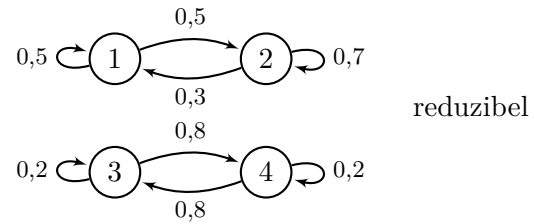
$$d_j := \text{ggT}\{n \in \mathbb{N} : p_{jj}^{(n)} > 0\}.$$

Ist die Menge leer, so setzen wir $d_j = \infty$. Ein Zustand heißt aperiodisch, wenn seine Periode 1 ist. Sind alle Zustände aperiodisch, so nennen wir die ganze Markov-Kette aperiodisch.

Beispiel 4.21.



$$\text{d) } P = \begin{pmatrix} 0,5 & 0,5 & 0 & 0 \\ 0,3 & 0,7 & 0 & 0 \\ 0 & 0 & 0,2 & 0,8 \\ 0 & 0 & 0,8 & 0,2 \end{pmatrix}$$



Definition 4.22. Wir sagen, dass ein Zustand i in $n \in \mathbb{N}$ Schritten zu Zustand j führt, wenn $p_{ij}^{(n)} > 0$. Dies notieren wir mit $i \rightsquigarrow j[n]$. Wenn ein $n \in \mathbb{N}$ mit $i \rightsquigarrow j[n]$ existiert, schreiben wir $i \rightsquigarrow j$. Wir sagen, dass i mit j kommuniziert, wenn gilt $i \rightsquigarrow j$ und $j \rightsquigarrow i$. Wir schreiben dann $i \longleftrightarrow j$.

Satz 4.23. Es gelte $i \longleftrightarrow j$ und sei d_i bzw. d_j die Periode von i bzw. j . Dann gilt

$$d_i = d_j$$

BEWEIS. Sei n so, dass $p_{jj}^{(n)} > 0$. Wähle m und ℓ so, dass $p_{ij}^{(m)} > 0$ und $p_{ij}^{(\ell)} > 0$. Dann gilt $p_{ii}^{(m+\ell)} > 0$ und somit teilt d_i bereits $(m+\ell)$. Weiter gilt $p_{ii}^{(m+\ell+n)} > 0$. Damit teilt d_i auch $(m+\ell+n)$. Also teilt d_i bereits n und somit gilt $d_i \leq d_j$. Aufgrund der Symmetrie in i und j folgt $d_j \leq d_i$ und somit die Gleichheit $d_i = d_j$. \square

Bemerkung 4.24. Für den Rest des Kapitels sei E eine endliche Menge.

Proposition 4.25. Für eine irreduzible, aperiodische Markov-Kette mit Übergangsmatrix $P = (p_{ij})_{i,j \in E}$ und endlichem Zustandsraum E gilt

$$\exists n \in \mathbb{N} \forall i, j \in E : p_{ij}^{(n)} > 0$$

BEWEIS (SKIZZE). Es genügt zu beweisen, dass für ein periodischen Zustand i eine natürliche Zahl n_0 existiert, so dass $p_{ii}^{(n)} > 0$ für alle $n \geq n_0$.

Für je zwei natürliche Zahlen $a, b \in \mathbb{N}$ gibt es ein $n_0 \in \mathbb{N}$, so dass gilt: Bezeichnet $d := \text{ggT}(a, b)$ den größten gemeinsamen Teiler von a und b , so gibt es für alle $n \in \mathbb{N}$, $n \geq n_0$ nichtnegative Zahlen $x, y \in \mathbb{N}_0$ mit $nd = xa + yb$. Wegen $p_{ii}^{(ax+by)} \geq (p_{ii}^{(a)})^x (p_{ii}^{(b)})^y$ folgt daraus unmittelbar: Gilt für ein $a, b \in \mathbb{N}$, dass sowohl $p_{ii}^{(a)} > 0$ als auch $p_{ii}^{(b)} > 0$, so gilt auch $p_{ii}^{(nd)} > 0$ für alle $n \geq n_0$. Aus der Aperiodizität des Zustands i folgt andererseits, dass es Werte a_0, \dots, a_k geben muss mit $p_{ii}^{(a_i)} > 0$ und der Eigenschaft, dass für $d_1 = \text{ggT}(a_0, a_1)$ und $d_i = \text{ggT}(d_{i-1}, a_i)$ gilt $d_1 > d_2 > \dots > d_k = 1$. Damit lässt sich die Behauptung folgern. \square

Definition 4.26. Eine irreduzible und aperiodische Markov-Kette mit endlichem Zustandsraum nennt man ergodisch.

4.2 Ergodisches Verhalten

In aktuellen Abschnitt werden wir einige wichtige Aussagen zum Langzeitverhalten einer Klasse von endlichen Markovketten herleiten.

Definition 4.27. Sei $(X_n)_{n \in \mathbb{N}_0}$ eine Markov-Kette mit endlichem Zustandsraum E und Übergangsmatrix $P = (p_{ij})_{i,j \in E}$. Eine Verteilung α auf E (d. h. $\alpha: E \rightarrow [0, 1]$, $\sum_{i \in E} \alpha(i) = 1$) heißt

stationäre Verteilung oder **invariantes Maß** für die Markovkette $(X_n)_{n \in \mathbb{N}_0}$ beziehungsweise für P , wenn

$$\alpha(j) = \sum_{i \in E} \alpha(i) \cdot p_{ij} \quad \forall j \in E$$

Fasst man α als Zeilenvektor auf, so gilt: $\alpha \cdot P = \alpha$. Also beschreibt α einen Linkseigenvektor der stochastischen Matrix P zum Eigenwert 1.

Bemerkung 4.28. Man beachte, dass gilt

$$\alpha = \alpha \cdot P = \alpha \cdot P^2 = \alpha \cdot P^3 = \dots$$

Nutzt man dies erhält man folgende Eigenschaft

$$\mathbb{P}^\alpha(X_0 = j) = \alpha(j) = (\alpha \cdot P^n)_j = \mathbb{P}^\alpha(X_n = j).$$

Satz 4.29. Für jede ergodische, endliche Markov-Kette mit Übergangsmatrix P existiert genau eine stationäre Verteilung α . Außerdem gilt für jede Startverteilung ρ auf E

$$\lim_{n \rightarrow \infty} \mathbb{P}^\rho(X_n = j) = \alpha(j) \quad \forall j \in E$$

BEWEIS.

1. Schritt:

Wir messen den Abstand zwischen zwei Verteilungen ρ_1 und ρ_2 auf E durch

$$\|\rho_1 - \rho_2\| := \sum_{j \in E} |\rho_1(j) - \rho_2(j)|$$

Dann gilt

$$\begin{aligned} \|\rho_1 \cdot P - \rho_2 \cdot P\| &= \sum_{j \in E} \left| \sum_{i \in E} \rho_1(i) \cdot p_{ij} - \sum_{i \in E} \rho_2(i) \cdot p_{ij} \right| \\ &\leq \sum_{i \in E} \sum_{j \in E} |\rho_1(i) - \rho_2(i)| \cdot p_{ij} \\ &= \sum_{i \in E} |\rho_1(i) - \rho_2(i)| = \|\rho_1 - \rho_2\| \end{aligned}$$

Nach der Voraussetzung der Ergodizität existiert ein k und ein $\delta > 0$, so dass für alle $i, j \in E$

$$p_{ij}^{(k)} > \frac{\delta}{|E|}$$

Ohne Einschränkung können wir annehmen, dass $\delta > 1$ gilt. Folglich ist dann

$$P^k \geq \delta \cdot U \quad \text{komponentenweise}$$

wobei U eine stochastische Matrix mit $u_{ij} = \frac{1}{|E|}$ für alle $i, j \in E$.

Somit ist die Matrix $V := (1 - \delta)^{-1} \cdot \underbrace{(P^k - \delta U)}_{\geq 0}$ ebenfalls eine stochastische Matrix.

$$\begin{aligned} \sum_{j \in E} V_{ij} &= \frac{1}{1 - \delta} \cdot \sum_{j \in E} (P^k - \delta U)_{ij} = \frac{1}{1 - \delta} \cdot \left(\sum_{j \in E} p_{ij}^{(k)} - \delta \cdot \sum_{j \in E} u_{ij} \right) \\ &= \frac{1}{1 - \delta} \cdot \left((1 - \delta) \cdot \underbrace{\left(\frac{1}{|E|} + \dots + \frac{1}{|E|} \right)}_{|E| \text{-mal}} \right) = \frac{1}{1 - \delta} \cdot (1 - \delta) = 1 \end{aligned}$$

Es gilt

$$P^k = \delta U + (1 - \delta) V,$$

das heißt weiter, dass

$$\begin{aligned} \|\rho_1 P^k - \rho_2 P^k\| &\leq \delta \|\rho_1 U - \rho_2 U\| + (1 - \delta) \|\rho_1 V - \rho_2 V\| \\ (\rho_1 U)(j) &= \sum_{i \in E} \rho_1(i) \cdot u_{ij} = \left(\sum_{i \in E} \rho_1(i) \right) \frac{1}{|E|} = \frac{1}{|E|} = (\rho_2 U)(j) \\ \|\rho_1 P^k - \rho_2 P^k\| &\leq (1 - \delta) \|\rho_1 V - \rho_2 V\| \leq (1 - \delta) \|\rho_1 - \rho_2\|. \end{aligned}$$

Für $n \geq k$:

$$\|\rho_1 P^n - \rho_2 P^n\| \leq \|\rho_1 P^{km} - \rho_2 P^{km}\| \leq (1 - \delta)^m \|\rho_1 - \rho_2\|, \quad m := \left\lfloor \frac{n}{k} \right\rfloor. \quad (*)$$

Angenommen ρ_1 und ρ_2 sind zwei stationäre Verteilungen, dann gilt

$$\|\rho_1 - \rho_2\| = \|\rho_1 P^n - \rho_2 P^n\| \leq (1 - \delta)^m \|\rho_1 - \rho_2\|.$$

Somit gilt bereits $\rho_1 = \rho_2$ und wir haben die Eindeutigkeit gezeigt.

2. Schritt:

Für eine beliebige Verteilung ρ betrachte

$$(\rho P^n)_{n \in \mathbb{N}_0} \subset \{\alpha \in \mathbb{R}^{|E|} : \alpha(i) \geq 0 \ \forall i \in E, \ \sum_{i=1}^n \alpha(i) = 1\}$$

Wegen Kompaktheit findet man eine Teilfolge $(n_k)_{k \in \mathbb{N}}$ und eine Verteilung α , so dass

$$\lim_{k \rightarrow \infty} \rho P^{n_k} = \alpha$$

Hauptaussage aus Schritt 2 mit $\rho_1 = \rho$ und $\rho_2 = \rho P$

$$\alpha = \lim_{k \rightarrow \infty} \rho P^{n_k} = \lim_{k \rightarrow \infty} \rho_2 P^{n_k} \lim_{k \rightarrow \infty} \rho P^{n_k+1} = \lim_{k \rightarrow \infty} \rho P^{n_k} \cdot P = \alpha P$$

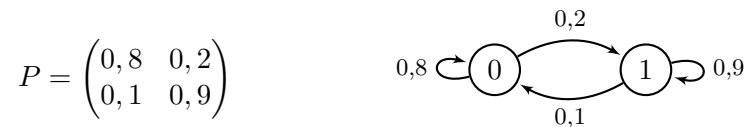
$\implies \alpha$ stationär

Weitere Aussagen folgen aus (*) mit $\rho_1 = \rho$ beliebiger Verteilung und $\rho_2 = \alpha$

$$\underbrace{\|\rho P^n - \alpha P^n\|}_{=\|\rho P^n - \alpha\|} \leq (1 - \delta)^m \|\rho - \alpha\|$$

\implies Beh.

□

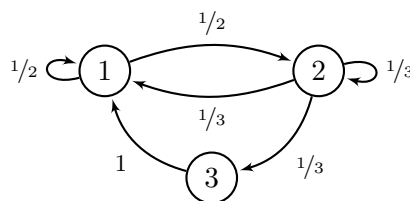
Beispiel 4.30.

$$\begin{pmatrix} \alpha_0 & \alpha_1 \end{pmatrix} \begin{pmatrix} 0,8 & 0,2 \\ 0,1 & 0,9 \end{pmatrix} = \begin{pmatrix} \alpha_0 & \alpha_1 \end{pmatrix} \iff \begin{matrix} \alpha_0 = 0,8\alpha_0 + 0,1\alpha_1 \\ \alpha_1 = 0,2\alpha_0 + 0,9\alpha_1 \end{matrix} \implies \alpha = \begin{pmatrix} \alpha_0 & 2\alpha_0 \end{pmatrix}$$

Wegen $\alpha_0 + \alpha_1 = 1 \implies \alpha_0 = \frac{1}{3}, \alpha_1 = \frac{2}{3}$

Beispiel 4.31.

$$E = \{1, 2, 3\}, P = \begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/3 & 1/3 & 1/3 \\ 1 & 0 & 0 \end{pmatrix}$$



$P^3 > 0$ in jeder Komponenten \implies Markov-Kette ergodisch

Man zeigt: $\alpha P = \alpha$ für $\alpha = (1/2 \quad 3/8 \quad 1/8)$

Aus 4.26 folgt:

$$P^n \xrightarrow{n \rightarrow \infty} \begin{pmatrix} 1/2 & 3/8 & 1/8 \\ 1/2 & 3/8 & 1/8 \\ 1/2 & 3/8 & 1/8 \end{pmatrix}$$

Definition 4.32. Zur Erinnerung: Eine $n \times n$ Matrix $P = (p_{ij})_{1 \leq i, j \leq n}$ heißt stochastisch, falls alle Einträge p_{ij} nicht-negativ und alle Zeilensummen gleich 1 sind, also

$$\sum_{j=1}^n p_{ij} = 1 \quad \forall i = 1, \dots, n$$

Sind zusätzlich alle Spaltensummen gleich 1, also

$$\sum_{i=1}^n p_{ij} = 1 \quad \forall j = 1, \dots, n$$

so heißt P **doppeltstochastisch**.

Bemerkung 4.33.

- Jede stochastische Matrix kann als Übergangsmatrix einer Markov-Kette interpretiert werden.
- Symmetrische stochastische Matrizen sind doppeltstochastisch.

Stationäre Verteilungen von doppeltstochastischen Matrizen sind leicht zu bestimmen.

Lemma 4.34. Ist P eine doppeltstochastische $n \times n$ Matrix, dann ist $\pi = \left(\frac{1}{n}, \dots, \frac{1}{n}\right)$ eine stationäre Verteilung für P . Es gilt also

$$\pi P = \pi.$$

BEWEIS.

Für $0 \leq k < n$ gilt:

$$(\pi P)_k = \sum_{i=0}^{n-1} \pi_i \cdot p_{ik} = \frac{1}{n} \cdot \underbrace{\sum_{i=0}^{n-1} p_{ik}}_{=1} = \frac{1}{n} = \pi_k$$

□

Für doppeltstochastische Matrizen ist die Bestimmung einer invarianten Verteilung somit einfach. Satz 4.26 und Lemma 4.31 liefern die folgende Aussage:

Satz 4.35. Für jede ergodische, endliche Markov-Kette mit doppeltstochastischer $n \times n$ Übergangsmatrix gilt unabhängig vom Startzustand ρ

$$\lim_{k \rightarrow \infty} \mathbb{P}^\rho(X_k = j) = \frac{1}{n}.$$

Eine weitere interessante Klasse von Markovketten mit besser analysierbarem Verhalten sind die reversiblen Markovketten.

Definition 4.36. Es sei E eine endliche Menge und $P = (p_{ij})_{i,j \in E}$ eine stochastische Matrix. Eine Verteilung α auf E heißt **reversible** Verteilung für die Matrix P , wenn

$$\boxed{\forall i, j \in E : \alpha_i p_{ij} = \alpha_j p_{ji}}$$

gilt.

Lemma 4.37. Es sei E eine endliche Menge und $P = (p_{ij})_{i,j \in E}$ eine stochastische Matrix. Eine Verteilung α auf E sei reversibel für die Matrix P , dann ist α invariant für P .

BEWEIS. Unter Ausnutzung der geforderten Reversibilität rechnen wir

$$\sum_{i \in E} \alpha_i p_{ij} = \sum_{i \in E} \alpha_j p_{ji} = \alpha_j \sum_{i \in E} p_{ji} = \alpha_j.$$

Die zeigt die behauptete Invarianz. □

Es ist häufig bedeutend leichter, reversible Verteilungen zu berechnen als stationäre Verteilungen. Allerdings existieren reversible Verteilungen im Unterschied zu stationären NICHT für beliebige ergodische Markovketten.

Wir schließen mit einigen Beispielen.

Beispiel 4.38 (Modell einer einfachen Warteschlange). Wir betrachten ein einfaches Warteschlangenmodell in diskreter Zeit. Zu jedem Zeitschritt tritt genau eins der folgenden Ereignisse auf

- Wenn in der Warteschlange weniger als n Kunden anstehen, dann stellt sich mit Wahrscheinlichkeit λ ein neuer Kunde in die Schlange.

- Wenn die Warteschlange nicht leer ist, dann wird mit Wahrscheinlichkeit μ der vorne in der Schlange stehende Kunde bedient.
- Mit Wahrscheinlichkeit $\lambda + \mu$ tritt kein neues Ereignis ein.

Betrachten wir die Zahl X_t der Kunden in der Warteschlange zur Zeit t , dann ist bereits aus der Beschreibung ersichtlich, dass $(X_t)_{t \geq 0}$ eine endliche Markovkette beschreibt. Die Übergangsmatrix hat die folgenden nicht-trivialen Einträge

$$\begin{aligned}
 P_{i,i+1} &= \lambda && \text{wenn } i < n \\
 P_{i,i-1} &= \mu && \text{wenn } i > 0 \\
 P_{i,i-1} &= \begin{cases} 1 - \lambda & \text{wenn } i = 0 \\ 1 - \lambda - \mu & \text{wenn } 1 \leq i \leq n-1 \\ 1 - \mu & \text{wenn } i = n. \end{cases}
 \end{aligned}$$

Die Markovkette ist irreduzibel, endlich und aperiodisch, insbesondere existiert eine eindeutig bestimmte stationäre Verteilung α . Zur Berechnung der Verteilung α betrachte man das assoziierte Gleichungssystem

$$\begin{aligned}
 \alpha_0 &= (1 - \lambda)\alpha_0 + \mu\alpha_1 \\
 \alpha_i &= \lambda\alpha_{i-1} + (1 - \lambda - \mu)\alpha_i + \mu\alpha_{i+1} \quad (1 \leq i \leq n-1) \\
 \alpha_n &= \lambda\alpha_{n-1} + (1\mu)\alpha_n.
 \end{aligned}$$

Man kann leicht nachprüfen, dass

$$\tilde{\alpha}_i = \tilde{\alpha}_0 \left(\frac{\lambda}{\mu} \right)^i$$

das Gleichungssystem löst. Die Normierungsbedingung $\sum_{i=0}^n \alpha_i = 1$ liefert dann

$$\alpha_0 = \frac{1}{\sum_{i=0}^n \left(\frac{\lambda}{\mu} \right)^i}$$

sowie

$$\alpha_i = \frac{\left(\frac{\lambda}{\mu} \right)^i}{\sum_{j=0}^n \left(\frac{\lambda}{\mu} \right)^j}$$

für $1 \leq i \leq n$

Beispiel 4.39. Gegeben sei ein Kartenstapel. Wir mischen, indem wir zwei Karten vertauschen. Dies wird n -mal wiederholt.

Modellierung: Identifiziere die Karten mit Zahlen $[n] := \{1, \dots, n\}$. Die Zustandsmenge S sei die Menge aller Permutationen von $[n]$.

Seien σ und $\rho \in S$. Nach Definition ist $p_{\sigma,\rho} > 0$, genau dann, wenn es $i, j \in [n]$ mit $i \neq j$ gibt, so dass

$$\rho(k) = \begin{cases} \sigma(j) & k = i \\ \sigma(i) & k = j \\ \sigma(k) & k \notin \{i, j\} \end{cases}$$

Da man „zurücktauschen“ kann gilt: $p_{\sigma,\rho} = p_{\rho,\sigma} \implies$ Übergangsmatrix P ist doppeltstochastisch. Die Kette ist auch ergodisch, da jede Permutation erreicht werden kann und sie aperiodisch ist. Wegen 4.32 folgt nun für jede Startverteilung ρ ,

$$\lim_{n \rightarrow \infty} \mathbb{P}^\rho(X_n = \sigma) = \alpha(\sigma) = \frac{1}{|S|}.$$

Ohne die Voraussetzung der Aperiodizität gilt noch

Satz 4.40. Eine irreduzible Markovkette mit endlichem Zustandsraum E besitzt eine eindeutige stationäre Verteilung α und es gilt

$$\alpha_j = \frac{1}{\mathbb{E}_j[\tau_j]}$$

für alle $j \in E$.

BEWEIS (SKIZZE). Wir zeigen zunächst dass es einen Wahrscheinlichkeitsvektor α gibt mit $\alpha = \alpha P$. Hierzu betrachten wir für einen Zustand $i \in E$ die Folge $(\alpha^n)_{n \in \mathbb{N}}$ bestehend aus den Wahrscheinlichkeitsvektoren

$$\alpha_j^n := \frac{1}{n} \sum_{k=0}^{n-1} p_{ij}^{(k)}, \quad j \in E.$$

Weil die Menge

$$\{(x_i)_{i \in E} \subset [0, 1]^{|E|} \mid \sum_{i \in E} x_i = 1\}$$

eine kompakte Teilmenge des $\mathbb{R}^{|E|}$ ist, existiert ein Zufallsvektor $\alpha = (\alpha_i)_{i \in E}$ und eine konvergente Teilfolge $(\alpha^{n_k})_{n_k \in \mathbb{N}}$ mit $\alpha^{n_k} \rightarrow \alpha$ für $k \rightarrow \infty$. Wie im Ergodensatz zeigt man die Invarianz von α : Zunächst gilt

$$\sum_{i \in S} \alpha_i^n = 1$$

für jedes n und somit ist α eine Verteilung. Ausserdem gilt

$$\lim_{k \rightarrow \infty} \frac{1}{n_k} \sum_{l=1}^{n_k} P^l = \lim_{k \rightarrow \infty} \frac{1}{n_k} \sum_{l=0}^{n_k-1} P^l,$$

woraus sich die Invarianz von α folgern lässt. Die Erwartungswerte $h_{ij} := \mathbb{E}_i[\tau_j]$ existieren, wie man leicht sieht (Übung). Für jeden Zustand $i \in E$ gelten dann die Gleichungen

$$\alpha_i h_{ij} = \alpha_i (1 + \sum_{k \neq j} p_{ik} h_{kj}).$$

Hierbei beachte wieder, dass mit Hilfe einer Ein-Schritt-Analyse

$$\begin{aligned} h_{jj} &= \mathbb{E}_j[\tau_j] = \mathbb{E}_j[\tau_j; X_1 = j] + \sum_{k \neq j} \mathbb{E}_j[\tau_j; X_1 = k] \\ &= p_{jj} + \sum_{k \neq j} [p_{jk} + p_{jk} \mathbb{E}_k[\tau_j]]. \end{aligned}$$

Addition dieser Gleichungen gibt

$$\begin{aligned} \alpha_j h_{jj} + \sum_{i \neq j} \alpha_i h_{ij} &= 1 + \sum_{i \in E} \sum_{k \neq j} \alpha_i p_{ik} h_{kj} \\ &= 1 + \sum_{k \neq j} h_{kj} \sum_{i \in E} \alpha_i p_{ik} = 1 + \sum_{k \neq j} \alpha_k h_{kj} \end{aligned}$$

Wegen $h_{jj} > 0$ ist somit $\alpha_j = \frac{1}{h_{jj}}$. □

4.3 Anwendungen

In diesem Abschnitt geben wir einen oberflächlichen Ausblick auf die vielfältigen Anwendungen der Theorie der Markovketten.

4.3.1 Markov-Chain-Monte Carlo

Zunächst stellen wir exemplarisch die Markov-Chain-Monte Carlo Method dar. Wir sind bereits auf die Monte-Carlo Methode kurz eingegangen. Diese erlaubt es, durch Erzeugung unabhängiger identisch gemäß einer Dichte f verteilten Zufallsvariablen X_1, X_2, \dots, X_N Kenngrößen wie z.B. den Erwartungswert $\int_{-\infty}^{\infty} x f(x) dx$ durch

$$\frac{1}{N} \sum_{i=1}^N X_i$$

zu approximieren. Oft ist es jedoch nicht oder nur schwer möglich, Zufallsvariablen gemäß einer gegebenen Dichte f zu generieren, weil z.B. die Normierungskonstante nicht analytisch berechenbar ist. Die Markov-Chain-Monte-Carlo Methode liefert hier einen möglichen Ausweg. Wir betrachten den Fall eines endlichen Zustandsraums E und sei π ein Wahrscheinlichkeitsmaß auf E für $x \in E$:

$$\pi(x) = \frac{e^{-\beta H(x)}}{Z} \in [0, 1]$$

$$Z = \sum_{x \in E} e^{-\beta H(x)}$$

Die Normierungskonstante Z ist analytisch häufig nicht berechenbar und wir kennen somit die Dichte nur bis auf den Normierungsfaktor. Betrachte das Problem eine Zufallsvariable Y mit Verteilung π zu simulieren, um statistische Kenngrößen approximativ zu berechnen. Die Monte-Carlo Methode ist hier häufig nicht realisierbar. Eine mögliche Herangehensweise ist folgende: Erzeuge Markov-Kette $(X_n)_{n \geq 0}$ mit Zustandsraum E und Übergangsmatrix P , so dass

$$\pi P = \pi.$$

Wenn die Markov-Kette irreduzibel und aperiodisch ist, gilt

$$\lim_{n \rightarrow \infty} \mathbb{E}[f(X_n)] = \mathbb{E}_\pi[f] = \sum_{x \in E} f(x) \pi(x).$$

Folglich lässt sich

$$\sum_{x \in E} f(x) \pi(x)$$

mit Hilfe der Markovkette $(X_n)_{n \geq 0}$ approximativ berechnen.

Diese Herangehensweise wird **Markov-Ketten-Monte-Carlo-Methode (MCMC-Methode)** genannt. (für weiterführende Information siehe Persi Diaconis, *The Markov-Chain Monte-Carlo Revolution*, Bull. Amer. Math. Soc. 46 (2009), 179–205)

Konstruktion einer Metropolis-Kette

Es sei $(q_{ij})_{i,j \in E}$ die Übergangsmatrix einer beliebigen irreduziblen und aperiodischen Markov-Kette mit endlichem Zustandsraum E .

Definition. Definiere eine stochastische Matrix P wie folgt

$$p_{xy} = \begin{cases} q_{xy} \cdot \min\left(1, \frac{\pi(y) \cdot q_{xy}}{\pi(x) \cdot q_{yx}}\right) & \text{falls } x \neq y, q_{xy} \neq 0 \\ 0 & \text{falls } x \neq y, q_{xy} = 0 \\ 1 - \sum_{z \neq x} p_{xz} & \text{falls } x = y \end{cases}$$

Behauptung. Es gilt $\pi \cdot P = \pi$, d. h. π ist eine stationäre Verteilung für P .

In der Tat sieht man anhand einer einfachen Rechnung leicht, dass die Metropolis-Kette reversibel bzgl. π ist. Wie wir gesehen haben, konvergiert eine aperiodische und irreduzible Markovkette mit endlichem Zustandsraum gegen die stationäre Verteilung. In den Anwendungen sind nicht-asymptotische Abschätzungen und explizite Fehlerschranken notwendig. Diese hängen stark von der konkreten Aufgabenstellung ab und sind in der Regel nicht leicht herzuleiten.

4.3.2 Stochastische Optimierung

Wir formulieren zunächst eine mögliche Problemstellung. Es sei f eine Funktion auf der Knotenmenge Ω eines der Einfachheit halber regulären Graphen. Suche Knoten $x \in \Omega$ mit

$$f(x) = \max_{y \in \Omega} f(y)$$

Ein „Hill-Climb“ ist ein Algorithmus, der auf folgende Art und Weise versucht Maxima zu lokalisieren.

Ist x ein Nachbar von y mit $f(x) < f(y)$, dann gehe zu y .

Ein solcher Algorithmus bleibt in *lokalen* Maxima stehen!

Fixiere $\lambda \geq 1$ und sei

$$\pi_\lambda(x) = \frac{\lambda^{f(x)}}{Z_\lambda}, \quad \text{wobei } \sum_{x \in \Omega} \lambda^{f(x)} =: Z_\lambda$$

π_λ ist ein Wahrscheinlichkeitsmaß auf Ω .

Konstruiere nun die Metropolis-Kette zu π_λ auf Ω . Beachte wiederum, dass hierbei Kenntnis der Normierung Z_λ nicht notwendig ist.

Ein Übergang von x nach y wird (falls $f(x) < f(y)$) mit Wahrscheinlichkeit

$$\lambda^{-(f(x)-f(y))}$$

akzeptiert.

Beobachtung. $\Omega^* = \{x \in \Omega : f(x) = f^* = \max_{y \in \Omega} f(y)\}$

$$\lim_{\lambda \rightarrow \infty} \pi_\lambda(x) = \frac{\lambda^{f(x)}/\lambda^{f^*}}{|\Omega^*| + \sum_{y \in \Omega \setminus \Omega^*} \frac{\lambda^{f(y)}}{\lambda^{f^*}}} = \frac{\mathbf{1}_{\Omega^*}(x)}{|\Omega^*|}$$

4.3.3 Randomisierte Algorithmen

Für interessante Analysen randomisierter Algorithmen verweisen wir auf Buch von Mitzenmacher und Upfal. Wir beschränken uns darauf, mit Hilfe eines sehr einfachen Verfahrens die Fragestellungen und die Methodik exemplarisch aufzuzeigen und die Verbindung zu den hier erlernten Konzepten zu verdeutlichen.

Seien Y_1, Y_2, \dots, Y_r unabhängige und identisch verteilte Zufallsvariablen mit Werten in \mathbb{R} und Dichte f (d.h. $\mathbb{P}(Y_i \geq y) = \int_{-\infty}^y f(z) dz$).

Ein möglicher Algorithmus:

```

1:  $M := Y_0$ 
2: for  $i = 2$  to  $r$  do
3:   if  $Y_i > M$  then
4:      $M = Y_i$ 
5: return  $M$ 

```

Es bezeichne I_{j+1} die Zeit des j -ten Austausches des aktuellen Maximums, d. h.

$$I_1 = 1$$

$$I_{j+1} = \begin{cases} \min\{i > I_j : Y_i > Y_{I_j}\} & \text{falls } \{i > I_j : Y_i > Y_{I_j}\} \neq \emptyset \\ I_j & \text{sonst} \end{cases}$$

Sei nun

$$J := \min\{j \geq 1 : I_{j+1} = I_j\}$$

Dann ist $J - 1$ die Anzahl der durchgeführten Austausche und kann somit als Maß für die „Komplexität“ betrachtet werden. Untersuche $\mathbb{E}[J]$:

Führe die Folge von Zufallsvariablen $(X_n)_{n \geq 0}$ mit $X_0 = 0$ und $X_j = \text{Rang}(Y_{I_j})$, $j \geq 1$ ein. $(X_j)_{j \geq 0}$ ist eine Markov-Kette mit Zustandsraum $E = \{0, 1, \dots, r\}$ und Übergangsmatrix

$$p_{xy} = \begin{cases} (r-x)^{-1} & \text{falls } x < y \\ 1 & \text{falls } x = y = r \\ 0 & \text{sonst} \end{cases}$$

Ist nämlich das Ereignis $\{X_0 = x_0, X_1 = x_1, \dots, X_n = x_n\}$ gegeben, dann wissen wir

- Ein Vergleich mit dem Element vom Rang x_n hat stattgefunden.
- Ein Vergleich mit dem Elementen vom Rang x_{n+1}, \dots, r stehen noch aus. Von den übrigen dieser $r - x_n$ Elemente ziehe eines mit gleicher Wahrscheinlichkeit.

Beachten, dass $J = \min\{j \geq 1 : X_j = r\} = \tau_r \implies \mathbb{E}[J] = \mathbb{E}^0[\tau_r]$, $e(x) = \mathbb{E}^x[\tau_r]$. Es gilt

$$e(x) = 1 + (Pe)(x), \quad x < r$$

$$= 1 + \frac{1}{r-x} \cdot (e(x+1) + \dots + e(r))$$

$$e(r) = 0, \quad e(r-1) = 1 + \frac{1}{r-x} \cdot e(r) = 1, \quad e(r-2) = \dots$$

Es ergibt sich

$$e(x) = \sum_{j=1}^{r-x} \frac{1}{j} \implies \mathbb{E}[J] = \sum_{j=1}^r \frac{1}{j} \approx \log(r).$$

4.3.4 Modellierung in den Naturwissenschaften

Markovketten oder allgemeiner Markovprozesse spielen in vielen Bereichen der Naturwissenschaften eine fundamentale Rolle. In diesem Abschnitt gehen wir auf das einfache Wright-Fischer-Modell aus der Populationsgenetik ein. Betrachte genetisches Material mit zwei Ausprägungen A oder a in einer Population von N Individuen. Jede Generation hat N Individuen und jedes dieser Lebewesen „sucht sich“ unabhängig aus der vorherigen Generation ein zufälliges Elternteil und übernimmt dessen Ausprägung.

$Y_n = \#$ Individuen mit Ausprägung A in Generation n

Wir interessieren uns im folgenden für die Wahrscheinlichkeit

$$\lim_{n \rightarrow \infty} \mathbb{P}^x(Y_n = N).$$

Die Folge $(Y_n)_{n \geq 0}$ bildet eine Markov-Kette mit Zustandsraum $E = \{0, \dots, N\}$. Die Zustände 0 und N sind absorbierend. Sei die Übergangsmatrix $P = (p_{xy})_{x,y \in E}$ zu $(Y_n)_{n \geq 0}$. Man sieht leicht, dass

$$p_{xy} = \binom{N}{y} \cdot \left(\frac{x}{N}\right)^y \cdot \left(\frac{N-x}{N}\right)^{N-y}$$

Behauptung.

$$\underbrace{\lim_{n \rightarrow \infty} \mathbb{P}^x(Y_n = N)}_{h_n(x) = \mathbb{P}^x(\tau_N < \infty)} = \frac{x}{N}$$

Aufgrund unserer Analyse von Absorptionswahrscheinlichkeiten in Satz 3.13 wissen wir, dass die Gleichungen

$$Ph_N(x) = h_N(x), \quad h_N(N) = 1, \quad h_N(0) = 0$$

erfüllt sind. Zum Nachweis der Behauptung definieren wir zunächst $h(x) = \frac{x}{N}$. Dann gilt

$$\begin{aligned} Ph(x) &= \sum_{y=0}^N p_{xy} h(y) = \sum_{y=0}^N \binom{N}{y} \cdot \left(\frac{x}{N}\right)^y \cdot \left(\frac{N-x}{N}\right)^{N-y} \cdot \frac{y}{N} \\ &= \frac{1}{N} \cdot \sum_{y=0}^N y \binom{N}{y} \left(\frac{x}{N}\right)^y \left(\frac{N-x}{N}\right)^{N-y} = \frac{1}{N} \cdot \mathbb{E}[S] \end{aligned}$$

wobei S Bernoulli-verteilt mit $p = \frac{x}{N}$ und $n = N$

$$\mathbb{E}[S] = \frac{x}{N} \cdot N \implies Ph(x) = \frac{x}{N} = h(x).$$

Insbesondere folgern wir, dass h eine invariante Funktion für die Markovkette ist. Wir zeigen nun dass $h_N = h$. Hierfür betrachte $g := h_N - h$ und $m := \max_{0 \leq y \leq N} g(y)$. Angenommen, es wäre $m > 0$, dann gäbe es ein $0 < x < N$ mit $g(x) = m$. Es folgt

$$0 = g(x) - (Pg)(x) = \sum_{y=0}^N p_{xy}(m - g(y))$$

und somit $g(y) = m$ für alle y , was aber für $y = 0$ nicht erfüllt ist. Somit ist $m = 0$. Dasselbe Argument für das Minimum liefert die Behauptung $g = 0$.

Regression

5.1 Einfache lineare Regression

Betrachte ein Zufallsexperiment, dessen Ergebnisse nicht nur vom Zufall abhängen, sondern zusätzlich noch von einer erklärenden Variablen.

Beispiel 5.1 (Agrarwissenschaftliches Experiment). Einfluss der Luftfeuchtigkeit auf Tomatenernte im Gewächshaus.

Betrachte die Zufallsvariable Y als Ergebnis der Experimente. Mit x bezeichne den Wert der erklärenden Größe und mit ε bezeichne eine $\mathcal{N}(0, \sigma^2)$ -verteilte Zufallsvariable.

Das durch die Gleichung

$$Y = \alpha + \beta \cdot x + \varepsilon, \quad \alpha, \beta \in \mathbb{R}$$

beschriebene Modell heißt einfaches lineares Regressionsmodell.

Beachte, dass

$$\mathbb{E}[Y] = \mathbb{E}[\alpha + \beta \cdot x + \varepsilon] = \alpha + \beta \cdot x + \mathbb{E}[\varepsilon] = \alpha + \beta \cdot x$$

zwischen dem Erwartungswert von Y und der erklärenden Variable ein linearer Zusammenhang besteht.

In den zufälligen Term ε wird alles aufgenommen, was wir bei unserem Experiment nicht festlegen können. (Bodenqualität, genetische Variation, ...)

Die durch

$$y = \alpha + \beta x$$

beschriebene Gerade heißt Regressionsgerade. Die Regressionskoeffizienten sind dabei zu bestimmen.

Führe bei verschiedenen Werten x_1, \dots, x_n der erklärenden Variable unabhängige Experimente durch. Bezeichne Y_1, \dots, Y_n die zugehörigen Zufallsvariablen.

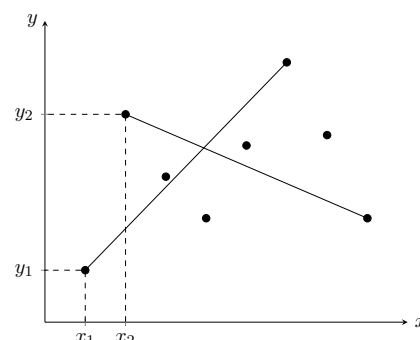
$$Y_1 = \alpha + \beta x_1 + \varepsilon_1$$

$$Y_2 = \alpha + \beta x_2 + \varepsilon_2$$

$$\vdots$$

$$Y_n = \alpha + \beta x_n + \varepsilon_n$$

$\varepsilon_1, \dots, \varepsilon_n$ unabh. und $\mathcal{N}(0, \sigma^2)$ -verteilt



Beobachtung I: Für jedes $i \in \{1, \dots, n\}$ hat die Zufallsvariable Y_i eine $\mathcal{N}(\alpha + \beta x_i, \sigma^2)$ -Verteilung, d. h. sie hat die Dichte

$$\varphi_i(z) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z - \alpha - \beta x_i)^2}{2\sigma^2}}$$

Beobachtung II: Wegen der Unabhängigkeit von Y_1, \dots, Y_n ist die gemeinsame Dichte von Y_1, \dots, Y_n gegeben durch

$$\varphi(z_1, \dots, z_n) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z_i - \alpha - \beta x_i)^2}{2\sigma^2}}$$

Die log-Likelihood-funktion ist

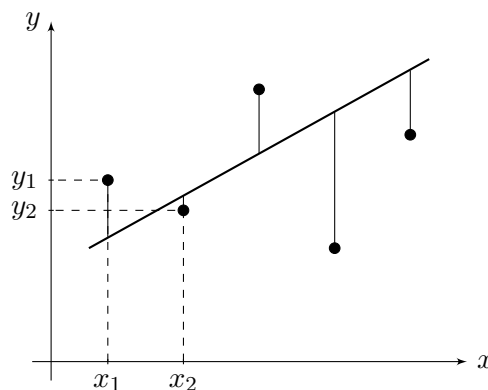
$$\ell_{Y_1, \dots, Y_n}(\alpha, \beta, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2$$

Man denke sich σ^2 als fest vorgegeben und bekannt.

Um α und β nach der Maximum-Likelihood-Methode zu schätzen, muss ℓ_{Y_1, \dots, Y_n} maximiert werden.

Die Maximierung der log-Likelihoodfunktion erfordert die Minimierung von

$$Q(\alpha, \beta) = \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2$$



„Methode der kleinsten Quadrate“

Satz 5.2. Die kleinste Quadrate-Schätzer für α und β sind gegeben durch

i) $\hat{\alpha} = \bar{z} - \hat{\beta} \bar{x}$

ii) $\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(z_i - \bar{z})}{\sum_{i=1}^n (x_i - \bar{x})^2}$

Beide Schätzer sind erwartungstreu, d. h.

$$\mathbb{E}[\hat{\alpha}] = \alpha \quad \text{und} \quad \mathbb{E}[\hat{\beta}] = \beta$$

Ihre Varianz ist gegeben durch

$$\text{Var}(\hat{\alpha}) = \sigma^2 \frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{und} \quad \text{Var}(\hat{\beta}) = \sigma^2 \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

BEWEIS.

a) Bestimme kleinste Quadrate Schätzer. Leite nach α bzw. β ab und setze gleich 0

$$\sum_{i=1}^n (z_i - \alpha - \beta x_i) = 0, \quad \sum_{i=1}^n (z_i - \alpha - \beta x_i) x_i = 0$$

Mit den Abkürzungen

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

folgt aus der ersten Gleichung

$$\hat{\alpha} = \bar{y} - \beta \bar{x}$$

Diese eingesetzt in die zweite Gleichung liefert

$$\sum_{i=1}^n (y_i - \bar{y} - \beta(x_i - \bar{x})) x_i = 0$$

also

$$\hat{\beta} = \frac{\sum_{i=1}^n (y_i - \bar{y}) x_i}{\sum_{i=1}^n (x_i - \bar{x}) x_i} = \frac{\sum_{i=1}^n (z_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x}) z_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

denn

$$\begin{aligned} \sum_{i=1}^n (z_i - \bar{z})(x_i - \bar{x}) &= \sum_{i=1}^n (z_i - \bar{z}) x_i - \sum_{i=1}^n (z_i - \bar{z}) \bar{x} \\ &= \sum_{i=1}^n (z_i - \bar{z}) x_i - \bar{x} \sum_{i=1}^n (z_i - \bar{z}) \\ &= \sum_{i=1}^n (z_i - \bar{z}) x_i - \underbrace{\bar{x} \sum_{i=1}^n (z_i - n\bar{z})}_{=0} \end{aligned}$$

b) $\hat{\alpha}, \hat{\beta}$ sind erwartungstreu.

$$\begin{aligned} \mathbb{E}[\hat{\beta}] &= \mathbb{E} \left[\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sum_{i=1}^n (x_i - \bar{x}) Y_i \right] = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \mathbb{E} \left[\sum_{i=1}^n (x_i - \bar{x}) Y_i \right] \\ &= \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sum_{i=1}^n (x_i - \bar{x}) \cdot \mathbb{E}[Y_i] = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \underbrace{\sum_{i=1}^n (x_i - \bar{x}) \cdot (\alpha + \beta x_i)}_{= \alpha \sum_{i=1}^n (x_i - \bar{x}) + \beta \sum_{i=1}^n (x_i - \bar{x}) x_i} \\ &= \beta \cdot \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sum_{i=1}^n (x_i - \bar{x}) = \beta \cdot \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sum_{i=1}^n \underbrace{(x_i - \bar{x})(x_i - \bar{x})}_{(x_i - \bar{x})^2} \\ &= \beta \end{aligned}$$

$$\text{c) } \text{Var}(\hat{\beta}) = \sigma^2 \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\begin{aligned} \text{Var}(\hat{\beta}) &= \text{Var} \left(\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sum_{i=1}^n (x_i - \bar{x}) Y_i \right) \\ &= \left(\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)^2 \cdot \text{Var} \left(\sum_{i=1}^n (x_i - \bar{x}) Y_i \right) \\ &\stackrel{\text{unabh.}}{=} \frac{1}{(\sum_{i=1}^n (x_i - \bar{x})^2)^2} \cdot \sum_{i=1}^n \text{Var}((x_i - \bar{x}) Y_i) \\ &= \frac{1}{(\sum_{i=1}^n (x_i - \bar{x})^2)^2} \cdot \sum_{i=1}^n (x_i - \bar{x})^2 \text{Var}(Y_i) \\ &= \sigma^2 \frac{1}{(\sum_{i=1}^n (x_i - \bar{x})^2)^2} \cdot \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= \sigma^2 \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned} \quad \square$$

Beispiel 5.3. Zeitschrift Focus:

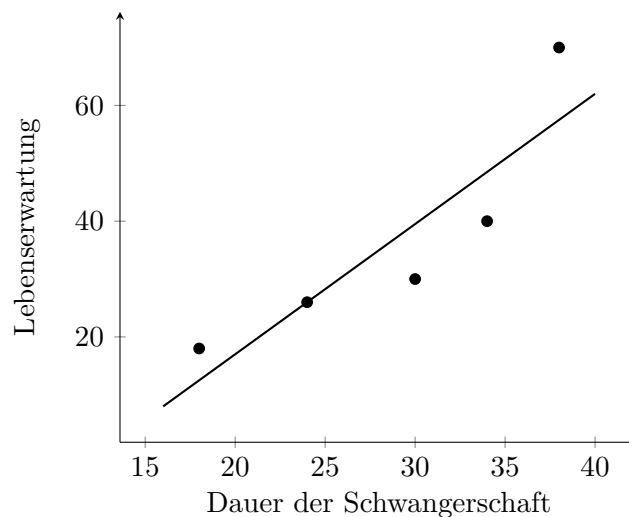
	Dauer der Schwangerschaft (x)	Lebenserwartung (y)	xy	x^2	y^2
Lemur	18	18	324	324	324
Makak	24	26	624	576	676
Gibbon	30	30	900	900	900
Schimpanse	34	40	1.360	1.156	1.600
Mensch	38	70	2.660	1.444	4.900
Summe	144	184	5.868	4.400	8.400

Durch Einsetzen in die obigen Formeln erhält man

$$\hat{\alpha} = -28 \quad \text{und} \quad \hat{\beta} = 2,25$$

Daraus ergibt sich die folgende Regressionsgerade

$$y = -28 + 2.25 \cdot x$$



Satz 5.4. Es gelten:

$$\hat{\alpha} \sim \mathcal{N}\left(\alpha, \sigma^2 \frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}\right) \quad \text{und} \quad \hat{\beta} \sim \mathcal{N}\left(\beta, \sigma^2 \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)$$

Bemerkung 5.5. Je größer die Streuung s_x^2 der x -Werte ist, desto genauer ist die Schätzung für β , wobei $s_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$. Je größer n ist, desto genauer ist die Schätzung für α und β .

Unter der Normalverteilungsannahme an die Fehler ε_i ist der Kleinste-Quadrate-Schätzer ein ML-Schätzer.

Satz 5.6.

- 1) Ein erwartungstreuer Schätzer für σ^2 ist

$$\hat{\sigma}^2 := \frac{1}{n-2} \cdot \sum_{i=1}^n \hat{\varepsilon}_i^2 = \frac{1}{n-2} \cdot \sum_{i=1}^n (y_i - (\hat{\alpha} - \hat{\beta}x_i))^2$$

- 2) Unter der Normalverteilungsannahme gilt

$$\frac{1}{\hat{\sigma}^2} \cdot \sum_{i=1}^n \hat{\varepsilon}_i^2 \sim \chi_{n-2}^2, \quad (\hat{\alpha}, \hat{\beta}) \text{ und } \hat{\sigma}^2 \text{ unabh.}$$

- 3) Unter der Normalverteilungsannahme gilt

$$\frac{\hat{\beta} - \beta}{\hat{\sigma}_{\hat{\beta}}} \sim t_{n-2} \quad \text{mit } \hat{\sigma}_{\hat{\beta}} := \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

$$\frac{\hat{\alpha} - \alpha}{\hat{\sigma}_{\hat{\alpha}}} \sim t_{n-2} \quad \text{mit } \hat{\sigma}_{\hat{\alpha}} := \sqrt{\hat{\sigma}^2 \frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}}$$

- 4) Konfidenzintervall zum Niveau $1 - \gamma$

$$[\hat{\beta} - \hat{\sigma}_{\hat{\beta}} \cdot t_{1-\frac{\gamma}{2}; n-2}; \hat{\beta} + \hat{\sigma}_{\hat{\beta}} \cdot t_{1-\frac{\gamma}{2}; n-2}] \quad \text{für } \beta$$

$$[\hat{\alpha} - \hat{\sigma}_{\hat{\alpha}} \cdot t_{1-\frac{\gamma}{2}; n-2}; \hat{\alpha} + \hat{\sigma}_{\hat{\alpha}} \cdot t_{1-\frac{\gamma}{2}; n-2}] \quad \text{für } \alpha$$

wobei $t_{1-\frac{\gamma}{2}; n-2}$ das $1 - \frac{\gamma}{2}$ -Quantil der t -Verteilung mit $n - 2$ Freiheitsgraden ist.

5.2 Lineare Regression

Wir nehmen an, dass wir n verschiedene $k + 1$ -Tupel von Messwerten haben

$$\begin{aligned}(x_{11}, x_{12}, \dots, x_{1k}, y_1) &= (\vec{x}_1, y_1) \\ &\vdots \\ (x_{n1}, x_{n2}, \dots, x_{nk}, y_n) &= (\vec{x}_n, y_n)\end{aligned}$$

Führe das allgemeine lineare Modell ein, welches wie folgt definiert ist.

$$y(\vec{x}) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k,$$

das die Messwerte „näherungsweise“ beschreiben soll.

EINSCHUB. (Differentiation in mehreren Variablen)

Seien

$$\vec{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix} \quad \text{und} \quad f(\vec{\beta}) = f(\beta_0, \dots, \beta_k)$$

wobei f eine „ $k + 1$ -dimensionale“ Funktion.

Ist f differenzierbar, so kann der Gradient berechnet werden:

$$\frac{\partial f}{\partial \vec{\beta}}(\vec{\beta}) = \begin{pmatrix} \frac{\partial f}{\partial \beta_0}(\vec{\beta}) \\ \vdots \\ \frac{\partial f}{\partial \beta_k}(\vec{\beta}) \end{pmatrix} =: \nabla f(\vec{\beta})$$

Für einen $k + 1$ -dimensionalen Vektor $\vec{c} = \begin{pmatrix} c_0 \\ \vdots \\ c_k \end{pmatrix}$ definiere das Skalarprodukt von $\vec{c}, \vec{\beta}$ durch

$$f(\vec{\beta}) = \underbrace{\sum_{n=0}^k c_n \beta_n}_{c_0 \beta_0 + \dots + c_k \beta_k} = \langle \vec{c}, \vec{\beta} \rangle_{\mathbb{R}^k}.$$

Dann gilt

$$\frac{\partial f}{\partial \vec{\beta}}(\vec{\beta}) = \begin{pmatrix} \frac{\partial f}{\partial \beta_0}(\vec{\beta}) \\ \vdots \\ \frac{\partial f}{\partial \beta_k}(\vec{\beta}) \end{pmatrix} = \begin{pmatrix} c_0 \\ \vdots \\ c_k \end{pmatrix} = \vec{c}.$$

Sei nun noch $A = (a_{ij})_{i,j=0,\dots,k}$ eine symmetrische Matrix, d. h. $a_{ij} = a_{ji} \forall i, j$ und sei

$$f(\vec{\beta}) = \langle \vec{\beta}, A \cdot \vec{\beta} \rangle_{\mathbb{R}^k} = \sum_{i,j=0}^k a_{ij} \beta_i \beta_j \quad (*)$$

Lemma. Für die obige in (*) definierte Funktion gilt:

$$\nabla f(\vec{\beta}) = 2A\vec{\beta}$$

BEWEIS.

$$\begin{aligned}
 f(\vec{\beta}) &= \sum_{i,j=0}^k a_{ij} \beta_i \beta_j = \sum_{i=0}^k a_{ii} \beta_i^2 + \sum_{i \neq j} a_{ij} \beta_i \beta_j \\
 &= \sum_{i=0}^k a_{ii} \beta_i^2 + 2 \cdot \sum_{i < j} a_{ij} \beta_i \beta_j \\
 \frac{\partial f}{\partial \beta_0}(\vec{\beta}) &= 2 \cdot a_{00} \cdot \beta_0 + 2 \cdot \sum_{j=1}^k a_{0j} \cdot \beta_j = 2 \cdot a_0^\top \vec{\beta}
 \end{aligned}$$

Wobei $a_0 = A \cdot e_0$ für einen $k+1$ -dim. Vektor e_0 mit $e_0 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$. Damit ist a_0 der erste Spaltenvektor von A .

$$\Rightarrow \frac{\partial f}{\partial \vec{\beta}} = 2A\vec{\beta}$$

□

Kommen wir zurück zum allgemeinen linearen Modell: $y(\vec{x}) = \beta_0 + \beta_1 x_1 + \cdots \beta_k x_k$.

Auch hier können wir das Prinzip der kleinsten Quadrate anwenden. Wir erhalten

$$S(\vec{\beta}) = \sum_{i=0}^n (y_i - (\beta_0 + \beta_1 x_{i1} + \cdots \beta_k x_{ik}))^2.$$

Bestimme das Minimum der Funktion S .

Führe zunächst folgende Objekte ein

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{k1} & \cdots & x_{kk} \end{pmatrix} \quad \text{„Datenmatrix“}$$

$$X \cdot \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{k1} & \cdots & x_{kk} \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix} = \begin{pmatrix} \beta_0 + \beta_1 x_{11} + \cdots \beta_k x_{1k} \\ \vdots \\ \beta_0 + \beta_1 x_{k1} + \cdots \beta_k x_{kk} \end{pmatrix}$$

Satz 5.7. Gegeben sei eine Datenmatrix X und der Datenvektor $y = \begin{pmatrix} y_0 \\ \vdots \\ y_k \end{pmatrix}$. Dann gilt für die Minimalstelle $\hat{\vec{\beta}}$ von S , falls sie existiert und eindeutig ist, folgende Formel:

$$\hat{\vec{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_k \end{pmatrix} = (X^\top X)^{-1} \cdot X^\top \vec{y},$$

sofern $(X^\top X)$ invertierbar.

Beispiel 5.8. Säureresistenz einer Plastiksorte in Abhängigkeit der Konzentration eines Zusatzstoffes:

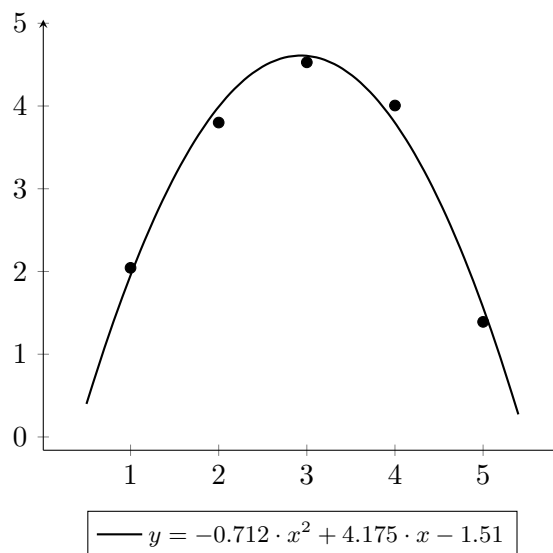
x	1	2	3	4	5
y	2,044	3,800	4,528	4,006	1,391

Ansatz: $y = \beta_0 + \beta_1 x + \beta_2 x^2$ (s. unten)

$$\vec{y} = \begin{pmatrix} 2.044 \\ 3.800 \\ 4.528 \\ 4.006 \\ 1.391 \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x & x^2 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \\ 1 & 5 & 25 \end{pmatrix}, \quad X^T \vec{y} = \begin{pmatrix} 15,771 \\ 46,212 \\ 156,884 \end{pmatrix}$$

$$X^T X = \begin{pmatrix} 5 & 15 & 55 \\ 15 & 55 & 225 \\ 55 & 225 & 979 \end{pmatrix}, \quad (X^T X)^{-1} = \begin{pmatrix} 4,6 & -3,3 & 0,5 \\ -3,3 & 2,671 & -0,428 \\ 0,5 & -0,428 & 0,071 \end{pmatrix}$$

$$\Rightarrow \hat{\vec{\beta}} = (X^T X)^{-1} \cdot X^T \vec{y} = \begin{pmatrix} -1,51 \\ 4,175 \\ -0,712 \end{pmatrix}$$



Bemerkung 5.9. Sei $n > k$.

$$X: \mathbb{R}^{k+1} \rightarrow \mathbb{R}^n$$

Gilt die Voraussetzung $\text{Rang}(X) = k + 1$, dann ist $X^T X$ invertierbar.

Ang. es existiert \vec{c} mit $(X^T X) \cdot \vec{c} = \vec{0} \Rightarrow \underbrace{\langle \vec{c}, (X^T X) \vec{c} \rangle}_{=0} = \|X \vec{c}\|^2 \Rightarrow X \vec{c} = \vec{0} \Rightarrow \vec{c} = \vec{0}$.

BEWEIS (SATZ 5.7). Es gilt

$$\begin{aligned} S(\vec{\beta}) &= \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}))^2 \\ &= (\vec{y} - X \vec{\beta})^T \cdot (\vec{y} - X \vec{\beta}) = (\vec{y}^T - (X \vec{\beta})^T) (\vec{y} - X \vec{\beta}) \\ &= \vec{y}^T \vec{y} - (X \vec{\beta})^T \cdot \vec{y} - \vec{y}^T (X \vec{\beta}) + (X \vec{\beta})^T (X \vec{\beta}) \\ &= \vec{y}^T \vec{y} - 2 \cdot \vec{\beta}^T (X^T \vec{y}) + \vec{\beta}^T (X^T X) \cdot \vec{\beta} \end{aligned}$$

Notwendige Bedingung für das Minimum:

$$\frac{\partial S}{\partial \vec{\beta}}(\vec{\beta}) = \nabla S(\vec{\beta}) = \vec{0}$$

$$\frac{\partial}{\partial \vec{\beta}} \vec{y}^\top \vec{y} = \vec{0}, \quad \frac{\partial}{\partial \vec{\beta}} (2 \cdot \vec{\beta}^\top (X^\top \vec{y})) = 2X^\top \vec{y}, \quad \frac{\partial}{\partial \vec{\beta}} (\underbrace{\vec{\beta} (X^\top X) \vec{\beta}}_{\text{symmetrisch}}) = 2(X^\top X) \vec{\beta}.$$

Setze nun

$$\nabla S(\vec{\beta}) = \vec{0} \implies (X^\top X) \vec{\beta} = X^\top \vec{y} \iff \vec{\beta} = (X^\top X)^{-1} X^\top \vec{y}.$$

□

Definition 5.10. Seien $Y_i: \Omega \rightarrow \mathbb{R}$ für $i = 1, 2, \dots, n$ gegebene Zufallsvariablen auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$. Dann heißt

$$\vec{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} : \Omega \rightarrow \mathbb{R}^n$$

ein n -dimensionaler Zufallsvektor.

Weiter sei

$$\mathbb{E}[\vec{Y}] = \begin{pmatrix} \mathbb{E}[Y_1] \\ \vdots \\ \mathbb{E}[Y_n] \end{pmatrix}$$

Satz 5.11. Sei \vec{Y} ein n -dimensionaler Zufallsvektor und sei $\vec{Z} = \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix}$ ein weiterer n -dimensionaler Zufallsvektor und sei $C = (c_{ij})$ eine Matrix mit reellen Einträge.

Dann gelten:

- 1) $\mathbb{E}[C \cdot \vec{Y}] = C \cdot \mathbb{E}[\vec{Y}]$
- 2) $\mathbb{E}[\vec{Y} + \vec{Z}] = \mathbb{E}[\vec{Y}] + \mathbb{E}[\vec{Z}]$

BEWEIS.

- 1) Es gilt

$$\begin{aligned} \mathbb{E}[C \cdot \vec{Y}] &= \mathbb{E} \left[\begin{pmatrix} \sum_{k=1}^n c_{1k} \cdot Y_k \\ \vdots \\ \sum_{k=1}^n c_{nk} \cdot Y_k \end{pmatrix} \right] = \begin{pmatrix} \mathbb{E}[\sum_{k=1}^n c_{1k} \cdot Y_k] \\ \vdots \\ \mathbb{E}[\sum_{k=1}^n c_{nk} \cdot Y_k] \end{pmatrix} \\ &= \begin{pmatrix} \sum_{k=1}^n c_{1k} \cdot \mathbb{E}[Y_k] \\ \vdots \\ \sum_{k=1}^n c_{nk} \cdot \mathbb{E}[Y_k] \end{pmatrix} = C \cdot \begin{pmatrix} \mathbb{E}[Y_1] \\ \vdots \\ \mathbb{E}[Y_n] \end{pmatrix} \end{aligned}$$

- 2) Es gilt

$$\mathbb{E}[\vec{Y} + \vec{Z}] = \mathbb{E} \left[\begin{pmatrix} Y_1 + Z_1 \\ \vdots \\ Y_n + Z_n \end{pmatrix} \right] = \begin{pmatrix} \mathbb{E}[Y_1 + Z_1] \\ \vdots \\ \mathbb{E}[Y_n + Z_n] \end{pmatrix} = \begin{pmatrix} \mathbb{E}[Y_1] + \mathbb{E}[Z_1] \\ \vdots \\ \mathbb{E}[Y_n] + \mathbb{E}[Z_n] \end{pmatrix}$$

□

Erinnerung.

$$\vec{Y} = X\vec{\beta} + \vec{\varepsilon}, \quad \vec{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Wir stellen folgende Annahmen an ε_i :

$$i) \mathbb{E}[\varepsilon_i] = 0 \quad ii) \text{Var}(\varepsilon_i) = \sigma^2$$

für $i = 1, 2, \dots, n$. Weiter seien $\varepsilon_1, \dots, \varepsilon_n$ unabhängig.

Satz 5.12. Im linearen Modell $\vec{Y} = X\vec{\beta} + \vec{\varepsilon}$ gilt, dass der Schätzer $\hat{\vec{\beta}}$ erwartungstreu ist.

BEWEIS.

$$\begin{aligned} \mathbb{E}[\hat{\vec{\beta}}] &= \mathbb{E}[(X^\top X)^{-1} \cdot X^\top \vec{Y}] = (X^\top X)^{-1} \cdot X^\top \mathbb{E}[\vec{Y}] \\ &= (X^\top X)^{-1} \cdot X^\top \underbrace{\mathbb{E}[X\vec{\beta} + \vec{\varepsilon}]}_{\mathbb{E}[X\vec{\beta}] + \mathbb{E}[\vec{\varepsilon}]} = (X^\top X)^{-1} (X^\top X) \vec{\beta} \\ &= \vec{\beta} \end{aligned}$$

□

Definition 5.13. Sei $\vec{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$ ein n -dim. Zufallsvektor. Die Varianz $\text{Var}(\vec{Y})$ ist eine $n \times n$ -Matrix mit den Einträgen

$$c_{ij} = \text{Cov}(Y_i, Y_j) = \mathbb{E}[(Y_i - \mathbb{E}[Y_i])(Y_j - \mathbb{E}[Y_j])]$$

Dabei gilt insbesondere

$$c_{ii} = \mathbb{E}[(Y_i - \mathbb{E}[Y_i])^2] = \text{Var}(Y_i).$$

Diese Matrix heißt Varianz-Kovarianz-Matrix.

Rechenregel.

$$\text{Var}(M \cdot \vec{Y}) = M \cdot \text{Var}(\vec{Y}) \cdot M^\top$$

Satz 5.14. Es gilt

$$\text{Var}(\hat{\vec{\beta}}) = \sigma^2 (X^\top X)^{-1}.$$

BEWEIS. Wir wissen: $\hat{\vec{\beta}} = (X^\top X)^{-1} X^\top \vec{Y}$

Beachte, dass

$$\text{Cov}(Y_i, Y_j) = \begin{cases} \sigma^2 & i = j \\ 0 & i \neq j \end{cases}$$

$$\begin{aligned} \text{Var}(\hat{\vec{\beta}}) &= \text{Var}((X^\top X)^{-1} X^\top \vec{Y}) = (X^\top X)^{-1} X^\top \cdot \text{Var}(\vec{Y}) \cdot ((X^\top X)^{-1} X^\top)^\top \\ &= (X^\top X)^{-1} X^\top \cdot \sigma^2 I \cdot X (X^\top X)^{-1} = \sigma^2 (X^\top X)^{-1}. \end{aligned}$$

□

A Tabelle zur Standardnormalverteilung

Die folgende Tabelle enthält die Werte der Verteilungsfunktion der Standardnormalverteilung.

Die Tabelle ist zu lesen: $\Phi(1,11) = 0,8665$, dabei gilt: $\Phi(-z) = 1 - \Phi(z)$ für $z \geq 0$.

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990

Ausgewählte Quantile z_γ der Standardnormalverteilung:

γ	0.750	0.800	0.850	0.900	0.950	0.975	0.990	0.995	0.999	0.9995
z_γ	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.290

B Tabelle zur Student- t -Verteilung

Die folgende Tabelle enthält die Werte der Quantile $t_{1-\alpha,n}$ für die Student- t -Verteilung.

Die Tabelle ist wie folgt zu lesen: Sei $n = 50$, $1 - \alpha = 0,999 \implies t_{0,999,50} = 3.261$.

Wie aus der Zeile $n = \infty$ zu entnehmen, gilt $t_{1-\alpha,n} = z_{1-\alpha}$ für $n \rightarrow \infty$. Als Faustregel gilt für $n > 30$: $t_{1-\alpha,n} \approx z_{1-\alpha}$ als Approximation, wobei $z_{1-\alpha}$ das $(1 - \alpha)$ -Quantil der Standardnormalverteilung.

$n \backslash 1 - \alpha$	0.75	0.8	0.9	0.95	0.975	0.99	0.995	0.999
1	1.000	1.376	3.078	6.314	12.706	31.821	63.657	318.31
2	0.816	1.061	1.886	2.920	4.303	6.965	9.925	22.327
3	0.765	0.978	1.638	2.353	3.182	4.541	5.841	10.215
4	0.741	0.941	1.533	2.132	2.776	3.747	4.604	7.173
5	0.727	0.920	1.476	2.015	2.571	3.365	4.032	5.893
6	0.718	0.906	1.440	1.943	2.447	3.143	3.707	5.208
7	0.711	0.896	1.415	1.895	2.365	2.998	3.499	4.785
8	0.706	0.889	1.397	1.860	2.306	2.896	3.355	4.501
9	0.703	0.883	1.383	1.833	2.262	2.821	3.250	4.297
10	0.700	0.879	1.372	1.812	2.228	2.764	3.169	4.144
11	0.697	0.876	1.363	1.796	2.201	2.718	3.106	4.025
12	0.695	0.873	1.356	1.782	2.179	2.681	3.055	3.930
13	0.694	0.870	1.350	1.771	2.160	2.650	3.012	3.852
14	0.692	0.868	1.345	1.761	2.145	2.624	2.977	3.787
15	0.691	0.866	1.341	1.753	2.131	2.602	2.947	3.733
16	0.690	0.865	1.337	1.746	2.120	2.583	2.921	3.686
17	0.689	0.863	1.333	1.740	2.110	2.567	2.898	3.646
18	0.688	0.862	1.330	1.734	2.101	2.552	2.878	3.610
19	0.688	0.861	1.328	1.729	2.093	2.539	2.861	3.579
20	0.687	0.860	1.325	1.725	2.086	2.528	2.845	3.552
21	0.686	0.859	1.323	1.721	2.080	2.518	2.831	3.527
22	0.686	0.858	1.321	1.717	2.074	2.508	2.819	3.505
23	0.685	0.858	1.319	1.714	2.069	2.500	2.807	3.485
24	0.685	0.857	1.318	1.711	2.064	2.492	2.797	3.467
25	0.684	0.856	1.316	1.708	2.060	2.485	2.787	3.450
26	0.684	0.856	1.315	1.706	2.056	2.479	2.779	3.435
27	0.684	0.855	1.314	1.703	2.052	2.473	2.771	3.421
28	0.683	0.855	1.313	1.701	2.048	2.467	2.763	3.408
29	0.683	0.854	1.311	1.699	2.045	2.462	2.756	3.396
30	0.683	0.854	1.310	1.697	2.042	2.457	2.750	3.385
50	0.679	0.849	1.299	1.676	2.009	2.403	2.678	3.261
60	0.679	0.848	1.296	1.671	2.000	2.390	2.660	3.232
∞	0.674	0.842	1.282	1.645	1.960	2.326	2.576	3.090