Mathematische Grundlagen der Datenanalyse

Lars Grüne
Mathematisches Institut
Fakultät für Mathematik und Physik
Universität Bayreuth
95440 Bayreuth
lars.gruene@uni-bayreuth.de
http://num.math.uni-bayreuth.de

Vorlesungsskript

Zweite Auflage

Wintersemester 2024/2025

Vorwort

Dieses Skript ist im Rahmen einer gleichnamigen Vorlesung entstanden, die im Wintersemester 2024/2025 an der Universität Bayreuth gehalten wird. Es handelt sich um die zweite Auflage, in der einige Fehler der ersten Auflage behoben wurden. Ich bedanke mich bei allen, die mir Fehler in der ersten Auflage des Skripts mitgeteilt haben.

Bei der Ausarbeitung von Kapitel 2–9 des Skripts war mir das im Springer Verlag erschienene Buch *Algorithmische Mathematik* von Helmut Harbrecht und Michael Multerer (online unter https://doi.org/10.1007/978-3-642-41952-2) eine große Hilfe. Zudem bedanke ich mich bei meinem Kollegen Andreas Christmann für die Bereitstellung seines Skripts zu der Vorlesung "Statistik für Studierende der Informatik", das mir bei einigen über die genannte Referenz hinausgehenden Themen sehr gute Dienste geleistet hat.

Bayreuth, Februar 2025 Lars Grüne

Inhaltsverzeichnis

Kapitel 1

Einführung

Diese Vorlesung gibt eine Einführung in die mathematischen Grundlagen aus dem Gebiet der Stochastik, die benötigt werden, um Algorithmen und Methoden zur Datenanalyse zu verstehen. Um diese Motivation zu illustrieren, werden ausgewählte Algorithmen zur Datenanalyse in den letzten beiden Kapiteln 11 und 12 der Vorlesung behandelt. Da die Algorithmen neben der Stochastik auch Konzepte aus der linearen Algebra benötigen, werden diese in Kapitel 10 wiederholt.

Des Großteil der Vorlesung, Kapitel 2–9, widmet sich aber den Konzepten der Stochastik, die von Grund auf und so weit im Rahmen des Umfangs dieser Vorlesung möglich mathematisch rigoros eingeführt und mit vielen Beispielen erläutert werden. Ziel ist es, Sprache und Denkweise der Stochastik so weit zu vermitteln, dass die zu Grunde liegenden Ideen der an Ende der Vorlesung und in weiteren Vorlesungen behandelten Algorithmen verständlich werden.

Kapitel 2

Messbare Räume und Rechenregeln

2.1 Mengen

In der Wahrscheinlichkeitstheorie spielen *Mengen* eine wichtige Rolle, weil sie die möglichen Ereignisse zusammenfassen. Eine vollständge formale Definition von Mengen benötigen wir an dieser Stelle nicht. Es genügt die intuitive Vorstellung von einer Menge als einer Zusammenfassung von paarweise verschiedenen Objekten, den *Elementen* der Menge. Die Elemente werden durch geschweifte Klammern zu einer Menge zusammengefasst, wie z.B. in

$$\{1, 2, 3, 4, 5, 6\}, \quad \{\text{Kopf}, \text{Zahl}\}, \quad \{1, 2, 3, \ldots\} = \mathbb{N}$$

($\mathbb{N}$ ist die Menge der natürlichen Zahlen.) Wenn a ein Element einer Menge A ist, so schreiben wir $a \in A$, ansonsten $a \notin A$.

Mit einem senkrechten Strich können Bedingungen an die Elemente angegeben werden, z.B.

$$\{n \in \mathbb{N} \mid n \text{ ungerade}\} = \{1, 3, 5, 7, \ldots\}.$$

Manche Mengen benötigen eine kompliziertere Definition, wie z.B. die Menge $\mathbb{R}$ der reellen Zahlen, deren Definition wir hier nicht angeben, die wir aber nichtsdestotrotz verwenden (informell kann man sich $\mathbb{R}$ als die Menge aller endlichen und unendlichen Dezimalzahlen vorstellen). Mit $\emptyset$ oder $\{\}$ wird die *leere Menge* bezeichnet.

In der Wahrscheinlichkeitstheorie bezeichnet man die Grundmenge eines Wahrscheinlichkeitsexperiments mit Ω ("Omega") und ihre Elemente mit ω (das ist ein kleines "omega", kein "w"). Enthält Ω nur endlich viele Elemente, so bezeichnen wir deren Anzahl mit $|\Omega|$. Die erste Menge aus dem obigen Beispiel könnte also z.B. die Ergebnisse eines Würfelwurfes beschreiben, die zweite Menge die Ergebnisse eines Münzwurfs.

Die folgenden Mengenoperationen sollten bekannt sein, werden aber der Vollständigkeit halber kurz wiederholt.

- Eine Menge A ist eine Teilmenge von Ω, falls alle Elemente $\omega \in A$ auch Elemente in Ω sind, also $\omega \in \Omega$ erfüllen. Wir schreiben in diesem Fall $A \subset \Omega$. Falls A keine

Teilmenge von Ω ist, schreiben wir $A \not\subset \Omega$

Beispiel: $\{1,2,3,4\} \subset \mathbb{N}$, aber $\{-1,0,1,2,3\} \not\subset \mathbb{N}$.

- Die Menge aller Teilmengen von Ω ist die *Potenzmenge* $\mathcal{P}(\Omega)$, formal

$$\mathcal{P}(\Omega) := \{A \,|\, A \subset \Omega\}.$$

Beispiel: $\Omega = \{\text{Kopf}, \text{Zahl}\} \Rightarrow \mathcal{P}(\Omega) = \big\{\{\}, \{\text{Kopf}\}, \{\text{Zahl}\}, \{\text{Kopf}, \text{Zahl}\}\big\}$

- Für eine Teilmenge $A \subset \Omega$ bezeichnet

$$A^c := \{\omega \in \Omega \,|\, \omega \notin A\}$$

das *Komplement* von A

Beispiele:

  - $\Omega = \{1,2,3,4,5,6\}$, $A = \{1,2,3\} \Rightarrow A^c = \{4,5,6\}$
  - Für $A = \Omega$ gilt $A^c = \emptyset$

- Die *Vereinigungsmenge* von $A, B \subset \Omega$ ist definiert durch

$$A \cup B := \{\omega \in \Omega \,|\, \omega \in A \text{ oder } \omega \in B\}$$

und die *Schnittmenge* als

$$A \cap B := \{\omega \in \Omega \,|\, \omega \in A \text{ und } \omega \in B\}.$$

Beispiel: $\Omega = \{1,2,3,4,5,6\}$, $A = \{2,4,6\}$, $B = \{1,2,3\} \Rightarrow A \cup B = \{1,2,3,4,6\}$, $A \cap B = \{2\}$.

- Die *Differenzmenge* von $A, B \subset \Omega$ ist definiert als

$$A \setminus B := \{\omega \in A \,|\, \omega \notin B\} = A \cap B^c.$$

Beispiel: $\Omega = \{1,2,3,4,5,6\}$, $A = \{2,4,6\}$, $B = \{1,2,3\} \Rightarrow A \setminus B = \{4,6\}$.

Der folgende Satz fasst wichtige Rechenregeln für Mengen zusammen. Die Beweise ergeben sich für die allermeisten Aussagen direkt aus den Definitionen und werden daher hier nicht betrachtet. Für Punkt 4) geben wir im nachfolgenden Satz einen Beweis für eine verallgemeinerte Version.

Satz 2.1 Sei Ω eine Menge und seien $A, B, C \subset \Omega$. Dann gelten

1) $A \cup B = B \cup A, \quad A \cap B = B \cap A$ (Kommmutativgesetz)

2) $(A \cup B) \cup C = A \cup (B \cup C), \quad (A \cap B) \cap C = A \cap (B \cap C)$ (Assoziativgesetz)

3) $(A \cup B) \cap C = (A \cap C) \cup (B \cap C), \quad (A \cap B) \cup C = (A \cup C) \cap (B \cup C)$ (Distributivgesetz)

4) $(A \cup B)^c = A^c \cap B^c, \quad (A \cap B)^c = A^c \cup B^c$ (de Morgan'sche Regeln)

5) $A \cup \emptyset = A, \quad A \cup \Omega = \Omega, \quad A \cap \emptyset = \emptyset, \quad A \cap \Omega = A.$

Die de Morgan'schen Regeln können auf endlich oder sogar abzählbar unendlich viele[1] Mengen A_i erweitert werden.

Satz 2.2 Sei Ω eine Menge und seien $A_i \subset \Omega$, $i \in \mathbb{N}$. Dann gilt für jedes $n \in \mathbb{N}$

$$(A_1 \cup A_2 \cup \ldots \cup A_n)^c = A_1^c \cap A_2^c \cap \ldots \cap A_n^c$$

sowie

$$\left(\bigcup_{i=1}^{\infty} A_i \right)^c = \bigcap_{i=1}^{\infty} A_i^c.$$

Beide Aussagen gelten analog für die vertauschten Operationen $\cup$ und $\cap$.

Beweis: Es genügt, den Fall für unendlich viele A_i zu beweisen, da der Fall der endlich vielen Mengen als Spezialfall davon mit $A_{n+1} = A_{n+2} = \ldots = \emptyset$ geschrieben werden kann.

Eine Gleichheit von zwei Mengen beweist man üblicherweise am einfachsten, wenn man die beiden Inklusionen $\subset$ und $\supset$ separat beweist. So gehen wir hier auch vor.

" $(\bigcup_{i=1}^{\infty} A_i)^c \subset \bigcap_{i=1}^{\infty} A_i^c$ ": Hierfür müssen wir zeigen, dass jedes $\omega \in (\bigcup_{i=1}^{\infty} A_i)^c$ auch in $\bigcap_{i=1}^{\infty} A_i^c$ liegt. Die Relation $\omega \in \bigcup_{i=1}^{\infty} A_i$ gilt genau dann, wenn es mindestens ein $i \in \mathbb{N}$ gibt mit $\omega \in A_i$. Die Relation $\omega \in (\bigcup_{i=1}^{\infty} A_i)^c$ gilt also genau dann, wenn es kein solches i gibt, wenn also $\omega \notin A_i$ gilt für alle $i \in \mathbb{N}$. Dies wiederum bedeutet, dass $\omega \in A_i^c$ gilt für all $i \in \mathbb{N}$ und damit gilt auch $\omega \in \bigcap_{i=1}^{\infty} A_i^c$.

" $(\bigcup_{i=1}^{\infty} A_i)^c \supset \bigcap_{i=1}^{\infty} A_i^c$ ": Nun müssen wir umgekehrt zeigen, dass für jedes $\omega \in \bigcap_{i=1}^{\infty} A_i^c$ auch $\omega \in (\bigcup_{i=1}^{\infty} A_i)^c$ gilt. Sei dazu $\omega \in \bigcap_{i=1}^{\infty} A_i^c$. Dann muss ω in allen Mengen A_i^c enthalten sein, also ist ω in keiner der Mengen A_i enthalten. Damit liegt ω auch nicht in $\bigcup_{i=1}^{\infty} A_i$, also liegt es im Komplement dieser Menge, also $\omega \in (\bigcup_{i=1}^{\infty} A_i)^c$. □

2.2 Zufallsexperimente und Zufallsereignisse

Wir wollen nun die Mengen Ω dazu verwenden, um die Ergebnisse von Zufallsexperimenten zu modellieren. Dabei verwenden wir die folgenden Begriffe.

- Ein *Zufallsexperiment* ist ein Experiment mit zufälligem Ausgang unter klar definierten Bedingungen. Dies kann z.B. der Wurf einer Münze oder eines Würfels sein oder auch die mit einem zufälligen Messfehler behaftete Messung einer Größe in einem physikalischen oder chemischen Versuch, z.B. beim Wiegen einer Substanz oder beim Messen einer Temperatur.

- Die *möglichen Ergebnisse* des Zufallsexperiments werden in der Menge Ω zusammengefasst. Beispiele finden sich in Beispiel 2.3.

[1] Man spricht von *abzählbar unendlich* vielen Objekten, wenn diese mit Zahlen aus $\mathbb{N}$ durchnummeriert werden können.

- Eine *Realisierung* des Zufallsexperiments entspricht der Durchführung des Experiments. Mathematisch modellieren wir dies mit dem Ziehen eines zufälligen ω aus Ω. Dieses ω wird dann ebenfalls *Realisierung* genannt.

Beispiel 2.3 Hier einige Beispiele für die Wahl von Ω in verschiedenen Zufallsexperimenten.

- Münzwurf: $\Omega = \{\text{Kopf}, \text{Zahl}\}$
- Werfen eines Würfels: $\Omega = \{1, 2, 3, 4, 5, 6\}$
- Zweimaliges Werfen eines Würfels: hier schreiben wir die Ergebnisse als Paare aus dem Ergebnis des ersten und des zweiten Wurfs, also

$$\begin{aligned}\Omega = \{&(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),(2,1),(2,2),(2,3),(2,4),(2,5),(2,6),\\ &(3,1),(3,2),(3,3),(3,4),(3,5),(3,6),(4,1),(4,2),(4,3),(4,4),(4,5),(4,6),\\ &(5,1),(5,2),(5,3),(5,4),(5,5),(5,6),(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\}.\end{aligned}$$

- Bei Messfehlern kann im Prinzip jede reelle Zahl als Ergebnis des Experiments herauskommen, also $\Omega = \mathbb{R}$. Wenn man a priori Informationen über die Genauigkeit der Messvorrichtung hat, dann kann man diesen eventuell einschränken, also $\Omega = [a, b]$ für geeignete Werte $a, b \in \mathbb{R}$ mit $a < b$.

□

Das Ergebnis eines Zufallsexperiments ist naturgemäß unsicher. Man unterscheidet bei der Einschätzung von Unsicherheit zwischen *epistemischer* und *aleatorischer* Unsicherheit. Die epistemische Unsicherheit entsteht auf Grund von ungenauen oder mangelnden Informationen. Die Messfehler aus dem letzten Teil von Beispiel 2.3 sind ein typischer Beispiel hierfür. Die aleatorische Unsicherheit entsteht auf Grund von tatsächlichen Zufälligkeiten, die sich auch durch genauere oder mehr Informationen nicht beseitigen lassen. Das Werfen von Würfeln und Münzen — jedenfalls in der idealisierten Form — fällt in diese Kategorie.

Definition 2.4 Sei ein Zufallsexperiment mit möglichen Ergebnissen Ω gegeben. Dann wird eine Teilmenge $A \subset \Omega$ *zufälliges Ereignis* genannt. Falls eine Realisierung $\omega \in A$ erfüllt, so sagen wir, dass das Ereignis A *eingetreten* ist.

Ein Ereignis A, das aus genau einem $\omega \in \Omega$ besteht, heißt *Elementarereignis*. Das Ereignis $A = \Omega$ heißt *sicheres Ereignis* und das Ereignis $A = \emptyset$ heißt *unmögliches Ereignis*. □

Beispiel 2.5

- Beim Werfen eines Würfels ist das Ereignis "es wird eine gerade Zahl gewürfelt" durch $A = \{2, 4, 6\}$ gegeben.
- Beim zweimaligen Werfen eines Würfel ist das Ereignis "die Augensumme ist größer oder gleich 10" durch $A = \{(4,6),(5,5),(5,6),(6,4),(6,5),(6,6)\}$ gegeben.

- Beim Messen einer Größe ist das Ereignis "der Messfehler ist im Betrag kleiner als 10^{-2}" durch $\Omega = [-10^{-2}, 10^{-2}]$ gegeben.

□

In der mathematischen Modellierung eines Zufallsexperiments wird man in der Regel nicht alle möglichen Ereignisse betrachten sondern nur eine Auswahl davon. Dies dient in der Regel dazu, die Komplexität des Modells zu verringern. Ist Ω endlich, so kann dadurch z.B. die Anzahl der möglichen Ereignisse verringert werden auf die, die für den betrachteten Fall wichtig sind. Interessiert uns z.B. beim zweimaligen Werfen eines Würfels nur, ob die Augensumme gerade oder ungerade ist, brauchen wir auch nur diese beiden Ereignisse definieren. Ist Ω unendlich, so gibt es bereits unendlich viele Elementarereignisse und dadurch eine noch viel größere Menge an Ereignissen. Hier möchte man sich i.d.R. auf Ereignismengen einschränken, auf denen man noch sinnvoll rechnen kann. Dies ist z.B. bei der Definition der Borel'schen σ-Algebra der Fall, die uns in Definition 6.1 begegnen wird.

Ganz beliebig kann die Menge der "erlaubten" Ereignismengen aber nicht gewählt werden. Im Folgenden werden wir die nötige Struktur dieser Menge von Mengen klären. Zunächst übersetzen wir dazu einige Mengenoperationen und -relationen in die Sprache der Ereignisse.

Definition 2.6 Gegeben seien zwei Ereignismengen $A, B \subset \Omega$. Dann sagen wir:

- Das Ereignis A *oder* B tritt ein, wenn $\omega \in A \cup B$.
- Das Ereignis A *und* B tritt ein, wenn $\omega \in A \cap B$.
- Das Ereignis A *impliziert* das Ereignis B (oder das Ereignis B *folgt aus* dem Ereignis A), wenn $A \subset B$.
- Die Ereignisse A und B sind *unvereinbar*, wenn $A \cap B = \emptyset$.
- Das Ereignis A^c ist das *Gegenereignis* zum Ereignis A. Es tritt genau dann ein, wenn A nicht eintritt.

□

Beispiel 2.7 Betrachte das zweimalige Würfeln und die beiden Ereignisse A "der erste Würfel zeigt eine 6" sowie B "die Augensumme ist größer oder gleich 7". Dann ist

$$A = \{(6,1), (6,2), (6,3), (6,4), (6,5), (6,6)\}$$

und

$$\begin{aligned} B = \{&(1,6), (2,5), (2,6), (3,4), (3,5), (3,6), (4,3), (4,4), (4,5), (4,6), \\ &(5,2), (5,3), (5,4), (5,5), (5,6), (6,1), (6,2), (6,3), (6,4), (6,5), (6,6)\}. \end{aligned}$$

Offenbar ist $A \subset B$, also folgt das Ereignis B aus dem Ereignis A. Das ist auch anschaulich einleuchtend, denn wenn der erste Wurf eine 6 ist, beträgt die Augensumme immer mindestens 7. □

Genau wie bei den de Morgan'schen Gesetzen können die oder- und und-Ereignisse auch für abzählbar unendlich viele Mengen mittels

$$\omega \in \bigcup_{i=1}^{\infty} A_i \quad \text{bzw.} \quad \omega \in \bigcap_{i=1}^{\infty} A_i$$

definiert werden.

Die Menge $\mathcal{A}$ aller möglichen Ereignisse A soll nun erfüllen, dass sie alle Gegenereignisse und alle oder-Ereignisse enthält. Eine solche Menge von Mengen nennt man σ-Algebra (gesprochen sigma-Algebra). Sie ist formal wie folgt definiert:

Definition 2.8 Sei Ω eine Ergebnismenge. Dann heißt eine Menge $\mathcal{A}$ von Teilmengen $A \subset \Omega$ *σ-Algebra*, wenn gilt

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

2) $A \in \mathcal{A} \;\Rightarrow\; A^c \in \mathcal{A}$

3) $A_i \in \mathcal{A}$ für alle $i \in \mathbb{N} \;\Rightarrow\; \bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$

Das Paar $(\Omega, \mathcal{A})$ heißt dann *messbarer Raum* und die Mengen $A \in \mathcal{A}$ *messbare Mengen*. □

Beachte, dass 3) auch für endlich viele Mengen $A_i \in \mathcal{A}$, $i = 1, \ldots, n$ gilt, indem man die verbleibenden Mengen als $A_{n+1} = A_{n+2} = \ldots = \emptyset$ wählt.

Aus 1) und 2) folgt, dass auch $\emptyset = \Omega^c \in \mathcal{A}$ gilt und dass aus 2) und 3) folgt, dass auch für die und-Ereignisse

$$\bigcap_{i=1}^{\infty} A_i \in \mathcal{A}$$

gilt, denn:

$$A_i \in \mathcal{A} \overset{2)}{\Rightarrow} A_i^c \in \mathcal{A} \overset{3)}{\Rightarrow} \bigcup_{i=1}^{\infty} A_i^c \in \mathcal{A} \overset{\text{de Morgan}}{\Longrightarrow} \left(\bigcap_{i=1}^{\infty} A_i\right)^c \in \mathcal{A} \overset{2)}{\Rightarrow} \bigcap_{i=1}^{\infty} A_i \in \mathcal{A}.$$

Genau wie in Punkt 3) der Definition gilt dies auch für endlich viele Mengen. Zusammen mit 2) folgt damit schließlich, dass für $A_1, A_2 \in \mathcal{A}$ auch $A_1 \setminus A_2 = A_1 \cap A_2^c \in \mathcal{A}$ gilt.

Man kann den Sinn der Bedingungen aus dieser Definition plausibel machen, wenn man sich überlegt, welche weiteren Ereignisse man aus direkt beobachtbaren Ereignissen logisch ableiten kann. Nehmen wir zur Illustration den Wurf eines Würfels und nehmen wir an, dass wir das exakte Ergebnis nicht sehen können, dass uns aber jemand sagt, ob eine gerade Zahl geworfen wurde oder nicht und ob eine Augenzahl größer oder gleich 5 geworfen wurde oder nicht. Wir können aus den Angaben also direkt ermitteln, ob die Ereignisse $A_1 = \{2, 4, 6\}$ und $A_2 = \{5, 6\}$ eingetreten sind. Damit können wir aber auch ermitteln ob $A_1 \cup A_2$ und $A_1 \cap A_2$ eingetreten ist. Ebenso können wir ermitteln, ob die Gegenereignisse

eingetreten sind und ob überhaupt gewürfelt wurde, also ob das Ereignis Ω eingetreten ist. Die durch logische Schlüsse ableitbaren Informationen über eingetretene Ereignisse sind also alle wieder in jeder σ-Algebra enthalten, die A_1 und A_2 enthält. Die Bedingungen an die σ-Algebra $\mathcal{A}$ stellen also sicher, dass $\mathcal{A}$ die gesamten Ereignisse, die aus direkt beobachtbaren Ereignissen $A \in \mathcal{A}$ logisch ableitbar sind, ebenfalls enthält. Wenn man die zur Verfügung stehenden Information als "Messungen" in einem Experiment interpretiert, sind die Mengen $A \in \mathcal{A}$ also genau die Ereignisse, deren Gültigkeit man aus den Messungen ableiten kann. Dies erklärt den Begriff "messbar" in Definition 2.8.

Falls $\mathcal{A}$ unendlich viele Ereignisse enthält, so stellt die Tatsache, dass 3) auch für Vereinigungen abzählbar unendlich vieler Mengen gelten muss, sicher, dass auch Ereignisse, die sich aus abzählbar unendlich vielen direkten Ereignissen ableiten lassen, wieder in $\mathcal{A}$ liegen.

Kennt man nun die direkt beobachtbaren Ereignisse, stellt sich die Frage, ob es eine σ-Algebra gibt, die *genau* die Ereignisse enthält, die man aus diesen ableiten kann (und nicht mehr!). Der folgende Satz zeigt, dass dies der Fall ist.

Satz 2.9 Sei Ω eine Ergebnismenge und $\mathcal{S} \subset \mathcal{P}(\Omega)$ eine Menge von Ereignissen. Dann ist

$$\sigma(\mathcal{S}) := \bigcap_{\substack{\mathcal{A} \subset \mathcal{P}(\mathcal{A}) \\ \mathcal{S} \subset \mathcal{A},\, \mathcal{A}\,\text{ist}\,\sigma-\text{Algebra}}} \mathcal{A}$$

die kleinste σ-Algebra mit $\mathcal{S} \subset \sigma(\mathcal{S})$. Sie wird die *von* $\mathcal{S}$ *erzeugte* σ*-Algebra* genannt.

Beweis: Wir weisen die drei Bedingungen aus Definition 2.8 nach. Dazu definieren wir

$$\mathcal{M} := \{\mathcal{A} \subset \mathcal{P}(\Omega) \,|\, \mathcal{S} \subset \mathcal{A},\, \mathcal{A}\,\text{ist}\,\sigma-\text{Algebra}\}.$$

Damit ist $\sigma(\mathcal{S}) = \bigcap_{\mathcal{A} \in \mathcal{M}} \mathcal{A}$. Weil $\mathcal{S} \in \mathcal{P}(\Omega)$ gilt und $\mathcal{P}(\Omega)$ eine σ-Algebra ist, gilt $\mathcal{P}(\Omega) \in \mathcal{M}$. Damit enthält $\mathcal{M}$ mindestens eine σ-Algebra.

1) Für alle $\mathcal{A} \in \mathcal{M}$ gilt 1), also ist $\Omega \in \mathcal{A}$. Damit liegt Ω in $\sigma(\mathcal{S})$ und es folgt 1) für $\sigma(\mathcal{S})$.

2) Sei $A \in \sigma(\mathcal{S})$. Dann gilt $A \in \mathcal{A}$ für alle $\mathcal{A} \in \mathcal{M}$. Wegen 2) ist dann auch $A^c \in \mathcal{A}$ für alle $\mathcal{A} \in \mathcal{M}$ und damit $A^c \in \sigma(\mathcal{S})$. Also gilt 2) für $\sigma(\mathcal{S})$.

3) Seien schließlich $A_i \in \sigma(\mathcal{S})$ für $i \in \mathbb{N}$. Dann gilt für alle $\mathcal{A} \in \mathcal{M}$ die Relation $A_i \in \mathcal{A}$ für alle $n \in \mathbb{N}$, und weil jedes $\mathcal{A}$ 3) erfüllt, folgt $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$ für alle $\mathcal{A} \in \mathcal{M}$. Damit folgt $\bigcup_{i=1}^{\infty} A_i \in \sigma(\mathcal{S})$, also 3). □

Der Satz gibt leider keine Konstruktionsvorschrift für $\sigma(S)$ an. Eine solche gibt es im allgemeinen auch nicht. Für einfache Beispiele kann man $\sigma(S)$ aber direkt aus den Bedingungen konstruieren, wie das folgende Beispiel zeigt.

Beispiel 2.10 Betrachte wieder das Werfen eines Würfels mit $\Omega = \{1, 2, 3, 4, 5, 6\}$. Wir nehmen an, dass wir das Ergebnis unseres Wurfs nicht sehen können, dass uns aber jemand sagt, ob die Augenzahl gerade ist und ob sie ≥ 5 ist. Direkt entscheiden können wir also, ob die Ereignisse

$$S = \big\{\{2, 4, 6\}, \{5, 6\}\big\}$$

vorliegen. Aus Bedingung 1) aus Definition 2.8 folgt nun zunächst, dass Ω in $\sigma(S)$ liegen muss. Aus 2) folgt dann, dass $\Omega^c = \emptyset$, $\{2,4,6\}^c = \{1,3,5\}$ und $\{5,6\}^c = \{1,2,3,4\}$ in $\sigma(S)$ liegen müssen.

Mit 3) und der Beobachtung, dass auch Schnitte von Mengen wieder in der σ-Algebra liegen müssen, folgt nun, dass auch

$$\{2,4,6\} \cup \{5,6\} = \{2,4,5,6\}, \quad \{2,4,6\} \cup \{1,2,3,4\} = \{1,2,3,4,6\},$$

$$\{5,6\} \cup \{1,3,5\} = \{1,3,5,6\}, \quad \{1,3,5\} \cup \{1,2,3,4\} = \{1,2,3,4,5\}$$

$$\{2,4,6\} \cap \{5,6\} = \{6\}, \quad \{2,4,6\} \cap \{1,2,3,4\} = \{2,4\},$$

$$\{5,6\} \cap \{1,3,5\} = \{5\}, \quad \{1,3,5\} \cap \{1,2,3,4\} = \{1,3\}$$

in $\sigma(S)$ liegen (nicht aufgeführte Kombinationen führen auf bereits genannte Mengen).

Wendet man nun 2) und 3) auf die neu entstandenen Mengen an, so liefern nur die Vereinigungen

$$\{6\} \cup \{1,3\} = \{1,3,6\}, \quad \{2,4\} \cup \{5\} = \{2,4,5\}$$

Mengen hervor, die noch nicht aufgetreten sind. Weitere Vereinigungen, Schnitte und Komplementbildungen bringen keine neuen Mengen hervor. Wir erhalten damit

$$\begin{aligned} \sigma(S) \quad = \quad & \big\{\{\emptyset,\{5\},\{6\},\{1,3\},\{2,4\},\{5,6\} \\ & \{1,3,5\},\{1,3,6\},\{2,4,5\},\{2,4,6\} \\ & \{1,2,3,4\},\{1,3,5,6\},\{2,4,5,6\}, \\ & \{1,2,3,4,5\},\{1,2,3,4,6\},\Omega\}. \end{aligned}$$

□

Kapitel 3

Wahrscheinlichkeitsverteilungen und Kombinatorik

Nachdem wir nun Ereignisse eingeführt haben, wollen wir als nächstes formalisieren, mit welcher Wahrscheinlichkeit ein Ereignis als Ergebnis eines Zufallsexperiments herauskommen kann. Um die dafür notwendigen Bedingungen zu motivieren, betrachten wir zunächst die Wahrscheinlichkeiten, die durch Experimente empirisch bestimmt werden können.

Definition 3.1 Es sei $A \in \mathcal{A}$ ein Ereignis eines Zufallsexperiments, das n mal unabhängig (also sich gegenseitig nicht beeinflussend) durchgeführt wird.

(a) Mit $h_n(A)$ bezeichnen wir die Anzahl der Experimente, in denen das Ereignis A eingetreten ist. Die Größe $h_n(A)$ heißt die *absolute Häufigkeit*.

(b) Mit

$$H_n(A) := \frac{h_n(A)}{n}$$

bezeichnen wir die *relative Häufigkeit* von A. □

Man überlegt sich leicht, dass $H_n(A)$ immer zwischen 0 und 1 liegt, weil $h_n(A)$ nicht kleiner als 0 und nicht größer als n sein kann. Der folgende Satz gibt weitere Eigenschaften von $H_n(A)$ an.

Satz 3.2 Sei $(\Omega, \mathcal{A})$ ein messbarer Raum (vgl. Definition 2.8). Dann gilt für alle $A \in \mathcal{A}$, dass $0 \leq H_n(A) \leq 1$ und dass $H_n(\Omega) = 1$, und für zwei unvereinbare Ereignisse $A, B \in \mathcal{A}$ (also $A \cap B = \emptyset$) gilt

$$H_n(A \cup B) = H_n(A) + H_n(B).$$

Beweis: Die absolute Häufigkeit $h_n(A)$ ist immer größer oder gleich 0 und immer kleiner oder gleich n, also liegt $H_n(A)$ zwischen 0 und 1.

Das Ereignis Ω tritt bei jedem Experiment ein, also ist $h_n(\Omega) = n$ und damit $H_n(\Omega) = 1$.

Für zwei unvereinbare Ereignisse A und B können wir die Ergebnisse $\omega_1, \ldots, \omega_n$ der n Zufallsexperimente *eindeutig* aufteilen in drei Gruppen:

$$(1)\ \omega_i \in A, \qquad (2)\ \omega_i \in B, \qquad (3)\ \omega_i \text{ liegt weder in } A \text{ noch in } B,$$

denn ein ω_i, das sowohl in A als auch in B liegt, kann es wegen $A \cap B = \emptyset$ nicht geben. Dann ist $h_n(A)$ gerade die Anzahl der ω_i aus (1) und $h_n(B)$ die Anzahl der ω_i aus (2). Da zudem $h_n(A \cup B)$ gerade die Anzahl der ω_i aus (1) und (2) ist, folgt $h_n(A \cup B) = h_n(A) + h_n(B)$ und damit

$$H_n(A \cup B) = \frac{h_n(A \cup B)}{n} = \frac{h_n(A) + h_n(B)}{n} = \frac{h_n(A)}{n} + \frac{h_n(B)}{n} = H_n(A) + H_n(B).$$

□

Oft ist es so, dass die Zahl $H_n(A)$ für $n \to \infty$ gegen einen festen Wert konvergiert. Betrachtet man beim Würfeln mit einem idealen Würfel z.B. das Ereignis $A = \{1\}$, so sollte $H_n(A)$ gegen $1/6$ konvergieren. Grafisch kann man dies gut erkennen, wenn man die Ergebnisse der Zufallsexperimente in einem *Histogramm* darstellt. Dies ist die grafische Darstellung der absoluten Häufigkeiten $h_n(A)$ eines Zufallsexperiments als Säulendiagramm. Abbildung 3.1 zeigt das Histogramm (auf Basis einer Computersimulation des Würfelns) für die Ereignisse $A = \{\omega\}$, $\omega = 1, 2, \ldots, 6$ für $n = 100, 10\,000$ und $1\,000\,000$ Würfe. Die Tatsache, dass alle Zahlen gleich wahrscheinlich sind und die relativen Häufigkeiten gegen $1/6$ konvergieren, ist hier optisch gut zu erkennen.

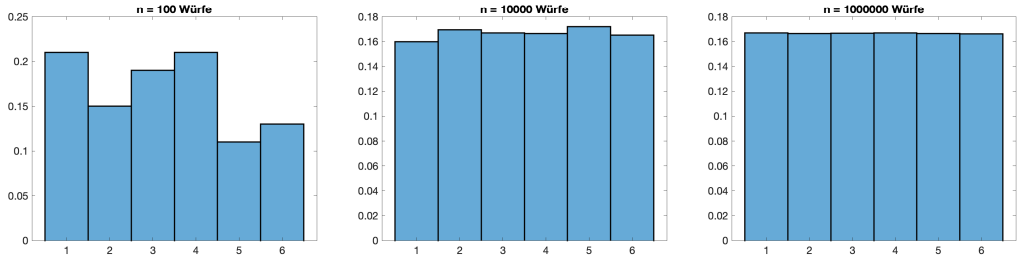


Abbildung 3.1: Histogramme für das Werfen eines Würfels mit $n = 100, 10\,000, 1\,000\,000$ (von links nach rechts)

Wenn $H_n(A)$ konvergiert, so nennen wir den Grenzwert die *Wahrscheinlichkeit* $\mathbb{P}(A)$ *von* A, formal

$$\mathbb{P}(A) := \lim_{n \to \infty} H_n(A). \tag{3.1}$$

Die Wahrscheinlichkeit, beim Würfeln eine gerade Zahl zu werfen, ist also gerade $\mathbb{P}(A) = 1/2$.

3.1 Wahrscheinlichkeitsverteilungen

Wir wollen die Wahrscheinlichkeiten von allen möglichen Ereignissen $A \in \mathcal{A}$ nun abstrakt definieren. Die folgende Definition gibt — motiviert durch die Eigenschaften von $H_n(A)$ — die Bedingungen an, die $\mathbb{P}$ erfüllen muss.

Definition 3.3 Es sei $(\Omega, \mathcal{A})$ ein messbarer Raum. Eine Abbildung $\mathbb{P} : \mathcal{A} \to [0, 1]$ heißt *Wahrscheinlichkeitsverteilung*, falls gilt:

(A1) $\mathbb{P}(A) \geq 0$ für alle $A \in \mathcal{A}$ *Positivität*

(A2) $\mathbb{P}(\Omega) = 1$ *Normierung*

(A3) Für jede Folge von Ereignissen $A_i \in \mathcal{A}$, $i \in \mathbb{N}$, die *paarweise unvereinbar sind* (also $A_i \cap A_j = \emptyset$ für $i \neq j$ erfüllen) gilt

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i). \qquad \textit{σ-Additivität}$$

Das Tupel $(\Omega, \mathcal{A}, \mathbb{P})$ heißt dann *Wahrscheinlichkeitsraum* und für $A \in \mathcal{A}$ nennen wir $\mathbb{P}(A)$ die *Wahrscheinlichkeit von A.* □

Aus den soeben definierten Eigenschaften folgen weitere Eigenschaften von $\mathbb{P}$, die im folgenden Satz zusammengefast sind.

Satz 3.4 Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. Dann gilt

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

(2) Für jede endliche Folge von paarweise unvereinbaren Ereignissen $A_i \in \mathcal{A}$, $i = 1, \ldots, n$ gilt

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

(3) Die Wahrscheinlichkeit für das Komplement oder Gegenereignis A^c für ein $A \in \mathcal{A}$ ist gegeben durch $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.

(4) Für alle $A, B \in \mathcal{A}$ gilt die Additionsformel

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

(5) Bilden die Ereignisse $A_i \in \mathcal{A}$ eine *Zerlegung* von Ω, d.h. sind die A_i paarweise unvereinbar und erfüllen $\bigcup_{i=1}^{\infty} A_i = \Omega$, so folgt

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

Beweis: (1) Die Ereignisse $A_1 = \Omega$ und $A_i = \emptyset$ für $i \geq 2$ sind paarweise unvereinbar, also folgt

$$1 \overset{(A2)}{=} \mathbb{P}(\Omega) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) \overset{(A3)}{=} \sum_{i=1}^{\infty} \mathbb{P}(A_i) = 1 + \sum_{i=2}^{\infty} \mathbb{P}(A_i).$$

Also muss $\sum_{i=2}^{\infty} \mathbb{P}(A_i) = 0$ gelten und damit $\mathbb{P}(\emptyset) = P(A_2) = 0$.

(2) Setzen wir $A_i = \emptyset$ für $i \geq n + 1$, so folgt die Aussage aus (A3).

(3) Es gilt $A \cup A^c = \Omega$ und $A \cap A^c = \emptyset$. Mit (A2) und (2) folgt

$$\mathbb{P}(A) + \mathbb{P}(A^c) \overset{(2)}{=} \mathbb{P}(A \cup A^c) = \mathbb{P}(\Omega) \overset{(A2)}{=} 1$$

und daraus die Behauptung.

(4) Wir setzen $C := B \setminus A \subset B$. Dann gilt mit den Rechenregeln für Mengenoperationen

$$A \cup C = A \cup \underbrace{(B \setminus A)}_{=B \cap A^c} = (A \cup B) \cap \underbrace{(A \cup A^c)}_{=\Omega} = A \cup B$$

und

$$A \cap C = A \cap \underbrace{(B \setminus A)}_{=B \cap A^c} = B \cap \underbrace{(A \cap A^c)}_{=\emptyset} = \emptyset.$$

Also sind A und C unvereinbar und es folgt

$$\mathbb{P}(A \cup B) = \mathbb{P}(A \cup C) \overset{(2)}{=} \mathbb{P}(A) + \mathbb{P}(C).$$

Analog folgt für $D := A \cap B$, dass $B = C \cup D$ und $C \cap D = \emptyset$ und damit

$$\mathbb{P}(B) \overset{(2)}{=} \mathbb{P}(C) + \mathbb{P}(D) = \mathbb{P}(C) + \mathbb{P}(A \cap B) \quad \Rightarrow \quad \mathbb{P}(C) = \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

Zusammen ergibt sich

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

also die Behauptung.

(5) Die Aussage folgt aus

$$1 \overset{(A2)}{=} \mathbb{P}(\Omega) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) \overset{(A3)}{=} \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

□

Für eine aufsteigende Folge von Mengen $A_1 \subset A_2 \subset \ldots \subset \Omega$ kann man mittels

$$\lim_{i \to \infty} A_i := \bigcup_{i=1}^{\infty} A_i$$

einen mengenwertigen Grenzwert definieren. Als Beispiel betrachte $\Omega = \mathbb{R}$ und die Intervalle $A_i = (1/i, 2]$. Die Null liegt in keinem dieser Intervalle, jede Zahl $x > 0$ und $x \leq 2$ liegt aber in dem Intervall A_i mit $i > 1/x$. Folglich ist $\lim_{n \to \infty} A_i = (0, 2]$. Analog kann man für eine absteigende Folge von Mengen $\Omega \supset A_1 \supset A_2 \supset \ldots$ einen Grenzwert definieren mittels

$$\lim_{i \to \infty} A_i := \bigcap_{i=1}^{\infty} A_i.$$

Für solche Folgen gilt nun der folgende Satz.

Satz 3.5 Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. Für jede aufsteigende und jede absteigende Folge von Mengen $A_i \in \mathcal{A}$, $i \in \mathbb{N}$, gilt

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\left(\lim_{n\to\infty} A_n\right).$$

Beweis: Im aufsteigenden Fall definiere $B_1 := A_1$ und $B_i := A_i \setminus A_{i-1}$ für alle $i \geq 2$. Dann sind die B_i paarweise unvereinbar und es gilt $\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} B_i$. Mit (A3) folgt

$$\begin{aligned}
\mathbb{P}\left(\lim_{n\to\infty} A_n\right) &= \mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(B_i) = \lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{P}(B_i) \\
&= \lim_{n\to\infty} \underbrace{\left[P(A_1) + \sum_{i=2}^{n} (\mathbb{P}(A_i) - \mathbb{P}(A_{i-1}))\right]}_{=\mathbb{P}(A_i)} = \lim_{n\to\infty} \mathbb{P}(A_i).
\end{aligned}$$

Im absteigenden Fall ist $A_1^c \subset A_2^c \subset \ldots$ eine aufsteigende Folge mit $(\bigcup_{i=1}^{\infty} A_i^c)^c = \bigcap_{i=1}^{\infty} A_i$. Für diese ist die Aussage des Satzes bereits bewiesen. Damit folgt

$$\begin{aligned}
\mathbb{P}\left(\lim_{n\to\infty} A_n\right) &= \mathbb{P}\left(\left(\bigcup_{i=1}^{\infty} A_i^c\right)^c\right) = 1 - \mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i^c\right) \\
&= 1 - \lim_{n\to\infty} \mathbb{P}(A_n^c) = 1 - \underbrace{\lim_{n\to\infty} (1 - \mathbb{P}(A_n))}_{=1-\lim_{n\to\infty} \mathbb{P}(A_n)} = \lim_{n\to\infty} \mathbb{P}(A_n).
\end{aligned}$$

□

Falls Ω nur endlich viele Elemente $|\Omega| = n$ besitzt, so lässt sich $\mathbb{P}(A)$ leicht aus $\mathbb{P}(\{\omega\})$ für $\omega \in \Omega$ berechnen.

Satz 3.6 Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum mit $|\Omega| = n$.

(a) Dann gilt für alle $A \in \mathcal{A}$ mit $A = \{\omega_1, \ldots, \omega_k\}$, $k = |A|$

$$\mathbb{P}(A) = \sum_{i=1}^{k} \mathbb{P}(\{\omega_i\}).$$

(b) Falls alle $\omega \in \Omega$ gleich wahrscheinlich sind, also $\mathbb{P}(\{\omega\}) = \frac{1}{n}$ für alle $\omega \in \Omega$ (man spricht dann von einem *Laplace-Modell*), so gilt für alle $A \in \mathcal{A}$

$$\mathbb{P}(A) = \frac{|A|}{n}.$$

Beweis: Aussage (a) folgt aus $(A3)$ und der Tatsache, dass die Ereignisse $A_i = \mathbb{P}(\{\omega_i\})$ für die $\omega_i \in A$ paarweise unvereinbar sind. Aussage (b) folgt dann aus

$$\mathbb{P}(A) = \sum_{i=1}^{k} \mathbb{P}(\{\omega_i\}) = \sum_{i=1}^{k} \frac{1}{n} = \frac{k}{n} = \frac{|A|}{n}.$$

□

Für einen idealen Würfel lässt z.B. der einmalige Würfelwurf mit einem Laplace-Modell auf $\Omega = \{1, 2, 3, 4, 5, 6\}$ mit $n = 6$ modellieren.

3.2 Kombinatorik

In diesem Abschnitt betrachten wir einige Aussagen für den speziellen Fall, dass Ω eine endliche Menge ist. Kombinatorik ist — etwas salopp gesagt — die Lehre vom "geschickten Abzählen" verschiedener möglicher Kombinationen und darum geht es in der Regel, wenn man Wahrscheinlichkeiten auf endlichen Mengen errechnen will.

Unser Grundmodell ist in diesem Abschnitt eine Urne, die eine Menge Ω von unterschiedlichen Objekten enthält, die gezogen werden können. Man kann sich z.B. verschiedenfarbige oder unterschiedlich nummerierte Kugeln vorstellen. Wichtig ist, dass alle Objekte unterschiedlich aussehen, also keine zwei oder mehr gleichartige Kugeln darunter sind. Wir untersuchen zuerst die Frage, wie viele unterschiedliche Ergebnisse es geben kann, wenn man nacheinander m Objekte zieht. Das Ergebnis hängt natürlich davon ab, ob das jeweils gezogene Objekt vor dem nächsten Ziehen zurückgelegt wird, oder nicht. Die Folge $(\omega_1, \ldots, \omega_m)$ der gezogenen Objekte nennen wir *Variation* der Länge m. Wir erinnern für den folgenden Satz an die Definition der *Fakultät* $p! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot p$ für $p \in \mathbb{N}$.

Satz 3.7 Sei Ω eine Menge mit $|\Omega| = n$.

(1) Die Anzahl der unterschiedlichen Variationen der Länge m mit Zurücklegen ist gegeben durch n^m.

(2) Die Anzahl der unterschiedlichen Variationen der Länge m ohne Zurücklegen ist gegeben durch $\frac{n!}{(n-m)!}$.

Beweis: (1) In jedem Zug gibt es gerade n verschiedene Möglichkeiten. Also kann die Variation $(\omega_1, \ldots, \omega_m)$ an jeder Stelle n verschiedene Einträge haben, was auf $n \cdot n \cdot \ldots \cdot n = n^m$ Möglichkeiten führt.

(2) Im ersten Zug haben wir hier n Möglichkeiten, im zweiten dann noch $n-1$ im dritten $n-2$ und so weiter und im m-ten und letzten Zug dann $n-m+1$ Möglichkeiten. Dies führt auf $n \cdot (n-1) \cdot \ldots \cdot (n-m+1)$ Möglichkeiten. Wegen

$$\begin{aligned} \frac{n!}{(n-m)!} &= \frac{1 \cdot 2 \cdot 3 \cdot \ldots \cdot (n-m) \cdot (n-m+1) \cdot \ldots \cdot n}{1 \cdot 2 \cdot 3 \cdot \ldots \cdot (n-m)} \\ &= \frac{1 \cdot 2 \cdot 3 \cdot \ldots \cdot (n-m)}{1 \cdot 2 \cdot 3 \cdot \ldots \cdot (n-m)} (n-m+1) \cdot \ldots \cdot n = (n-m+1) \cdot \ldots \cdot n \end{aligned}$$

ist das genau die Anzahl aus (2). □

Bei der gerade abgezählten Anzahl der Variationen spielt die Reihenfolge der Ziehungen eine Rolle. Die Variation $(3, 6, 2)$ ist also eine andere Variation als $(6, 2, 3)$. Wenn die Reihenfolge keine Rolle spielt (wie z.B. bei der Ziehung der Lottozahlen oder bei der Zusammenstellung von Übungsgruppen zu einer Vorlesung) spricht man von *Kombinationen*. Die

Anzahl der unterschiedlichen Kombinationen ist natürlich geringer als die Anzahl der Variationen. Der folgende Satz gibt an, wie groß sie genau ist. Wir verwenden hier für $p, q \in \mathbb{N}$ mit $p > q$ die Schreibweise (gesprochen "p über q")

$$\binom{p}{q} = \frac{p!}{q!(p-q)!}.$$

Diese Größe wird *Binomialkoeffizient* genannt.

Satz 3.8 Sei Ω eine Menge mit $|\Omega| = n$.

(1) Die Anzahl der unterschiedlichen Kombinationen der Länge m mit Zurücklegen ist gegeben durch $\binom{n+m-1}{m}$.

(2) Die Anzahl der unterschiedlichen Kombinationen der Länge m ohne Zurücklegen ist gegeben durch $\binom{n}{m}$.

Beweis: Wir beweisen zunächst (2), weil wir (1) dann darauf zurückführen können: Die Anzahl der Variationen ohne Zurücklegen beträgt nach Satz 3.7(2) gerade $\frac{n!}{(n-m)!}$. Die Anzahl der Kombinationen erhält man daraus, indem man sich überlegt, wie viele Variationen jeweils zu der gleichen Kombination gehören. Wegen des nicht-Zurücklegens enthält jede Kombination m verschiedene Elemente. Diese können wir nun an die verschiedenen Stellen der zugehörigen Variation stellen. Es gibt dabei m Möglichkeiten für das Element an der ersten Stelle, $m-1$ an der zweiten usw., so dass wir insgesamt auf $m!$ Möglichkeiten kommen. Es gibt also $m!$ Variationen von jeder Kombination, d.h. wir müssen wir die Zahl der Variationen durch $m!$ teilen und erhalten so

$$\frac{n!}{(n-m)!m!} = \binom{n}{m}$$

Möglichkeiten.

(1) Man könnte hier auf die Idee kommen, den Beweis von (2) zu übernehmen und Satz 3.7(2) durch Satz 3.7(1) zu ersetzen. Das geht aber schief, weil die Anzahl der Variationen von einer Kombination beim Ziehen mit Zurücklegen nicht mehr für alle Kombinationen identisch ist. Z.B. gibt es für die Kombination $(1,2,3)$ sechs mögliche Variationen ($(1,2,3)$, $(1,3,2)$, $(2,1,3)$, $(2,3,1)$, $(3,1,2)$, $(3,2,1)$), für die Kombination $(1,1,1)$ aber nur die eine Variation $(1,1,1)$. Wir formulieren das Problem daher in ein Problem ohne Zurücklegen um und wenden dann Teil (2) an.

Wir nummerieren dazu die Objekte in der Menge Ω und notieren uns die Nummern der m gezogenen Elemente in aufsteigender Reihenfolge, was auf eine Folge $1 \le a_1 \le a_2 \le \ldots \le a_m \le n$ führt. Nun addieren wir zu jedem Eintrag a_i die Zahl $i-1$ und erhalten so die Folge $b_i = a_i + i - 1$, also $b_1 = a_1$, $b_2 = a_2 + 1$, $b_3 = a_3 + 2$,..., $b_m = a_m + m - 1$. Diese Folge hat nun wegen $b_{i+1} = a_{i+1} + i \ge a_i + i > a_i + i - 1 = b_i$ keine zwei gleichen Elemente. Man überlegt sich nun, dass mit diesem Vorgehen jede mögliche aufsteigend angeordnete Folge von m paarweise verschiedenen Zahlen zwischen 1 und $n+m-1$ entstehen kann. Im Fall $m = 3$ wird z.B. die Folge $(1,2,5)$ dadurch erzeugt, dass man die Elemente mit den

Nummern 3, 1 und 1 zieht. Dies führt zu $a_1 = 1$, $a_2 = 1$ und $a_3 = 3$, damit zu $b_1 = 1+0 = 1$, $b_2 = 1 + 1 = 2$ und $b_3 = 3 + 2 = 5$. Die gesuchte Anzahl ist also gerade gleich der Anzahl der Kombinationen für dieses Problem, mit $n + m - 1$ statt n Elementen.

Diese Anzahl sagt uns aber gerade der schon bewiesene Teil (2) des Satzes, wenn wir n durch $n + m + 1$ ersetzen. Dies führt auf die behaupteten

$$\binom{n+m-1}{m}$$

Möglichkeiten. □

Wir illustrieren die Anwendung der Sätze für Laplace-Modelle mit zwei Beispielen.

Beispiel 3.9 (Lottozahlen) Beim Lotto gibt es den Höchstgewinn, wenn 6 richtige Zahlen aus 49 möglichen getippt werden. Auf die Reihenfolge der getippten bzw. gezogenen Zahlen kommt es dabei nicht an. Zudem werden die gezogenen Zahlen vor dem Ziehen der nächsten Zahl nicht zurückgelegt.

Wir sind also in der Situation von Satz 3.8(2) mit $n = 49$ und $m = 6$. Die Anzahl der möglichen Kombinationen beträgt

$$\binom{49}{6} = \frac{49!}{6!(49-6)!} = 13\,983\,816.$$

Legt man ein Laplace-Modell zu Grunde, nimmt also an, dass alle Kombinationen gleichwahrscheinlich sind, so liegt die Chance auf 6 Richtige bei ein mal Tippen nach Satz 3.6(b) bei

$$\mathbb{P}(A) = \frac{1}{13\,983\,816} \approx 0,0000000715 = 0,00000715\%.$$

□

Beispiel 3.10 (Geburtstagsparadoxon) Wie groß ist die Wahrscheinlichkeit, dass unter einer Gruppe von n anwesenden Personen in einem Raum zwei Personen am gleichen Tag Geburtstag haben? Wir nehmen dabei für das Wahrscheinlichkeitsmodell an, dass niemand in der Gruppe am 29. Februar Geburtstag hat und dass die übrigen 365 Tage des Jahres alle gleich wahrscheinlich sind.

Das betrachtete Ereignis ist also

$$A = \text{“Mindestens zwei Personen haben am gleichen Tag Geburtstag”}.$$

Tatsächlich wird die Rechnung leichter, wenn man das Gegenereignis betrachtet, also

$$A^c = \text{“Alle Geburtstage sind verschieden”}.$$

Dann können wir die gesuchte Wahrscheinlichkeit mittels Satz 3.4(3) als

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

berechnen. Dass die Modellierung mit dem Gegenereignis einfacher wird, ist sehr oft in der Stochastik so, weswegen es sich bei einem konkreten Problem immer empfiehlt, probehalber auch das Gegenereignis zu betrachten. Der tiefere Grund dafür wird am Ende von Abschnitt 5.2 erläutert.

Für das stochastische Modell müssen wir die zufällige Auswahl von m Geburtstagen in ein Urnenmodell übersetzen. Dazu wählen wir eine Urne mit 365 Kugeln, für jedes Datum eine. Jeder zufällige Geburtstag entspricht dann dem Ziehen einer Kugel und da zwei Personen am gleichen Tag Geburtstag haben können, müssen wir mit Zurücklegen ziehen. Wir verwenden für die Modellierung hier die Variationen, weil sich beim Übergang zu den Kombinationen das Verhältnis zwischen Ergebnissen mit und ohne übereinstimmende Geburtstage ändert (vergleiche die Anmerkung zu Beginn des Beweises von Satz 3.8(1)). Die Menge Ω der möglichen Variationen enthält nach Satz 3.7 gerade 365^m Elemente. Die Menge der Variationen in A^c kann man gerade durch das Ziehen ohne Zurücklegen bestimmen, denn dadurch vermeiden wir genau die Variationen, in denen ein Geburtstag zweimal vorkommt. Nach Satz 3.7 enthält A^c also gerade $\frac{365!}{(365-m)!} = 365 \cdot 364 \cdot \ldots \cdot (365 - m + 1)$ Elemente. Satz 3.6(b) liefert also eine Wahrscheinlichkeit von

$$\mathbb{P}(A) = 1 - \mathbb{P}(A^c) = 1 - \frac{365 \cdot 364 \cdot \ldots \cdot (365 - m + 1)}{365^m}.$$

Einsetzen ergibt die Wahrscheinlichkeiten

$$\begin{aligned} m = 10: &\quad \mathbb{P}(A) = 0.1169 = 11,69\% \\ m = 20: &\quad \mathbb{P}(A) = 0.4114 = 41,14\% \\ m = 22: &\quad \mathbb{P}(A) = 0.4757 = 47,54\% \\ m = 23: &\quad \mathbb{P}(A) = 0.5073 = 50,73\% \\ m = 30: &\quad \mathbb{P}(A) = 0.7063 = 70,63\%. \end{aligned}$$

Bereits bei 23 Personen ist die Chance, dass zwei Personen am gleichen Tag Geburtstag haben, also größer als 50%. Bei 30 Personen ist die Chance bereits größer als 70%. Weil das viel wahrscheinlicher ist, als man intuitiv erwarten würde, wenn man 365 Tage auf 23 oder 30 Personen aufteilt, ist dies als das Geburtstagsparadoxon bekannt. □

Wir sehen: Zufälle kommen in manchen realen Situationen häufiger vor, als man es intuitiv erwarten würde. Die Suche nach einem tieferen Grund für ein häufig auftretendes Phänomen ist in diesem Fall vergebens. Zielführender ist die Berechnung der genauen Wahrscheinlichkeit für das Phänomen.

Kapitel 4

Bedingte Wahrscheinlichkeit und die Bayes'sche Formel

4.1 Definition und Grundlagen

Die bedingte Wahrscheinlichkeit formalisiert die Idee, dass uns das Eintreten eines Ereignisses bei der Realisierung eines Zufallsexperiments Informationen über das Eintreten eines anderen Ereignisses (für dieselbe Realisierung) liefert.

Beispiel 4.1 Wir werfen einen Würfel zweimal und betrachten die Ereignisse

$$A = \{\text{die Augensumme ergibt eine 3}\} \quad \text{und} \quad B = \{\text{der erste Wurf ergibt eine 1}\}.$$

Die Augensumme 3 ergibt sich für die 36 gleich wahrscheinlichen Möglichkeiten aus Beispiel 2.3 bei genau 2 Paaren, nämlich bei $(1,2)$ und bei $(2,1)$. Nach Satz 3.6 ist die Wahrscheinlichkeit von A also $\mathbb{P}(B) = 2/36 = 1/18 = 0.0\bar{5}$.

Wenn wir nun aber wissen, dass das Ereignis B eingetreten ist, dann sind nur noch die 6 Paare $(1,1)$, $(1,2)$, $(1,3)$, $(1,4)$, $(1,5)$, $(1,6)$ aus Beispiel 2.3 möglich. Unter diesen gibt es genau eines, nämlich $(1,2)$, das die Augensumme 3 besitzt. Also ist die Wahrscheinlichkeit für A unter der Voraussetzung, dass B eingetreten ist, gleich $1/6 = 0,1\bar{6}$. Man nennt dies die *bedingte Wahrscheinlichkeit von A unter der Bedingung B*, geschrieben als $\mathbb{P}(A \,|\, B)$. □

Um zu einer allgemeinen Formel für $\mathbb{P}(A \,|\, B)$ zu kommen, betrachten wir zuerst wieder die relativen Häufigkeiten aus Definition 3.1. Die relative Häufigkeit für das Ereignis "A gilt unter der Bedingung B", geschrieben als $H_n(A \,|\, B)$, ist dann gerade die Anzahl der Ereignisse, bei denen A und B eintreten (im Beispiel 4.1 nur das Paar $(1,2)$, also 1), geteilt durch die Anzahl der Ereignisse, bei denen B eintritt (im Beispiel 4.1 die 6 Paare, die vorne eine 1 stehen haben). Formal:

$$H_n(A \,|\, B) = \frac{h_n(A \cap B)}{h_n(B)} = \frac{\frac{h_n(A \cap B)}{n}}{\frac{h_n(B)}{n}} = \frac{H_n(A \cap B)}{H_n(B)}.$$

Wenn wir nun also wieder die Wahrscheinlichkeit des Ereignisses als Grenzwert dieser Größen für $n \to \infty$ definieren, so ergibt sich die folgende Definition für die bedingte Wahrscheinlichkeit.

Definition 4.2 Es sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $A, B \in \mathcal{A}$ mit $\mathbb{P}(B) > 0$. Die *bedingte Wahrscheinlichkeit von A unter der Bedingung B* ist definiert als

$$\mathbb{P}(A \,|\, B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \tag{4.1}$$

□

Im Beispiel 4.1 ist $B = \{(1,1), (1,2), (1,3), (1,4), (1,5), (1,6)\}$ und $A \cap B = \{(1,2)\}$. Die Wahrscheinlichkeiten dieser Ereignisse sind $\mathbb{P}(B) = 6/36 = 1/6$ und $\mathbb{P}(A \cap B) = 1/36$. Also ergibt sich

$$\mathbb{P}(A \,|\, B) = \frac{1/36}{6/36} = \frac{1}{6},$$

was identisch mit dem durch Abzählen der Elementarereignisse erhaltenen Wert in Beispiel 4.1 ist.

Der folgende Satz fasst einige Rechenregeln für bedingte Wahrscheinlichkeiten zusammen.

Satz 4.3 Es sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $A, B, C \in \mathcal{A}$ mit $\mathbb{P}(C) > 0$. Dann gilt

1) $\mathbb{P}(C \,|\, C) = 1$

2) $\mathbb{P}(A \,|\, C) = 1 - \mathbb{P}(A^c \,/\, C)$

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

Beweis: 1) Es gilt

$$\mathbb{P}(C \,|\, C) = \frac{\mathbb{P}(\overbrace{C \cap C}^{=C})\}\{\mathbb{P}(C)}{=} \frac{\mathbb{P}(C)}{\mathbb{P}(C)} = 1.$$

2) Nach Satz 2.1 3) und 5) gilt

$$(A \cap C) \cup (A^c \cap C) = (\underbrace{A \cup A^c}_{=\Omega}) \cap C = C$$

und damit mit Satz 3.4 (2)

$$\begin{aligned} 1 &= \frac{\mathbb{P}(C)}{\mathbb{P}(C)} = \frac{\mathbb{P}\big((A \cap C) \cup (A^c \cap C)\big)}{\mathbb{P}(C)} = \frac{\mathbb{P}(A \cap C) + \mathbb{P}(A^c \cap C)}{\mathbb{P}(C)} \\ &= \frac{\mathbb{P}(A \cap C)}{\mathbb{P}(C)} + \frac{\mathbb{P}(A^c \cap C)}{\mathbb{P}(C)} = \mathbb{P}(A \,|\, C) + \mathbb{P}(A^c \,|\, C) \end{aligned}$$

3) Mit Satz 3.4 (4) gilt

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

□

Beispiel 4.4 Sie suchen in Ihrer WG nach Ihrem Handy (dessen Akku mal wieder leer ist). Am wahrscheinlichsten erscheint es Ihnen, dass das Handy in Ihrem Zimmer, in der Küche oder im Bad liegt. Sie schätzen die Wahrscheinlichkeit, dass das Handy in einem dieser Räume liegt (dass Sie es also nicht z.B. in der Uni-Bibliothek liegen lassen haben) auf 0,8. Dabei schätzen Sie die Wahrscheinlichkeiten, in welchem der drei Räume in Ihrer WG liegt, als gleich hoch ein. Ihr Zimmer und die Küche haben Sie schon vergeblich durchsucht. Wie groß ist nun die Wahrscheinlichkeit, dass sich das Handy im Bad befindet?

Wir nummerieren die Räume mit $i = 1 \mathrel{\hat{=}}$ Ihr Zimmer, $i = 2 \mathrel{\hat{=}}$ Küche und $i = 3 \mathrel{\hat{=}}$ Bad und setzen

$$A_i = \{\text{das Handy befindet sich in Raum } i\}.$$

Zur Beantwortung der obigen Frage müssen wir also die bedingte Wahrscheinlichkeit von A_3 unter der Bedingung, dass A_1 und A_2 *nicht* eingetreten sind, berechnen, also die bedingte Wahrscheinlichkeit von A_3 unter der Bedingung, dass $(A_1 \cup A_2)^c$ eingetreten sind, d.h. $\mathbb{P}(A_3 \,|\, (A_1 \cup A_2)^c)$.

Die Ereignisse A_1 bis A_3 sind nach Annahme gleichwahrscheinlich, also $\mathbb{P}(A_1) = \mathbb{P}(A_2) = \mathbb{P}(A_3) =: p$ und weil die Ereignisse unvereinbar sind, folgt

$$\mathbb{P}(A_1 \cup A_2 \cup A_3) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \mathbb{P}(A_3) = 3p = 0{,}8,$$

woraus $p = 0{,}8/3 = 0{,}2\overline{6}$ folgt. Aus der Unvereinbarkeit der A_i folgt $A_i \subset A_j^c$ für alle $i \neq j$ und damit

$$\begin{aligned}\mathbb{P}(A_3 \,|\, (A_1 \cup A_2)^c) &= \mathbb{P}(A_3 \,|\, A_1^c \cap A_2^c) = \frac{\mathbb{P}(A_3 \cap A_1^c \cap A_2^c)}{\mathbb{P}(A_1^c \cap A_2^c)} \\ &= \frac{\mathbb{P}(A_3)}{1 - \mathbb{P}(A_1 \cup A_2)} = \frac{p}{1 - 2p} \approx 0.5714.\end{aligned}$$

Mit einer Wahrscheinlichkeit von ca. 57,14% finden Sie Ihr Handy also im Bad. □

4.2 Multiplikationsregeln

Umstellen der Definition der bedingten Wahrscheinlichkeit liefert wegen $A \cap B = B \cap A$ direkt die Identitäten

$$\mathbb{P}(A \cap B) = \mathbb{P}(A \,|\, B) \cdot \mathbb{P}(B) = \mathbb{P}(B \,|\, A) \cdot \mathbb{P}(A). \tag{4.2}$$

Beispiel 4.5 Bei einer anstehenden Klausur ist bekannt, dass sie mit einer Wahrscheinlichkeit von 0,8 bestanden wird. Diejenigen, die die Klausur bestehen, haben mit einer Wahrscheinlichkeit von 0,4 die Note "gut" oder "sehr gut". Mit welcher Wahrscheinlichkeit erreicht man damit die Note "gut" oder "sehr gut"?

Definieren wir die Ereignisse

$$A = \{\text{Die Note ist "gut" oder "sehr gut"}\} \text{ und } B = \{\text{die Klausur wird bestanden}\},$$

so kennen wir der Angabe die bedingte Wahrscheinlichkeit $\mathbb{P}(A \,|\, B) = 0{,}4$ und die Wahrscheinlichkeit der Bedingung $\mathbb{P}(B) = 0{,}8$ und wollen daraus die unbedingte Wahrscheinlichkeit $\mathbb{P}(A)$ berechnen. Da man mit "gut" oder "sehr gut" auf jeden Fall bestanden hat, gilt $A \cap B = A$ und wir können mit Formel (4.2) berechnen:

$$\mathbb{P}(A) = \mathbb{P}(A \cap B) = \mathbb{P}(A \,|\, B) \cdot \mathbb{P}(B) = 0{,}4 \cdot 0.8 = 0.32.$$

Die Formel (4.2) lässt sich wie folgt auf mehr als zwei Mengen erweitern. □

Satz 4.6 Es sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $A_1, A_2, \ldots, A_n \in \mathcal{A}$ mit $\mathbb{P}(A_1 \cap A_2 \cap \ldots \cap A_{n-1}) > 0$. Dann gilt

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

Beweis: Weil $A_1 \cap A_2 \cap \ldots \cap A_{n-1} \subset A_1 \cap \ldots \cap A_m$ gilt für alle $m < n$, folgt

$$\mathbb{P}(A_1 \cap \ldots \cap A_m) \geq \mathbb{P}(A_1 \cap A_2 \cap \ldots \cap A_{n-1}) > 0.$$

Damit können alle bedingten Wahrscheinlichkeiten in der Formel auch gebildet werden (man sagt, sie sind *wohldefiniert*). Die Formel beweisen wir nun mit einer Technik, die *vollständige Induktion* heißt. Man beweist dazu zuerst, dass die Formel für $n = 2$ gilt. Dies gilt hier, weil die behauptete Formel für $n = 2$ gerade (4.2) mit $A = A_1$ und $B = A_2$ ist.

Dann beweist man, dass, wenn die Formel für n gilt, sie auch für $n + 1$ (an Stelle von n) gilt. Wenn man das weiß, kann man aus der bereits gezeigten Gültigkeit für $n = 2$ die Gültigkeit für $n = 3$ folgern, daraus die Gültigkeit für $n = 4$, daraus die für $n = 5$ usw., und damit für alle $n \geq 2$.

Um zu zeigen, dass die Formel für $n + 1$ gilt, falls sie für n gilt, nehmen wir an, dass sie für n gilt (dies ist die sogenannte *Induktionsannahme*). Dann verwenden wir zunächst (4.2) mit $A = A_{n+1}$ und $B = A_1 \cap \ldots \cap A_n$ und erhalten so

$$\begin{aligned} \mathbb{P}(A_1 \cap \ldots \cap A_{n+1}) &= \mathbb{P}\big(A_{n+1} \cap (A_1 \cap \ldots \cap A_n)\big) \\ &= \mathbb{P}(A_{n+1} \,|\, A_1 \cap \ldots \cap A_n)\mathbb{P}(A_1 \cap \ldots \cap A_n) \\ &= \mathbb{P}(A_1 \cap \ldots \cap A_n)\mathbb{P}(A_{n+1} \,|\, A_1 \cap \ldots \cap A_n) \end{aligned}$$

Die Induktionsannahme besagt nun, dass für den ersten Term in diesem Produkt

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

gilt und damit

$$\begin{aligned}&\mathbb{P}(A_1 \cap \ldots \cap A_{n+1})\\ &= \mathbb{P}(A_1) \cdot \mathbb{P}(A_2 \,|\, A_1) \cdots \mathbb{P}(A_n \,|\, A_1 \cap \ldots \cap A_{n-1}) \cdot \mathbb{P}(A_{n+1} \,|\, A_1 \cap \ldots \cap A_n)\end{aligned}$$

also gerade die behauptete Gleichheit für $n+1$ an Stelle von n. □

Die Formel aus Satz 4.6 lässt sich schön mittels eines Baumdiagramms darstellen, siehe Abbildung 4.1.

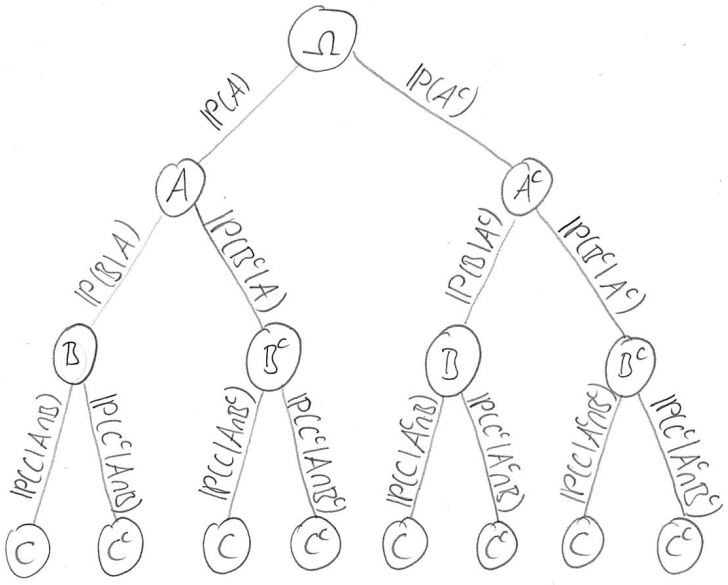


Abbildung 4.1: Darstellung der Formel aus Satz 4.6 als Baum

Die Wahrscheinlichkeit, zu einem Knoten im Baum zu gelangen, kann durch Aufmultiplizieren der Werte bestimmt werden, die vom Knoten Ω aus zu der Menge führen. Im Baum in Abbildung 4.1 entspricht der Ast ganz links also der Wahrscheinlichkeit $\mathbb{P}(A \cap B \cap C)$, der zweite Ast von links der Wahrscheinlichkeit $\mathbb{P}(A \cap B^c \cap C)$ und so weiter. Aus diesen Werten kann man dann durch Addieren die Wahrscheinlichkeiten $\mathbb{P}(C)$ errechnen: Da für die Wahrscheinlichkeit $\mathbb{P}(C)$ wegen Satz 3.4(2) die Gleichung

$$\mathbb{P}(C) = \mathbb{P}(A \cap B \cap C) + \mathbb{P}(A^c \cap B \cap C) + \mathbb{P}(A \cap B^c \cap C) + \mathbb{P}(A^c \cap B^c \cap C). \tag{4.3}$$

Die Wahrscheinlichkeit $\mathbb{P}(C)$ ergibt sich also durch Addieren der Wahrscheinlichkeiten aller Ästen, in denen "C" ganz unten steht.

Situationen, in denen bedingte Wahrscheinlichkeiten eine Rolle spielen, können unserer Intuition manchmal widersprechen. Ein bekanntes Beispiel ist das Ziegenproblem aus einer amerikanischen Fernsehshow im folgenden Beispiel.

Beispiel 4.7 In einer Spielshow hat die Kandidatin die Möglichkeit, zwischen drei verschlossenen Türen mit Gewinnen zu wählen. Hinter zwei Türen befinden sich jeweils eine Ziege, hinter einem Tor ein Auto. Nachdem sich die Kandidatin für eine Tür entschieden

hat, öffnet der Moderator eine der anderen Türen, und zwar immer eine, hinter der sich eine Ziege befindet. Nun hat die Kandidatin die Möglichkeit, ihre Wahl zu ändern. Die Frage ist, ob dies ihre Chancen auf den Gewinn des Autos erhöht.

Um das Problem mathematisch zu formulieren, definieren wir für $i = 1, 2, 3$ die folgenden sechs Ereignisse

$$\begin{aligned} A_i &= \{\text{Kandidatin wählt Tür } i\} \\ B_i &= \{\text{Moderator wählt Tür } i\} \end{aligned}$$

Mit i^* bezeichnen wir die Nummer der Tür, hinter der das Auto steht. Wir betrachten im Folgenden den Fall $i^* = 1$. Die anderen Fälle können völlig analog analysiert werden und führen auf das gleiche Ergebnis, weswegen wir sie nicht weiter betrachten.

Da die Kandidatin am Anfang keine Information hat, wählt sie die Tür mit gleicher Wahrscheinlichkeit aus den drei Möglichkeiten. Es gilt also

$$\mathbb{P}(A_1) = \mathbb{P}(A_2) = \mathbb{P}(A_3) = \frac{1}{3}.$$

Wenn die Kandidatin Tür 1 wählt, also das Ereignis A_1 eintritt, so hat der Moderator die Möglichkeit, Tür 2 oder Tür 3 zu öffnen. Wir nehmen an, dass er hierbei keine Präferenzen hat, dass also

$$\mathbb{P}(B_2 \,|\, A_1) = \mathbb{P}(B_3 \,|\, A_1) = \frac{1}{2} \text{ und } \mathbb{P}(B_1 \,|\, A_1) = 0$$

gilt (das Ergebnis ändert sich nicht, falls er eine Vorliebe für eine der beiden Türen hat).

Wenn die Kandidatin hingegen Tür 2 oder Tür 3 wählt, hat der Moderator nur die Möglichkeit, Tür 3 bzw. Tür 2 zu öffnen. Es gilt also

$$\mathbb{P}(B_3 \,|\, A_2) = 1 \text{ und } \mathbb{P}(B_1 \,|\, A_2) = \mathbb{P}(B_2 \,|\, A_2) = 0$$

sowie

$$\mathbb{P}(B_2 \,|\, A_3) = 1 \text{ und } \mathbb{P}(B_1 \,|\, A_3) = \mathbb{P}(B_3 \,|\, A_3) = 0.$$

Nun haben wir alle Wahrscheinlichkeiten beisammen, um das Problem als Baumdiagramm darzustellen. Wir stellen dabei ein Diagramm für den Fall, dass sich die Kandidatin nicht umentscheidet und ein Diagramm für den Fall, dass sie sich umentscheidet, auf. Abbildung 4.2 stellt die beide Bäume dar, oben ohne und unten mit Umentscheidung. Äste mit Wahrscheinlichkeit 0 sind der Übersichtlichkeit halber nicht eingezeichnet.

Unten an den Bäumen sind die aufmultiplizierten Wahrscheinlichkeiten für die einzelnen Ereignisse angegeben. Man sieht: Ohne Umentscheidung ist die Wahrscheinlichkeit für das Ereignis A_1, also den Gewinn des Autos $\mathbb{P}(A_1) = 1/6 + 1/6 = 1/3 = 0{,}\bar{3}$. Mit Umentscheidung beträgt die Wahrscheinlichkeit für das Ereignis A_1 gerade $\mathbb{P}(A_1) = 1/3 + 1/3 = 2/3 = 0{,}\bar{6}$. Das Umentscheiden verdoppelt also die Gewinnwahrscheinlichkeit, was möglicherweise nicht ganz intuitiv ist. Im konkreten Fall kann es für die Kandidatin natürlich trotzdem die falsche Entscheidung sein, sich umzuentscheiden, aber dies ist mit einer geringeren Wahrscheinlichkeit der Fall, als wenn sie ihre Entscheidung nicht ändert. □

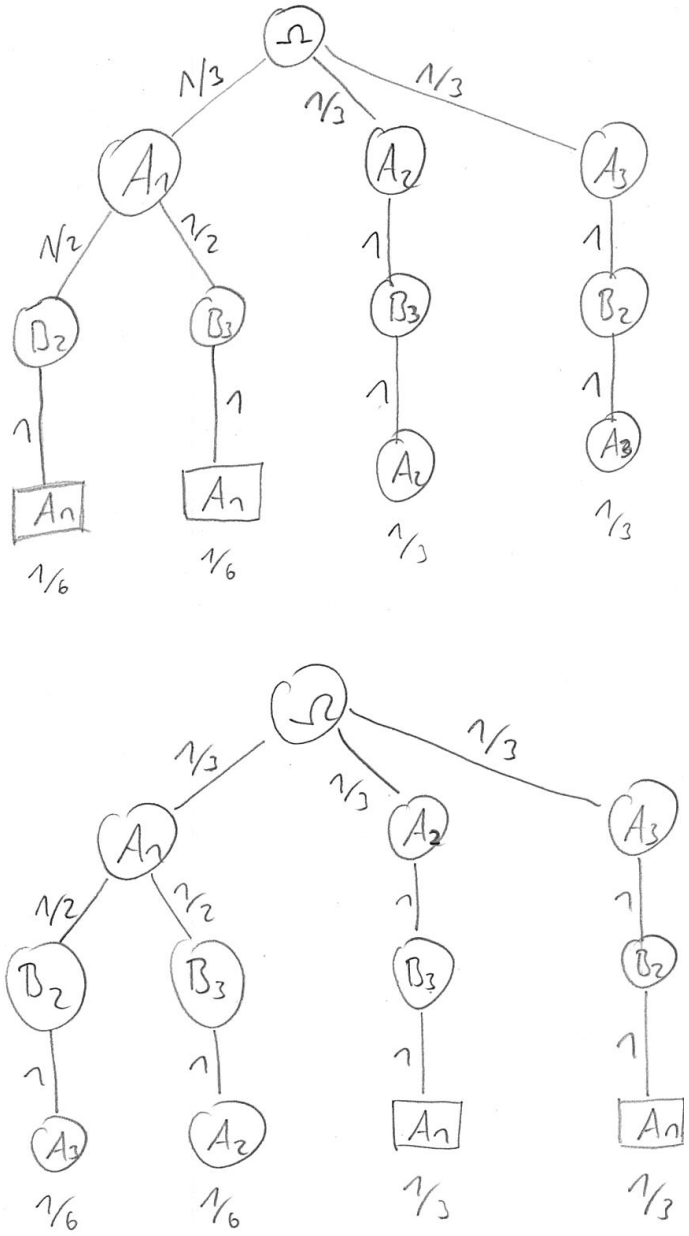


Abbildung 4.2: Baumdiagramme für das Ziegenproblem, oben ohne Umentscheidung der Kandidatin, unten mit

4.3 Die Bayes'sche Formel

Wir haben in Formel (4.3) bereits für drei Mengen gesehen, dass wir die unbedingte Wahrscheinlichkeit eines Ereignisses aus den bedingten Wahrscheinlichkeiten, die durch die Äste

des Diagramms repräsentiert werden, darstellen kann. Der folgende Satz zeigt dies für eine beliebige, möglicherweise sogar unendliche Zahl von Mengen.

Satz 4.8 (Gesetz der totalen Wahrscheinlichkeit) Es sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und B_i, $i \in \mathbb{N}$, eine Zerlegung von Ω, vgl. Satz 3.4(5). Dann gilt für alle $A \in \mathcal{A}$

$$\mathbb{P}(A) = \sum_{\substack{i=1 \\ \mathbb{P}(B_i)>0}}^{\infty} \mathbb{P}(A \,|\, B_i)\mathbb{P}(B_i).$$

Beweis: Auf Grund der Zerlegungseigenschaft der B_i gilt

$$A = A \cap \Omega = A \cap \left(\bigcup_{i=1}^{\infty} B_i \right) = \bigcup_{i=1}^{\infty} (A \cap B_i)$$

und die Mengen $A \cap B_i$ sind paarweise unvereinbar. Mit (4.2) folgt

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \underbrace{\mathbb{P}(A \cap B_i)}_{=0 \text{ falls } \mathbb{P}(B_i)=0} = \sum_{\substack{i=1 \\ \mathbb{P}(B_i)>0}}^{\infty} \mathbb{P}(A \cap B_i) = \sum_{\substack{i=1 \\ \mathbb{P}(B_i)>0}}^{\infty} \mathbb{P}(A \,|\, B_i)\mathbb{P}(B_i).$$

□

Mit diesem Satz können wir die berühmte Bayes'sche Formel herleiten.

Satz 4.9 (Bayes'sche Formel) Unter den Voraussetzungen von Satz 4.8 gilt im Fall $\mathbb{P}(A) > 0$ die *Bayes'sche Formel*

$$\mathbb{P}(B_j \,|\, A) = \frac{\mathbb{P}(A \,|\, B_j)\mathbb{P}(B_j)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A \,|\, B_j)\mathbb{P}(B_j)}{\displaystyle\sum_{\substack{i=1 \\ \mathbb{P}(B_i)>0}}^{\infty} \mathbb{P}(A \,|\, B_i)\mathbb{P}(B_i)}$$

für alle $j \in \mathbb{N}$.

Beweis: Aus der rechten Gleichung von (4.2) folgt mit $B = B_j$

$$\mathbb{P}(A \,|\, B_j)\mathbb{P}(B_j) = \mathbb{P}(B_j \,|\, A)\mathbb{P}(A).$$

Dividieren durch $\mathbb{P}(A)$ liefert die erste Gleichung. Mit Satz 4.8 folgt die zweite Gleichung. □

Beachte, dass die Summe endlich wird, wenn nur endlich viele der B_i ungleich der leeren Menge $\emptyset$ sind. Die Formel gilt also analog, wenn wir n Mengen $B_1, \ldots, B_n$ betrachten, wenn wir "∞" in der Summe durch "n" ersetzen.

Die übliche Interpretation dieser Formel ist wie folgt: Die Wahrscheinlichkeiten für die Ereignisse B_i sind Vorabinformationen (engl. *prior distribution* oder kurz *prior*), die üblicherweise aus Statistiken über längere Zeiträume gewonnen werden. Nun tritt ein Ereignis

B^* ein, das man nicht identifizieren kann. Man kennt aber Zusatzinformationen über das eingetretene Ereignis, die durch ein Ereignis A modelliert werden, sowie — ebenfalls wieder aus Statistiken — für alle i die bedingten Wahrscheinlichkeiten, dass A eintritt unter der Bedingung B_i. Durch diese Zusatzinformationen kann man nun neue Wahrscheinlichkeiten ausrechnen, mit der $B^* = B_i$ gilt (engl. *posterior distribution* oder kurz *posterior*). Das folgende Beispiel illustriert dies.

Beispiel 4.10 In einem Produktionsprozess werden Bauteile hergestellt. Von diese haben nach einer langjährigen Statistik 90% eine hohe Qualität, 8% können noch als zweite Wahl verkauft werden und 2% sind Ausschuss. Dies ist die Vorabinformation. Für jedes Werkstück wird nun nach der Produktion ein einfacher Test durchgeführt, der bei Bauteilen hoher Qualität mit 98% "in Ordnung", bei Bauteilen mit mittlerer Qualität mit 50% in Ordnung und bei Bauteilen mit geringer Qualität mit 10% Wahrscheinlichkeit "in Ordnung" ausgibt, und ansonsten jeweils "nicht in Ordnung". Das Ergebnis dieses Tests ist nun die Zusatzinformation.

Die Frage lautet: Wie hoch ist die Wahrscheinlichkeit, dass ein Bauteil, das vom Test als "in Ordnung" klassifiziert ist, tatsächlich von hoher Qualität ist?

Dazu setzen wir

$$\begin{aligned} B_1 &= \{\text{Bauteil ist von hoher Qualität}\} \\ B_2 &= \{\text{Bauteil ist 2. Wahl}\} \\ B_3 &= \{\text{Bauteil ist Ausschuss}\} \\ A &= \{\text{Test ergibt "in Ordnung"}\}. \end{aligned}$$

Gesucht ist nun die Wahrscheinlichkeit $\mathbb{P}(B_1 \mid A)$, also dass das Bauteil von hoher Qualität ist (B_1) unter der Bedingung, dass der Test "in Ordnung" ergibt (A). Bekannt sind die Wahrscheinlichkeiten

$$\mathbb{P}(B_1) = 0{,}9, \quad \mathbb{P}(B_2) = 0.08, \quad \mathbb{P}(B_3) = 0{,}02$$

$$\mathbb{P}(A \mid B_1) = 0{,}98, \quad \mathbb{P}(A \mid B_2) = 0{,}5, \quad \mathbb{P}(A \mid B_3) = 0{,}1.$$

Damit erhalten wir mit der Bayes'schen Formel für $j = 1$

$$\begin{aligned} \mathbb{P}(B_1 \mid A) &= \frac{\mathbb{P}(A \mid B_1)\mathbb{P}(B_1)}{\mathbb{P}(A \mid B_1)\mathbb{P}(B_1) + \mathbb{P}(A \mid B_2)\mathbb{P}(B_2) + \mathbb{P}(A \mid B_3)\mathbb{P}(B_3)} \\ &= \frac{0{,}98 \cdot 0{,}9}{0{,}98 \cdot 0{,}9 + 0{,}5 \cdot 0{,}08 + 0{,}1 \cdot 0{,}02} \\ &= \frac{0{,}882}{0{,}882 + 0{,}04 + 0{,}002} = \frac{0{,}882}{0{,}924} = 0{,}9545. \end{aligned}$$

Mit Hilfe des Tests kann der Anteil der Bauteile von hoher Qualität also von 90% auf 95,45% erhöht werden. □

Größenordnungsmäßig erscheint die Wahrscheinlichkeit von 95,45%, mit der als "in Ordnung" erkannte Bauteile tatsächlich in Ordnung sind, sinnvoll, denn den größten Anteil

der Werkstücke von minderer Qualität bilden die Werkstücke zweiter Wahl, von denen ja durch den Test 50% gefunden werden. Die Fehlerquote sollte sich also ungefähr halbieren (von 10% auf 5%), was sie auch tut. Die genaue verbleibende Fehlerquote von 4,55% ($= 100\% - 95,45\%$) ist aber alles andere als offensichtlich und kann nur erhalten werden, wenn die Qualität des Tests in allen drei Fällen berücksichtigt wird zusammen mit der Wahrscheinlichkeit, dass diese Fälle auftreten. Diese Zusammenhänge zwischen den Wahrscheinlichkeiten der Vorabinformationen und den Wahrscheinlichkeiten der Zusatzinformationen richtig herauszuarbeiten ist gerade die Stärke der Bayes'schen Formel.

Kapitel 5

Stochastische Unabhängigkeit

5.1 Motivation

Die bedingte Wahrscheinlichkeit zeigt, dass die Information, dass ein Ereignis B eingetreten ist, die Wahrscheinlichkeit des Ereignisses A ändern kann, dass also $\mathbb{P}(A \mid B) \neq \mathbb{P}(A)$ ist. Es kann aber auch passieren, dass die Kenntnis, dass B eingetreten ist, an der Wahrscheinlichkeit von A nichts ändert, dass also $\mathbb{P}(A \mid B) = \mathbb{P}(A)$ ist. Das folgende Beispiel veranschaulicht beide Situationen.

Als Beispiel verwenden wir wieder einmal das Würfeln mit einem idealen Würfel. Wir betrachten die drei Ereignisse

$$\begin{aligned} A &= \{\text{Es wird eine gerade Zahl gewürfelt}\} &&= \{2,4,6\} \\ B &= \{\text{Es wird eine Zahl größer oder gleich 4 gewürfelt}\} &&= \{4,5,6\} \\ C &= \{\text{Es wird eine Zahl größer oder gleich 3 gewürfelt}\} &&= \{3,4,5,6\}. \end{aligned}$$

Die Wahrscheinlichkeiten der Ereignisse erhält man leicht durch Abzählen: $\mathbb{P}(A) = 1/2$, $\mathbb{P}(B) = 1/2$, $P(C) = 2/3$.

Wenn man aber nun einmal würfelt und weiß, dass Ereignis A eingetreten ist, dann ändert sich die Wahrscheinlichkeit von Ereignis B: Es sind jetzt nur noch die drei Zahlen $\{2,4,6\}$ möglich, jede mit Wahrscheinlichkeit $1/3$. Von diesen liegen zwei in B und eine nicht. Unter der Kenntnis, dass Ereignis A eingetreten ist, ändert sich die Wahrscheinlichkeit von B von $\mathbb{P}(B) = 1/2$ zu $\mathbb{P}(B \mid A) = 2/3$. In diesem Fall sagt man, dass A und B *stochastisch abhängig* sind

Für die Wahrscheinlichkeit von C hat die Kenntnis, dass A eingetreten ist, allerdings keine Auswirkungen. Zwar verringert sich die Menge der möglichen Zahlen in C auf $\{4,6\}$, dies sind aber immer noch $2/3$ der gesamten Möglichkeiten $\{2,4,6\}$. Damit ergibt sich $\mathbb{P}(C \mid A) = \mathbb{P}(C) = 2/3$. Hier sagt man, dass A und C *stochastisch unabhängig* sind.

5.2 Definition

Offenbar liegt stochastische Unabhängigkeit von zwei Ereignissen $A, B \in \mathcal{A}$ genau dann vor, wenn $\mathbb{P}(A \mid B) = \mathbb{P}(A)$ gilt. Dies als Definition zu verwenden, ist aber unpraktisch, weil

wir zur Definition von $\mathbb{P}(A \mid B)$ immer $\mathbb{P}(B) > 0$ voraussetzen müssen, was die Mengen, die wir betrachten können, einschränkt. Aus der Definition von $\mathbb{P}(A \mid B)$ folgt im Fall $\mathbb{P}(B) > 0$ aber

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

Der Ausdruck ganz rechts ist also für $\mathbb{P}(B) > 0$ äquivalent zu $\mathbb{P}(A \mid B) = \mathbb{P}(A)$, ergibt aber auch Sinn, $\mathbb{P}(B) = 0$ ist. Dies führt auf die folgende Definition.

Definition 5.1 Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. Dann heißen zwei Ereignisse $A, B \in \mathcal{A}$ *stochastisch unabhängig*, falls

$$P(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$$

gilt. □

Stochastische Unabhängigkeit überträgt sich auf die Gegenereignisse A^c und B^c, wie der folgende Satz zeigt.

Satz 5.2 Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $A, B \in \mathcal{A}$ stochastisch unabhängig. Dann sind auch A^c und B^c, A und B^c sowie A^c und B stochastisch unabhängig.

Beweis: Es gilt

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

Also sind A und B^c stochastisch unabhängig. Die anderen Fälle werden analog bewiesen. □

Stochastische Unabhängigkeit kann man auf mehr als zwei oder sogar abzählbar viele Ereignisse $A_i \in \mathcal{A}$, $i \in \mathbb{N}$, verallgemeinern.

Definition 5.3 Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. Dann heißen $A_i \in \mathcal{A}$, $i \in J \subset \mathbb{N}$, *stochastisch unabhängig*, wenn für jedes $m \in \mathbb{N}$ und jede Auswahl $A_{i_1}, A_{i_2}, \ldots, A_{i_m}$ von m Ereignissen mit paarweise verschiedenen $i_k \in J$ gilt:

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_m}) = P(A_{i_1}) \cdot \mathbb{P}(A_{i_2}) \cdots \mathbb{P}(A_{i_m}).$$

□

Das folgende Beispiel zeigt, dass es nicht genügt, wenn die Mengen paarweise stochastisch unabhängig sind, denn es kann trotzdem eine Auswahl von $m > 2$ Mengen geben, für die die Gleichheit aus Definition 5.3 nicht gilt.

Beispiel 5.4 Sei $\Omega = \{1,2,3,4\}$ mit Laplace-Wahrscheinlichkeit

$$\mathbb{P}(\{1\}) = \mathbb{P}(\{2\}) = \mathbb{P}(\{3\}) = \mathbb{P}(\{4\}) = \frac{1}{4}.$$

Dann gilt für $A = \{1,2\}$, $B = \{1,3\}$, $C = \{2,3\}$, dass

$$\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C) = \frac{1}{2}.$$

Dann gilt

$$\begin{aligned}
\mathbb{P}(A \cap B) &= \mathbb{P}(\{1\}) = \frac{1}{4} = \mathbb{P}(A)\mathbb{P}(B) \\
\mathbb{P}(A \cap C) &= \mathbb{P}(\{2\}) = \frac{1}{4} = \mathbb{P}(A)\mathbb{P}(C) \\
\mathbb{P}(B \cap C) &= \mathbb{P}(\{3\}) = \frac{1}{4} = \mathbb{P}(B)\mathbb{P}(C),
\end{aligned}$$

also sind A, B, C paarweise stochastisch unabhängig. Wegen

$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(\emptyset) = 0 \neq \frac{1}{8} = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$$

sind die drei Ereignisse zusammen aber nicht stochastisch unabhängig. □

5.3 Vereinigung stochastisch unabhängiger Ereignisse

Eine wichtige Folgerung aus der stochastischen Unabhängigkeit ist die folgende Formel, die aus der De Morgan'schen Regel

$$\bigcup_{i=1}^{n} A_i = \left(\bigcap_{i=1}^{n} A_i^c \right)^c$$

folgt und für $n \in \mathbb{N}$ ebenso wie für $n = \infty$ gilt:

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i \right) = 1 - \mathbb{P}\left(\bigcap_{i=1}^{n} A_i^c \right) = 1 - \prod_{i=1}^{n} \mathbb{P}(A_i^c).$$

Diese Regel ist der tiefere Grund dafür, warum es oft sinnvoll ist, für die Berechnung gewisser Wahrscheinlichkeiten zum Gegenereignis überzugehen (wie in Beispiel 3.10 bereits erwähnt). Als Beispiel dafür betrachte die Wahrscheinlichkeit, bei drei mal Würfeln mindestens eine 6 zu würfeln. Wir definieren dazu die stochastisch unabhängigen[1] Ereignisse

$$A_i = \{\text{der } i\text{-te Wurf ergibt eine } 6\}$$

[1] Dass die A_i tatsächlich stochastisch unabhängig sind folgt aus der Tatsache, dass bei einem idealen Würfel jedes Elementarereignis (p,q,r), wobei $p,q,r \in \{1,\ldots,6\}$ die Ergebnisse der drei Würfe sind, gleich wahrscheinlich ist.

für $i = 1, 2, 3$. Das Ereignis "Mindestens einer der drei Würfe ergibt eine 6", also "der erste Wurf ergibt eine 6 *oder* der zweite Wurf ergibt eine 6 *oder* der dritte Wurf ergibt eine 6" ist dann nach Definition 2.6 gegeben durch

$$\bigcup_{i=1}^{3} A_i.$$

Direkt ist die Wahrscheinlichkeit dieses Ereignisses schwer zu berechnen. Geht man aber zu den Gegenereignissen

$$A_i^c = \{\text{der } i\text{-te Wurf ergibt eine 1, 2, 3, 4 oder 5}\}$$

über, für die offenbar $\mathbb{P}(A_i^c) = 5/6$ gilt, so kann man die obigen Formel anwenden und erhält

$$\mathbb{P}\left(\bigcup_{i=1}^{3} A_i\right) = 1 - \prod_{i=1}^{3} \mathbb{P}(A_i^c) = 1 - \left(\frac{5}{6}\right)^3 = 1 - \frac{125}{216} \approx 0{,}4213.$$

Kapitel 6

Zufallsvariablen

6.1 Definition

In vielen Fällen möchte man das Ergebnis eines Zufallsexperiments als einen Zahlenwert darstellen. Man könnte dies nun dadurch erreichen, dass man Ω als Menge dieser Zahlen wählt. Das führt aber zu Problemen, wenn man nicht nur einen solchen Zahlenwert betrachten möchte sondern mehrere und deren Zusammenhang untersuchen möchte.

Als Beispiel betrachte das zweimalige Würfeln eines (wie immer idealen) Würfels, bei dem uns die Augensumme als Ergebnis interessiert. Die naheliegende Art, dieses Zufallsexperiment zu modellieren, ist mittels der Menge $\Omega = \{1, \dots, 6\} \times \{1, \dots, 6\}$, vgl. Beispiel 2.3. Wollte man nun die Augensumme direkt als Ergebnisse in einer neuen Menge $\widetilde{\Omega} = \{2, \dots, 12\}$ modellieren, so müsste man jedes Paar $(p, q) \in \Omega$ durch seine Augensumme $p+q \in \widetilde{\Omega}$ ersetzen. Das lässt sich machen, wenn man die Wahrscheinlichkeitsverteilung auf $\widetilde{\Omega}$ richtig definiert. Da die Mengen Ω und $\widetilde{\Omega}$ nun aber nicht formal miteinander verknüpft sind, verliert man die Information, welches Augenpaar eine gewisse Augensumme hervorgerufen hat. Möchte man jetzt zugleich noch eine weitere abgeleitete Größe betrachten (z.B. die Anzahl der gewürfelten geraden Zahlen, also 0, 1, oder 2), so ist das nicht mehr möglich, wenn man nur die Ergebnisse in $\widetilde{\Omega}$ kennt.

Daher hat sich eine andere Vorgehensweise durchgesetzt, nämlich das Zuweisen von Zahlenwerten durch eine Abbildung $X : \Omega \to \mathbb{R}$, die jedem $\omega \in \Omega$ eine reelle Zahl zuordnet. Dadurch erhält man die gesuchten Größen zusätzlich als Bild von X, ohne Ω durch eine andere Menge ersetzen zu müssen und so Informationen zu verlieren. Durch die Abbbildung X kann man dann z.B. Ereignisse der Form

$$A = \{\omega \in \Omega \,|\, X(\omega) \text{ nimmt Werte im Bereich } [a, b] \text{ an}\}$$

definieren, die wir im Folgenden oft kurz als

$$\{X \in [a, b]\} \quad \text{oder} \quad \{a \leq X \leq b\} \tag{6.1}$$

schreiben werden. Im Augensummenbeispiel wäre X definiert durch

$$X((p, q)) := p + q$$

und das Ereignis $A = \{X \geq 11\}$ lautet ausführlich geschrieben $A = \{(5,6),(6,5),(6,6)\}$.

Ziel ist es jetzt natürlich, die Wahrscheinlichkeit $\mathbb{P}(A)$ für die mittels X definierten Ereignisse zu bestimmen. Damit das möglich ist, muss man sicher stellen, dass die Menge A auch in $\mathcal{A}$ liegt. Dies motiviert die folgende Definition.

Definition 6.1 Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. Dann heißt $X : \Omega \to \mathbb{R}$ *Zufallsvariable* über $(\Omega, \mathcal{A}, \mathbb{P})$, wenn

$$\{\omega \in \Omega \,|\, X(\omega) \leq x\} \in \mathcal{A}$$

gilt für alle $x \in \mathbb{R}$. □

Die Auswahl der Ereignisse in Definition 6.1 erscheint zunächst recht eingeschränkt und es ist nicht offensichtlich, dass die dort formulierte Bedingung auch z.B. $\{\omega \in \Omega \,|\, X(\omega) \in [a,b]\} \in \mathcal{A}$ sicher stellt. Das ist aber tatsächlich der Fall. Man kann beweisen (was wir hier nicht durchführen werden), dass aus Definition 6.1

$$\{\omega \in \Omega \,|\, X(\omega) \in B\} \in \mathcal{A} \quad \text{für alle} \quad B \in \mathcal{B} = \sigma\{(-\infty, x] \,|\, x \in \mathbb{R}\}))$$

folgt. $\mathcal{B}$ ist dabei die von den halboffenen Intervallen $(-\infty, x]$ erzeugte σ-Algebra, die z.B. alle offenen Intervalle (a,b) und alle abgeschlossenen Intervalle $[a,b]$ (und noch viele weitere Teilmengen von $\mathbb{R}$) enthält. $\mathcal{B}$ heißt *Borel'sche σ-Algebra* und die Mengen $B \in \mathcal{B}$ werden als Borel-Mengen bezeichnet.

6.2 Diskrete Zufallsvariablen

Wir werden uns jetzt zunächst mit sogenannten *diskreten* Zufallsvariablen befassen. Dies sind Zufallsvariablen, die nur endlich oder abzählbar unendlich viele Werte annehmen können. Das Gegenstück dazu sind die *kontinuierlichen* oder *stetigen* Zufallsvariablen, die wir in Kapitel 9 betrachten werden. Das obige Beispiel der Augensumme beim zweimaligen Würfeln, dass durch die Zufallsvariable $X : \Omega \to \mathbb{R}$, $X((p,q)) := p + q$ modelliert wird, ist ein Beispiel für eine diskrete Zufallsvariable.

Wir wollen nun betrachten, wie sich die Wahrscheinlichkeiten, die durch $\mathbb{P}$ für Ereignisse aus $\mathcal{A}$ definiert sind, auf die Werte von X übertragen. Im diskreten Fall reicht es dazu, die einzelnen Werte von X zu betrachten. Wir erinnern dazu an die oben eingeführte Kurzschreibweise (6.1).

Definition 6.2 Für eine diskrete Zufallsvariable X mit den Werten $x_1, x_2, \ldots$ bezeichnet die Zuordnung

$$p_i := \mathbb{P}(\{X = x_i\}), \quad i \in I_X$$

die *Wahrscheinlichkeitsverteilung von X*. Mit $I_X \subset \mathbb{N}$ bezeichnen wir dabei die Indizes der möglichen Werte von X. □

Beachte, dass entweder $I_X = \{1, \ldots, n\}$ für ein endliches n oder $I_X = \mathbb{N}$ (oder $= \mathbb{N}_0$, wenn dies für das Nummerieren praktischer ist).

Wir müssen uns kurz überlegen, dass die Menge $\{X = x_i\}$, oder ausführlich geschrieben $\{\omega \in \Omega \,|\, X(\omega) = x_i\}$ tatsächlich ein Element von $\mathcal{A}$ ist, da wir nur dann $\mathbb{P}(\{X = x_i\})$ auswerten können. Wie oben gesagt, liegt $\{X \in [a, b]\}$ für alle abgeschlossenen Intervalle $[a, b]$ in $\mathcal{A}$. Setzen wir nun $a = b = x_i$, so erhalten wir $[a, b] = \{x_i\}$ und damit $\{X = x_i\} = \{X \in [a, b]\} \in \mathcal{A}$.

Beispiel 6.3 Im einführenden Beispiel der Augensumme beim zweimaligen Würfeln ergeben sich die möglichen Werte $x_1 = 2$, $x_2 = 3$, …, $x_{11} = 12$, also $I_X = \{1, \ldots, 11\}$. Ihre Wahrscheinlichkeiten p_i ergeben sich aus der Anzahl der möglichen Paare (p, q), für die $p + q = x_i$ gilt, multipliziert mit der Wahrscheinlichkeit $1/36$ für jedes Paar. Abzählen der Paare aus Beispiel 2.3 liefert die folgenden Werte:

$$
\begin{array}{lll}
x_1 = 2 & x_2 = 3 & x_3 = 4 \\
p_1 = \frac{1}{36} & p_2 = \frac{2}{36} = \frac{1}{18} & p_3 = \frac{3}{36} = \frac{1}{12} \\
x_4 = 5 & x_5 = 6 & x_6 = 7 \\
p_4 = \frac{4}{36} = \frac{1}{9} & p_5 = \frac{5}{36} & p_6 = \frac{6}{36} = \frac{1}{6} \\
x_7 = 8 & x_8 = 9 & x_9 = 10 \\
p_7 = \frac{5}{36} & p_8 = \frac{4}{36} = \frac{1}{9} & p_9 = \frac{3}{36} = \frac{1}{12} \\
x_{10} = 11 & x_{11} = 12 & \\
p_{10} = \frac{2}{36} = \frac{1}{18} & p_{11} = \frac{1}{36}. &
\end{array}
$$

□

Kennt man die Wahrscheinlichkeiten der Werte x_i, so kennt man die Wahrscheinlichkeiten von X für alle Intervalle, wie der folgende Satz zeigt.

Satz 6.4 Sei $X : \Omega \to \mathbb{R}$ eine Zufallsvariable über dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$. Dann gelten

1) $0 \leq p_i \leq 1$ für alle $i \in I_X$

2) $\displaystyle\sum_{i \in I_X} p_i = 1$

3) $\displaystyle\mathbb{P}(\{X \in [a, b]\}) = \sum_{\substack{i \in I_X \\ a \leq x_i \leq b}} p_i.$

Beweis: Übungsaufgabe.

Man kann also $\mathbb{P}(\{X \in [a, b]\})$ für alle Intervalle $[a, b]$ berechnen, wenn man die Werte p_i kennt. Es muss weder bekannt sein, was $\mathbb{P}(A)$ für beliebige Mengen $A \in \mathcal{A}$ ist, noch, welche $\omega \in \Omega$ nun genau von X auf ein x_i abgebildet werden. Aus diesem Grunde wird,

wenn man mit Zufallsvariablen arbeitet, üblicherweise weder die Abbildung X noch das zu Grunde liegende Wahrscheinlichkeitsmaß $\mathbb{P}$ explizit angegeben. Statt dessen werden nur die p_i, also die Wahrscheinlichkeitsverteilung von X an Hand der zu modellierenden Situation festgelegt. Das folgende Beispiel illustriert dieses Vorgehen für eine oft verwendete Verteilung.

Beispiel 6.5 (Geometrische Verteilung) Wir betrachten ein Glücksrad, das mit Wahrscheinlichkeit $0 < p < 1$ bei einmaligem Drehen einen Hauptgewinn ergibt. Die Zufallsvariable X soll nun die Anzahl der Fehlversuche bis zum Hauptgewinn darstellen. Offenbar ist die Wertemenge gerade gleich $\mathbb{N}_0$, denn es kann ja beliebig lange dauern, bis der Hauptgewinn eintritt. Es ist also $x_i = i$ für alle $i \in I_X = \mathbb{N}_0$.

Es sei nun das Ereignis

$$A_i := \{\text{Hauptgewinn beim } i\text{-ten Drehen}\}.$$

Offenbar können die A_i als unabhängig voneinander angesehen werden, da ein Gewinn im i-ten Drehen keinen Aufschluss über einen Gewinn beim j-ten Drehen für $j \neq i$ gibt. Wir können also die Formel aus Definition 5.1 verwenden. Dies ergibt die folgenden Wahrscheinlichkeiten für die Werte x_i:

$$\begin{array}{lcllcll} \mathbb{P}(X=0) & = & \mathbb{P}(A_1) & & & = & p \\ \mathbb{P}(X=1) & = & \mathbb{P}(A_1^c \cap A_2) & = & \mathbb{P}(A_1^c)\cdot\mathbb{P}(A_2) & = & (1-p)p \\ \mathbb{P}(X=2) & = & \mathbb{P}(A_3^c \cap A_2^c \cap A_3) & = & \mathbb{P}(A_1^c)\cdot\mathbb{P}(A_2^c)\cdot\mathbb{P}(A_3) & = & (1-p)^2 p \end{array}$$

und so weiter, woraus sich die allgemeine Formel

$$p_i = \mathbb{P}(X = i) = (1-p)^i p$$

für alle $i \in \mathbb{N}_0$ ergibt. Diese Wahl der Wahrscheinlichkeiten p_i definiert die sogenannte *geometrische Verteilung*. Wie in Satz 6.4 bewiesen, gilt

$$\sum_{i \in I_X} p_i = \sum_{i=1}^{\infty} (1-p)^i p = p \sum_{i=1}^{\infty} (1-p)^i = p \frac{1}{1-(1-p)} = p\frac{1}{p} = 1.$$

Setzen wir als konkreten Zahlenwert z.B. $p = 1/12$ ein (das Glücksrad hat 12 Felder, von denen eines den Hauptgewinn gibt), und wollen die Wahrscheinlichkeit ermitteln, dass bei 10 mal Drehen mindestens einmal der Hauptgewinn herauskommt, so erhalten wir

$$\mathbb{P}(\{X \in [0,9]\}) = \sum_{i=0}^{9} p_i = p \sum_{i=0}^{9} (1-p)^i = p \frac{1-(1-p)^{10}}{1-(1-p)} = 1 - (1-p)^{10} \approx 0.5811.$$

□

Beachte, wie der Zugang über die Zufallsvariablen hier die Modellierung vereinfacht: Wir müssen weder Ω festlegen (was hier schon etwas komplizierter wäre) noch durch eine geeignete Festlegung von $\mathbb{P}$ auf Ω sicher stellen, dass die A_i tatsächlich stochastisch unabhängig sind.

6.3 Verteilungsfunktionen

Für diskrete Zufallsvariablen liefern die p_i also alle Informationen, die man braucht, um Wahrscheinlichkeiten für die Werte von X zu berechnen. Für kontinuierliche Zufallsvariablen wird das nicht so leicht sein. Hierfür ist eine weitere, im diskreten Fall leicht aus den p_i abzuleitende Darstellung vorteilhafter.

Definition 6.6 Sei $X : \Omega \to \mathbb{R}$ eine Zufallsvariable über dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$. Dann heißt die durch

$$F(x) := \mathbb{P}(\{X \leq x\})$$

gegebene Funktion $F : \mathbb{R} \to \mathbb{R}$ die *Verteilungsfunktion* von X. □

Man sieht mit Satz 6.4 3) leicht, dass F für diskrete Zufallsvariablen durch

$$F(x) = \sum_{\substack{i \in I_X \\ x_i \leq x}} p_i$$

gegeben und damit eine Treppenfunktion ist, deren Sprungstellen gerade an den x_i liegen, wo sie um p_i nach oben springen. Abbildung 6.1 zeigt ein Beispiel einer solchen Funktion.

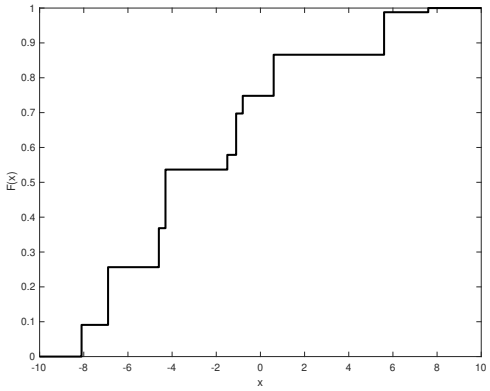


Abbildung 6.1: Beispiele für eine Verteilungfunktionen einer diskreten Zufallsvariablen

Für allgemeine Zufallsvariablen gilt der folgende Satz.

Satz 6.7 Sei $X : \Omega \to \mathbb{R}$ eine Zufallsvariable über dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ und F ihre Verteilungsfunktion. Dann gilt

1) $\lim_{x \to -\infty} F(x) = 0$

2) $\lim_{x \to \infty} F(x) = 1$

3) F ist (nicht notwendigerweise streng) monoton wachsend, d.h.

$$x < y \quad \Rightarrow \quad F(x) \leq F(y).$$

4) F ist rechtsseitig stetig, d.h.

$$\lim_{y \searrow x} F(y) = F(x).$$

Beweis: 1) Sei $x_n \to -\infty$ für $n \to \infty$. Indem wir Folgenglieder mit $x_{n+1} > x_n$ weglassen, können wir annehmen, dass $x_{n+1} \leq x_n$ gilt für alle n. Dann gilt für die Ereignisse $A_n := \{X \leq x_n\}$, dass $A_{n+1} \subset A_n$. Sie bilden also eine absteigende Folge und es gilt $\lim_{n\to\infty} A_n = \bigcap_{i=1}^{\infty} = \emptyset$ (vgl. die Definitionen vor Satz 3.5). Satz 3.5 sagt dann, dass

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\left(\lim_{n\to\infty} A_n\right) = \mathbb{P}(\emptyset) = 0.$$

2) Analog zu 1) mit $x_n \to \infty$ und den aufsteigenden Ereignissen $A_n := \{X \leq x_n\}$ mit $\lim_{n\to\infty} A_n = \Omega$.

3) Für $x < y$ gilt $\{X \leq x\} \subset \{X \leq y\}$, also $\{X \leq y\} = \{X \leq x\} \cup (\{X \leq y\} \setminus \{X \leq x\})$. Die Aussage folgt dann aus Satz 3.4(4) mit $A = \{X \leq x\}$ und $B = (\{X \leq y\} \setminus \{X \leq x\})$, weil für diese Mengen $A \cap B = \emptyset$ gilt.

4) Sei[1] $y_n \searrow x$ für $n \to \infty$. Dann gilt für die absteigende Folge von Ereignissen $A_n := \{X \leq y_n\}$, dass $\lim_{n\to\infty} A_n = \{X \leq x\}$. Die Behauptung folgt nun wie in 1) mit Satz 3.5. □

6.4 Erwartungswert

Der Erwartungswert ist der Mittelwert einer Zufallsvariablen, wenn das Zufallsexperiment "unendlich oft" durchgeführt wird.

Betrachten wir als Beispiel die Augensumme von zwei mal Würfeln. Wenn wir dies n mal mit Ergebnissen y_1 bis y_n durchführen, so ist der Mittelwert nach n Experimenten gegeben durch

$$M_n := \frac{y_1 + y_2 + \ldots + y_n}{n} = \frac{1}{n} \sum_{k=1}^{n} y_k.$$

Wir können die Summe aber auch anders bilden, indem wir die absoluten und relativen Häufigkeiten $h_n(x_i)$ und $H_n(x_i)$ für die einzelnen Ergebnisse $x_1 = 2$ bis $x_{11} = 12$ verwenden. Die Zahl 2 kommt unter den y_k gerade $h_n(2)$-mal vor, die Zahl 3 $h_n(3)$-mal und so weiter. Damit können wir die obige Summe als

$$\frac{y_1 + y_2 + \ldots + y_n}{n} = \frac{h_n(2) \cdot 2 + h_n(3) \cdot 3 + \ldots h_n(12) \cdot 12}{n} = \sum_{i=1}^{11} x_i \frac{h_n(x_i)}{n} = \sum_{i=1}^{11} x_i H_n(x_i)$$

schreiben. Für unendlich viele Experimente müssen wir nun den Grenzwert für n gegen ∞ betrachten. Mit (3.1) erhalten wir

$$\lim_{n\to\infty} M_n = \lim_{n\to\infty} \sum_{i=1}^{11} x_i H_n(x_i) = \sum_{i=1}^{11} \lim_{n\to\infty} x_i H_n(x_i) = \sum_{i=1}^{11} x_i \mathbb{P}(\{X = x_i\}).$$

[1] $y_n \searrow x$ bedeutet: $y_n \to x$ für $n \to \infty$ und $y_{n+1} \leq y_n$ für alle $n \in \mathbb{N}$.

Im allgemeinen Fall ist dies für diskrete Zufallsvariablen genau die Definition des Erwartungswerts.

Definition 6.8 Sei $X : \Omega \to \mathbb{R}$ eine diskrete Zufallsvariable über dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ mit $p_i = \mathbb{P}(\{X = x_i\})$ mit $\sum_{i \in I_X} |x_i| p_i < \infty$. Dann heißt

$$\mathbb{E}[X] := \sum_{i \in I_X} x_i p_i = \sum_{i \in I_X} x_i \mathbb{P}(\{X = x_i\})$$

der *Erwartungswert* von X. Ist $\sum_{i \in I_X} |x_i| p_i < \infty$ nicht erfüllt, so besitzt X keinen Erwartungswert. □

Beachte, dass $\sum_{i \in I_X} |x_i| p_i < \infty$ immer erfüllt ist, falls I_X endlich ist, also X nur endlich viele Werte annimmt.

Im Falle der geometrischen Verteilung gilt, weil alle $x_i \geq 0$ sind,

$$\sum_{i=0}^{\infty} |x_i| p_i = \sum_{i=0}^{\infty} p_i x_i = \sum_{i=0}^{\infty} p_i i = p(1-p) \sum_{i=0}^{\infty} i(1-p)^{i-1} = p(1-p) \frac{1}{(1-(1-p))^2} = \frac{1-p}{p} < \infty$$

und damit existiert der Erwartungswert und erfüllt $\mathbb{E}[X] = \frac{1-p}{p}$. Man hat also im Mittel $(1-p)/p$ Fehlversuche vor einem Hauptgewinn im Glücksrad.

Satz 6.9 Sei $X, Y : \Omega \to \mathbb{R}$ diskrete Zufallsvariablen über dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$. Dann gilt:

1) Falls $\mathbb{E}[X]$ existiert, so gilt für alle $a, b \in \mathbb{R}$

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b. \tag{6.2}$$

2) Falls $\mathbb{E}[X]$ und $\mathbb{E}[Y]$ existieren, so gilt

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]. \tag{6.3}$$

3) Falls $\mathbb{E}[X]$ und $\mathbb{E}[Y]$ existieren und $X(\omega) \leq Y(\omega)$ für alle $\omega \in \Omega$, so gilt

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

Beweis: 1) Es gilt

$$\mathbb{E}[aX + b] = \sum_{i \in I_X} (ax_i + b) p_i = a \underbrace{\sum_{i \in I_X} x_i p_i}_{= \mathbb{E}[X]} + b \underbrace{\sum_{i \in I_X} p_i}_{=1} = a\mathbb{E}[X] + b.$$

2) Seien x_i und Y_j, $i \in I_X$, $j \in I_Y$, die diskreten Werte von X und Y. Wir setzen $A_i := \{X = x_i\}$ und $B_j := \{Y = y_j\}$ und betrachten die Menge von Ereignissen $\{A_i \cap B_j \,|\, i \in I_X, j \in I_Y\} \subset \mathcal{A}$. Dies sind abzählbar viele Mengen, die wir als C_k, $k \in \mathbb{N}$ durchnummerieren

können. Auf jeder Menge C_k nehmen X und Y nun konstante Werte an, die wir mit $\tilde{x}_k$ und $\tilde{y}_k$ bezeichnen. Mit $\tilde{p}_k = \mathbb{P}(C_k)$ gilt

$$\mathbb{E}[X] + \mathbb{E}[Y] = \sum_{k\in\mathbb{N}} \tilde{x}_k\tilde{p}_k + \sum_{k\in\mathbb{N}} \tilde{y}_k\tilde{p}_k = \sum_{k\in\mathbb{N}} (\tilde{x}_k + \tilde{y}_k)\tilde{p}_k = \mathbb{E}[X+Y].$$

3) Unter der Voraussetzung gilt $X(\omega) - Y(\omega) \leq 0$, damit auch $\mathbb{E}[X-Y] \leq 0$, also nach 1) und 2)

$$\mathbb{E}[X] - \mathbb{E}[Y] = \mathbb{E}[X] + \mathbb{E}[-Y] = \mathbb{E}[X-Y] \leq 0 \;\Rightarrow\; \mathbb{E}[X] \leq \mathbb{E}[Y].$$

□

Beachte, dass aus 1) und 2) auch

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y] \tag{6.4}$$

folgt.

Satz 6.10 Sei $X : \Omega \to \mathbb{R}$ eine diskrete Zufallsvariable über dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ mit $p_i = \mathbb{P}(\{X = x_i\})$ und $g : \mathbb{R} \to \mathbb{R}$ eine stetige Funktion. Falls $\mathbb{E}[X]$ und $\mathbb{E}[g(X)]$ existieren, so gilt

$$\mathbb{E}[g(X)] = \sum_{i\in I_X} g(x_i)p_i.$$

Beweis: Für die diskrete Zufallsvariable $Y = g(X)$ gilt

$$\mathbb{E}[Y] = \sum_{j\in I_Y} y_j\mathbb{P}(\{Y = y_j\}) = \sum_{j\in I_Y} y_j \left(\sum_{\substack{i\in I_X \\ g(x_i)=y_j}} p_i \right) = \sum_{\substack{i\in I_X, j\in I_Y \\ g(x_i)=y_j}} y_j p_i = \sum_{i\in I_X} g(x_i)p_i.$$

□

6.5 Varianz

Betrachte die diskreten Zufallsvariablen X und Y mit Werten $X \in \{0, 5, 10\}$ und $Y \in \{4, 5, 6\}$, jeweils mit Wahrscheinlichkeit $1/3$. Man rechnet leicht nach, dass beide Zufallsvariablen den Erwartungswert $\mathbb{E}[X] = \mathbb{E}[Y] = 5$ besitzen. Betrachtet man aber zufällige Realisierungen, so sind diese bei X z.B. von der Form $0, 0, 10, 5, 5, 10, 0, 10, 5, 5, \ldots$ und bei Y z.B. von der Form $4, 4, 6, 5, 5, 6, 4, 6, 5, 5, \ldots$. Die Zufallsvariable X hat also Werte, die typischerweise viel weiter vom Erwartungswert entfernt liegen als bei Y. Dies ist im Erwartungswert aber gar nicht sichtbar. Dieser gibt also an, um welchen Wert sich die Realisierungen bewegen, aber nicht, wie weit die Realisierungen im Mittel vom Erwartungswert entfernt sind. Um diesen Aspekt zu erfassen, führen wir nun die Varianz und die Standardabweichung ein.

Definition 6.11 Sei $X : \Omega \to \mathbb{R}$ eine diskrete Zufallsvariable über dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ mit $p_i = \mathbb{P}(\{X = x_i\})$. Falls $\mathbb{E}[X^2]$ existiert, so heißt die Größe $\sigma^2 := \mathbb{V}[X]$ mit

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

die *Varianz* von X und $\sigma = \sqrt{\mathbb{V}[X]}$ die *Standardabweichung* von X. □

Beispiel 6.12 Für X und Y aus dem Beispiel vor Definition 6.11 gilt $\mathbb{E}[X] = \mathbb{E}[Y] = 5$. Daher nimmt $(X - \mathbb{E}[X])^2$ die Werte $(-5)^2 = 25$, $0^2 = 0$ und $5^2 = 25$, jeweils mit Wahrscheinlichkeit $1/3$ an. Die Varianz von X beträgt also $\sigma^2 = \mathbb{V}[X] = 1/3 \cdot 25 + 1/3 \cdot 0 + 1/3 \cdot 25 = 50/3 = 16{,}\bar{6}$ und die Standardabweichung $\sigma = \sqrt{\mathbb{V}[X]} \approx 4{,}08$. Die Zufallsvariable $(Y - \mathbb{E}[Y])^2$ hingegen nimmt mit Wahrscheinlichkeit jeweils $1/3$ die drei Werte $(-1)^2 = 1$, $0^2 = 0$ und $1^2 = 1$ an. Ihre Varianz beträgt also $\sigma^2 = \mathbb{V}[Y] = 1/3 \cdot 1 + 1/3 \cdot 0 + 1/3 \cdot 1 = 2/3 = 0{,}\bar{6}$ und die Standardabweichung $\sigma = \sqrt{\mathbb{V}[Y]} \approx 0.816$. Die geringere "Streuung" von Y um den Mittelwert ist also deutlich zu sehen.

Als weiteres Beispiel betrachte die Zufallsvariable Z mit den drei Werten $Z \in \{0, 5, 10\}$, jetzt aber mit Wahrscheinlichkeiten $\mathbb{P}(\{Z = 0\}) = 0.1$, $\mathbb{P}(\{Z = 5\}) = 0.8$ und $\mathbb{P}(\{Z = 10\}) = 0.1$. Die Werte sind also identisch mit denen von X, allerdings werden die Werte 0 und 10 viel seltener angenommen. Eine typische Folge von Realisierungen wäre z.B. $5, 5, 5, 5, 0, 5, 5, 10, 5, 5, 5, 5, 0, \ldots$. Hier beträgt der Erwartungswert wiederum $\mathbb{E}[X] = 5$, die Varianz beträgt $\sigma^2 = \mathbb{V}[Z] = 0.1 \cdot 25 + 0.8 \cdot 0 + 0.1 \cdot 25 = 0.2 \cdot 25 = 5$ und die Standardabweichung $\sigma = \sqrt{\mathbb{V}[Z]} \approx 2,236$. Die Varianz wird also nicht nur geringer, wenn die streuenden Werte näher am Erwartungswert liegen, sondern auch, wenn vom Erwartungswert weit entfernte Werte seltener auftreten. □

Wegen $|x_i| \leq 1 + x_i^2$ folgt aus $\mathbb{E}[X^2] = \sum_{i \in I_X} x_i^2 p_i < \infty$, dass auch $\sum_{i \in I_X} |x_i| p_i < \infty$. Die Bedingung für die Existenz der Varianz impliziert also die Existenz des Erwartungswerts. Beachte, dass die Umkehrung nicht gilt. Betrachte z.B. die Zufallsvariable, die den Wert $x_i = i$ für $i \in \mathbb{N}$ mit Wahrscheinlichkeit $p_i = c/i^3$ annimmt, wobei $c \approx 1/1{,}2021$ so gewählt ist, dass die p_i sich zu 1 aufsummieren. Dann gilt

$$\sum_{i=1}^{\infty} |x_i| p_i = \sum_{i=1}^{\infty} i \frac{c}{i^3} = \sum_{i=1}^{\infty} \frac{c}{i^2} = \frac{c\pi^2}{6} < \infty,$$

weswegen der Erwartungswert von X existiert, aber

$$\mathbb{E}[X^2] = \sum_{i=1}^{\infty} x_i^2 p_i = \sum_{i=1}^{\infty} i^2 \frac{c}{i^3} = \sum_{i=1}^{\infty} \frac{c}{i} = \infty,$$

weswegen der Erwartungswert von X^2 nicht existiert.

Die folgende Formel für die Varianz ist oft leichter auszurechnen.

Satz 6.13 Sei $X : \Omega \to \mathbb{R}$ eine diskrete Zufallsvariable über dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$, für die $\mathbb{E}[X^2]$ existiert. Dann gilt

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

Beweis: Mit Ausmultiplizieren gilt

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

Da $\mathbb{E}[X^2]$ und damit auch $\mathbb{E}[X]$ existiert, folgt mit (6.4) und weil $\mathbb{E}[X] \in \mathbb{R}$ gilt

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

□

Dies zeigt insbesondere, dass aus der Existenz von $\mathbb{E}[X^2]$ die Ungleichung $\mathbb{V}[X] < \infty$ folgt, was die Begründung für diese Bedingung in Definition 6.11 ist.

Lemma 6.14 Sei $X : \Omega \to \mathbb{R}$ eine diskrete Zufallsvariable über dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$. Dann gilt für alle $a, b \in \mathbb{R}$

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

Beweis: Dies folgt aus der Rechnung

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

□

Beispiel 6.15 Wir berechnen die Varianz $\mathbb{V}[X]$ der geometrischen Verteilung. Wir verwenden dabei die bereits bekannten Formeln $p_i = p(1-p)^i$ und $\mathbb{E}[X] = \frac{1-p}{p}$ sowie die Summenformel

$$\sum_{i=0}^{\infty} i(i-1)x^{i-2} = \frac{2}{(1-x)^3},$$

die wir hier nicht beweisen. Damit gilt

$$\begin{aligned}\mathbb{E}[X^2] &= \sum_{i=0}^{\infty} i^2 p(1-p)^{i-1} = \sum_{i=0}^{\infty} i(i-1) \underbrace{p(1-p)^i}_{=p(1-p)^2(1-p)^{i-2}} + \underbrace{\sum_{i=0}^{\infty} ip(1-p)^i}_{=\mathbb{E}[X]=\frac{1-p}{p}} \\ &= p(1-p)^2 \frac{2}{(1-(1-p))^3} + \frac{1-p}{p} = \frac{2(1-p)^2}{p^2} + \frac{1-p}{p}.\end{aligned}$$

Es folgt mit Satz 6.13

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

□

Wir haben die Varianz damit motiviert, dass sie angibt, wie weit die Realisierungen einer Zufallsvariablen um den Erwartungswert streuen, bzw. wie oft das passiert. Der folgende Satz zeigt, dass diese Anschauung auch mathematisch gerechtfertigt ist. Er gibt die Wahrscheinlichkeit an, dass die Werte der Zufallsvariablen außerhalb eines symmetrischen Intervalls mit Breite $2\varepsilon > 0$ um den Erwartungswert $\mathbb{E}[X]$ liegen und zeigt, dass diese für eine feste Breite ε um so kleiner ist, je kleiner die Varianz $\mathbb{V}[X]$ ist.

Satz 6.16 (Tschebyscheff'sche Ungleichung) Sei $X : \Omega \to \mathbb{R}$ eine diskrete Zufallsvariable über dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$, deren Varianz existiere. Dann gilt für alle $\varepsilon > 0$ die Abschätzung

$$\mathbb{P}(\{|X - \mathbb{E}[X]| \geq \varepsilon\}) \leq \frac{\mathbb{V}[X]}{\varepsilon^2}, \tag{6.5}$$

was äquivalent ist zu

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

Beweis: Es gilt

$$\begin{aligned}\mathbb{V}[X] &= \sum_{i \in I_X} (x_i - \mathbb{E}[X])^2 p_i \geq \sum_{\substack{i \in I_X \\ |x_j - \mathbb{E}[X]| \geq \varepsilon}} \underbrace{(x_i - \mathbb{E}[X])^2}_{\geq \varepsilon^2} p_i \\ &\geq \varepsilon^2 \sum_{\substack{i \in I_X \\ |x_j - \mathbb{E}[X]| \geq \varepsilon}} p_i = \varepsilon^2 \mathbb{P}(\{|X - \mathbb{E}[X]| \geq \varepsilon\}).\end{aligned}$$

Dividieren durch ε^2 liefert dann Ungleichung (6.5). Die äquivalente Formulierung erhält man, indem man ε in (6.5) überall durch $\varepsilon\sigma$ ersetzt, beide Seiten der Ungleichung im Argument von $\mathbb{P}$ durch σ teilt und auf der rechten Seite σ^2 gegen $\mathbb{V}[X]$ kürzt. □

Bemerkung 6.17 (i) Mit exakt dem gleichen Beweis kann man (6.5) mit "$<$" an Stelle von (beiden) "$\leq$" beweisen.

(ii) Indem man zum Gegenereignis übergeht, sieht man sofort, dass auch

$$P(\{|X - \mathbb{E}[X]| < \varepsilon\}) = 1 - \mathbb{P}(\{|X - \mathbb{E}[X]| \geq \varepsilon\}) \geq 1 - \frac{\mathbb{V}[X]}{\varepsilon^2}$$

gilt. □

Kapitel 7

Beispiele für diskrete Verteilungen

Wir haben mit der geometrischen Verteilung aus Beispiel 6.5 bereits eine diskrete Verteilung kennengelernt. Eine weitere, sehr einfache Verteilung ist die Gleichverteilung, für die $p_i = 1/n$ ist, wobei n die (endliche) Anzahl der Elemente in I_X angibt. Diese haben wir bereits vor der Einführung der Zufallsvariablen im Zusammenhang mit dem Laplace-Modell in Satz 3.6(b) kennengelernt. In diesem Kapitel stellen wir drei weitere diskrete Verteilungen vor.

7.1 Binomialverteilung

Die Binomialverteilung beschreibt, wie oft ein Ereignis $A \in \mathcal{A}$ eintritt, wenn wir n unabhängige Zufallsexperimente durchführen, bei denen das Ereignis jeweils mit Wahrscheinlichkeit p eintritt. Die Situation ist also die gleiche wie bei der geometrischen Verteilung in Beispiel 6.5, nur dass wir uns jetzt nicht fragen, wie groß die Wahrscheinlichkeit ist, dass der erste Hauptgewinn nach n Fehlversuchen auftritt, sondern die Wahrscheinlichkeit, dass in n Versuchen gerade k-mal der Hauptgewinn erzielt wird.

Wir definieren dazu für $i = 1, \ldots, n$ die Zufallsvariable

$$X_i = \begin{cases} 1, & A \text{ tritt im } i\text{-ten Versuch ein} \\ 0, & A \text{ tritt im } i\text{-ten Versuch nicht ein.} \end{cases}$$

Dann gibt die Zufallsvariable $S_n := X_1 + X_2 + \ldots, X_n$ gerade die Anzahl der Ereignisse A bei n Experimenten an. Wir wollen nun die Wahrscheinlichkeiten $\mathbb{P}(\{S_n = k\})$ und den Erwartungswert $\mathbb{E}(S_n)$ sowie die Varianz $\mathbb{V}[S_n]$ bestimmen. Dazu betrachten wir zunächst die einzelnen X_i. Für diese gilt

$$\mathbb{P}(\{X_i = 1\}) = \mathbb{P}(A) = p \quad \text{und} \quad P(\{X_i = 0\}) = \mathbb{P}(A^c) = 1 - p,$$

also $\mathbb{E}[X_i] = 1 \cdot p + 0 \cdot (1-p) = p$ und $\mathbb{V}[X_i] = (1-p)^2 \cdot p + p^2(1-p) = p - 2p^2 + p^3 + p^2 - p^3 = p - p^2 = p(1-p)$. Damit folgt

$$\mathbb{E}[S_n] = \sum_{i=1}^{n} \mathbb{E}[X_i] = np \quad \text{sowie} \quad \mathbb{V}[S_n] = \sum_{i=1}^{n} \mathbb{V}[X_i] = np(1-p).$$

Zur Berechnung von $\mathbb{P}(\{S_n = k\})$ betrachte die Tupel $(x_1, x_2, \ldots, x_n) \in \{0,1\}^n$ mit $\sum_{i=1}^n x_i = k$, die gerade einer Folge von Experimenten mit k Hauptgewinnen entsprechen. Die Wahrscheinlichkeit, dass das Ergebnis einem solchen Tupel entspricht, beträgt wegen der Unabhängigkeit der Zufallsexperimente

$$\mathbb{P}(\{X_1 = x_1\} \cap \{X_2 = x_2\} \cap \ldots \cap \{X_n = x_n\}) = \prod_{i=1}^n \mathbb{P}(\{X_i = x_i\}) = p^k(1-p)^{n-k}.$$

Jedes solche Tupel ist nun eindeutig durch die k paarweise verschiedenen Indizes $i_1, \ldots, i_k$ bestimmt, für die $x_{i_j} = 1$ gilt. Diese werden ohne Zurücklegen aus der Menge $\{1, \ldots, n\}$ gezogen, also gibt es nach Satz 3.8(2) gerade $\binom{n}{k}$ verschiedene Folgen dieser Art. Es folgt also

$$\begin{aligned}\mathbb{P}(\{S_n = k\}) &= \mathbb{P}\left(\left\{ X_1 = x_1\} \cap \{X_2 = x_2\} \cap \ldots \cap \{X_n = x_n\} \;\middle|\; x_i \in \{0,1\}, \sum_{i=1}^n x_i = k \right\}\right) \\ &= \binom{n}{k} p^k (1-p)^{n-k}.\end{aligned}$$

Definition 7.1 Eine diskrete Zufallsvariable $X : \Omega \to \mathbb{R}$ auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ heißt *binomialverteilt* mit Parametern $n \in \mathbb{N}$ und $p \in [0,1]$, falls sie die Wahrscheinlichkeitsfunktion

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

besitzt für alle $k \in \{0, 1, \ldots, n\}$. Wir schreiben dann $X \sim \mathrm{Bin}(n, p)$. □

Beispiel 7.2 Ein idealer Würfel wird 20 mal geworfen. Wie groß ist die Wahrscheinlichkeit, mindestens 2 Sechser zu würfeln? Offenbar erfüllt die Zufallsvariable S, die die Anzahl der geworfenen Sechser angibt, gerade $S \sim \mathrm{Bin}(20, 1/6)$. Es folgt also

$$\begin{aligned}\mathbb{P}(\{S \geq 2\}) &= 1 - \mathbb{P}(\underbrace{\{S < 2\}}_{=\{S \leq 1\}}) \\ &= 1 - \mathbb{P}(\{X = 0\}) - \mathbb{P}(\{X = 1\}) \\ &= 1 - \binom{20}{0}\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^{20} - \binom{20}{1}\left(\frac{1}{6}\right)^1\left(\frac{5}{6}\right)^{19} \approx 0.8696.\end{aligned}$$

Die Wahrscheinlichkeit beträgt also fast 87%. □

7.2 Poisson-Verteilung

Die Poisson-Verteilung beschreibt die Anzahl X_t von zufälligen Ereignissen, die bis zu einer gegebenen Zeit t eintreten. Sie wird zu Modellierung in vielen Bereichen benutzt, z.B. für Verbindungsanfragen in Computernetzwerken, für Ansteckungsereignisse in Infektionsmodellen oder für das Vorrücken eines Fahrzeugs im stop-and-go Verkehr. Die grundlegenden Modellannahmen sind:

- Das Auftreten eines Ereignisses hängt nur von der verstrichenen Zeit ab.
- Die Anzahl der Ereignisse im Zeitintervall $[t_0, t_1]$ ist stochastisch unabhängig von der Anzahl der Ereignisse im Zeitintervall $[t_2, t_3]$ für alle $t_0 < t_1 < t_2 < t_3$.
- Für immer kleinere $t > 0$ gilt mit immer höherer Genauigkeit[1] $\mathbb{P}(\{X_t = 1\}) \approx \mu t$. Der Parameter $\mu > 0$ heißt dabei die *Intensität* des Prozesses.

Aus diesen Annahmen kann man die folgende Verteilung berechnen.

Definition 7.3 Eine diskrete Zufallsvariable $X_t : \Omega \to \mathbb{R}$ auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ heißt *Poisson-verteilt* mit Parametern $t > 0$ und $\mu > 0$, wenn

$$\mathbb{P}(\{X_t = k\}) = \frac{(\mu t)^k}{k!} \exp(-\mu t)$$

gilt für alle $k \in \mathbb{N}$. Oft wird $\lambda = \mu t$ gesetzt. Wir schreiben dann $X \sim \pi_{\mu t}$ bzw. $X \sim \pi_\lambda$. □

Beispiel 7.4 Wir wollen die Wahrscheinlichkeit berechnen, dass in einem Netzwerkknoten innerhalb von 15 Minuten mindestens 3 und höchstens 7 Verbindungsanfragen eingehen. Für die gewählte Zeiteinheit "Minuten" sei die Intensität $\mu = 1/3$, also $\lambda = \mu t = 1/3 \cdot 15 = 5$. Es gilt also $X_{15} \sim \pi_{\mu t} = \pi_5$. Damit folgt

$$\mathbb{P}(\{3 \leq X_{15} \leq 7\}) = \sum_{k=3}^{7} \frac{5^k}{k!} \exp(-5) \approx 0.742.$$

□

Für den Erwartungswert einer Zufallsvariablen $X_t \sim \pi_\lambda$ gilt

$$\mathbb{E}[X_t] = \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} \exp(-\lambda) = \lambda \exp(-\lambda) \underbrace{\sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}}_{= \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = \exp(\lambda)} = \lambda \underbrace{\exp(-\lambda) \exp(\lambda)}_{=1} = \lambda.$$

Also ist $\lambda = \mu t$ gerade die erwartete Anzahl von Ereignissen in einem Zeitintervall der Länge t, d.h. μ ist die erwartete Anzahl von Ereignissen pro Zeiteinheit.

Wegen

$$\begin{aligned} \mathbb{E}[X_t^2] &= \sum_{k=0}^{\infty} k^2 \frac{\lambda^k}{k!} \exp(-\lambda) = \left(\sum_{k=0}^{\infty} k(k-1) \frac{\lambda^k}{k!} \exp(-\lambda) \right) + \underbrace{\sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} \exp(-\lambda)}_{= \mathbb{E}[X_t] = \lambda} \\ &= \lambda^2 \exp(-\lambda) \underbrace{\sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!}}_{= \sum_{k=1}^{\infty} \frac{\lambda^k}{k!} = \exp(\lambda)} + \lambda = \lambda^2 \exp(-\lambda) \exp(\lambda) + \lambda = \lambda^2 + \lambda \end{aligned}$$

gilt für die Varianz von $X_t \sim \pi_\lambda$

$$\mathbb{V}[X_t] = \mathbb{E}[X_t^2] - (\mathbb{E}[X_t])^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

[1] Mathematisch präzise: $\mathbb{P}(\{X_t = 1\}) = \mu t + o(|t|)$

7.3 Hypergeometrische Verteilung

Wir erinnern an das Lottobeispiel aus Beispiel 3.9. Wir können dieses abstrakt auch wie folgt formulieren: In einer Urne befinden sich 49 Kugeln. Von diesen sind 6 schwarz (die getippten) und 43 weiß (die nicht getippten). Die Zufallsvariable, die die Anzahl der richtig getippten Zahlen angibt, ist dann

$$X = \text{“Anzahl der gezogenen schwarzen Kugeln”}.$$

Verallgemeinert man dieses Spiel nun auf insgesamt $N \in \mathbb{N}$ (statt 49) Kugeln, von denen $M \in \mathbb{N}$, $M \leq N$ (statt 6) schwarz sind, so erhält man für X mit der gleichen Argumentation wie in Beispiel 3.9 beim Ziehen von n Kugeln die Verteilung

$$\mathbb{P}(\{X = m\}) = \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}} \tag{7.1}$$

für $m \in \{0, 1, \ldots, \min\{n, M\}\}$.

Definition 7.5 Eine diskrete Zufallsvariable $X : \Omega \to \mathbb{R}$ auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ heißt *hypergeometrisch verteilt*, falls sie die Wahrscheinlichkeitsverteilung (7.1) besitzt mit Parametern n, N und M. Wir schreiben dann $X \sim H(n, N, M)$. □

Für den Erwartungswert und die Varianz gilt mit längeren Rechnungen (für die wir auf [1, Satz 13.45] verweisen)

$$\mathbb{E}[X] = n\frac{M}{N} \quad \text{und} \quad \mathbb{V}[X] = n\frac{M}{N}\left(1 - \frac{M}{N}\right)\frac{N-n}{N-1}.$$

Kapitel 8

Stochastische Unabhängigkeit von Zufallsvariablen

8.1 Definition und erste Folgerungen

Wir haben im letzten Kapitel gesehen, dass die Bilder der Zufallsvariablen eine Wahrscheinlichkeitsverteilung definieren, mit der wir ohne Kenntnis des zu Grunde liegenden Ω und $\mathbb{P}$ rechnen können. Es ist daher sinnvoll, das Konzept der stochastischen Unabhängigkeit aus Definition 5.3 direkt für Zufallsvariablen zu definieren, ohne Ω und $\mathbb{P}$ explizit zu verwenden.

Definition 8.1 Seien $X_i : \Omega \to \mathbb{R}$, $i \in J \subset \mathbb{N}$, Zufallsvariablen über dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$. Dann heißen die X_i *stochastisch unabhängig*, wenn für jede Auswahl von $m \in \mathbb{N}$ verschiedenen Zufallsvariablen $X_{i_1}, \ldots, X_{i_m}$ mit $i_1, \ldots, i_m \in J$ gilt, dass

$$\mathbb{P}\left(\bigcap_{j=1}^{m} \{X_{i_j} \leq x_j\}\right) = \prod_{j=1}^{m} \mathbb{P}(\{X_{i_j} \leq x_j\}) \tag{8.1}$$

gilt für alle $x_1, \ldots, x_m \in \mathbb{R}$. □

Bemerkung 8.2 Wie bereits bei der Definition von Zufallsvarablen reicht es, die Bedingung für Ereignisse der Form $\{X_{i_j} \leq x_j\} = \{X_{i_j} \in (\infty, x_j]\}$ zu fordern. Man kann dann beweisen, dass sie für alle Ereignisse der Form $\{X_{i_j} \in B_j\}$ für beliebige Borel-Mengen $B_1, \ldots, B_m \in \mathcal{B}$ erfüllt ist, was wir hier aber nicht durchführen werden. □

Beispiel 8.3 Es sei $\Omega = \{1, 2, 3, 4\}$ mit $\mathbb{P}(\{\omega\}) = 1/4$ für alle $\omega \in \Omega$. Wir betrachten die Zufallsvariablen $X, Y, Z : \Omega \to \mathbb{R}$ gegeben durch

$$X(\omega) = \begin{cases} 0, & \omega \in \{1,2\} \\ 1, & \omega \in \{3,4\}, \end{cases} \qquad Y(\omega) = \begin{cases} 0, & \omega \in \{1,3\} \\ 1, & \omega \in \{2,4\}, \end{cases} \qquad Z(\omega) = \begin{cases} 0, & \omega \in \{3,4\} \\ 1, & \omega \in \{1,2\}. \end{cases}$$

Dann gilt

$$\mathbb{P}(\{X \leq x\}) = \begin{cases} \mathbb{P}(\emptyset) & = 0, & x < 0 \\ \mathbb{P}(\{1,2\}) & = 1/2, & 0 \leq x < 1, \\ \mathbb{P}(\Omega) & = 1, & x \geq 1, \end{cases}$$

$$\mathbb{P}(\{Y \leq x\}) = \begin{cases} \mathbb{P}(\emptyset) & = 0, & x < 0 \\ \mathbb{P}(\{1,3\}) & = 1/2 & 0 \leq x < 1, \\ \mathbb{P}(\Omega) & = 1, & x \geq 1, \end{cases}$$

und

$$\mathbb{P}(\{Z \leq x\}) = \begin{cases} \mathbb{P}(\emptyset) & = 0, & x < 0 \\ \mathbb{P}(\{3,4\}) & = 1/2 & 0 \leq x < 1, \\ \mathbb{P}(\Omega) & = 1, & x \geq 1. \end{cases}$$

Zudem gilt

$$\mathbb{P}(\{X \leq x_1\} \cap \{Y \leq x_2\}) = \begin{cases} \mathbb{P}(\emptyset) & = 0, & x_1 < 0 & \text{und } x_2 < 0 \\ \mathbb{P}(\emptyset) & = 0, & 0 \leq x_1 < 1 & \text{und } x_2 < 0 \\ \mathbb{P}(\emptyset) & = 0, & x_1 > 1 & \text{und } x_2 < 0 \\ \mathbb{P}(\emptyset) & = 0, & x_1 < 0 & \text{und } 0 \leq x_2 < 1 \\ \mathbb{P}(\{1\}) & = 1/4, & 0 \leq x_1 < 1 & \text{und } 0 \leq x_2 < 1 \\ \mathbb{P}(\{1,3\}) & = 1/2, & x_1 > 1 & \text{und } 0 \leq x_2 < 1 \\ \mathbb{P}(\emptyset) & = 0, & x_1 < 0 & \text{und } x_2 > 1 \\ \mathbb{P}(\{1,2\}) & = 1/2, & 0 \leq x_1 < 1 & \text{und } x_2 > 1 \\ \mathbb{P}(\Omega) & = 1, & x_1 > 1 & \text{und } x_2 > 1 \end{cases}$$

In allen Fällen prüft man leicht nach, dass

$$\mathbb{P}(\{X \leq x_1\} \cap \{Y \leq x_2\}) = \mathbb{P}(\{X \leq x_1\})\mathbb{P}(\{Y \leq x_2\})$$

gilt. Folglich sind X und Y stochastisch unabhängig. Für X und Z hingegen gilt

$$\mathbb{P}(\{X \leq x_1\} \cap \{Z \leq x_2\}) = \begin{cases} \mathbb{P}(\emptyset) & = 0, & x_1 < 0 & \text{und } x_2 < 0 \\ \mathbb{P}(\emptyset) & = 0, & 0 \leq x_1 < 1 & \text{und } x_2 < 0 \\ \mathbb{P}(\emptyset) & = 0, & x_1 > 1 & \text{und } x_2 < 0 \\ \mathbb{P}(\emptyset) & = 0, & x_1 < 0 & \text{und } 0 \leq x_2 < 1 \\ \mathbb{P}(\emptyset) & = 0, & 0 \leq x_1 < 1 & \text{und } 0 \leq x_2 < 1 \\ \mathbb{P}(\{3,4\}) & = 1/2, & x_1 > 1 & \text{und } 0 \leq x_2 < 1 \\ \mathbb{P}(\emptyset) & = 0, & x_1 < 0 & \text{und } x_2 > 1 \\ \mathbb{P}(\{1,2\}) & = 1/2, & 0 \leq x_1 < 1 & \text{und } x_2 > 1 \\ \mathbb{P}(\Omega) & = 1, & x_1 > 1 & \text{und } x_2 > 1 \end{cases}$$

Für $0 \leq x_1 < 1$ und $0 \leq x_2 < 1$ gilt daher

$$\mathbb{P}(\{X \leq x_1\} \cap \{Z \leq x_2\}) = 0 \neq 1/4 = \mathbb{P}(\{X \leq x_1\})\mathbb{P}(\{Z \leq x_2\}).$$

Also sind X und Z nicht stochastisch unabhängig. □

Bemerkung 8.4 Die stochastische Unabhängigkeit von Zufallsvariablen überträgt sich wie folgt auf die Verteilungsfunktion F aus Definition 6.6. Definiert man die gemeinsame Verteilungsfunktion als

$$F_{X_{i_1},X_{i_2},\ldots,X_{i_m}}(x_1,x_2,\ldots,x_m) := \mathbb{P}\left(\bigcap_{j=1}^{m}\{X_{i_j} \leq x_j\}\right),$$

so ist Gleichung (8.1) äquivalent zu

$$F_{X_{i_1},X_{i_2},\ldots,X_{i_m}}(x_1,x_2,\ldots,x_m) = \prod_{j=1}^{m} F_{X_{i_j}}(x_j)$$

für alle $x_1,\ldots,x_m \in \mathbb{R}$, wobei $F_{X_{i,j}}$ die Verteilungsfunktion von X_{i_j} gemäß Definition 6.6 bezeichnet. □

Unabhängigkeit überträgt sich bei der Verkettung von Zufallsvariablen mit weiteren Funktionen, wie der folgende Satz zeigt.

Satz 8.5 Seien $X_i : \Omega \to \mathbb{R}$, $i \in J \subset \mathbb{N}$, stochastisch unabhängige Zufallsvariablen über dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ und $g : \mathbb{R} \to \mathbb{R}$ eine stetige Funktion. Dann sind $g(X_i) : \Omega \to \mathbb{R}$, $i \in J \subset \mathbb{N}$, ebenfalls wieder stochastisch unabhängige Zufallsvariablen.

Beweisskizze: Wir betrachten die Mengen $B_j := \{x \in X_{i_j}(\Omega) \,|\, g(x) \leq x_j\}$. Diese Mengen sind Borel Mengen, liegen also in $\mathcal{B}$ (was wir hier nicht beweisen). Damit gilt nach Bemerkung 8.2

$$\begin{aligned}\mathbb{P}\left(\bigcap_{j=1}^{m}\{g(X_{i_j} \leq x_j\}\right) &= \mathbb{P}\left(\bigcap_{j=1}^{m}\{X_{i_j} \in B_j\}\right)\\ &= \prod_{j=1}^{m}\mathbb{P}(\{X_{i_j} \in B_j\}) = \prod_{j=1}^{m}\mathbb{P}(\{g(X_{i_j} \leq x_j\}),\end{aligned}$$

was gerade die Bedingung der Unabhängigkeit (8.1) zeigt. □

8.2 Erwartungswert und Varianz

Der folgende Satz zeigt die Konsequenz der Unabhängigkeit für Erwartungswert und Varianz.

Satz 8.6 Seien $X_i : \Omega \to \mathbb{R}$, $i \in J \subset \mathbb{N}$, stochastisch unabhängige diskrete Zufallsvariablen über dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$.

(a) Falls die Erwartungswerte $\mathbb{E}(X_i)$ existieren für alle $i \in J$, so gilt für jede Auswahl von $m \in \mathbb{N}$ verschiedenen Zufallsvariablen $X_{i_1},\ldots,X_{i_m}$ mit $i_1,\ldots,i_m \in J$, dass

$$\mathbb{E}[X_{i_1} \cdot X_{i_2} \cdots X_{i_m}] = \mathbb{E}[X_{i_1}] \cdot \mathbb{E}[X_{i_2}] \cdots \mathbb{E}[X_{i_m}].$$

(b) Falls die Erwartungswerte $\mathbb{E}[X_i^2]$ existieren für alle $i \in J$, so gilt für jede Auswahl von $m \in \mathbb{N}$ verschiedenen Zufallsvariablen $X_{i_1}, \ldots, X_{i_m}$ mit $i_1, \ldots, i_m \in J$ sowie Zahlen $a_1, \ldots, a_m$, dass

$$\mathbb{V}[a_1 X_{i_1} + a_2 X_{i_2} + \ldots + a_m X_{i_m}] = a_1^2 \mathbb{V}[X_{i_1}] + a_2^2 \mathbb{V}[X_{i_2}] + \ldots + a_m^2 \mathbb{V}[X_{i_m}].$$

Beweis (a) Mit der Definition 6.8 des Erwartungswerts und (8.1) gilt

$$\begin{aligned}
\mathbb{E}[X_{i_1} \cdot X_{i_2} \cdots X_{i_m}] &= \sum_{k_1 \in I_{X_{i_1}}} \cdots \sum_{k_m \in I_{X_{i_m}}} x_{k_1} \cdots x_{k_m} \mathbb{P}\left(\bigcap_{j=1}^{m} \{X_{i_j} = x_{k_j}\} \right) \\
&= \sum_{k_1 \in I_{X_{i_1}}} \cdots \sum_{k_m \in I_{X_{i_m}}} x_{k_1} \cdots x_{k_m} \prod_{j=1}^{m} \mathbb{P}(\{X_{i_j} = x_{k_j}) \\
&= \left(\sum_{k_1 \in I_{X_{i_1}}} x_{k_1} \mathbb{P}(\{X_{i_1} = x_{k_1}) \right) \cdots \left(\sum_{k_m \in I_{X_{i_m}}} x_{k_m} \mathbb{P}(\{X_{i_m} = x_{k_m}) \right) \\
&= E[X_{i_1}] \cdot \mathbb{E}[X_{i_2}] \cdots \mathbb{E}[X_{i_m}].
\end{aligned}$$

(b) Nach Satz 6.13 gilt

$$\mathbb{V}[a_1 X_{i_1} + a_2 X_{i_2} + \ldots + a_m X_{i_m}] = \mathbb{E}\left[(a_1 X_{i_1} + \ldots + a_m X_{i_m})^2\right] - \left(\mathbb{E}[a_1 X_{i_1} + \ldots + a_m X_{i_m}]\right)^2.$$

Ausmultiplizieren und Anwenden von 6.9 liefert

$$\begin{aligned}
&\mathbb{V}[a_1 X_{i_1} + a_2 X_{i_2} + \ldots + a_m X_{i_m}] \\
&= \sum_{j=1}^{m} a_j^2 \mathbb{E}[X_{i_j}^2] - a_j^2 (\mathbb{E}[X_{i_j}])^2 + \sum_{k \neq l} (a_k a_l \underbrace{\mathbb{E}[X_{i_k} X_{i_l}]}_{\substack{= \mathbb{E}[X_{i_k}] \mathbb{E}[X_{i_l}]) \\ \text{wegen Unabhängigkeit}}} - a_k a_l \mathbb{E}[X_{i_k}] \mathbb{E}[X_{i_l}]) \\
&= \sum_{j=1}^{m} a_j^2 \mathbb{E}[X_{i_j}^2] - a_j^2 (\mathbb{E}[X_{i_j}])^2 \;=\; a_1^2 \mathbb{V}[X_{i_1}] + a_2^2 \mathbb{V}[X_{i_2}] + \ldots + a_m^2 \mathbb{V}[X_{i_m}].
\end{aligned}$$

Beispiel 8.7 Wir setzen Beispiel 8.3 fort. Es gilt $\mathbb{E}[X] = \mathbb{E}[Y] = \mathbb{E}[Z] = 1/2$.

$$XY(\omega) := X(\omega)Y(\omega) = \begin{cases} 0, & \omega \in \{1,2,3\} \\ 1, & \omega \in \{4\}, \end{cases} \qquad XZ(\omega) := X(\omega)Z(\omega) = 0 \text{ für alle } \omega \in \Omega.$$

Damit gilt $\mathbb{E}[XY] = 1/4$ und $\mathbb{E}[X]\mathbb{E}[Y] = 1/2 \cdot 1/2 = 1/4$. Die Werte stimmen also wie erwartet überein, weil X und Y ja nach Beispiel 8.3 stochastisch unabhängig sind.

Für X und Z gilt hingegen $\mathbb{E}[XZ] = 0$ und $\mathbb{E}[X]\mathbb{E}[Z] = 1/2 \cdot 1/2 = 1/4$. Die Werte stimmen also nicht überein, was wegen der fehlenden stochastischen Unabhängigkeit auch nicht zu erwarten war. □

Achtung: Aus $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ folgt nicht unbedingt stochastische Unabhängigkeit von X und Y, die Umkehrung von Satz 8.6(a) gilt also i.A. nicht. Ein Beispiel sind die Zufallsvariablen $X : \Omega \to \mathbb{R}$, die die drei Werte -1, 0 und 1 jeweils mit Wahrscheinlichkeit $1/3$ annimmt und die Zufallsvariable $Y = X^2$. Für diese gilt $\mathbb{E}[X] = 0$, $\mathbb{E}[Y] = 2/3$ und wegen $XY = X$ auch $\mathbb{E}(XY) = 0$. Damit folgt $\mathbb{E}[XY] = 0 = \mathbb{E}[X]\mathbb{E}[Y]$. Trotzdem sind X und Y nicht stochastisch unabhängig, denn

$$\mathbb{P}(\{X \leq -1\}) = 1/3, \;\; \mathbb{P}(\{Y \leq 0.5\}) = 1/3 \;\; \text{aber} \;\; \mathbb{P}(\{X \leq -1\} \cap \{Y \leq 0.5\}) = 0,$$

weil aus $X \leq -1$ ja $Y \geq 1$ folgt, das Ereignis $\{X \leq -1\} \cap \{Y \leq 0.5\}$ also nie eintreten kann und daher die Wahrscheinlichkeit 0 besitzt.

8.3 Kovarianz

Auch wenn die Gleichung $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ nicht zur Überprüfung der stochastischen Unabhängigkeit verwendet werden kann, so kann man aus der Tatsache, wie "gut" die Gleichung erfüllt ist, doch wichtige Informationen über den Zusammenhang von X und Y erhalten. Dies ist die Grundlage für die folgende Größe, wie Satz 8.9 im Anschluss zeigt.

Definition 8.8 Seien $X, Y : \Omega \to \mathbb{R}$ diskrete Zufallsvariablen über dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$, für die die Erwartungswerte $\mathbb{E}[X]$, $\mathbb{E}[Y]$ und $\mathbb{E}[XY]$ existieren. Dann definieren wir die *Kovarianz* von X und Y als

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

□

Aus der Definition folgt sofort $\mathrm{Cov}(X, X) = \mathbb{V}[X]$. Der Begriff der Kovarianz erweitert die Varianz also auf zwei Zufallsvariablen. Der folgende Satz zeigt — als Analogon zu Satz 6.13, den Zusammenhang der Kovarianz und der Gleichung $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.

Satz 8.9 Seien $X, Y : \Omega \to \mathbb{R}$ diskrete Zufallsvariablen über dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$, für die die Erwartungswerte $\mathbb{E}[X]$, $\mathbb{E}[Y]$ und $\mathbb{E}[XY]$ existieren. Dann gilt

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

Beweis: Es gilt

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

□

Die Interpretation der Kovarianz ist wie folgt:

- Ist $\mathrm{Cov}(X,Y) > 0$, so nimmt X mit hoher Wahrscheinlichkeit genau dann hohe Werte (also Werte größer als $\mathbb{E}[X]$) an, wenn Y hohe Werte annimmt und umgekehrt. X und Y sind also tendenziell gleichläufig.
- Ist $\mathrm{Cov}(X,Y) < 0$ so verhalten sich X und Y tendenziell gegenläufig: X nimmt mit hoher Wahrscheinlichkeit hohe Werte an, wenn Y niedrige Werte annimmt und umgekehrt.
- Ist $\mathrm{Cov}(X,Y) \approx 0$, so kann man von den Werten von X nur mit geringer Wahrscheinlichkeit auf die Werte von Y schließen und umgekehrt.
- Falls X und Y stochastisch unabhängig sind, so folgt $\mathrm{Cov}(X,Y) = 0$.

Beispiel 8.10 Wir betrachten X und Z aus Beispiel 8.3. Wie in Beispiel 8.7 bereits berechnet, gilt $\mathbb{E}[X] = \mathbb{E}[Z] = 1/2$ und $\mathbb{E}[XZ] = 0$. Also folgt $\mathrm{Cov}(X,Z) = -1/4$. Die Zufallsvariablen X und Y sollten sich also tendenziell gegenläufig verhalten, d.h. X nimmt hohe Werte an wenn Z niedrige Werte annimmt und umgekehrt. Ein Blick auf die Definition von X und Z in Beispiel 8.3 zeigt, dass genau dies der Fall ist. □

Grafisch kann man die Kovarianz anschaulich darstellen, wenn man die Realisierungen $(X(\omega), Y(\omega)) \in \mathbb{R}^2$ als Punkte in ein zweidimensionales Koordinatensystem einträgt. Abbildung 8.1 zeigt diese Darstellung für zwei Zufallsvariablen mit positiver Kovarianz (links), Kovarianz gleich Null (Mitte) und negativer Kovarianz (rechts).

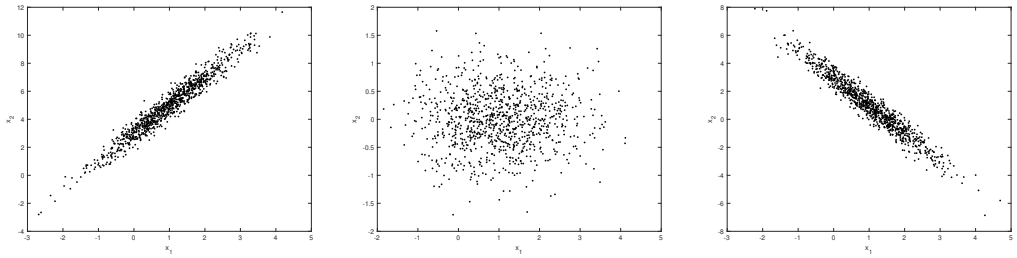


Abbildung 8.1: Realisierungen von zwei Zufallsvariablen mit positiver Kovarianz (links), Kovarianz gleich Null (Mitte) und negativer Kovarianz (rechts), dargestellt als Punkte in der zweidimensionalen Ebene

Betrachtet man mehr als zwei Zufallsvariablen, also $X_1, \ldots, X_n$ mit $n \geq 3$, so kann man die Kovarianzen $\mathrm{Cov}(X_i, X_j)$ paarweise berechnen. Für $i = j$ ergibt sich dann gerade die Varianz von X_i. Man erhält so n^2 Werte, die man in eine Kovarianzmatrix

$$\mathrm{Cov}(X) = \begin{pmatrix} \mathrm{Cov}(X_1,X_1) & \mathrm{Cov}(X_1,X_2) & \ldots & \mathrm{Cov}(X_1,X_n) \\ \mathrm{Cov}(X_2,X_1) & \mathrm{Cov}(X_2,X_2) & \ldots & \mathrm{Cov}(X_2,X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n,X_1) & \mathrm{Cov}(X_n,X_2) & \ldots & \mathrm{Cov}(X_n,X_n) \end{pmatrix}, \qquad X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \tag{8.2}$$

schreiben kann. Dabei haben wir $X_1, \ldots, X_n$ der leichteren Notation wegen zu einem Vektor X von Zufallsvariablen, einem sogenannten *Zufallsvektor* zusammengefasst. Beachte, dass auf der Diagonalen von $\mathrm{Cov}(X)$ gerade die Varianzen von X_1 bis X_n stehen.

Bemerkung 8.11 Ein verwandtes Konzept zur Kovarianz ist das der *Korrelation*, die definiert ist als

$$\mathrm{Corr}(X,Y) := \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathbb{V}[X]\mathbb{V}[Y]}}.$$

Die Korrelation ist eine normierte Version der Kovarianz, die nur Werte zwischen -1 und 1 annehmen kann. Sie nimmt den Wert 1 an, wenn X und Y genau gleichläufig sind und -1, wenn X und Y genau gegenläufig sind (wie dies gerade für X und Z in Beispiel 8.3 der Fall ist). Die Korrelation ist also leichter zu interpretieren, kann aber dazu führen, dass eine betragsmäßig sehr kleine (und damit praktisch irrelevante) Kovarianz $\mathrm{Cov}(X,Y) \approx 0$ überbewertet wird. □

8.4 Das schwache Gesetz der großen Zahlen

Wir haben den Erwartungswert als Grenzwert des empirischen (= durch n Zufallsexperimente bestimmten) Mittelwerts einer Zufallsvariable für $n \to \infty$ definiert. Mit Hilfe der Varianz können wir nun berechnen, wie nahe der empirische Mittelwert nach n Durchführungen des Zufallsexperiments am Erwartungswert liegt. Natürlich kann man hier keine festen, vom Zufall unabhängigen Schranken angeben. Es ist ja theoretisch auch bei einem idealen, nicht gefälschten Würfel möglich, 1000 Sechser nacheinander zu würfeln, wodurch wir den Mittelwert 6 erhalten, der weit vom Erwartungswert 3,5 entfernt liegt. Dass dies passiert, ist aber extrem wenig wahrscheinlich und genau diese Wahrscheinlichkeit wird durch den folgenden Satz—das sogenannte *schwache Gesetz der großen Zahlen*—abgeschätzt.

Satz 8.12 Seien $X_i : \Omega \to \mathbb{R}$, $i \in \mathbb{N}$, stochastisch unabhängige diskrete Zufallsvariablen über dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ mit Erwartungswert $\mu = \mathbb{E}[X_i]$ und Varianzen $\sigma_i^2 = \mathbb{V}[X_i] \leq M$ für ein $M \in \mathbb{R}$. Setzen wir

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

so gilt für alle $\varepsilon > 0$, dass

$$\mathbb{P}(\{|\overline{X}_n - \mu| \geq \varepsilon\}) \leq \frac{M}{\varepsilon^2 n} \to 0 \text{ für } n \to \infty.$$

Beweis: Wegen Satz 6.9 gilt

$$\mathbb{E}[\overline{X}_n] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu.$$

Mit Satz 8.6(b) gilt zudem

$$\mathbb{V}[\overline{X}_n] = \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2 \leq \frac{1}{n^2}\sum_{i=1}^{n} M = \frac{M}{n}.$$

Damit folgt die Behauptung aus der Tschebyscheff'schen Ungleichung (6.5) angewendet auf $X = \overline{X}_n$. □

Die Bedingung $|\overline{X}_n - \mu| \geq \varepsilon$ kann äquivalent auch als

$$\overline{X}_n \notin [\mu - \varepsilon, \mu + \varepsilon] \text{ mit Gegenereignis } \overline{X}_n \in [\mu - \varepsilon, \mu + \varepsilon]$$

geschrieben werden. Das Gesetz der großen Zahlen gibt also eine Schranke für die Wahrscheinlichkeit an, dass der empirische Mittelwert im Intervall $[\mu - \varepsilon, \mu + \varepsilon]$ mit Breite 2ε liegt. Beachte den Zusammenhang zwischen n und ε. Wenn wir die Intervallbreite ε halbieren möchten, die Schranke $\frac{M}{\varepsilon^2 n}$ an die Wahrscheinlichkeit dabei nicht verändern wollen, so müssen wir n vervierfachen, weil $(\varepsilon/2)^2(4n) = (\varepsilon^2/4)4n = \varepsilon^2 n$. Um die Abweichung von $\overline{X}_n$ vom Erwartungswert μ zu halbieren, müssen wir den Stichprobenumfang n also vervierfachen.

Beispiel 8.13 Beim Würfeln mit einem Würfel beträgt der Erwartungswert der Augensumme

$$\mu = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = \frac{1}{6}21 = 3{,}5$$

und die Varianz

$$\sigma^2 = \frac{1}{6}(2{,}5^2 + 1{,}5^2 + 0{,}5^2 + 0{,}5^2 + 1{,}5^2 + 2{,}5^2) = \frac{1}{6}17{,}5 = 2{,}91\bar{6}.$$

Nach 10000 Würfen erfüllt die Wahrscheinlichkeit, dass der Mittelwert um mehr als 0,1 vom Erwartungswert abweicht

$$\mathbb{P}(\{|\overline{X}_n - \mu| \geq 0{,}1\}) \leq \frac{2{,}91\bar{6}}{0{,}1^2 10000} = 0{,}02916,$$

ist also kleiner oder gleich knapp 3%. □

Das schwache Gesetz der großen Zahlen zeigt also, dass der Erwartungswert durch n Realisierungen von X_i geschätzt[1] werden kann. Haben wir also n Realisierungen $x_i = X_i(\omega)$, $i = 1, \ldots n$ durch Daten aus Experimenten gegeben, so gilt

$$\frac{1}{n}\sum_{i=1}^{n} x_i \approx \mathbb{E}[X_1] \tag{8.3}$$

mit hoher Wahrscheinlichkeit.

Da auch die Varianz und die Kovarianz ja nur spezielle Erwartungswerte sind, können diese analog durch Realisierungen der Zufallsvariablen

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])^2 \quad \text{und} \quad \frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])(Y_i - \mathbb{E}[Y_i])$$

geschätzt werden, falls $\mathbb{V}[X_i]$ bzw. $\mathrm{Cov}(X_i, Y_i)$ nicht von i abhängen. Haben wir n Realisierungen $x_i = X_i(\omega)$ und ggf. $y_i = Y_i(\omega)$ aus experimentellen Daten, so gilt also

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \mathbb{E}[X_i])^2 \approx \mathbb{V}[X_1] \quad \text{und} \quad \frac{1}{n}\sum_{i=1}^{n}(x_i - \mathbb{E}[X_i])(y_i - \mathbb{E}[Y_i]) \approx \mathrm{Cov}(X_i, Y_i). \tag{8.4}$$

Da dabei $\mathbb{E}[X_i]$ und $\mathbb{E}[Y_i]$ i.A. unbekannt sind, müssen diese vorher mittels (8.3) geschätzt werden. Mit der zweiten Formel kann man insbesondere die Einträge der Kovarianzmatrix (8.2) schätzen. Wir werden später in Beispiel 10.1 sehen, dass man das mit etwas linearer Algebra auch kompakter darstellen kann.

[1] "Geschätzt" bedeutet hier, dass mit hoher Wahrscheinlichkeit ein guter Näherungswert erzielt wird.

8.5 Die Monte-Carlo Simulation

Das schwache Gesetz der großen Zahlen ist die Grundlage für einen beliebten Simulationsalgorithmus zur näherungsweisen Berechnung von Erwartungswerten. Nehmen wir an, wir haben einen Algorithmus, mit dem wir (schnell) sehr viele Zufallsexperimente durchführen können. Dann können wir viele $X_i(\omega)$ realisieren und so eine Realisierung von $\overline{X}_n(\omega)$ berechnen. Nach dem schwachen Gesetz der großen Zahlen gilt dann für große n mit hoher Wahrscheinlichkeit, dass

$$\overline{X}_n(\omega) \approx \mu.$$

Dieses Vorgehen nennt man *Monte-Carlo Simulation*. Abbildung 8.2 zeigt den Verlauf zweier solcher Simulationen mit zunehmender Anzahl von Zufallsexperimenten n auf der x-Achse und $\overline{X}_n(\omega)$ auf der y-Achse. Die roten Punkte gehören zu einer Zufallsvariablen mit geringerer Varianz, die blauen Punkte zu einer Zufallsvariablen mit höherer Varianz. Beide haben den gleichen Erwartungswert μ, der als gestrichelte Linie eingezeichnet ist.

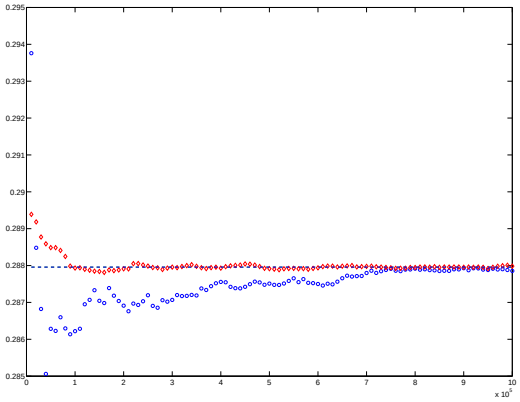


Abbildung 8.2: Beispiele für Monte-Carlo Simulationen

Ebenso wie Erwartungswerte kann man mit den Formeln (8.4) natürlich auch Varianzen und Kovarianzen mit simulierten Zufallsexperimenten näherungsweise berechnen.

Kapitel 9

Stetige Zufallsvariablen und Verteilungen

9.1 Dichtefunktionen

Für diskrete Zufallsvariablen ist die Verteilungsfunktion

$$F(x) = \mathbb{P}(\{X \leq x\})$$

eine Treppenfunktion, d.h. sie ist konstant bis auf eine diskrete Menge von Sprüngen. Für stetige Zufallsvariablen ist diese Funktion stetig, d.h. sie hat überhaupt keine Sprünge. Abbildung 9.1 zeigt den Unterschied dieser beiden Fälle.

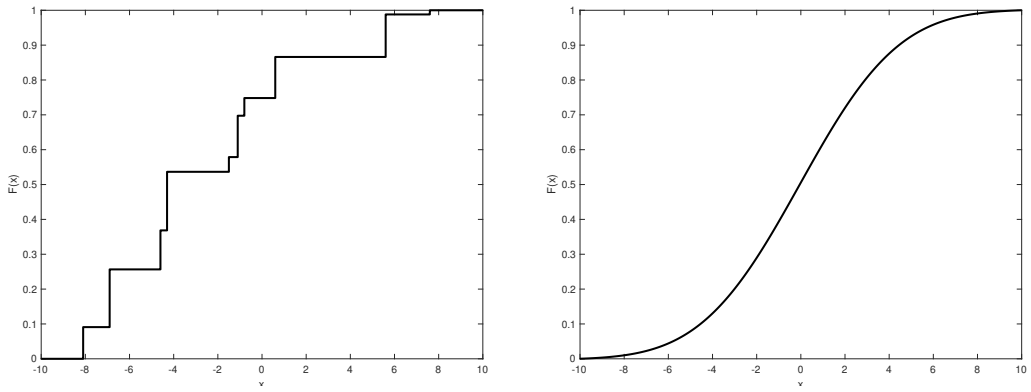


Abbildung 9.1: Beispiele für Verteilungfunktionen für diskrete (links) und stetige (rechte) Zufallsvariablen

Tatsächlich verlangen wir noch etwas mehr[1], nämlich dass sich F als ein Integral über eine integrierbare Funktion f, die sogenannte *Dichtefunktion* schreiben lässt.

[1] Jede Funktion, die als Integral geschrieben werden kann, ist stetig, aber nicht jede stetige Funktion kann als Integral geschrieben werden.

Definition 9.1 Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $X : \Omega \to \mathbb{R}$ eine Zufallsvariable. Dann heißt X *stetige Zufallsvariable*, wenn eine integrierbare Funktion $f : \mathbb{R} \to [0, \infty)$ existiert mit $\int_{-\infty}^{\infty} f(x)dx = 1$ und

$$F(y) = \int_{-\infty}^{y} f(x)dx.$$

Die Funktion f heißt dann *Dichtefunktion* von X. □

Weil F stetig ist, gilt für $a \leq b$

$$\begin{aligned}\mathbb{P}(\{a \leq X \leq b\}) &= \mathbb{P}(\{X \leq b\}) - \mathbb{P}(\{X < a\}) = \mathbb{P}(\{X \leq b\}) - \lim_{\hat{a} \nearrow a} \mathbb{P}(\{X \leq \hat{a}\}) \\ &= F(b) - \lim_{\hat{a} \nearrow a} F(\hat{a}) \;=\; F(b) - F(a).\end{aligned}$$

Daraus folgt

$$\mathbb{P}(\{a \leq X \leq b\}) = F(b) - F(a) = \int_{-\infty}^{b} f(x)dx - \int_{-\infty}^{a} f(x)dx = \int_{a}^{b} f(x)dx.$$

Im Spezialfall $a = b$ ergibt sich so

$$\mathbb{P}(\{X = a\}) = \mathbb{P}(\{a \leq X \leq a\}) = \int_{a}^{a} f(x)dx = 0.$$

Für eine stetige Zufallsvariable X hat das Ereignis "X nimmt den Wert a an" also immer die Wahrscheinlichkeit 0, egal welchen Wert a annimmt. Das klingt erst einmal paradox, denn man könnte meinen, dass dann alle Ereignisse die Wahrscheinlichkeit 0 besitzen. Dies ist aber nicht so. Das Paradoxon ergibt sich aus der Tatsache, dass es in jedem Teilintervall $[a, b] \subset \mathbb{R}$ mit $a < b$ überabzählbar viele Zahlen $x \in [a, b]$ gibt. Auch wenn die Wahrscheinlichkeit $\mathbb{P}(\{X = x\})$ für jedes einzelne $x \in [a, b]$ gleich 0 ist, kann die Wahrscheinlichkeit $\mathbb{P}(\{X \in [a, b]\}) = \mathbb{P}(\{a \leq X \leq a\})$ deswegen trotzdem positiv sein.

Falls die Dichtefunktion f in einem Punkt $x \in \mathbb{R}$ stetig ist, so ist F als Integral über eine stetige Funktion in x differenzierbar und nach dem Hauptsatz der Differential- und Integralrechnung gilt

$$F'(x) = f(x).$$

Bemerkung 9.2 (i) Betrachtet man die bedingte Wahrscheinlichkeit (vgl. Definition 4.2), dass eine Zufallsvariable kleiner als x ist, also die Größe $\mathbb{P}(\{X \leq y\} \,|\, B)$ für ein beliebiges $B \in \mathcal{A}$, so kann man auch für diese eine Verteilungsfunktion und ggf. eine Dichtefunktion definieren. Wir schreiben diese als

$$\mathbb{P}(\{X \leq y\} \,|\, B) = F(y \,|\, B) = \int_{-\infty}^{y} f(x \,|\, B)dx.$$

(ii) Für unabhängige stetige Zufallsvariablen $X_1, \ldots, X_n$ folgt aus Bemerkung 8.4, dass für jede Auswahl von m Zufallsvariablen $X_{i_1}, \ldots, X_{i_m}$ die Dichtefunktionen $f_{X_{i_j}}$, $j = 1, \ldots, m$ und die gemeinsame Dichtefunktion $f_{X_{i_1}, \ldots, X_{i_m}}$ die Gleichung

$$f_{X_{i_1}, \ldots, X_{i_m}}(x_1, \ldots, x_m) = \prod_{j=1}^{m} f_{X_{i_j}}(x_j)$$

erfüllen. □

9.2 Erwartungswert und Varianz

Wir wollen nun einige Konzepte von den diskreten auf die stetigen Zufallsvariablen übertragen. Die Faustregel dabei lautet, dass die mit p_i gewichtete Summe durch das mit $f(x)$ gewichtete Integral ersetzt wird, was wir nach den folgenden Definitionen begründen.

Definition 9.3 Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $X : \Omega \to \mathbb{R}$ eine stetige Zufallsvariable, für die das Integral

$$\int_{-\infty}^{\infty} |x| f(x) dx$$

existiert. Dann heißt

$$\mathbb{E}[X] := \int_{-\infty}^{\infty} x f(x) dx$$

der *Erwartungswert* von X. Ansonsten sagen wir, dass der Erwartungswert von X nicht existiert.

Allgemein definieren wir für eine stetige Funktion $g : \mathbb{R} \to \mathbb{R}$ den Erwartungswert von $g(X)$ als

$$\mathbb{E}[g(X)] := \int_{-\infty}^{\infty} g(x) f(x) dx,$$

wiederum unter der Bedingung, dass das Integral $\int_{-\infty}^{\infty} |g(x)| f(x) dx$ existert. □

Die Tatsache, dass aus der Summe $\mathbb{E}[X] = \sum_{i \in I_X} x_i p_i$ nun ein Integral wird, lässt sich wie folgt begründen:

Wir teilen die reelle Achse in nur am Rand überlappende Intervalle $J_i^\varepsilon = [a_i^\varepsilon, a_i^\varepsilon + \varepsilon]$ der Breite $\varepsilon > 0$ auf. Dann beträgt die Wahrscheinlichkeit, dass eine Realisierung von X in das Intervall J_i^ε fällt, gerade

$$\mathbb{P}(\{X \in J_i^\varepsilon\}) = \int_{a_i^\varepsilon}^{a_i^\varepsilon + \varepsilon} f(x) dx.$$

Betrachten wir nun die diskrete Zufallsvariable

$$X^\varepsilon(\omega) = a_i^\varepsilon \text{ falls } X(\omega) \in [a_i, \varepsilon, a_i^\varepsilon + \varepsilon).$$

Für diese gilt offenbar $0 \leq X(\omega) - X^\varepsilon(\omega) \leq \varepsilon$ für alle $\omega \in \Omega$, also konvergiert X^ε gegen X für $\varepsilon \to 0$. Eine sinnvolle Definition von $\mathbb{E}[X]$ sollte also $\mathbb{E}[X^\varepsilon] \to \mathbb{E}[X]$ erfüllen für $\varepsilon \to 0$.

Dies gilt genau für die Definition von $\mathbb{E}[X]$ aus Definition 9.3, denn für p_i gilt

$$p_i = \mathbb{P}(\{X^\varepsilon = a_i^\varepsilon\}) = \mathbb{P}(\{X \in [a_i^\varepsilon, a_i^\varepsilon + \varepsilon)\}) = \mathbb{P}(\{X \in J_i^\varepsilon\}) = \int_{a_i^\varepsilon}^{a_i^\varepsilon + \varepsilon} f(x) dx$$

und damit

$$\mathbb{E}[X^\varepsilon] = \sum_{i \in I_X} x_i p_i = \sum_{i \in I_X} a_i^\varepsilon \int_{a_i^\varepsilon}^{a_i^\varepsilon + \varepsilon} f(x) dx = \sum_{i \in I_X} \int_{a_i^\varepsilon}^{a_i^\varepsilon + \varepsilon} a_i^\varepsilon f(x) dx.$$

Also folgt

$$0 \leq \mathbb{E}[X] - \mathbb{E}[X^\varepsilon] = \underbrace{\int_{-\infty}^{\infty} x f(x)dx}_{= \sum_{i \in I_X} \int_{a_i^\varepsilon}^{a_i^\varepsilon + \varepsilon} x f(x)dx} - \sum_{i \in I_X} \int_{a_i^\varepsilon}^{a_i^\varepsilon + \varepsilon} a_i^\varepsilon f(x)dx = \sum_{i \in I_X} \int_{a_i^\varepsilon}^{a_i^\varepsilon + \varepsilon} \underbrace{x - a_i^\varepsilon}_{\leq \varepsilon} f(x)dx$$

$$\leq \sum_{i \in I_X} \int_{a_i^\varepsilon}^{a_i^\varepsilon + \varepsilon} \varepsilon f(x)dx = \varepsilon \int_{-\infty}^{\infty} f(x)dx = \varepsilon \to 0$$

für $\varepsilon \to 0$ und folglich $\mathbb{E}[X^\varepsilon] \to \mathbb{E}[X]$ für $\varepsilon \to 0$. Die Integralformel für $\mathbb{E}[X]$ aus Definition 9.3 geht also für $\varepsilon \to 0$ aus der Summenformel für $\mathbb{E}[X^\varepsilon]$ hervor.

Mit dieser Definition des Erwartungswerts können wir die Definitionen der Varianz und Kovarianz

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

und

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

direkt auf stetige Zufallsvariablen erweitern. Alle bisher gemachten Aussagen über diskrete Zufallsvariablen, wie z.B. die Tschebyscheff-Ungleichung oder das schwache Gesetz der großen Zahlen, gelten damit analog für stetige Zufallsvariablen (allerdings müssen die gegebenen Beweise eventuell an die neue Definition angepasst werden).

Wir geben im Rest dieses Kapitels einige wichtige stetige Verteilungen an.

9.3 Gleichverteilung

Bei der *Gleichverteilung* sind alle Werte von X in einem Intervall $[a, b]$, $a < b$ gleichwahrscheinlich. Das bedeutet, dass $\mathbb{P}(\{X \in [a, b]\}) = 1$ und, für $a \leq c < d \leq b$, dass $\mathbb{P}(\{X \in [c, d]\}) = \frac{d-c}{b-a}$ ist. Dies ist genau dann der Fall, wenn die Dichtefunktion durch

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \in [a, b] \\ 0, & \text{sonst} \end{cases} \tag{9.1}$$

gegeben ist. Die Verteilungsfunktion lautet

$$F(y) = \int_{-\infty}^{y} f(x)dx = \begin{cases} \int_{-\infty}^{y} 0\,dx & = 0, \quad y < a \\ \int_{-\infty}^{a} 0\,dx + \int_{a}^{y} \frac{1}{b-a} dy & = \frac{x-a}{b-a}, \quad y \in [a, b] \\ \int_{-\infty}^{a} 0\,dx + \int_{a}^{b} \frac{1}{b-a} dy + \int_{b}^{y} 0\,dx = \frac{b-a}{b-a} & = 1, \quad y > b \end{cases}$$

Satz 9.4 Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $X : \Omega \to \mathbb{R}$ eine gleichverteilte Zufallsvariable. Dann gilt

$$\mathbb{E}[X] = \frac{a+b}{2} \quad \text{und} \quad \mathbb{V}[X] = \frac{(b-a)^2}{12}.$$

Beweis: Nach Definition 9.3 und (9.1) gilt

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xf(x)dx = \int_a^b \frac{x}{b-a}dx = \frac{x^2}{2(b-a)}\Big|_a^b = \frac{\overbrace{b^2-a^2}^{=(b+a)(b-a)}}{2(b-a)} = \frac{a+b}{2}.$$

Für X^2 gilt analog

$$\mathbb{E}[X^2] = \int_{-\infty}^{\infty} x^2 f(x)dx = \int_a^b \frac{x^2}{b-a}dx = \frac{x^3}{3(b-a)}\Big|_a^b = \frac{\overbrace{b^3-a^3}^{=(a^2+ab+b^2)(b-a)}}{3(b-a)} = \frac{a^2+ab+b^2}{3}.$$

Damit folgt für die Varianz

$$\mathbb{V}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{a^2+ab+b^2}{3} - \frac{(a+b)^2}{2^2} = \frac{a^2-2ab+b^2}{12} = \frac{(a-b)^2}{12}.$$

□

9.4 Exponentialverteilung

Die Exponentialverteilung mit Parameter $\lambda > 0$ ist beschrieben durch die Dichtefunktion

$$f(x) = \begin{cases} 0, & x \leq 0 \\ \lambda \exp(-\lambda x), & x > 0 \end{cases}$$

Berechnung von ergibt die Verteilungsfunktion

$$F(x) = \begin{cases} 0, & x \leq 0 \\ 1 - \exp(-\lambda x), & x > 0 \end{cases}$$

Für eine exponentialverteilte Zufallsvariable X schreiben wir $X \sim \text{Exp}(\lambda)$.

Die Exponentialverteilung hängt eng mit der Poisson-Verteilung aus Abschnitt 7.2 zusammen. Zur Erinnerung: die Poisson-Verteilung beschreibt die Anzahl X_t von zufälligen Ereignissen, die bis zu einer gegebenen Zeit t eintreten, mit

$$\mathbb{P}(\{X_t = k\}) = \frac{(\mu t)^k}{k!} \exp(-\mu t)$$

für alle $k \in \mathbb{N}$. Wir führen nun zusätzlich zu X_t die Zufallsvariable

$$T = \text{“Zeitabstand zwischen zwei zufälligen Ereignissen”}$$

ein. Der folgende Satz zeigt, wie die beiden Zufallsvariablen zusammenhängen.

Satz 9.5 Sei $X_t \sim \pi_{\mu t}$ eine Poisson-verteilte Zufallsvariable, die die Anzahl zufälliger Ereignisse angibt, die im Intervall $[0, t]$ eingetreten sind. Dann ist die zufällige Zeit T zwischen zwei Ereignissen exponentialverteilt mit $T \sim \text{Exp}(\mu)$.

Beweis: Weil die Anzahl der zufälligen Ereignisse in einem Teilintervall unabhängig von der Anzahl auf einem vorhergehenden Teilintervall ist, ist die Zeit zwischen zwei Ereignissen genau so verteilt wie die Wartezeit ab dem Zeitpunkt $t = 0$ bis zum ersten Ereignis. Wir können zum Beweis also die Verteilung dieser Wartezeit berechnen, die wir wieder mit T bezeichnen. Für diese Wartezeit gilt dann

$$F(t) = \mathbb{P}(\{T \leq t\}) = 1 - \mathbb{P}(\{T \geq t\}) = 1 - \mathbb{P}(\{X_t = 0\}) = 1 - \exp(-\mu t), \tag{9.2}$$

woraus die Behauptung folgt. □

Beispiel: In einem Netzwerkknoten, in dem die eingehenden Verbindungsanfragen Poisson-verteilt sind, gehen pro Stunde im Mittel 20 Anfragen ein. Wie groß ist die Wahrscheinlichkeit, dass zwischen zwei Anfragen 3 bis 6 Minuten Zeit liegen?

Zur Beantwortung berechnen wir zunächst das μ für die Poisson-Verteilung, wobei wir die Einheit der Zeit t in Minuten angeben. Dann muss gemäß Abschnitt 7.2 gelten

$$20 = \mathbb{E}[X_{60}] = \mu \cdot 60,$$

woraus $\mu = 1/3$ folgt. Damit erhalten wir

$$\mathbb{P}(\{3 \leq T \leq 6\}) = F(6) - F(3) = (1 - e^{-\mu 6}) - (1 - e^{-\mu 3}) = -e^{-2} + e^{-1} \approx 0{,}2325.$$

Satz 9.6 Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $X : \Omega \to \mathbb{R}$ eine exponentialverteilte Zufallsvariable mit Parameter λ. Dann gilt

$$\mathbb{E}[X] = \frac{1}{\lambda} \quad \text{und} \quad \mathbb{V}[X] = \frac{1}{\lambda^2}.$$

Beweis: Weil die Dichte für $x < 0$ gleich 0 ist, gilt mit $t = \lambda x$ und partieller Integration

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^\infty x\lambda \exp(-\lambda x)dx \;=\; \frac{1}{\lambda}\int_0^\infty t\exp(-t)dt \\
&= \frac{1}{\lambda}\left(t(-\exp(-t))\Big|_0^\infty - \int_0^\infty 1(-\exp(-t))dt\right) \\
&= \frac{1}{\lambda}\left(t(-\exp(-t))\Big|_0^\infty - \exp(-t)\Big|_0^\infty\right) \\
&= \frac{1}{\lambda}(0 - 0 + 0 - (-1)) \;=\; \frac{1}{\lambda}.
\end{aligned}$$

Eine ähnliche Rechnung ergibt

$$\begin{aligned}
\mathbb{E}[X^2] &= \int_0^\infty x^2\lambda \exp(-\lambda x)dx \;=\; \frac{1}{\lambda}\int_0^\infty (\lambda x)^2 \exp(-\lambda x)dx \;=\; \frac{1}{\lambda^2}\int_0^\infty t^2\exp(-t)dt \\
&= \frac{1}{\lambda^2}\left(t^2(-\exp(-t))\Big|_0^\infty - \int_0^\infty 2t(-\exp(-t))dt\right) \\
&= \frac{1}{\lambda^2}(0 + 2) \;=\; \frac{2}{\lambda^2}.
\end{aligned}$$

Damit folgt für die Varianz

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

□

9.5 Die Gauß- oder Normalverteilung

Die Gauß-Verteilung oder Normalverteilung ist eine der wichtigsten stetigen Verteilungen in der Stochastik. Einen Grund dafür werden wir im folgenden Abschnitt 9.6 sehen. Diese Verteilung wird sehr oft eingesetzt, um zufällige Abweichungen von Sollwerten zu modellieren, also Beobachtungs- oder Messfehler oder Fehler im Fertigungsprozess. Die Verteilung besitzt zwei Parameter μ und σ^2, die (wie die Notation nahelegt) den Erwartungswert und die Varianz der Verteilung angeben.

Definition 9.7 Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. Eine stetige Zufallsvariable $X : \Omega \to \mathbb{R}$ mit der Dichtefunktion

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

heißt *Gauß-verteilt* oder *normalverteilt* mit Parametern $\mu \in \mathbb{R}$ und $\sigma^2 > 0$, kurz geschrieben als $X \sim \mathcal{N}(\mu, \sigma^2)$. Im Fall $X \sim \mathcal{N}(0,1)$ heißt X *standardnormalverteilt*. In diesem Fall bezeichnen wir die Dichtefunktion mit

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right).$$

□

Die zugehörige Verteilungsfunktion

$$F(y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx$$

hat keine geschlossen Darstellung und muss i.d.R. mit Methoden der numerischen Integration berechnet werden. Abbildung 9.2 zeigt die Dichtefunktion f (links) und die Verteilungsfunktion F (rechts) für $X \sim \mathcal{N}(0,1)$.

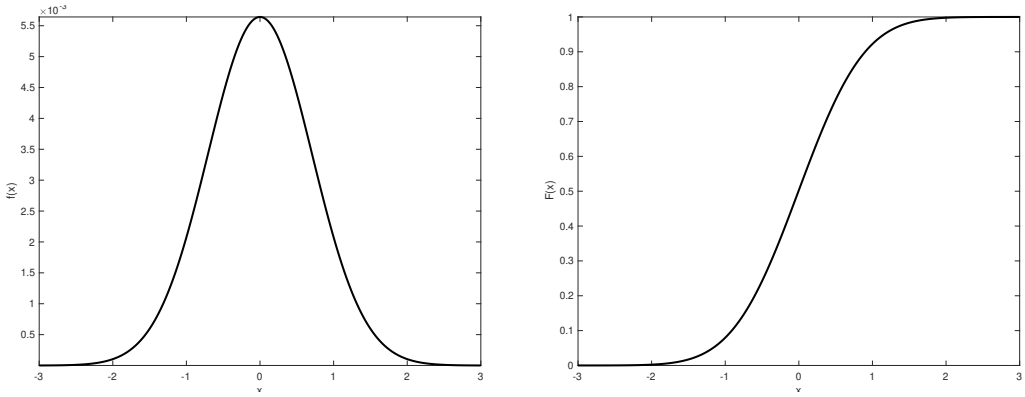


Abbildung 9.2: Dichtefunktion f (links) und Verteilungfunktionen F (rechte) für die Standardnormalverteilung

Wir berechnen nun $\mathbb{E}[X]$ und $\mathbb{V}[X]$ zunächst für standardnormalverteilte Zufallsvariablen $X \sim \mathcal{N}(0,1)$.

Satz 9.8 Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $X : \Omega \to \mathbb{R}$ eine standardnormalverteilte Zufallsvariable. Dann gilt

$$\mathbb{E}[X] = 0 \quad \text{und} \quad \mathbb{V}[X] = 1.$$

Beweis: Durch Nachrechnen sieht man, dass $\varphi'(x) = -x\varphi(x)$ gilt für alle $x \in \mathbb{R}$. Damit folgt

$$\int_0^\infty x\varphi(x)dx = \int_0^\infty -\varphi'(x)dx = -\varphi(x)\Big|_0^\infty = 0 - \frac{-1}{\sqrt{2\pi}} = \frac{1}{\sqrt{2\pi}}$$

und

$$\int_{-\infty}^0 x\varphi(x)dx = \int_{-\infty}^0 -\varphi'(x)dx = -\varphi(x)\Big|_{-\infty}^0 = \frac{-1}{\sqrt{2\pi}} - 0 = -\frac{1}{\sqrt{2\pi}}.$$

Damit gilt

$$\int_{-\infty}^\infty |x|f(x)dx = \int_{-\infty}^0 -x\varphi(x)dx + \int_0^\infty x\varphi(x)dx = -\frac{-1}{\sqrt{2\pi}} + \frac{1}{\sqrt{2\pi}} < \infty.$$

Der Erwartungswert $\mathbb{E}[X]$ existiert also und erfüllt

$$\int_{-\infty}^\infty xf(x)dx = \int_{-\infty}^0 x\varphi(x)dx + \int_0^\infty x\varphi(x)dx = \frac{-1}{\sqrt{2\pi}} + \frac{1}{\sqrt{2\pi}} = 0.$$

Wegen $\mathbb{E}[X] = 0$ gilt $\mathbb{V}[X] = \mathbb{E}[X^2]$ und deswegen mit partieller Integration mit $u(x) = x$ und $v(x) = x\varphi(x)$

$$\mathbb{V}[X] = \int_{-\infty}^\infty x^2\varphi(x)dx = \frac{1}{\sqrt{2\pi}}\left[-x\exp(-x^2/2)\Big|_{-\infty}^\infty + \int_{-\infty}^\infty 1 \cdot \exp(-x^2/2)dx\right] = 1.$$

Darin gilt $-x\exp(-x^2/2)\Big|_{-\infty}^\infty = 0$ und $\int_{-\infty}^\infty \exp(-x^2/2)dx = \sqrt{2\pi}$, was man durch eine längere Rechnung mit Polarkoordinaten zeigen kann (siehe [1, Beweis von Satz 14.17]). Also ist der Term in eckigen Klammern gleich $\sqrt{2\pi}$ und die Varianz damit wie behauptet gleich 1. □

Für normalverteiltes $X \sim \mathcal{N}(\mu, \sigma^2)$ mit beliebigem μ und σ können wir die Berechnung auf den standardnormalverteilten Fall zurückführen.

Satz 9.9 Falls $Y \sim \mathcal{N}(0, 1)$, $\mu \in \mathbb{R}$ und $\sigma > 0$, so gilt

$$X = \sigma Y + \mu \sim \mathcal{N}(\mu, \sigma^2).$$

Damit folgt $\mathbb{E}[X] = \mu$ und $\mathbb{V}[X] = \sigma^2$.

Umgekehrt gilt für $X \sim \mathcal{N}(\mu, \sigma^2)$, dass

$$Y = \frac{X - \mu}{\sigma^2} \sim \mathcal{N}(0, 1).$$

Beweis: Für das X aus dem Satz gilt

$$\mathbb{P}(\{X \leq b\}) = \mathbb{P}\left(\left\{Y \leq \frac{b-\mu}{\sigma}\right\}\right) = \int_{-\infty}^{\frac{b-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} \exp(-y^2/2)dy.$$

Mit der Substitution $x = \sigma y + \mu$ können wir fortfahren

$$= \int_{-\infty}^{b} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx.$$

Folglich ist $X \sim \mathcal{N}(\mu, \sigma^2)$. Die Behauptung für $\mathbb{E}[X]$ und $\mathbb{V}[X]$ ergibt sich aus

$$\mathbb{E}[X] = \mathbb{E}[\sigma Y + \mu] = \sigma\mathbb{E}[Y] + \mu = \sigma \cdot 0 + \mu = \mu$$

und

$$\mathbb{V}[X] = \mathbb{V}[\sigma Y + \mu] = \sigma^2\mathbb{V}[Y] = \sigma^2 \cdot 1 = \sigma^2.$$

Die umgekehrte Aussage folgt mit der analogen Rechnung in die entgegengesetzte Richtung. □

Die Transformation

$$Y = \frac{X-\mu}{\sigma},$$

die aus $X \sim \mathcal{N}(\mu, \sigma^2)$ das $Y \sim \mathcal{N}(0, 1)$ macht, wird *Standardisierung von* X genannt. Diese Transformation erzeugt tatsächlich auch für nicht Gauß'sche Zufallsvariablen eine neue Zufallsvariable mit Erwartungswert 0 und Varianz 1: Falls X eine Zufallsvariable mit Erwartungswert $\mathbb{E}[X]$ und Varianz $\mathbb{V}[X]$ ist, so ist

$$X^* = \frac{X - \mathbb{E}[X]}{\sqrt{\mathbb{V}[X]}} \tag{9.3}$$

eine Zufallsvariable mit

$$\mathbb{E}[X^*] = \mathbb{E}\left[\frac{X - \mathbb{E}[X]}{\sqrt{\mathbb{V}[X]}}\right] = \frac{\mathbb{E}[X] - \mathbb{E}[X]}{\sqrt{\mathbb{V}[X]}} = 0$$

und

$$\mathbb{V}[X^*] = \mathbb{V}\left[\frac{X - \mathbb{E}[X]}{\sqrt{\mathbb{V}[X]}}\right] = \mathbb{V}\left[\frac{X}{\sqrt{\mathbb{V}[X]}}\right] = \frac{1}{\sqrt{\mathbb{V}[X]}^2}\mathbb{V}[X] = 1.$$

Das folgende Beispiel zeigt ein typisches Anwendungsbeispiel der Gauß-Verteilung aus der Fertigung.

Beispiel 9.10 In einer Fertigung wird als Qualitätsmerkmal für ein Werkstück festgesetzt, dass die Abweichung von einer vorgegebenen Länge nicht mehr als 3,6mm betragen darf. Die im Herstellungsprozess entstehende Abweichung wird als normalverteilt mit Mittelwert 0mm und Standardabweichung σ =3mm modelliert (die Gültigkeit dieser Annahme muss in der Praxis natürlich durch Stichproben verifiziert werden). Wie viele Prozent der Werkstücke werden dann durchschnittlich mit der gewünschten Genauigkeit produziert?

Mit der Zufallsvariablen $X =$ "Abweichung von der vorgegebenen Länge" können wir dies berechnen. Nach der Modellannahme gilt $X \sim \mathcal{N}(0,\ 9)$. Also folgt

$$\begin{aligned}\mathbb{P}(\{|X| \leq 3{,}6\}) &= \mathbb{P}(\{X \leq 3{,}6\} \setminus \{X \leq -3{,}6\}) \\ &= F(3{,}6) - F(-3{,}6) \approx 0{,}88493 - 0{,}11507 = 0{,}76986,\end{aligned}$$

d.h. etwa 77% der Werkstücke erfüllen das Qualitätsmerkmal. □

9.6 Der zentrale Grenzwertsatz

Die Gauß-Verteilung ist deswegen so wichtig, weil sie in realen Situationen tatsächlich sehr häufig auftritt — zumindest näherungsweise. Der Grund dafür ist, dass reale "Zufälle" meist nicht auf eine Ursache zurückgehen (also auf "einmal würfeln"), sondern oft die Summe vieler Ursachen (also auf "sehr oft würfeln"). Das folgende Beispiel zeigt zunächst anschaulich, dass sich in diesem Fall näherungsweise eine Gauß-Verteilung ergibt.

Beispiel 9.11 Wir bestimmen die Wahrscheinlichkeiten der Augensummen bei n-maligen Würfeln durch Simulation. Dafür führen wir das n-malige Würfeln per Computersimulation für $n = 2, 5, 10, 20$ jeweils 100 000 mal durch. Wir zählen dann die Häufigkeit jeder gewürfeltem Augensumme und stellen diese als Histogramm dar. Dies liefert, wenn man die Skalen der Grafik entsprechend anpasst, eine Näherung der Dichtefunktion (was aus dem schwachen Gesetz der großen Zahlen angewendet auf die Zufallsvariablen $Y = 1$ falls $X \in [a, b]$ und $Y = 0$ falls $X \notin [a, b]$ folgt). Abbildung 9.3 stellt die Histogramme dar.

Während das Histogramm für $n = 2$ noch recht "dreieckig" aussieht (wie nach den Wahrscheinlichkeiten aus Beispiel 6.3 zu erwarten), lässt sich bereits bei $n = 5$ und noch viel klarer bei $n = 10$ und $n = 20$ die typische "Glockenkurve" der Dichtefunktion der Normalverteilung erkennen, vgl. Abb. 9.2(links).

Es scheint also, dass die Verteilung der Augensummen sich für wachsende n der Gauß-Verteilung annähert, und das, obwohl die Verteilung der Zahlen bei *einem* Wurf ja gar nicht Gauß-verteilt ist. □

Um dies formal zu fassen, müssen wir die Verteilung der Augensummen noch richtig normieren, da sich die Histogramme für wachsende n immer weiter nach rechts (und bei anderen Beispielen auch nach links) ausdehnen. Dazu verwenden wir die Standardisierung (9.3). Für Zufallsvariablen X_i, $i \in \mathbb{N}$ betrachten wir die summierten Zufallsvariablen und ihre Standardisierung, also

$$S_n = \sum_{i=1}^{n} X_i \quad \text{und} \quad S_n^* = \frac{S_n - \mathbb{E}[S_n]}{\sqrt{\mathbb{V}[S_n]}}.$$

Dann gilt der folgende *zentrale Grenzwertsatz*.

Satz 9.12 Seien $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und X_i, $i \in \mathbb{N}$, stochastisch unabhängige und identisch verteilte Zufallsvariablen, für die $\mathbb{E}[X_i]$ und $\mathbb{V}[X_i]$ existieren. Dann gilt für alle $B \in \mathcal{B}$ die Konvergenz

$$\lim_{n\to\infty} \mathbb{P}(\{S_n^* \in B\}) = \mathbb{P}(\{X \in B\})$$

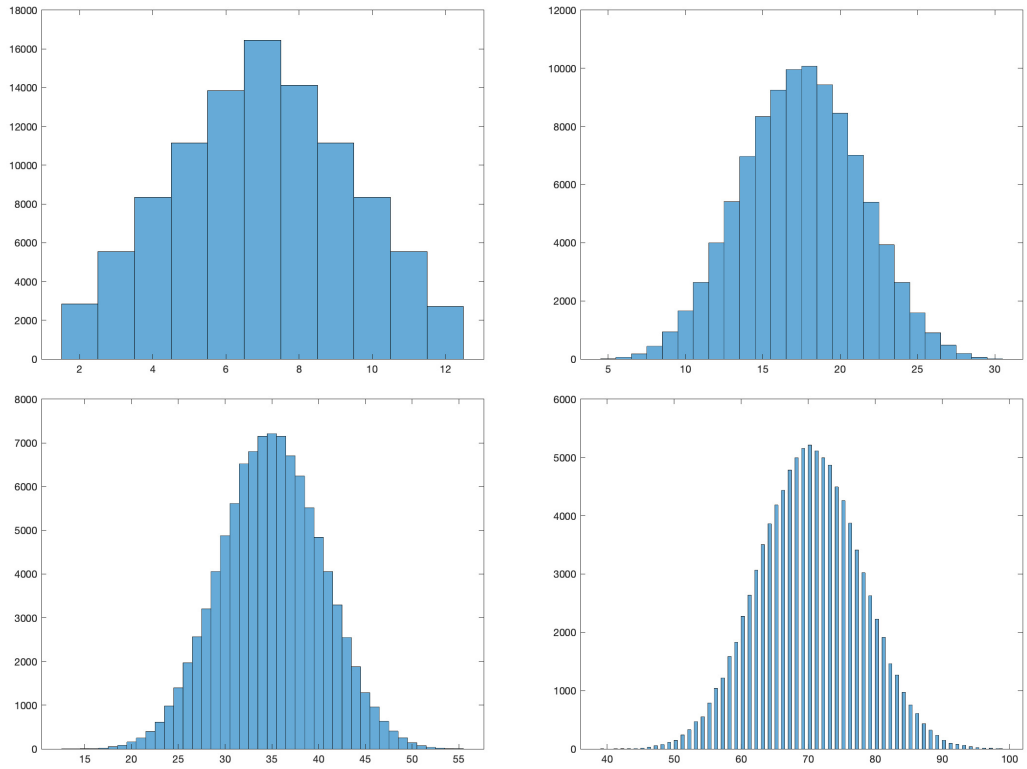


Abbildung 9.3: Histogramm für "Augensumme aus n mal Würfeln" für $n = 2$ (oben links), $n = 5$ (oben rechts), $n = 10$ (unten links), $n = 20$ (unten rechts), durchgeführt mit jeweils 100 000 Zufallsexperimenten

wobei $X \sim \mathcal{N}(0, 1)$ eine standardnormalverteilte Zufallsvariable ist. Man sagt dann, dass S_n^* *schwach gegen* X *konvergiert*.

Der Satz lässt sich auch unter schwächeren Bedingungen an die X_i beweisen, die wir hier aber nicht thematisieren wollen. Sein Beweis erfordert Techniken, die den Umfang dieser Vorlesung deutlich übersteigen, weswegen wir ihn hier nicht behandeln.

Betrachten wir die bereits aus dem schwachen Gesetz der großen Zahlen bekannte Zufallsvariable

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

und deren Standardisierung $\overline{X}_n^*$, so gilt für diese, dass $\mathbb{E}[\overline{X}_n] = \frac{1}{n}\mathbb{E}[S_n]$ und $\mathbb{V}[\overline{X}_n] = \frac{1}{n^2}\mathbb{V}[S_n]$. Damit folgt

$$\overline{X}_n^* = \frac{\overline{X}_n - \mathbb{E}[\overline{X}_n]}{\sqrt{\mathbb{V}[\overline{X}_n]}} = \frac{\frac{1}{n}S_n - \frac{1}{n}\mathbb{E}[S_n]}{\sqrt{\frac{1}{n^2}\mathbb{V}[S_n]}} = \frac{\frac{1}{n}(S_n - \mathbb{E}[S_n])}{\frac{1}{n}\sqrt{\mathbb{V}[S_n]}} = \frac{S_n - \mathbb{E}[S_n]}{\sqrt{\mathbb{V}[S_n]}} = S_n^*.$$

Der zentrale Grenzwertsatz gilt also ebenfalls für $\overline{X}_n^*$.

Wegen $\mathbb{E}[\overline{X}_n] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \mathbb{E}[X_1]$ und $\mathbb{V}[\overline{X}_i] = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}[X_i] = \frac{\mathbb{V}[X_1]}{n}$ folgt zudem

$$\sqrt{n}\frac{\overline{X}_n - \mathbb{E}[X_1]}{\sqrt{\mathbb{V}[X_1]}} = \frac{\overline{X}_n - \mathbb{E}[X_1]}{\sqrt{\frac{\mathbb{V}[X_1]}{n}}} = \frac{\overline{X}_n - \mathbb{E}[\overline{X}_n]}{\sqrt{\mathbb{V}[\overline{X}_n]}} = \overline{X}_n^* = S_n^*,$$

so dass wir S_n^* auch mit Erwartungswert und Varianz von X_1 berechnen können.

Der zentrale Grenzwertsatz ist einer der Gründe, warum die Gauß- oder Normalverteilung in der Praxis oft — aber nicht immer — als Modell gut geeignet ist.

Kapitel 10

Eine kurze Wiederholung wichtiger Begriffe aus der Linearen Algebra

Wir wiederholen in diesem Kapitel einige Begriffe aus der linearen Algebra, die im Prinzip aus der Ingenieurmathematik I bekannt sein sollten, hier aber noch einmal ins Gedächtnis gerufen werden sollen.

10.1 Vektoren

Mit $\mathbb{R}$ bezeichnen wir die reellen Zahlen und mit $\mathbb{R}^n$, $n \in \mathbb{N}$, die Menge der n-dimensionalen Vektoren über $\mathbb{R}$. Die Menge $\mathbb{R}^n$ ist ein sogenannter *Vektorraum*. Einen Vektor $x \in \mathbb{R}^n$ kann man als n-Tupel

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \tag{10.1}$$

mit n Einträgen $x_1, x_2, \ldots, x_n \in \mathbb{R}$ darstellen.

Hinter dieser Darstellung steckt das Konzept der *Basis* des Vektorraums $\mathbb{R}^n$. Eine Basis des $\mathbb{R}^n$ ist eine Menge $V = \{v_1, \ldots, v_n\}$ von Vektoren $v_i \in \mathbb{R}^n$, so dass jeder andere Vektor $x \in \mathbb{R}^n$ als Linearkombination

$$x = \tilde{x}_1 v_1 + \tilde{x}_2 v_2 + \ldots + \tilde{x}_n v_n = \sum_{i=1}^{n} \tilde{x}_i v_i \tag{10.2}$$

für geeignete $\tilde{x}_1, \ldots, \tilde{x}_n \in \mathbb{R}$ geschrieben werden kann. In der Basis V kann der Vektor x dann kurz als

$$\tilde{x} = \begin{pmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \vdots \\ \tilde{x}_n \end{pmatrix} \tag{10.3}$$

geschrieben werden. Im Spezialfall, dass V die sogenante Standardbasis

$$v_1 = e_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad v_2 = e_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \ldots, v_n = e_n = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}$$

ist, gilt

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = x = \tilde{x}_1 e_1 + \tilde{x}_2 e_2 + \ldots + \tilde{x}_n e_n = \begin{pmatrix} \tilde{x}_1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \tilde{x}_2 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \ldots + \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ \tilde{x}_n \end{pmatrix} = \begin{pmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \vdots \\ \tilde{x}_n \end{pmatrix}.$$

In der Standardbasis stimmen die Darstellungen (10.1) und (10.3) also überein, weswegen diese als Standard (daher der Name) verwendet wird, so lange nicht explizit etwas anderes erwähnt wird.

Wir haben oben bereits verwendet, dass Vektoren komponentenweise addiert werden, für $x, y \in \mathbb{R}^n$ gilt also

$$x + y = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_n + y_n \end{pmatrix}.$$

Zudem kann man Vektoren mit Skalaren $\lambda \in \mathbb{R}$ multiplizieren: es gilt

$$\lambda x = \lambda \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} \lambda x_1 \\ \lambda x_2 \\ \vdots \\ \lambda x_n \end{pmatrix}.$$

Man könnte Vektoren auch komponentenweise miteinander multiplizieren, aber für diese Operation gibt es kaum sinnvolle Anwendungen, weswegen sie i.d.R. nicht betrachtet wird. Sinnvoll hingegen ist die skalare Multiplikation von Vektoren $x, y \in \mathbb{R}^n$, die durch

$$\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i$$

gegeben ist. Beachte, dass das Ergebnis der skalaren Multiplikation, das sogennannte *Skalarprodukt* $\langle x, y \rangle$, ein Skalar, also eine Zahl in $\mathbb{R}$ und kein Vektor im $\mathbb{R}^n$ ist — daher auch der Name. Interpretiert man die n-Tupel als Koordinaten in einem Koordinatensystem und stellt die Vektoren als Ortsvektoren dar, also als Pfeile vom Nullpunkt zu dem durch die Vektorkoordinaten spezifizierten Punkt, so ist das Skalarprodukt genau dann gleich 0, wenn die Vektoren senkrecht aufeinander stehen.

Die (euklidische) Norm eines Vektors ist gegeben durch

$$\|x\|_2 = \sqrt{\langle x, x\rangle} = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

Diese Größe ist nach dem Satz des Pythagoras gerade die Länge des Ortsvektors x und sie ist genau dann gleich 0 wenn $x = 0$ ist, also wenn alle Einträge von x gleich 0 sind.

10.2 Matrizen

Matrizen bestehen aus zeilen- und spaltenweise angeordneten rellen Zahlen $a_{ij} \in \mathbb{R}$. Eine Matrix mit n Zeilen und m Spalten hat die Form

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix}.$$

Wir schreiben $A \in \mathbb{R}^{n\times m}$. Eine solche Matrix kann man mittels der Regel "Zeile mal Spalte" mit einem Vektor $x \in \mathbb{R}^m$ multiplizieren

$$Ax = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix} = \begin{pmatrix} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1m}x_m \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2m}x_m \\ \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nm}x_m \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^m a_{1i}x_i \\ \sum_{i=1}^m a_{2i}x_i \\ \vdots \\ \sum_{i=1}^m a_{ni}x_i \end{pmatrix}$$

und erhält so einen Vektor $Ax \in \mathbb{R}^n$. Über die Matrix-Vektor-Multiplikation definiert eine Matrix $A \in \mathbb{R}^{n\times m}$ also eine Abbildung (genauer: eine lineare Abbildung) von $\mathbb{R}^m$ nach $\mathbb{R}^n$. Analog kann man Matrizen $A \in \mathbb{R}^{n\times m}$ und $B \in \mathbb{R}^{m\times l}$ miteinander multiplizieren indem man A mit jeder Spalte von B multipliziert. So erhält man eine Matrix $AB \in \mathbb{R}^{n\times l}$.

Eine Matrix heißt quadratisch, falls die Anzahl der Zeilen gleich der Anzahl der Spalten ist, falls also $A \in \mathbb{R}^{n\times n}$ für ein $n \in \mathbb{N}$ ist. Eine spezielle quadratische Matrix ist die Einheitsmatrix $I_n \in \mathbb{R}^{n\times n}$ gegeben durch

$$I_n = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \ldots & 1 \end{pmatrix}.$$

Beachte, dass $Ix = x$ für alle $x \in \mathbb{R}^n$ und $IB = B$ für alle $B \in \mathbb{R}^{n\times n}$.

Eine Matrix $A \in \mathbb{R}^{n\times n}$ heißt *invertierbar*, falls eine Matrix $A^{-1} \in \mathbb{R}^{n\times n}$ existiert mit $A^{-1}A = I_n$ und $AA^{-1} = I_n$. Die Matrix A^{-1} heißt dann die *Inverse* zu A. Beachte, dass nicht alle Matrizen invertierbar sind. Beispielsweise gilt für $A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$, dass

$$BA = \begin{pmatrix} b_{11} + b_{12} & b_{11} + b_{12} \\ b_{21} + b_{22} & b_{21} + b_{22} \end{pmatrix}.$$

Die Einträge oben links und oben rechts stimmen also immer überein, egal wie man $B \in \mathbb{R}^{2\times 2}$ wählt. Es kann also nie die Einheitsmatrix I_2 herauskommen, in der oben links eine 1 und oben rechts eine 0 stehen müsste. Falls eine Matrix invertierbar ist, gibt es verschiedene Methoden, die Inverse zu berechnen, die wir hier aber nicht behandeln werden.

In der zu A transponierten Matrix A^T werden die Zeilen und Spalten miteinander vertauscht, also

$$A^T = \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{12} & a_{22} & \cdots & a_{n2} \\ \vdots & \vdots & \cdots & \vdots \\ a_{1m} & a_{2m} & \cdots & a_{nm} \end{pmatrix} \in \mathbb{R}^{m\times n}.$$

Eine quadratische Matrix heißt *symmetrisch*, wenn $A^T = A$. Ein Beispiel für eine symmetrische Matrix ist die Kovarianzmatrix aus 8.2, weil für die Kovarianz offenbar

$$\mathrm{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mathbb{E}[X_i]) \cdot (X_j - \mathbb{E}[X_j])] = \mathrm{Cov}(X_j, X_i)$$

gilt. Fasst man Vektoren als Matrizen mit einer Spalte auf, so kann man auch Vektoren transformieren

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad \Rightarrow \quad x^T = (x_1\ x_2\ \ldots\ x_n).$$

Damit kann man das Skalarpodukt schreiben als $\langle x, y\rangle = x^T y$ und quadratische Ausdrücke der Form $x^T Ax \in \mathbb{R}$ für $x \in \mathbb{R}^n$ und $A \in \mathbb{R}^{n\times n}$ definieren.

Für transponierte Matrizen (und analog für Vektoren) gilt die Rechenregel $(AB)^T = B^T A^T$.

Eine symmetrische Matrix $A \in \mathbb{R}^{n\times n}$ heißt dann *positiv semidefinit*, falls $x^T Ax \geq 0$ gilt für alle $x \in \mathbb{R}^n$ und *positiv definit*, falls $x^T Ax > 0$ gilt für alle $x \in \mathbb{R}^n$ mit $x \neq 0$ (hierbei bezeichnet "0" den Vektor im $\mathbb{R}^n$, dessen Einträge alle gleich Null sind).

Beispiel 10.1 Wir wollen eine Schätzung der Kovarianzmatrix (8.2) mit Hilfe der rechten Formel aus (8.4) berechnen. Da wir die tiefgestellten Indizes bereits zum Durchnummerieren der einzelnen Zufallsvariablen X_i in unserem Zufallsvektor

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$$

benötigen, nummerieren wir die Realisierungen dieser Zufallsvariablen nun mit hochgestelltem Index, also

$$x^k = \begin{pmatrix} x_1^k \\ x_2^k \\ \vdots \\ x_n^k \end{pmatrix} = \begin{pmatrix} X_1^k(\omega) \\ X_2^k(\omega) \\ \vdots \\ X_n^k(\omega) \end{pmatrix} = X^k(\omega)$$

mit $k = 1, \dots, m$ durch. Da X_i^k die Durchführung des k-ten Zufallsexperiments für die Zufallsvariable X_i modelliert, besitzen X_i und X_i^k die gleiche Verteilung und erfüllen $\mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(X_i^k, X_j^k)$ für alle $k = 1, \dots, m$.

Mit dieser Nummerierung wird die rechte Formel aus (8.4) zu

$$\mathrm{Cov}(X_i, X_j) \approx \frac{1}{m} \sum_{k=1}^{m} (x_i^k - \mathbb{E}[X_i])(x_j^k - \mathbb{E}[X_j]). \tag{10.4}$$

Wir setzen nun $\tilde{x}_i^k = x_i^k - \mathbb{E}[X_i]$ und definieren

$$\widetilde{X} = \begin{pmatrix} \tilde{x}^1 & \tilde{x}^2 & \dots & \tilde{x}^m \end{pmatrix} \quad \text{mit den Vektoren} \quad \tilde{x}^k = \begin{pmatrix} \tilde{x}_1^k \\ \tilde{x}_2^k \\ \vdots \\ \tilde{x}_n^k \end{pmatrix}.$$

$\widetilde{X}$ ist dann eine Matrix aus dem $\mathbb{R}^{n \times m}$ und es gilt

$$C := \widetilde{X}\widetilde{X}^T = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{pmatrix} \in \mathbb{R}^{n \times n}$$

mit $c_{ij} = \sum_{k=1}^{m} (x_i^k - \mathbb{E}[X_i])(x_j^k - \mathbb{E}[X_j]) \approx m \,\mathrm{Cov}(X_i, X_j)$. Bis auf eine Skalierung mit m ist C also gerade die Schätzung der Kovarianzmatrix $\mathrm{Cov}(X)$. Aus dieser Formel sieht man zudem, dass C (genau wie $\mathrm{Cov}(X)$) symmetrisch ist. Wegen

$$x^T C x = x^T \widetilde{X}\widetilde{X}^T x = (\widetilde{X}^T x)^T \widetilde{X}^T x = \langle \widetilde{X}^T x, \widetilde{X}^T x \rangle = \|\widetilde{X}^T x\|_2 \geq 0$$

ist C auch positiv semidefinit. Falls $m > n$ ist (was der übliche Fall ist, weil wir ja für eine gute Schätzung viele Daten benötigen) und die Matrix $\widetilde{X}$ vollen Rang besitzt (was bedeutet, dass die n Zeilen linear unabhängig sind, dass also die Daten für eine Zufallsvariable sich nicht als Linearkombination der Daten der anderen Zufallsvariablen darstellen lassen), so gilt $\widetilde{X}^T x \neq 0$ für alle $x \in \mathbb{R}^n \setminus \{0\}$ und damit

$$x^T C x = \|\widetilde{X}^T x\|_2 > 0.$$

Also ist C in diesem Fall positiv definit. □

10.3 Basiswechsel

Für eine gegebene Basis $V = \{v_1, \dots, v_n\}$ des $\mathbb{R}^n$ und einen Vektor $x \in \mathbb{R}^n$ kann man $\tilde{x}$ aus (10.3) wie folgt berechnen: Schreiben wir die Basisvektoren nebeneinander in eine Matrix

$$T = \begin{pmatrix} v_1 & v_2 & \dots & v_n \end{pmatrix} \in \mathbb{R}^{n \times n},$$

so können wir (10.2) kurz als $T\tilde{x} = x$ schreiben. Man kann nun beweisen, dass eine Matrix, deren Spalten eine Basis bilden, immer invertierbar ist, dass es also eine Inverse T^{-1} gibt. Für diese gilt dann

$$\tilde{x} = I_n \tilde{x} = T^{-1} T \tilde{x} = T^{-1} x.$$

Damit haben wir eine Formel[1] für $\tilde{x}$ gefunden.

Wenn man einen Vektor x in einer neuen Basis V mittels $\tilde{x}$ ausdrückt, nennt man dies einen Wechsel der Basis oder auch einen Koordinatenwechsel oder eine Koordinatentranformation. Die Matrix T heißt dann auch Transformationsmatrix. Man kann in der neuen Basis mit dem Vektor $\tilde{x}$ (und allen weiteren transformierten Vektoren) ganz genau so rechnen wie in der alten. Will man aber eine lineare Abbildung, die durch die Matrixmultiplikation Ax definiert ist, auf $\tilde{x}$ anwenden, so muss man A zuerst ebenfalls passend transformieren. Für diese transformierte Matrix $\widetilde{A}$ soll gelten, dass $\widetilde{A}\tilde{x} = T^{-1}Ax$ ist, d.h., dass die Multiplikation von $\tilde{x}$ mit $\widetilde{A}$ das gleiche bewirkt wie die Multiplikation von x mit A und nachfolgendem Basiswechsel. Wegen $x = T\tilde{x}$, folgt

$$\widetilde{A}\tilde{x} = T^{-1}Ax = T^{-1}AT\tilde{x}.$$

Damit dies für jedes beliebige $\tilde{x}$ gilt, muss

$$\widetilde{A} = T^{-1}AT$$

gelten. Dies ist also die Transformationsformel für Matrizen bei einem Basiswechsel.

10.4 Eigenwerte und Eigenvektoren

Besonders einfache Matrizen sind Diagonalmatrizen

$$\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) := \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \ldots & \lambda_n \end{pmatrix}.$$

Im Rahmen der Datenanalyse werden wir z.B. in Abschnitt 12.2 sehen, welchen Vorteil Diagonalmatrizen haben. Es wäre nun schön, wenn man für beliebige Matrizen eine Möglichkeit hätte, diese mit einer Transformationmatrix T und der eben hergeleiteten Formel $T^{-1}AT$ auf Diagonalgestalt bringen könnte. Dies geht zwar nicht für alle Matrizen, aber zumindest für eine wichtige Unterklasse. Der Schlüssel dazu sind die Eigenwerte und Eigenvektoren einer Matrix.

Definition 10.2 Ein $\lambda \in \mathbb{R}$ heißt *reeller Eigenwert* einer Matrix $A \in \mathbb{R}^{n\times n}$, falls ein zugehöriger reeller Eigenvektor $v \in \mathbb{R}^n$, $v \neq 0$, existiert, so dass

$$Av = \lambda v$$

gilt. □

[1] Die Gleichung $T\tilde{x} = x$ stellt ein lineares Gleichungssystem für $\tilde{x}$ dar. I.A. ist es algorithmisch einfacher, dieses lineare Gleichungssystem zu lösen, statt T^{-1} zu berechnen. Trotzdem ist die geschlossene Formel $\tilde{x} = T^{-1}x$ oft praktisch, weswegen wir sie hier angeben.

Ein reeller Eigenvektor ist also ein Vektor, der bei Multiplikation mit A auf ein (reelles) Vielfaches seiner selbst abgebildet wird. Reelle Matrizen besitzen immer komplexe Eigenwerte- und vektoren, aber nicht unbedingt reelle. Für eine wichtige Klasse von Matrizen ist bedeutet die Bedingung, dass λ und v reell sind, aber keine Einschränkung. Es gilt:

Satz 10.3 Eine symmetrische Matrix $A \in \mathbb{R}^{n\times n}$ besitzt genau n reelle Eigenwerte $\lambda_1, \ldots, \lambda_n$. Die zugehörigen Eigenvektoren $v_1, \ldots, v_n$ bilden eine Basis des $\mathbb{R}^n$ und stehen senkrecht aufeinander, d.h. es gilt $\langle v_i, v_j\rangle = 0$ für $i \neq j$. Zudem können die v_i so gewählt werden, dass $\|v_i\|_2 = 1$ gilt für alle $i = 1, \ldots, n$. Man sagt, die v_i bilden eine *Orthonormalbasis*.

Betrachtet man die von den v_i erzeugte Transformationsmatrix T, so

$$AT = (\lambda_1 v_1, \lambda_2 v_2, \ldots, \lambda_d v_d) = T\Lambda \qquad \text{mit } \Lambda = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_d \end{pmatrix}.$$

Es folgt also

$$T^{-1}AT = \Lambda,$$

d.h. T transformiert die Matrix A gerade in Diagonalform und die Einträge der entstehenden Diagonalmatrix sind gerade die Eigenwerte von A.

Eine Matrix, deren Spalten eine Orthonormalbasis bilden, heißt *orthogonal*. Die von den v_i aus Satz 10.3 erzeugte Transformationsmatrix T ist also eine orthogonale Matrix. Solche Matrizen haben einige schöne Eigenschaften. Zum einen gilt

$$T^TT = \begin{pmatrix} v_1^Tv_1 & v_1^Tv_2 & \cdots & v_1^Tv_n \\ v_2^Tv_1 & v_2^Tv_2 & \cdots & v_2^Tv_n \\ \vdots & \vdots & \cdots & \vdots \\ v_n^Tv_1 & v_n^Tv_2 & \cdots & v_n^Tv_n \end{pmatrix} = \begin{pmatrix} \langle v_1, v_1\rangle & \langle v_1, v_2\rangle & \cdots & \langle v_1, v_n\rangle \\ \langle v_2, v_1\rangle & \langle v_2, v_2\rangle & \cdots & \langle v_2, v_n\rangle \\ \vdots & \vdots & \cdots & \vdots \\ \langle v_n, v_1\rangle & \langle v_n, v_2\rangle & \cdots & \langle v_n, v_n\rangle \end{pmatrix}$$
$$= \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \ldots & 1 \end{pmatrix} = I_n.$$

Folglich ist $T^T = T^{-1}$, was bedeutet, dass die Inverse T^{-1} einfach durch Transponieren von T berechnet werden kann.

Zum anderen gilt

$$\langle Tx, Ty\rangle = (Tx)^TTy = x^T\underbrace{T^TT}_{=I_n}y = x^TI_ny = x^Ty = \langle x, y\rangle.$$

T "erhält" also das Skalarprodukt und damit auch die euklidische Norm, denn

$$\|Tx\|_2 = \sqrt{\langle Tx, Tx\rangle} = \sqrt{\langle x, x\rangle} = \|x\|_2.$$

Die Multiplikation mit einer orthogonalen Matrix ändert die Länge eines Vektors also nicht.

Kapitel 11

Lineare Regression

11.1 Problemstellung

Regressionsprobleme gehören zu den Standardaufgaben der mathematischen Datenanalyse. Um das Problem zu motivieren, betrachten wir zunächst die folgende Problemstellung: Gegeben seien Daten (t_i, y_i), $i = 1, \ldots, N$, bei denen wir der Einfachheit halber annehmen, dass t_i und y_i skalare Größen sind (...). Es könnte z.B. t_i eine Temperatur sein und y_i die Konzentration eines Stoffes in einer chemischen Reaktion, die bei dieser Temperatur abläuft.

Aufgrund von theoretischen Überlegungen (z.B. aufgrund eines zugrundeliegenden physikalischen Gesetzes) kennt man eine Funktion $f(t)$, für die $f(t_i) = y_i$ gelten sollte. Diese Funktion wiederum hängt aber nun von unbekannten Parametern $\theta_1, \ldots, \theta_K$ ab, die wir in dem Vektor $\theta = (\theta_1, \ldots, \theta_K)^T$ zusammenfassen. Wir schreiben $f(t; \theta)$, um diese Abhängigkeit zu betonen. Zum Beispiel könnte $f(t; \theta)$ durch

$$f(t;\theta) = \theta_1 + \theta_2 t \quad \text{oder} \quad f(t;\theta) = \theta_1 + \theta_2 t + \theta_3 t^2$$

gegeben sein. Im ersten Fall beschreibt die gesuchte Funktion eine Gerade, im zweiten eine Parabel. Wichtig wird dabei im Folgenden sein, dass die Funktion linear von den θ_i abhängt.

Wenn wir annehmen, dass die Funktion f das Experiment wirklich exakt beschreibt und keine Messfehler vorliegen, so könnten wir die Parameter θ_i durch Lösen des Gleichungssystems

$$\begin{aligned} f(t_1;\theta) &= y_1 \\ &\vdots \\ f(t_N;\theta) &= y_N \end{aligned} \tag{11.1}$$

berechnen. Nun nutzen wir aus, dass f linear von θ_i abhängt, also von der Form

$$f(t;\theta) = \sum_{j=1}^{K} \theta_j f_j(t)$$

ist. Im ersten obigen Beispiel ist $K = 2$ und $f_1(t) = 1$, $f_2(t) = t$. Im zweiten Beispiel ist $K = 3$, $f_1(t) = 1$, $f_2(t) = t$ und $f_3(t) = t^2$. Mit der Abkürzung $x_i = (f_1(t_i), \ldots, f_K(t_i))$

(beachte: x_i ist ein K-dimensionaler Zeilenvektor) können wir das Gleichungssystem (11.1) dann als lineares Gleichungssystem

$$\begin{aligned} x_1\theta &= y_1 \\ &\vdots \\ x_N\theta &= y_N \end{aligned} \tag{11.2}$$

schreiben. Definieren wir die Matrix

$$X = \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} \in \mathbb{R}^{N\times K},$$

so können wir dies kurz als $X\theta = y$ schreiben. In den obigen Beispielen gilt

$$X = \begin{pmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_N \end{pmatrix}, \theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \quad \text{bzw.} \quad X = \begin{pmatrix} 1 & t_1 & t_1^2 \\ \vdots & \vdots & \vdots \\ 1 & t_N & t_N^2 \end{pmatrix}, \theta = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix} \quad \text{und} \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}$$

Diese linearen Gleichungssysteme haben N Gleichungen (eine für jedes Wertepaar (x_i, y_i)) und K Unbekannte (nämlich gerade die unbekannten Parameter θ_j), wobei N üblicherweise sehr viel größer als K ist. Man sagt, dass das Gleichungsystem *überbestimmt* ist. Da Messwerte eines Versuchs praktisch immer mit Fehlern behaftet sind oder natürlichen Schwankungen unterliegen, ist es sicherlich zu optimistisch, anzunehmen, dass das Gleichungssystem $X\theta = y$ lösbar ist. Es wird nur dann lösbar, wenn man die Fehler, die wir mit ε_i bezeichnen, in das Gleichungssystem einbezieht, wenn wir also

$$\begin{aligned} x_1\theta + \varepsilon_1 &= y_1 \\ &\vdots \\ x_N\theta + \varepsilon_N &= y_N \end{aligned} \tag{11.3}$$

betrachten. Dies ist ein lineares Regressionsproblem in Standardform. Die Zeilenvektoren x_i heißen *Regressoren* und die Werte θ_j heißen *Regressionskoeffizienten* oder *-parameter*. Die Fehler ε_i sind unbekannt, weswegen wir das Problem in dieser Form nicht lösen können.

11.2 Maximum Likelihood Schätzung

Um die Parameter zu schätzen, verwenden wir die Methode der Maximum Likelihood Schätzung, die wir hier allgemein vorstellen und im folgenden Abschnitt auf das Regressionsproblem anwenden. Gegeben seien dazu Daten $y_1, \ldots, y_N$, die wir in dem Vektor $y = (y_1, \ldots, y_N)^T$ zusammenfassen. Die Modellierungsannahme ist nun, dass diese Daten Realisierungen von Zufallsvariablen Y_i sind, die unabhängig und identisch verteilt (also "i.i.d.") sind mit Wahrscheinlichkeitsdichte $f_Y(\cdot \mid \theta)$ (da alle Y_i die gleiche Dichte haben, brauchen wir keinen Index i an die Dichte zu schreiben). Der Vektor $\theta = (\theta_1, \ldots, \theta_K)$ bezeichnet dabei einen Vektor von Parametern, von denen die Verteilung von Y abhängt. Ziel ist jetzt, durch die Wahl von θ^* die Wahrscheinlichkeit zu maximieren (deswegen "Maximum Likelihood"), dass die y_i Realisierungen der Y_i sind.

Dazu stellen wir die gemeinsame Verteilungsfunktion von $Y_1, \ldots, Y_N$ auf und setzen die Daten $y_1, \ldots, y_N$ ein. Weil die Y_i unabhängig sind, ist die gemeinsame Verteilungsfunktion nach Bemerkung 9.2(ii) gerade das Produkt der einzelnen Verteilungsfunktionen, also gegeben durch

$$L(\theta, y) := \prod_{i=1}^{N} f_Y(y_i|\theta). \tag{11.4}$$

Das so definierte L wird "Likelihood Funktion" genannt. Die Wahscheinlichkeit, dass die y_i Realisierungen der Y_i sind, maximieren wir nun gerade dadurch, dass wir die Likelihood Funktion über θ maximieren. Wir suchen also den Parametervektor θ^*, für die $L(\theta^*, y)$ maximal wird. Dazu geht man wie üblich vor: Berechne eine Nullstelle der Ableitung $dL/d\theta$ und teste, ob die zweite Ableitung $d^2L/d\theta^2$ in der Nullstelle positiv definit ist (oder einfach positiv falls θ eindimensional ist).

Da die Ableitung von Funktionen, die durch ein Produkt gegeben sind, kompliziert zu berechnen sind, formuliert man den zu maximierenden Ausdruck mit einem Trick um. Statt $L(\theta, y)$ zu maximieren, maximieren wir

$$l(\theta, y) = \log(L(\theta, y)),$$

wobei log der natürliche Logarithmus (auch als ln geschrieben) ist. Die Funktion l heißt *log-Likelihood Funktion* und weil log eine monoton wachsende Funktion ist, maximiert θ die Funktion L genau dann, wenn es die Funktion l maximiert. Wir können also l statt L maximieren und erhalten das gleiche optimale θ. Weil nun $\log(a \cdot b) = \log(a) + \log(b)$ gilt, folgt

$$l(\theta, y) = \sum_{i=1}^{N} \log(f_Y(y_i|\theta)).$$

Nun ist also eine Summe von Termen zu maximieren, was i.A. eine viel einfachere Aufgabe als die Maximierung eines Produkts ist. Wir führen dies im übernächsten Abschnitt für eine konkrete Verteilung für die Y_i durch, die sich aus der Anwendung der Maximum Likelihood Methode auf das Regressionsproblem ergibt.

11.3 Regression mittels Maximum Likelihood

Im Regressionsproblem nehmen wir als Modellierungsannahme an, dass die Fehler ε_i Realisierungen von i.i.d. Zufallsvariablen E_i mit Wahrscheinlichkeitsdichte $f_E(\varepsilon)$ sind. Die für die Maximum Likelihood Methode benötigten Dichten der Y_i erhalten wir dann aus (11.3) für $i = 1, \ldots, n$ folgenden Gleichungen

$$x_i\theta + \varepsilon_i = y_i \quad \Rightarrow \quad \varepsilon_i = y_i - x_i\theta.$$

Wir können also ε_i in Abhängigkeit von y_i ausdrücken und diesen Ausdruck in die Dichte für E_i einsetzen. So erhalten wir $f_Y(y_i \,|\, \theta) = f_E(y_i - x_i\theta)$. Genau genommen hängt die Dichte für Y auch noch von den x_i ab, weil diese aber fest gegeben und keine zu optimierenden Parameter sind, schreiben wir sie nicht explizit als Argument von f_Y.

Wenden wir auf diese Dichtefunktion nun die Maximum Likelihood Methode an, so erhalten wir aus (11.4) die Likelihood Funktion

$$L(\theta, y) = \prod_{i=1}^{N} f_Y(y_i \,|\, \theta) = \prod_{i=1}^{N} f_E(y_i - x_i\theta) \tag{11.5}$$

und die log-Likelihood Funktion

$$l(\theta, y) = \sum_{i=1}^{N} \log(f_Y(y_i \,|\, \theta)) = \sum_{i=1}^{N} \log(f_E(y_i - x_i\theta)).$$

11.4 Explizite Lösung für Gauß-verteilte ε_i

Die Berechnung des maximierenden Parametervektors θ lässt sich (fast) explizit durchführen, wenn wir eine konkrete Annahme an die Verteilung der ε_i und damit an die Form der Funktion f_E stellen. Wir führen das für den Fall durch, dass die ε_i Gauß-verteilt sind mit Erwartungswert 0 und Varianz σ^2, also $\varepsilon_i \sim N(0, \sigma^2)$. In dem Fall gilt

$$f_E(\varepsilon) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\varepsilon}{\sigma}\right)^2\right),$$

also

$$f_Y(y_i \,|\, \theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y_i - x_i\theta}{\sigma}\right)^2\right)$$

und damit wegen $\log(\exp(x)) = x$

$$\log(f_Y(y_i \,|\, \theta)) = \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2}\left(\frac{y_i - x_i\theta}{\sigma}\right)^2.$$

Insgesamt ergibt sich für die zu maximierende log-Likelihood Funktion also

$$l(\theta, y) = N \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - x_i\theta)^2.$$

Will man nun den Parametervektor $\theta^\star \in \mathbb{R}^K$ bestimmen, für den l maximal wird, so sieht man, dass der erste Term gar nicht von θ abhängt und daher bei der Bestimmung des maximierenden $\theta^\star$ weggelassen werden kann. Ebenso hängt θ^* nicht vom Vorfaktor $1/(2\sigma^2)$ der Summe ab. Um $\theta^\star$ zu bestimmen, reicht es daher, den Ausdruck

$$-\sum_{i=1}^{N}(y_i - x_i\theta)^2$$

zu maximieren, den wir mit der 2-Norm $\|x\|_2 = \sqrt{\sum_{i=0}^{N} x_i^2}$ kurz als $-\|X\theta - y\|_2^2$ schreiben können. Aquivalent können wir auch den Ausdruck

$$\|X\theta - y\|_2^2 =: g(\theta)$$

minimieren, was wir im folgenden machen werden. Um eine Funktion $g(\theta)$ zu minimieren, muss man die Ableitung gleich Null setzen und zudem prüfen, dass die zweite Ableitung positiv definit ist. Die Ableitung von g — hier in Form des Gradienten — berechnet sich als

$$\nabla g(\theta) = 2X^T X\theta - 2X^T y.$$

Falls $N > K$ ist (was der Normalfall ist) und die Matrix X vollen Rang besitzt, ist $X^T X$ invertierbar und wir können $\theta^\star$ als

$$\theta^\star = \underbrace{(X^T X)^{-1} X^T}_{=:X^\dagger} y \tag{11.6}$$

schreiben. Hierbei heißt $X^\dagger$ die *Moore-Penrose Pseudoinverse* von X. Für konkrete Daten wird man allerdings nicht die Inverse $(X^T X)^{-1}$ berechnen, sondern das lineare Gleichungssystem $X^T X\theta^\star = X^T y$ lösen, weil das schneller geht (deswegen ist dies eine "fast" explizite Lösung: theoretisch kann man die Lösung explizit hinschreiben, praktisch muss man aber noch einen numerischen Algorithmus einsetzen). Die zweite Ableitung lautet $\nabla^2 g(\theta) = 2X^T X$, was eine positiv definite Matrix ist, falls X vollen Rang besitzt. Die Methode lässt sich auf den Fall erweitern, dass X nicht vollen Rang besitzt, allerdings entsteht X ja aus den Daten t_i und den Funktionen f_i, so dass man durch deren geeignete Wahl immer sicherstellen kann, dass X vollen Rang besitzt.

Interessant ist, dass die Lösung $\theta^\star$ gar nicht von der Varianz σ^2 der ε_i abhängt. Dies folgt aus der speziellen Form der Dichtefunktion der Gaußverteilung und ist bei anderen Verteilungen anders. Man kann aber genau wie das optimale $\theta^\star$ auch das optimale $\sigma^\star$ durch Maximierung von l berechnen[1] Damit kommt man auf

$$(\sigma^\star)^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - x_i\theta^*)^2.$$

Dies ist genau die Größe, die durch die Wahl von $\theta^\star$ minimiert wird. Für das Aufstellen von $f(t;\theta^*)$ ist diese Größe unerheblich. Sie gibt uns aber Informationen darüber, wie gut f zu den fehlerbehafteten Daten y_i passt. Ein großes σ^* bedeutet, dass entweder die Messfehler groß sind oder f nicht gut zu den exakten Daten passt (oder beides zugleich). Weil außerdem $(y_i - x_i\theta^*)^2 = (f(t_i;\theta^\star) - y_i)^2$ gilt, ist $(\sigma^\star)^2$ gerade die empirische Varianz $\frac{1}{N}\sum_{i=1}^{N}(f(t_i;\theta^\star) - y_i)^2$ der Abweichung von f von den Messwerten, die also von $\theta^\star$ minimiert wird. Daher wird die lineare Regression mit Gauß-verteilten Messfehlern auch die "kleinste Quadrate Methode" (engl. least squares method) genannt.

Besitzen die Gleichungssystems (11.2) und (11.3) für $\hat{y}_i = y_i - \varepsilon_i$ (d.h. für Messwerte ohne Messfehler) eine exakte Lösung $\hat{\theta}$ (was genau dann der Fall ist, wenn f exakt zu den Daten ohne Messfehler passt), so gilt

$$(\sigma^\star)^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - x_i\theta^*)^2 \leq \frac{1}{N}\sum_{i=1}^{N}(y_i - x_i\hat{\theta})^2 = \frac{1}{N}\sum_{i=1}^{N}\varepsilon_i^2. \tag{11.7}$$

[1]Wir hätten σ^2 auch als weitere Komponente θ_{K+1} in den Vektor θ aufnehmen können. Dann wäre bei der Maximierung über θ gerade $\theta^\star_{K+1} = (\sigma^\star)^2$ herausgekommen.

Sofern die ε_i Erwartungswert 0 besitzen, ist dies nach (8.4) für große N (also viele Daten) näherungsweise gerade die Varianz der Zufallsvariablen, die tatsächlich hinter den Werten ε_i stecken. Diese tatsächliche Varianz beschränkt also die Varianz der Zufallsvariablen E aus unserer Modellannahme.

Im Falle anderer Verteilungen für ε_i als der Gaußverteilung kommt man in Spezialfällen (wie z.B. der Exponential- oder der Poisson-Verteilung) auch noch auf explizite Ausdrücke für $\theta^\star$. Im Allgemeinen muss man bei anderen Verteilungen aber auf einen numerischen Optimierungsalgorithmus zurückgreifen, um $\theta^\star$ zu bestimmen.

Als Beispiel für die Regression betrachten wir die monatlichen Wetterdaten der Station Heinersreuth von Januar 1990 bis Dezember 2022, die unter https://www.wetterkontor.de/de/wetter/deutschland/monatswerte-station.asp?id=P175 abgerufen werden können. Abbildung 11.1 stellt die Daten grafisch dar. Gut ist hier z.B. der heiße Sommer 2004 oder der kalte Winter 2017 zu erkennen.

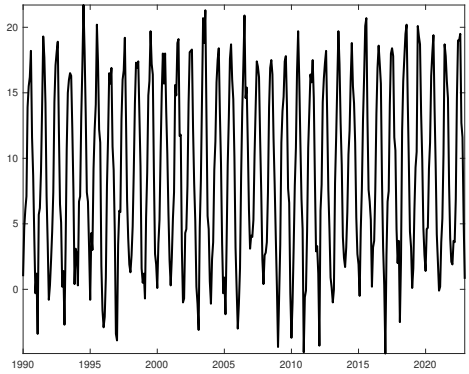


Abbildung 11.1: Monatliche Durchschnittestemperaturen Heinersreuth, Januar 1990 bis Dezember 2022

Nicht direkt zu erkennen ist allerdings die Klimaerwärmung. Durch eine lineare Regression kann man diese sichtbar machen. Wir machen dafür den Ansatz

$$f(t;\theta) = \theta_1 + \theta_2 t$$

mit Gauß-verteilten Messfehlern. Wir nehmen also an, dass die gegebenen Temperaturdaten um eine lineare Funktion in t als Mittelwert streuen und suchen die lineare Funktion, die im Maximum-Likelihood-Sinne am besten zu den Daten passt. Abbildung 11.2 zeigt das Ergebnis, links mit dem gleichen Koordinatenausschnitt wie in Abbildung 11.1, links als Zoom um die resultierende Gerade.

Ein Blick auf die rechte Grafik zeigt, dass die Mitteltemperatur in Heinersreuth in den letzten 32 Jahren tatsächlich um mehr als 0,9°C gestiegen ist.

Wenn man dieses Beispiel genauer ansieht, stellt man fest, dass die bisher gemachten Voraussetzungen nicht wirklich erfüllt sind, denn die deutlich sichtbaren jahreszeitlichen Schwankungen sind sicherlich nicht Gauß-verteilt. Eine aufwändigere Analyse zeigt aber, dass periodische Schwankungen durch die lineare Regression weitgehend “ausgemittelt”

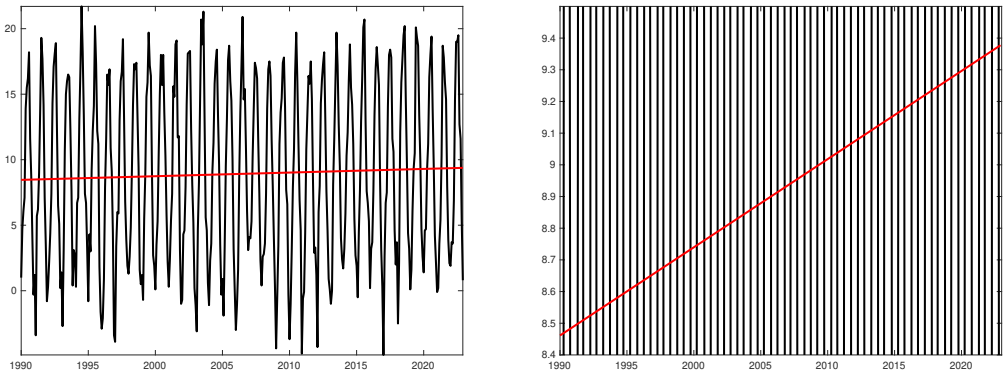


Abbildung 11.2: Lineare Regression zu den Wetterdaten aus Abbildung 11.1 mit Gerade (links Gesamtansicht, rechts Zoom)

werden und es reicht, wenn die verbleibenden Schwankungen Gauß-verteilt sind. Tatsächlich ergibt eine Wiederholung der linearen Regression mit Jahresmittelwerten, die nun keinerlei saisonalen Schwankungen mehr unterliegen, eine fast identische Regressionsgerade, wie Abbildung 11.3 zeigt. Periodische Schwankungen stören die lineare Regression also nur unwesentlich. Für eine stochastisch fundierte Analyse sollte man aber vorzugsweise die jährlich gemittelten Werte aus Abbildung 11.3 verwenden.

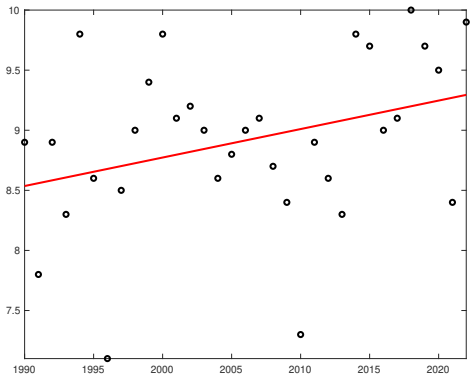


Abbildung 11.3: Jährliche Durchschnittstemperaturen Heinersreuth, 1990 bis 2022, mit Regressionsgerade

Interessant ist es, die Regression (mit den jährlich gemittelten Daten) mit einer Parabel statt einer Geraden durchzuführen, also mit dem Ansatz

$$f(t;\theta) = \theta_1 + \theta_2 t + \theta_3 t^2.$$

Das Ergebnis ist in Abbildung 11.4 zu sehen.

Hier ist deutlich zu erkennen, dass die mittlere Temperatur immer schneller ansteigt.

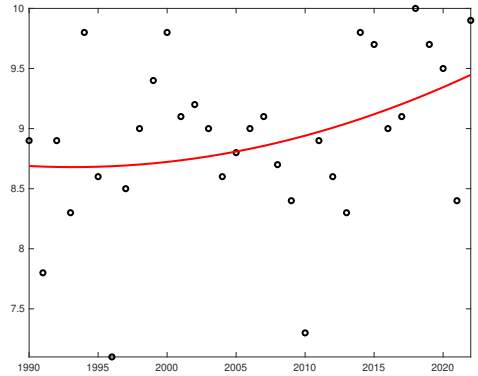


Abbildung 11.4: Lineare Regression zu den Wetterdaten aus Abbildung 11.3 mit Parabel

11.5 Regularisierung

In der Praxis möchte man sicher stellen, dass die im Vektor $\theta^\star$ gesammelten Parameter sich in sinnvollen Wertebereichen befinden. Einträge von $\theta^\star$ mit sehr großem Betrag lassen darauf schließen, dass die Funktion "mit Gewalt" an die Daten angepasst wurde, ein Effekt, den man mit "Overfitting" bezeichtet. Er tritt üblicherweise auf, wenn die Anzahl der Parameter in θ groß ist (was auch erklärt, dass wir den Effekt in der Temperaturregression nicht gesehen haben, da wir dort nur 2 bzw. 3 Parameter haben).

Ein üblicher Ansatz, um diesen Effekt zu vermeiden, ist, dass statt dem Ausdruck $\|X\theta - y\|_2^2$ der Ausdruck

$$\|X\theta - y\|_2^2 + \mu\|\theta\|_2^2 \tag{11.8}$$

minimiert wird, wobei $\mu > 0$ ein Gewichtsparameter ist (statt $\|\cdot\|_2^2$ können auch andere Ausdrücke mit Normen verwendet werden). Dieses Vorgehen nennt man *Tikhonov-Regularisierung*. Intuitiv ist klar, dass dieser Ansatz betragsmäßig sehr große Einträge von θ vermeidet (je größer μ ist, um so mehr), allerdings scheint er nicht so recht zur Herleitung des Minimierungsproblems mit der Maximum Likelihood Methode zu passen.

Das stimmt aber nicht, denn wir können die Annahme, dass θ keine betragsmäßig großen Einträge besitzt, auch stochastisch formulieren, indem wir annehmen, dass die θ_j Realisierungen von (von E_i unabhängigen) Zufallsvariablen Θ_j sind. Z.B. können wir annehmen, dass $\Theta_j \sim N(0, \rho^2)$ gilt. Dies wäre im Bayes'schen Sinne (vgl. Abschnitt 4.3) gerade die Vorabinformation (also die prior distribution) über Θ_j. Statt die Dichtefunktion der gemeinsamen Verteilung der Y_i wollen wir nun die Dichtefunktion der gemeinsamen Verteilung der Θ_j bedingt auf $Y_i = y_i$ über θ maximieren. Dieser Ansatz wird *Maximum A-posteriori Probability* (MAP) Methode genannt.

Um diese Dichtefunktion zu bestimmen, benötigen wir eine Formel für bedingte Dichtefunktionen. Diese ist gegeben durch

$$f_\Theta(\theta \,|\, y) = \frac{f_{(Y,\Theta)}(y,\theta)}{\hat{f}_Y(y)}$$

wobei $f_{(Y,\Theta)}$ die Dichtefunktion der gemeinsamen Verteilung von $\Theta = (\Theta_1, \dots, \Theta_K)^T$ und $Y = (Y_1, \dots, Y_N)^T$ bezeichnet und $\hat{f}_Y$ die Dichtefunktion der sogenannten Randverteilung von Y ist. Wie diese genau aussieht, ist tatsächlich nicht wichtig, da sie nicht von θ abhängt. Wenn wir nur den Zähler des Bruchs, also die gemeinsame Verteilungsfunktion von Θ und Y maximieren, erhalten wir also das gleiche θ^*. Wir müssen dafür nun noch einen Ausdruck für $f_{(Y,\Theta)}$ herleiten.

Diesen können wir erhalten, weil E_i und Θ_j per Annahme unabhängig voneinander sind. Daher ist die gemeinsame Dichtefuntion von (E_i, Θ_j) gegeben durch das Produkt

$$f_{E_i}(\varepsilon_i) f_{\Theta_j}(\theta_j).$$

Wegen $\varepsilon_i = y_i - x_i\theta$ ergibt sich die gemeinsame Dichtefunktion von (Y_i, Θ_j) also als

$$f_{E_i}(y_i - x_i\theta) f_{\Theta_j}(\theta_j).$$

Weil die Y_i und die E_j auch untereinander unabhängig sind, erhalten wir die gemeinsame Dichtefunktion von (Y, Θ) dann einfach durch Aufmultiplizieren, also als

$$\underbrace{\left(\prod_{i=1}^{N} f_{E_i}(y_i - x_i\theta)\right)}_{=L(\theta,y)} \left(\prod_{j=1}^{K} f_{\Theta_j}(\theta_j)\right)$$

mit L aus (11.5). Bezeichnen wir das rechte Produkt mit $g(\theta)$, dann ergibt sich in Gauß'schen Fall

$$L(\theta, y)g(\theta) = \left(\prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y_i - x_i\theta}{\sigma}\right)^2\right)\right)\left(\prod_{j=1}^{K} \frac{1}{\rho\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\theta_j}{\rho}\right)^2\right)\right).$$

Logarithmieren liefert analog zur Rechnung in Abschnitt 11.4

$$\begin{aligned}\ell(\theta; y, X) &= \underbrace{N \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) + K \log\left(\frac{1}{\rho\sqrt{2\pi}}\right)}_{=:C} - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - x_i\theta)^2 - \frac{1}{2\rho^2}\sum_{j=1}^{n}\theta_j^2 \\ &= C - \frac{1}{2\sigma^2}\|X\theta - y\|_2^2 - \frac{1}{2\rho^2}\|\theta\|_2^2.\end{aligned}$$

Da C wiederum unabhängig von θ ist, ist das Maximieren dieses Ausdrucks äquivalent zum Minimieren von (11.8) mit $\mu = \sigma^2/\rho^2$. Die Tikhonov-Regularisierung resultiert also aus dem MAP Ansatz, d.h. aus einer Art Bayes'schen Maximum Likelihood Ansatz, bei dem der Prior gerade die Annahme macht, dass die Einträge von θ Gauß-verteilt sind.

11.6 Anwendung: Dynamic Mode Decomposition

Betrachtet man eine Familie von Zufallsvariablen $X_t : \Omega \to \mathbb{R}^n$, die von der Zeit $t \in \mathbb{T}$ abhängen, so spricht man von einem *stochastischen Prozess*. Die Menge $\mathbb{T}$ kann dabei unterschiedliche Formen annehmen. Wir betrachten hier der Einfachkeit halber $\mathbb{T} =$

$\{0, 1, 2, \ldots, T\}$. Eine Realisierung $X_t(\omega)$, $t \in \mathbb{T}$ dieser Zufallsvariablen wird *Pfad* genannt. Auf stochastische Prozesse kann man alle Methoden anwenden, die man auch auf "einfache" Zufallsvariablen anwenden kann. Besonders interessiert ist man aber oft daran, Gesetzmäßigkeiten zu finden, die beschreiben, wie sich die Zufallsvariablen über die Zeit verändern bzw. wie sie über die Zeit hinweg zusammenhängen. Typische Fragen sind z.B.

- Sind alle X_t identisch verteilt? In diesem Fall nennt man X_t einen *stationären Prozess*
- Was sind die Kovarianzen und Korrelationen zwischen X_t und X_s für verschiedene Zeiten? Mann nennt diese Größen die *Autokovarianz* und *Autokorrelation*.
- Gibt es ein einfaches Gesetz, mit dem man X_{t+1} aus X_t, X_{t-1}, …, X_0 berechnen kann. Kann man X_{t+1} eventuell nur auf Basis von X_t und einer von X_t unabhängigen Zufallsgröße ε_t berechnen, ohne $X_{t-1}, \ldots, X_0$ zu Hilfe nehmen zu müssen? In diesem letzten Fall nennt man X_t einen *Markovprozess*.

Dynamic Mode Decomposition benutzt die lineare Regression, um eine Regel für die zeitliche Entwicklung von Markov Prozessen aus gegebenen Daten zu ermitteln. Gegeben sind hier Datenvektoren $v_t \in \mathbb{R}^n$ zu Zeiten $t \in \mathbb{T}$. Die Annahme ist nun, dass es eine Matrix $A \in \mathbb{R}^{n\times n}$ gibt, so dass

$$v_{t+1} = Av_t + \varepsilon_t$$

gilt für $t = 0, \ldots, T-1$, mit (nun vektorwertigen) Messfehlern $\varepsilon_t = (\varepsilon_{t,1}, \ldots, \varepsilon_{t,n})^T$. Bezeichnen wir mit $v_{t,i}$ den i-ten Eintrag des t-ten Vektors und mit a_i die i-te Zeile der Matrix A, so kann man die obige Gleichung ausführlich schreiben als

$$v_{t+1,i} = a_i v_t + \varepsilon_{t,i}, \quad i = 1, \ldots, n$$

Schreibt man dies für ein festes i für alle Zeiten $t = 0, \ldots, T-1$ auf, so ergibt sich

$$\begin{aligned} v_{1,i} &= a_i v_0 + \varepsilon_{0,i} \\ \vdots &\ \vdots\ \vdots \\ v_{T,i} &= a_i v_{T-1} + \varepsilon_{T-1,i}. \end{aligned}$$

Dies ist bis auf Transponieren der a_i und $v_{t,i}$ ein Problem der Form (11.3), also ein lineares Regressionsproblem in Standardform mit a_i^T an Stelle von θ. Löst man das Problem nun für $i = 1, \ldots, n$, so erhält man Werte für alle Zeilen a_i der Matrix A und damit für die Matrix A. Schreibt man die Vektoren v_t nebeneinander in Matrizen $V = (v_0, \ldots, v_{T-1})$ und $V' = (v_1, \ldots, v_M)$, so ergibt sich aus der Lösungsformel (11.6) für $\theta^\star$ die Formel

$$A = V'V^\dagger.$$

Die Gleichung $x_{t+1} = Ax_t$ ist dann die (im Sinne der Maximum-Likelihood Schätzung) beste Näherung für das dynamischen System, das die Daten v_t erzeugt hat. Setzt man $x_0 = v_0$, so gilt $v_t \approx x_t = Ax_{t-1} = A^2x_{t-2} = \ldots = A^tx_0 = A^tv_0$. Allerdings braucht man für diese Methode i.A. sehr hochdimensionale Datenvektoren, also großes n, weswegen auch A sehr hochdimensional ist, was weitere Rechnungen mit dem Näherungssystem erschwert.

Zur Abhilfe kann man die die Vektoren v_t, $t \in \mathbb{T}$ bestimmende Dynamik in ihre wesentlichen Bestandteile — die sogenannten *Moden* — zerlegen (daher der Name "Dynamic Mode

Decomposition" für diese Methode). Man berechnet dazu die Eigenwerte $\lambda_k \in \mathbb{C}$ und die zugehörigen Eigenvektoren $u_k \in \mathbb{C}^n$, also die Werte und Vektoren für die

$$Au_k = \lambda_k u_k$$

gilt. Schreibt man dann $v_0 = \sum_{k=1}^{n} b_k u_k$, so erhält man

$$v_t \approx A^t v_0 = \sum_{k=1}^{n} b_k \lambda_k^t u_k.$$

Man sieht hier, dass die Summanden für $|\lambda_k| < 1$ für größer werdende t immer weniger zur Summe beitragen (weil $|\lambda_k^t| = |\lambda_k|^t$ für $t \to \infty$ gegen 0 konvergiert) und daher vernachlässigt werden können. Wenn es nur wenige λ_k mit $|\lambda_k| \geq 1$ gibt, so kann man die in den v_t enthaltene Dynamik mit wenigen Moden näherungsweise mit guter Genauigkeit darstellen und kann das hochdimensionale System durch ein niedrigdimensionales approximieren.

11.7 Ausblick: Nichtlineare Regression und Neuronale Netze

Nicht immer ist es möglich, die Funktion f so zu wählen, dass sie linear von den Parametern θ_j abhängt. Ein Beispiel ist die Modellierung eines Populationswachstums, z.B. einer Bakterienpopulation in einer Laborumgebung oder auch der menschlichen Bevölkerung auf der Erde. Hier gibt es unterschiedliche Modelle, z.B. das exponentielle (= unbeschränkte) und das logistische (= beschränkte) Wachstumsmodell

$$f(t;\theta) = \theta_1 \exp(\theta_2 t), \qquad f(t;\theta) = \frac{\theta_3}{1 + \left(\frac{\theta_3}{\theta_1} - 1\right) \exp(-\theta_2 t)}.$$

Hier kann man analog zum linearen Fall einen Maximum-Likelihood-Ansatz machen und kommt dann, je nach Wahl der Verteilungen, auf Maximierungs- oder Minimierungsprobleme, die nun aber nichtlinear sind und auch für Gauß-verteilte ε_i i.A. keine analytisch berechenbare Lösung mehr haben. Man muss also numerische Lösungsverfahren verwenden und eine Möglichkeit dafür ist das Gauß-Newton-Verfahren, das in der Vorlesung "Numerische Mathematik für Informatik, Ingenieur- und Naturwissenschaften" behandelt wird. Dies minimiert Ausdrücke der Form

$$\left\| \begin{pmatrix} f(t_1;\theta) - y_1 \\ \vdots \\ f(t_N;\theta) - y_N \end{pmatrix} \right\|_2^2 \tag{11.9}$$

über θ.

Der gerade beschriebene Ansatz setzt voraus, dass auf Grund von theoretischen Überlegungen — also auf Grund einer mathematischen Modellierung — eine Idee vorliegt, wie die Funktion f aussehen könnte. Das ist aber nicht immer der Fall und wenn man nicht weiß,

welches f mit welchen Parametern zu den gegebenen Daten (t_i, y_i) passt, kann man sogenannte *universelle Approximatoren* verwenden. Eine derzeit sehr beliebte Klasse universeller Approximatoren sind die *Neuronalen Netze*, speziell in ihrer Form als tiefe Neuronale Netze. Wir betrachten hier als Beispiel die spezielle Klasse der sogenannten Feedforward Netze.

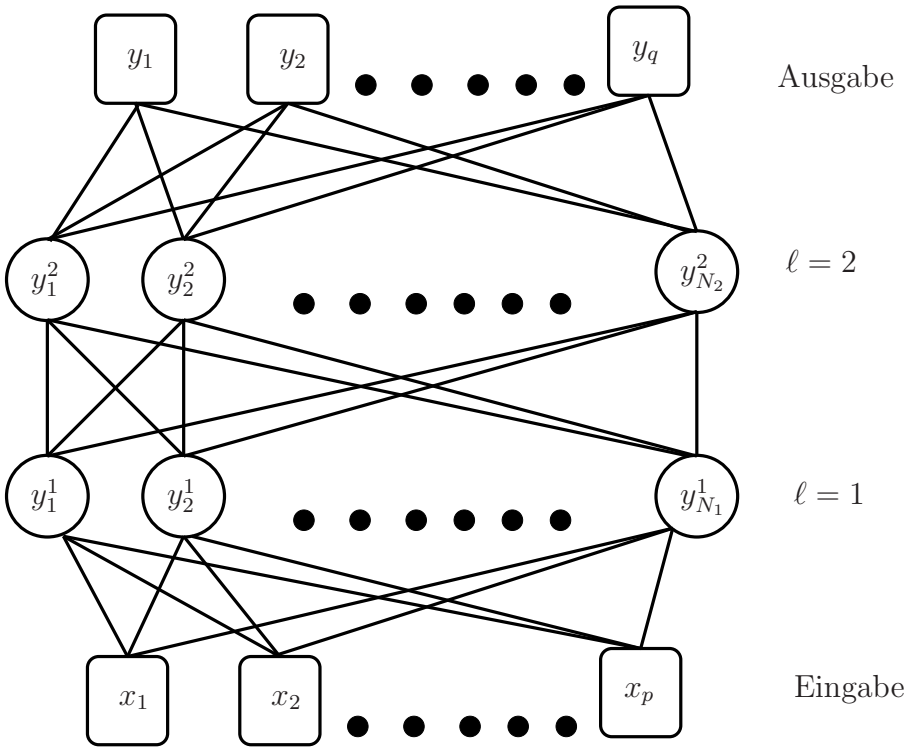


Abbildung 11.5: Schematische eines Feedforward Neuronalen Netzes mit 2 verdeckten Schichten

Ein Feedforward Neuronales Netz besteht aus

- einer Anzahl p an skalaren Eingabewerten — in Abbildung (11.5) die p Werte x_1 bis x_p
- einer Anzahl L an verdeckten Schichten von Neuronen, die ebenfalls wieder skalare Werte enthalten — in Abbildung (11.5) $L = 2$ Schichten für $\ell = 1$ und $\ell = 2$ mit den Werten y_1^1 bis $y_{N_1}^1$ in der Schicht für $\ell = 1$ und y_1^2 bis $y_{N_2}^2$ in der Schicht für $\ell = 2$
- einer Anzahl von q skalaren Ausgabewerten — in Abbildung (11.5) die q Werte y_1 bis y_p.

Die *Tiefe* ist gerade durch die Anzahl L der Schichten gegeben. Die Werte der Neuronen in einer jeden Schicht berechnen sich für ein Netz mit L verdeckten Schichten aus den Werten der vorhergehenden (in Abbildung (11.5) also der darunterliegenden) Schicht gemäß der

Regel

$$y_i^1 = \sigma_1\left(\sum_{k=1}^{p} a_k^{i,1} x_k + b_i^1\right)$$

für die erste Schicht,

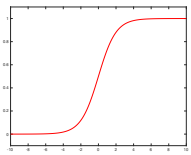
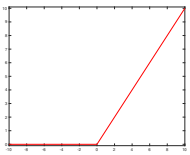
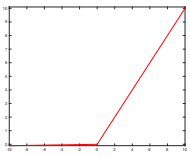
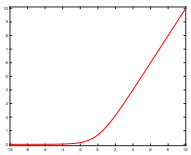
$$y_i^\ell = \sigma_\ell\left(\sum_{k=1}^{N_{\ell-1}} a_k^{i,\ell} y_k^{\ell-1} + b_i^\ell\right)$$

für alle weiteren verdeckten Schichten und

$$y_i = \sum_{k=1}^{N_L} a_k^i y_k^L + b_i$$

für aus Ausgabewerte. Die Funktionen $\sigma_\ell : \mathbb{R} \to \mathbb{R}$, $\ell = 1, \ldots, L$ sind die sogenannten *Aktivierungsfunktionen*. Typische Beispiele hierfür sind

Sigmoid	ReLU	leaky ReLU	Softplus
$\sigma(x) = \frac{1}{1+e^{-x}}$	$\sigma(x) = \begin{cases} x, & x \geq 0 \\ 0, & x < 0 \end{cases}$	$\sigma(x) = \begin{cases} x, & x \geq 0 \\ \varepsilon x, & x < 0 \end{cases}$	$\sigma(x) = \log(1 + e^x)$

Ein typischer Wert für ε in der leaky ReLU Funktion ist $\varepsilon = 0.01$. Die Werte $a_k^{i,\ell}$ bzw. a_k^i werden *Gewichte* des Neuronalen Netzes und die Werte b_i^ℓ bzw. b_i *Bias-Terme*. Diese stellen zusammen die Parameter des Neuronalen Netzes dar, die bestimmen, was für einen Ausgabevektor y ein Eingabevektor x bewirkt.

Letztendlich ist ein Neuronales Netz also wieder nichts anderes als eine Funktion der Form $y = f(x;\theta)$, wobei $x = (x_1, \ldots, x_p)$ (an Stelle des t in Abschnitt 11.1) nun gerade die p Eingabewerte des Netzes und $y = (y_1, \ldots, y_q)$ die q Ausgabewerte sind[2] und $\theta \in \mathbb{R}^N$ der Vektor ist, der alle Gewichte $a_k^{i,\ell}$ und a_k^i und alle Bias-Terme b_i^l und b_i des Netzes in einem großen Vektor zusammenfasst. Der gesuchte Parametervektor $\theta^\star$ wird nun wieder durch Lösen eines Optimierungproblems gefunden. Dies kann z.B. wieder die Minimierung von (11.9) (mit Daten y^i and Stelle von t_i) sein, es können aber auch viel kompliziertere sogenannte *Verlustfunktionen* minimiert werden, bei denen z.B. auch große Gewichte und Bias-Terme "bestraft" werden. Das Finden des optimalen $\theta^\star$ wird bei einem Neuronalen Netz auch "Trainieren des Netzes" genannt. Die dabei verwendeten Daten (x^i, y^i), $i = 1, \ldots, N$, nennt man dann "Trainingsdaten".

Bei der numerischen Berechnung der optimalen θ ist es wichtig, die Ableitung von f nach den Komponenten von θ leicht ausrechnen zu können. Einer der großen Vorteile der Neuronalen Netze gegenüber anderen universellen Approximatoren ist, dass es auf Grund der

[2] Beachte, dass auch die Aufgabenstellung in Abschnitt 11.1 und in der nichtlinearen Regression mit wenig Zusatzaufwand auf vektorwertige Probleme erweitert werden können. Der Unterschied bei den Neuronalen Netzen ist also nicht, dass x und y nun mehrdimensional sind sondern dass man keine Vorabinformationen aus einer Modellierung zum Aufstellen von f benötigt.

Netzstruktur einen effizienten rekursiven Algorithmus (die sogenannte *Backpropagation*) gibt, mit dem dies durchgeführt werden kann und der selbst bei sehr großen Netzen sehr schnell ist.

Da bei dem Ansatz über Neuronale Netze i.A. viel mehr Parameter im Vektor θ vorhanden sind als bei der klassischen linearen oder nichtlinearen Regression mit modelliertem f, benötigt man allerdings i.A. viel mehr Daten (x^i, y^i) als im klassischen Ansatz. Dadurch wird die Optimierung aber aufwändig und man muss geeignete Algorithmen verwenden (siehe den dritten Punkt in der nachfolgenden Liste).

Die stochastische Analyse ist bei Neuronalen Netzen nun an verschiedenen Stellen wichtig:

- Wie beim Maximum-Likelihood Ansatz kann man stochastische Methoden zur Herleitung von geeigneten *Verlustfunktionen* zur Bestimmung von $\theta^\star$ mittels Optimierung verwenden.
- Wie bei (11.7) kann man die Güte der Approximation für große Datenmengen mit stochastischen Methoden analysieren.
- Da der Rechenaufwand für die Optimierung von θ bei sehr großen Datenmengen zu groß wird, verwendet man in der Praxis sogenannte *stochastische Gradientenverfahren*, bei denen in jedem Rechenschritt nur eine zufällige Teilmenge der Daten (ein sogenannter *Minibatch*) für die Verbesserung von θ verwendet wird. Um sicherzustellen, dass der Algorithmus trotzdem (zumindest mit hoher Wahrscheinlichkeit) ein gutes $\theta^\star$ berechnet, werden wiederum stochastische Analysemethoden verwendet.

Wir führen dies in dieser einführenden Vorlesung nicht durch und verweisen dafür auf weitergehende, spezialisierte Lehrveranstaltungen.

Kapitel 12

Hauptkomponentenanalyse

Die *Hauptkomponentenanalyse* (englisch *principal component analysis* oder kurz *PCA*, in der Signalverarbeitung auch als *Karhunen-Loève-Transformation* bezeichnet) wurde von Karl Pearson im Jahr 1901 erfunden. Sie ist zunächst eine Methode, um die Struktur von Datensätzen zu erkennen. Sie hat vielfältige Anwendungen, von denen wir zwei in den letzten Abschnitten dieses Kapitels besprechen.

12.1 Idee der Methode

Im Gegensatz zur Regression, bei der die Daten immer in Paaren angeordnet waren, kommen die Daten nun als eine einzige (aber vektorwertige!) Größe, also x^i, $i = 1, \ldots, N$ mit $x^i \in \mathbb{R}^d$. Beispielsweise könnten die Komponenten von x^i Daten von den N Teilnehmenden dieser Vorlesung enthalten, wie Größe, Gewicht und Schuhgröße — in diesem Fall wäre jedes x_i ein dreidimensionaler Vektor. Für $d = 2$ und $d = 3$ kann man diese Daten als Punkte im $\mathbb{R}^2$ bzw. $\mathbb{R}^3$ darstellen. Abbildung 12.1 zeigt Beispieldaten in 2d und 3d.

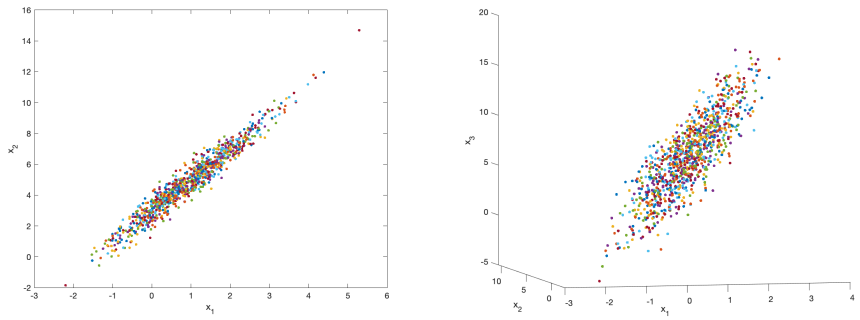


Abbildung 12.1: Daten in 2d und 3d

Ganz offensichtlich haben diese Daten eine erkennbare Struktur: Sie liegen nicht einfach "irgendwie" in der Ebene bzw. im Raum, sondern in der Nähe von Geraden bzw. Ebenen. Ziel der Hauptkomponentenanalyse ist es nun, die Vektoren, die die Geraden bzw. Ebenen (bzw. höherdimensionale Unterräume), entlang derer die Datensätze sich "vorrangig"

ausrichten, zu erkennen. Diese Vektoren sind die sogenannten *Hauptkomponenten*. Dabei wird insbesondere auch die Größe der Ausdehnung in Richtung der einzelnen Vektoren berechnet, womit man wichtige und unwichtigere Hauptkomponenten unterscheiden kann.

12.2 Spezialfall: Diagonale Kovarianzmatrix

Wie üblich modellieren wir den Sachverhalt so, dass die Komponenten x_j^i der Daten durch unabhängige und identisch verteilte Zufallsvariaben $X_j^i \sim X_j$ erzeugt werden und schreiben $X = (X_1, \ldots, X_d)^T$ für die vektorwertige Zufallsvariable. Wir nehmen zunächst einmal an, dass $\mathbb{E}[X] = 0$ und die Kovarianzmatrix $C = \mathrm{Cov}(X) \in \mathbb{R}^{d\times d}$ eine Diagonalmatrix ist, also

$$C = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_d \end{pmatrix}$$

mit $\lambda_k > 0$.

Der Werte von λ_j gibt nun gerade an, wie stark die Werte X^i in Richtung der x_j-Achse streuen. Genauer ist der erwartete quadrierte Abstand der Werte X^i von der x_j-Achse gerade gleich λ_j. Die Koordinatenachsen bilden in diesem Fall die Hauptkomponenten. Ist nun ein λ_j viel größer als alle anderen, so werden die Punkte X_j^i entlang der Geraden durch die 0 entlang x_j liegen. Sind zwei λ_j und λ_k viel größer als alle anderen, so werden die Punkte X^i nahe der Ebene liegen, die durch die Achsen x_j und x_k aufgespannt wird. Analog ergeben mehr als zwei große Eigenwerte einen entsprechend höherdimensionalen Unterraum, um den die Daten streuen. An den großen λ_j erkennt man also die Hauptkomponenten, um die die Daten streuen. Wir nennen diese im Folgenden die *dominierenden Hauptkomponenten*. Falls $E[X] \neq 0$ ist, funktioniert das genau so, wenn wir die Gerade, Ebene etc. nicht durch die 0 sondern durch $E[X]$ legen. Nach wie vor bestimmen die Hauptkomponenten die Richtung, für $E[X] \neq 0$ werden diese nur entsprechend verschoben. Wo genau man nun die Grenze für die dominierenden Hauptkomponenten zieht, d.h. welches λ_j man noch als (relativ) groß genug ansieht, um die zugehörige Koordinatenrichtung x_j zu den dominierenden Hauptkomponenten hinzuzunehmen, hängt wesentlich von der Anwendung und der Verteilung der Werte der λ_j ab. Oft gibt es sehr wenige λ_j, die sichtbar größer sind als die anderen, so dass sich die Aufteilung ganz natürlich ergibt. Wenn die λ_j ungefähr gleich groß sind, ist die Methode nicht sinnvoll anwendbar. In dieser Situation, die in Abbildung 12.2 in 2d dargestellt ist, gibt es aber auch keine ausgezeichnete Richtung, entlang der sich die Daten ausrichten. Die grundsätzliche Idee der Methode ist also bereits nicht anwendbar. Im Kontext der Dimensionsreduktion besprechen wir ein mathematisch präziseres Kriterium in Abschnitt 12.5.

12.3 Koordinatentransformation

Dieses direkte Ablesen der Hauptkomponenten aus C funktioniert leider nicht, wenn C keine Diagonalmatrix ist. Der Trick liegt jetzt darin, den Raum $\mathbb{R}^d$ durch eine lineare

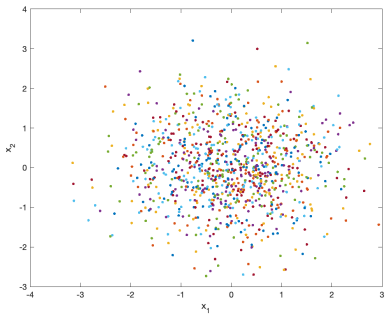


Abbildung 12.2: 2d Daten mit diagonaler Kovarianzmatrix $\mathrm{Cov}(X)$ und $\lambda_1 \approx \lambda_2$

Koordinatentransformation so zu transformieren, dass $C = \mathrm{Cov}(X) \in \mathbb{R}^{d\times d}$ in eine Diagonalmatrix umgewandelt wird. Wie in Abschnitt 10.4 gezeigt, können wir dies mit der Matrix $T = (v_1, v_2, \ldots, v_d)$ erreichen, die aus den Eigenvektoren v_i zu den Eigenwerten λ_i besteht, denn es gilt

$$T^{-1}CT = \Lambda = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_d \end{pmatrix}.$$

T ist also gerade die gesuchte Koordinatentransformation, die C in Diagonalform transformiert. Die Hauptkomponenten sind dann gerade die zu den größten Eigenwerten gehörigen Eigenvektoren v_k, bzw. die davon erzeugten und durch $E[X]$ gehenden Geraden, Ebenen oder höherdimensionalen Unterräume. Für die Frage, welche λ_k genau zu den dominierenden Hauptkomponenten gehören, gilt das am Ende von Abschnitt 12.2 Gesagte analog. In der Praxis nummeriert man die Eigenwerte λ_k und die zugehörigen v_k absteigend nach der Größe der λ_k (so dass λ_1 also der größte Eigenwert ist), so dass die Wichtigkeit der Hauptkomponenten mit k abnimmt.

12.4 Praktische Berechnung

In der praktischen Rechnung hat man natürlich keine abstrakten Zufallsvariablen X_j sondern konkrete Daten x_j^i, $i = 1, \ldots, N$. Die exakte Kovarianzmatrix wird dann durch die empirische Kovarianzmatrix ersetzt. Dazu berechnet man zuerst die empirischen Erwartungswerte

$$\mu_j = \frac{1}{N}\sum_{i=1}^{N} x_j^i$$

und damit die zentrierten Daten $\tilde{x}_j^i = x_j^i - \mu_j$. Die daraus resultierenden Vektoren $\tilde{x}^i = (\tilde{x}_1^i, \ldots, \tilde{x}_d^i)^T$ schreibt man nun in eine Matrix $\widetilde{X} = (\tilde{x}^1, \ldots, \tilde{x}^N) \in \mathbb{R}^{d\times N}$. Die Schätzung der Kovarianzmatrix ist dann bis auf Skalierung mit N nach Beispiel 10.1 gegeben durch $\widetilde{C} = \widetilde{X}\widetilde{X}^T$. Die Skalierung spielt hier allerdings keine Rolle, da es im Folgenden nur auf das Verhältnis der Eigenwerte λ_k zueinander ankommt, die sich durch eine andere Skalierung

nicht ändert. Mit dieser können wir nun λ_k und v_k empirisch — also aus den Daten — berechnen. Abbildung 12.3 zeigt die so berechneten Hauptkomponenten für die Daten aus dem einführenden Beispiel. Hier ergeben sich die Eigenwerte $\lambda_1 \approx 4845,5$, $\lambda_2 \approx 53,2$ bzw. $\lambda_1 \approx 17256$, $\lambda_2 \approx 1300$, $\lambda_3 \approx 48$. Die Länge der Vektoren v_k wurde hierbei mit $\sqrt{\lambda_k}/10$ skaliert, um die Größe der Eigenwerte mit abbilden zu können. Man erkennt leicht, dass die Richtung zusammen mit der Länge der Vektoren sehr gut zu der Streuung der Daten korrespondiert.

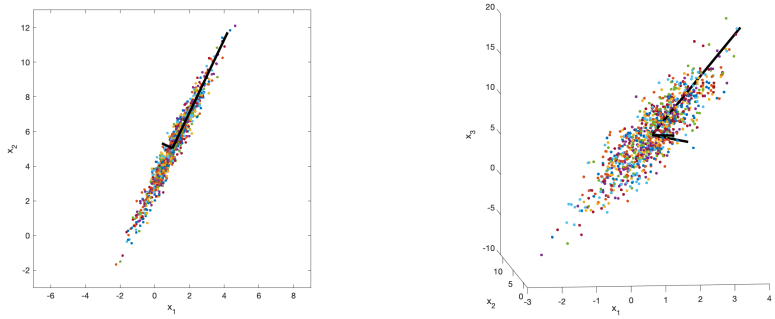


Abbildung 12.3: Daten in 2d und 3d mit Hauptkomponenten, skaliert mit $\sqrt{\lambda_k}/10$

Die Eigenwerte von $\widetilde{X}\widetilde{X}^T$ werden auch als die *Singulärwerte* von $\widetilde{X}^T$ genannt. Algorithmisch kann es effizienter sein, einen Algorithmus zur Berechnung der Singulärwerte von X^T zu verwenden als einen Algorithmus zur Berechnung der Eigenwerte von XX^T.

12.5 Anwendung: Dimensionsreduktion

In diesem Abschnitt werden wir die erste von zwei Anwendungen der Hauptkomponentenanalyse besprechen, die **Dimensionsreduktion**. Statt die d-dimensionalen Daten x^i, $i = 1, \ldots, N$ zu analysieren, kann man zunächst die Daten $y^i = T^{-1}(x^i - \mu)$ betrachten. Die quadrierten k-ten Komponenten dieser transformierten Datenvektoren sind im Mittel gleich λ_k, weswegen die zu (relativ)[1] kleinen λ_k gehörenden Daten Einträge (relativ) nahe bei 0 haben, die daher ohne großen Fehler auf 0 gesetzt werden bzw. einfach weggelassen werden können. Es reicht also, die zu den dominierenden Hauptkomponenten gehörigen Einträge von v^i zu speichern. Statistische Analysen, die auf diesen Daten durchgeführt werden, werden dann sehr ähnliche Ergebnisse liefern wie bei den vollständigen Daten, können aber wegen der geringeren Dimension viel schneller durchgeführt werden. Auch für grafische Darstellungen hochdimensionaler Daten wird dieses Vorgehen gerne genutzt. Die Transformation $y^i = T^{-1}(x^i - \mu)$ transformiert die Daten dabei gerade in die Koordinaten, für die die Erwartungswerte und die Kovarianzen Null sind, also gerade in die in Abschnitt 12.2 besprochene Situation. In Abbildung 12.4 ist dies exemplarisch für das zweidimensionale Beispiel gemacht. Hier sieht man gut, dass die y^i_2-Komponenten alle nahe bei 0 liegen.

[1] "Relativ" ist hier im Vergleich zu den großen λ_k zu verstehen.

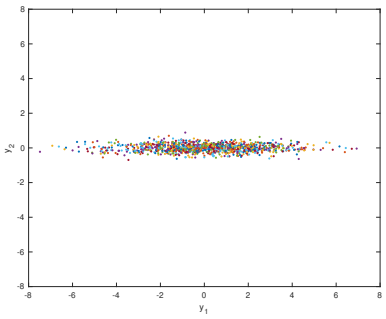


Abbildung 12.4: Transformierte Daten y^i in 2d

Um ein Kriterium festzulegen, nach dem man die dominierenden Hauptkomponenten festlegt, kann man sich ihren Beitrag zur Gesamtvarianz der Daten anschauen. Wir gehen hierzu noch einmal zurück zu den Zufallsvariablen X^i, die in unserem Modell der Verteilung der Daten zu Grunde liegen. Weil T und T^{-1} orthogonale Matrizen sind, gilt nach Abschnitt 10.4 mit $Y = T^{-1}(X - \mathbb{E}[X])$

$$\sum_{j=1}^{d} \mathbb{V}[X_j] = \mathbb{E}(\|X - \mathbb{E}[X]\|_2^2) = \mathbb{E}(\|T^{-1}(X - \mathbb{E}[X])\|_2^2) = \mathbb{E}(\|Y\|_2^2) = \sum_{j=1}^{d} \mathbb{V}[Y_j] = \sum_{j=1}^{d} \lambda_j,$$

weil die Diagonalmatrix Λ ja gerade die Varianzen der Y_j als Einträge besitzt. Die Summe der weggelassenen λ_j bestimmt also die Abnahme der Varianz. Möchte man also z.B. eine Reduktion haben, die 95% der Varianz erhält, so sollten die dominierenden Hauptkomponenten $v_1, \ldots, v_{k^*}$ so gewählt werden, dass $\sum_{j=1}^{k^*} \lambda_j \geq 0.95 \sum_{j=1}^{d} \lambda_j$ gilt.

Ein Nachteil dieses Verfahren ist, dass die Hauptkomponenten nicht direkt zu den in den Daten vorhandenen Merkmalen korrespondieren. Auch wenn man vielleicht nur 2 von 10 dominierenden Hauptkomponenten benötigt, so weiß man noch nicht, wie viele von den ursprünglichen Komponenten redundant sind. Unter Umständen kann man aber gewisse Komponenten der Originaldaten als unwesentlich identifizieren. Jede Komponente von y_j^i von y^i berechnet sich ja mittels $y_j^i = \hat{v}_j x^i$, wobei $\hat{v}_j$ die j-te Zeile von T^{-1} ist. Sind nun für ein gegebenes k die k-ten Einträge der $\hat{v}_j$, die zu den dominierenden Hauptkomponenten gehören, alle nahe bei 0, dann trägt die k-te Komponente der Vektoren x^i zu den dominierenden Hauptkomponenten nur sehr wenig bei und kann daher vernachlässigt werden, ohne dass sich die Varianz der Daten stark ändert.

12.6 Anwendung: Clusteranalyse

Die zweite Anwendung, die wir in diesem Abschnitt erläutern, ist die **Clusteranalyse**. Hier geht es darum, zu erkennen, ob die gegebene Datenmenge aus zwei (oder mehr) deutlich getrennten Teilmengen, den sogenannten *Clustern*, besteht. Ein Beispiel für eine solche Menge in 2d ist in Abbildung 12.5 zu sehen.

Da die Streuung der Daten in Richtung der Cluster (also von links unten nach rechts oben) offenbar am größten ist, sollte die Hauptkomponentenanalyse die Clusterrichtung als

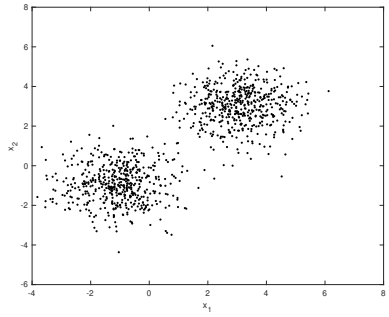


Abbildung 12.5: 2d Daten mit zwei Clustern

dominierende Hauptkomponente finden. Die in Abbildung 12.4 dargestellte Transformation sollte diese Richtung also auf die x_1-Achse transformieren, so dass die Cluster nach der Transformation nebeneinanderliegen. Abbildung 12.6 zeigt, dass genau dies passiert.

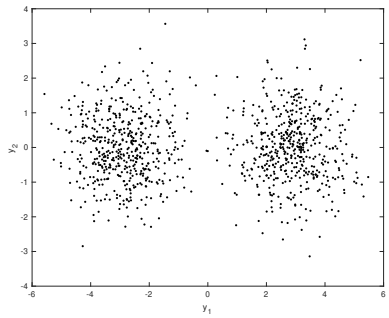


Abbildung 12.6: 2d Daten mit zwei Clustern, transformiert mittels V^{-1}

Nun ist es leicht, die Cluster algorithmisch zu unterscheiden: Die Punkte im linken Cluster haben eine negative x_1-Koordinate, die im rechten Cluster eine positive. Nutzt man diese Information, um die Punkt in den Originalkoordinaten einzufärben, so erhält man die farblich markierten Cluster in Abbildung 12.7.

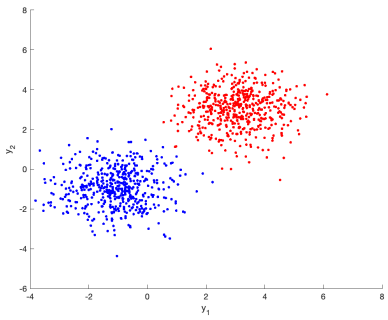


Abbildung 12.7: 2d Daten mit zwei mittels Hauptkomponentenanalyse erkannten und eingefärbten Clustern

Literaturverzeichnis

[1] H. Harbrecht and M. Multerer, *Algorithmische Mathematik*, Springer Berlin Heidelberg, 2022.