Skript zur Vorlesung

Wahrscheinlichkeit und Statistik

Departement Mathematik
ETH Zürich

13. Februar 2021

Inhaltsverzeichnis

0 Einleitung

Das Ziel dieser Vorlesung ist es, eine Einführung in die Wahrscheinlichkeitstheorie und Statistik zu geben. Diese zwei Gebiete zusammen werden als *Stochastik* bezeichnet. Die Stochastik ist die Lehre von der mathematischen Beschreibung und Untersuchung zufälliger Phänomene. Diese treten entweder in natürlicher Weise auf oder liefern eine adäquate idealisierte Beschreibung von sehr komplexen Situationen oder Vorgängen. Beispiele sind die Ankunft bestimmter Aufgaben in einem grossen Computernetzwerk, die Entwicklung des Wetters, das Wählerverhalten bei politischen Entscheidungen, die Entwicklung von Aktienkursen usw. Die Stochastik gibt einem mathematische Hilfsmittel, um solche und ähnliche Fragestellungen quantitativ zu erfassen und zu untersuchen.

Die zwei Teilgebiete der Stochastik ergänzen sich gegenseitig und bauen dabei in asymmetrischer Art aufeinander auf. In der *Wahrscheinlichkeitstheorie* konstruiert man mathematische Modelle für die interessierenden Fragestellungen und versucht, innerhalb dieser Modelle Aussagen über das Verhalten des betrachteten Systems zu machen. In der *Statistik* geht es hingegen darum, anhand von beobachteten Daten ein wahrscheinlichkeitstheoretisches Modell (inklusive allfälliger Parameter) zu identifizieren, das zu den gegebenen Daten passt.

Beispiel 1 (gezinkter Würfel). Eine Urne enthält gleich viele gewöhnliche und gezinkte Würfel; bei den letzteren ist die 6 durch eine 7 ersetzt. Ich ziehe zufällig einen Würfel und würfle damit. Bei jedem Wurf teile ich Ihnen nur mit, ob die gewürfelte Zahl gerade oder ungerade ist.

In der *Wahrscheinlichkeitstheorie* sollten Sie beispielsweise folgende Fragen beantworten können:

W-i) Mit welcher Wahrscheinlichkeit kommt eine gerade Zahl?

W-ii) Wie verhält sich die durchschnittliche Anzahl der in n Versuchen gewürfelten geraden Zahlen, z.B. für grosse n?

In der *Statistik* stellen Sie eher die folgenden Fragen:

S-i) Sie haben die Ergebnisse von n Versuchen. Hat der Würfel eine 7 oder nicht?

S-ii) Wie zuverlässig ist die Antwort auf S-i)? Können Sie auf dieser Grundlage eine Wette eingehen, oder brauchen Sie zum Beispiel noch mehr Daten?

$\diamond$

Ziel der Vorlesung ist es, für beide Teilgebiete die wichtigsten Grundlagen und Ideen zu entwickeln und bereitzustellen.

Teil I: Wahrscheinlichkeitstheorie

Ziel: Einführung von Grundbegriffen, Vermittlung von stochastischem Denken.

1 Wahrscheinlichkeiten

Grundidee: *Zufallsexperimente* sind Experimente, deren Ergebnisse nicht immer exakt vorausgesagt werden können. Wir möchten dafür ein mathematisches Modell.

Diese Art von Fragestellung taucht sehr oft auf. Manchmal geht es um *künstlich erzeugte* Experimente (Werfen einer Münze; Würfeln; Ziehung bei einer Lotterie), manchmal um *komplexe Phänomene* (Turbulenz; Wetter; Genetik; Kursentwicklungen an der Börse; Epidemien; Warteschlangen z.B. bei der Post oder in Computersystemen; usw.). Oft ist es auch nützlich, deterministische Fragen mit nicht bekannten Charakteristiken oder variablen Inputs als zufällige Experimente aufzufassen (Signalübertragung; Analyse von Sortieralgorithmen; usw.). In jedem Fall möchte man gewisse Fragestellungen quantitativ beschreiben und analysieren.

1.1 Grundbegriffe

Definition. Der *Ereignisraum* oder Grundraum $\Omega \neq \emptyset$ ist die Menge aller möglichen Ergebnisse des betrachteten Zufallsexperiments. Die Elemente $\omega \in \Omega$ heissen *Elementarereignisse* oder Ausgänge des Experiments.

Beispiel 2 (ein Würfelwurf). Beim Werfen eines Würfels ist $\Omega = \{1, 2, \dots, 6\}$. ◇

Beispiel 3 (zwei Münzwürfe). Wirft man eine Münze zweimal, so ist der Grundraum $\Omega = \{KK, KZ, ZK, ZZ\}$. ◇

Beispiel. Betrachtet man zu einem festen Zeitpunkt die Anzahl der Prozesse bei einem Server oder die Anzahl der Druckaufträge in einer Warteschlange, so ist ein möglicher Grundraum $\Omega = \{0, 1, 2, \dots\} = \mathbb{N}_0$. (In der Praxis hat man oft eine bekannte obere Schranke.) ◇

Beispiel. Für die Lebensdauer einer Glühbirne ist $\Omega = \{t : t \geq 0\} = [0, \infty) = \mathbb{R}_+$. ◇

Beispiel. Für die Entwicklung eines Aktienkurses ist eine mögliche Wahl der Funktionenraum $\Omega = \{\text{alle Funktionen: } [0, \infty) \to \mathbb{R}\}$. ◇

Definition. Die *Potenzmenge* von Ω, bezeichnet mit $\mathcal{P}(\Omega)$ oder 2^Ω, ist die Menge aller Teilmengen von Ω. Ein *prinzipielles Ereignis* ist eine Teilmenge $A \subseteq \Omega$, also eine Kollektion von Elementarereignissen. Die Klasse aller *(beobachtbaren) Ereignisse* ist $\mathcal{F}$; das ist eine Teilmenge der Potenzmenge von Ω.

Ist Ω endlich oder abzählbar, so wählt man oft als $\mathcal{F}$ die Potenzmenge 2^Ω; dann ist also jede Teilmenge von Ω ein beobachtbares Ereignis, und die Unterscheidung zwischen prinzipiellen und beobachtbaren Ereignissen wird hinfällig. Man spricht dann von einem diskreten Wahrscheinlichkeitsraum, und von Kapitel 1.2 an werden wir die wichtigsten Ideen und Begriffe zunächst in diesem einfacheren Rahmen entwickeln. Ist Ω überabzählbar (wie in den Beispielen 4 und 5), so muss man als $\mathcal{F}$ eine echte Teilklasse von 2^Ω nehmen; siehe [Williams, 2.3, S.42] für eine ausführliche Diskussion. In jedem Fall muss $\mathcal{F}$ gewisse Axiome erfüllen; es muss eine sogenannte σ-Algebra sein.

Bemerkungen. 1) Ein Mengensystem $\mathcal{F} \subseteq 2^\Omega$ heisst eine *σ-Algebra*, wenn i) $\Omega \in \mathcal{F}$ ist, ii) für jedes $A \in \mathcal{F}$ auch das Komplement $A^c \in \mathcal{F}$ ist , iii) für jede Folge $(A_n)_{n\in\mathbb{N}}$ mit

$A_n \in \mathcal{F}$ für alle $n \in \mathbb{N}$ auch die Vereinigung $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$ ist.

2) Die Potenzmenge 2^Ω von Ω ist offensichtlich eine σ-Algebra.

3) Dasselbe Experiment kann man in der Regel durch verschiedene $(\Omega, \mathcal{F})$ beschreiben, wie das nachfolgende Beispiel illustriert. ◇

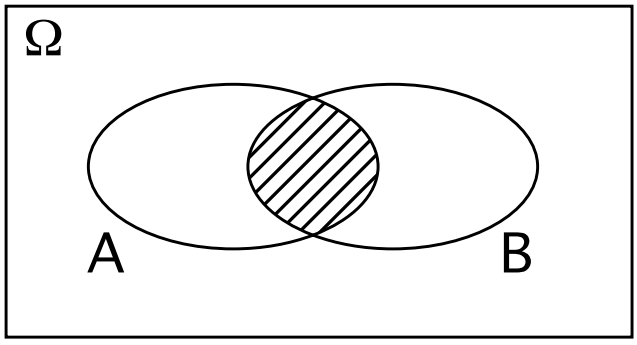
Beispiel. Jemand wirft einen Würfel, aber teilt uns statt der Augenzahl nur mit, ob die gewürfelte Zahl gerade oder ungerade ist.

Eine erste mögliche Beschreibung dieses Experiments wäre der Grundraum $\Omega_1 = \{G, U\}$ und $\mathcal{F}_1 = 2^{\Omega_1} = \{\emptyset, \Omega_1, \{G\}, \{U\}\}$. Dabei stehen offensichtlich G und U für gerade bzw. ungerade.

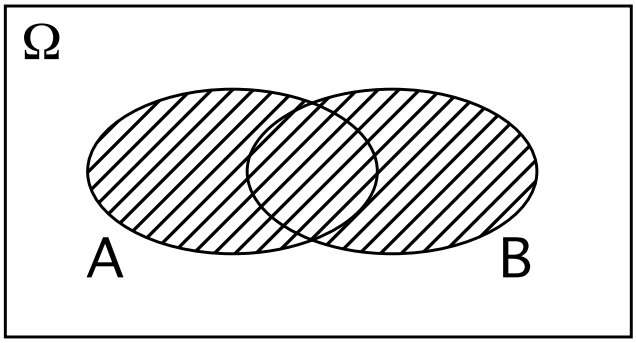
Eine zweite mögliche Beschreibung könnte mit $\Omega_2 = \{1, 2, \ldots, 6\}$ beginnen; dann darf aber nicht $\mathcal{F}_2 = 2^{\Omega_2}$ sein, weil wir die gewürfelte Augenzahl ja nicht beobachten können. Hier müsste man $\mathcal{F}_2 = \{\emptyset, \Omega_2, \{2, 4, 6\}, \{1, 3, 5\}\}$ wählen, um das Experiment aus unserer Sicht zu beschreiben. Man beachte, dass $\mathcal{F}_1$ und $\mathcal{F}_2$ dieselbe Anzahl von Mengen (also von beobachtbaren Ereignissen) enthalten. ◇

Wir gehen allgemein davon aus, dass wir bei unserem Zufallsexperiment jeweils genau einen Ausgang (ein Elementarereignis ω) erhalten. Generell sagen wir, dass das Ereignis A *eintritt*, falls das realisierte Elementarereignis ω in A liegt, d.h. $\omega \in A$. Mit Hilfe von Mengenoperationen können wir dann aus $A, B \in \mathcal{F}$ neue Ereignisse bilden:

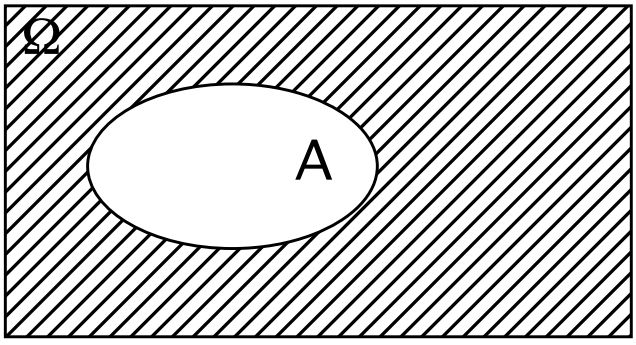
- $A \cap B$ *(Schnittmenge)* ist das Ereignis, dass A *und* B eintreten.



- $A \cup B$ *(Vereinigung)* ist das Ereignis, dass A *oder* B (oder beide) eintreten.



- A^c *(Komplement)* ist das Ereignis, dass A *nicht* eintritt.



Bemerkung. Dank der Axiome i), ii), iii) für eine σ-Algebra $\mathcal{F}$ liegen alle oben gebildeten Mengen auch wieder in $\mathcal{F}$. ◇

Definition. Ein *Wahrscheinlichkeitsmass* ist eine Abbildung $P : \mathcal{F} \to [0, 1]$, welche die nachfolgenden Axiome erfüllt. Für $A \in \mathcal{F}$ nennen wir $P[A] \in [0, 1]$ die Wahrscheinlichkeit (kurz WS), dass A eintritt. Die geforderten Grundregeln (*Axiome*) sind:

A0) $P[A] \geq 0$ für alle Ereignisse $A \in \mathcal{F}$.
(Das steckt implizit schon in $P : \mathcal{F} \to [0, 1]$.)

A1) $P[\Omega] = 1$.

A2) $P\left[\bigcup_{i=1}^{\infty} A_i\right] = \sum_{i=1}^{\infty} P[A_i]$, sofern die $A_i \in \mathcal{F}$ *paarweise disjunkt* sind, d.h. falls gilt $A_i \cap A_k = \emptyset$ für $i \neq k$.

Wir schreiben das kürzer als

$$P\left[\dot{\bigcup_{i=1}^{\infty}} A_i\right] = \sum_{i=1}^{\infty} P[A_i];$$

die Notation $\dot{\bigcup}$ steht dabei allgemein für eine Vereinigung von paarweise disjunkten Mengen.

Bemerkung. Auch A1) und A2) benutzen, dass $\mathcal{F}$ eine σ-Algebra ist, weil ja die Argumente von P Mengen aus $\mathcal{F}$ sein müssen. ◇

Aus den grundlegenden Axiomen A1) und A2) lassen sich weitere *Rechenregeln* herleiten. Falls Ω endlich oder abzählbar mit $\mathcal{F} = 2^{\Omega}$ ist, so wird das oft auf die Betrachtung von Elementarereignissen zurückgeführt. Die Herleitung aus den Axiomen ist aber sogar fast einfacher (und natürlich allgemeiner).

1) $P[A^c] = 1 - P[A]$.

⌜

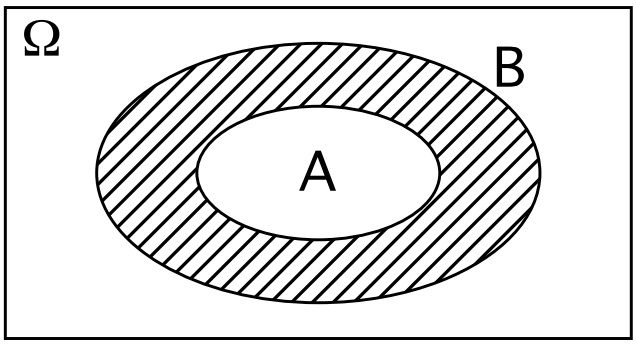
$\Omega = A \dot{\cup} A^c$, d.h. $\Omega = A \cup A^c$ und $A \cap A^c = \emptyset$, liefert nach A1) und A2)

$$1 = P[\Omega] = P[A] + P[A^c].$$

⌟

2) $P[\emptyset] = 0$, denn $\emptyset = \Omega^c$ und $P[\Omega] = 1$.

3) Für $A \subseteq B$ gilt $P[A] \leq P[B]$.



Mit $B \cap A^c = B \setminus A$ *(B minus A)* ist wegen $A \subseteq B$

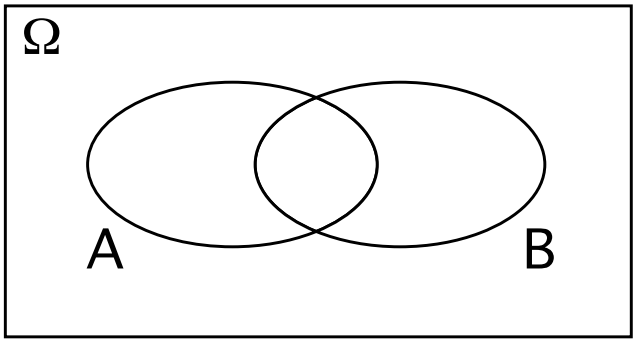
$$B = (B \cap A) \dot\cup (B \cap A^c) = A \dot\cup (B \setminus A),$$

wobei wir wieder $\dot\cup$ für eine disjunkte Vereinigung schreiben. Also ist nach A2)

$$P[B] = P[A] + P[B \setminus A] \geq P[A].$$

4) Für beliebige (nicht unbedingt disjunkte) A, B gilt die allgemeine *Additionsregel*

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



⌜

Wir zerlegen disjunkt wie folgt:

$$A = (A \cap B) \,\dot\cup\, (A \cap B^c) = (A \cap B) \,\dot\cup\, (A \setminus B),$$
$$B = (B \cap A) \,\dot\cup\, (B \cap A^c) = (A \cap B) \,\dot\cup\, (B \setminus A)$$

und

$$\begin{aligned} A \cup B &= \big((A \cup B) \cap (A \cap B)\big) \,\dot\cup\, \big((A \cup B) \cap (A \cap B)^c\big) \\ &= (A \cap B) \,\dot\cup\, (A \cap B^c) \,\dot\cup\, (B \cap A^c) \\ &= (A \cap B) \,\dot\cup\, (A \setminus B) \,\dot\cup\, (B \setminus A). \end{aligned}$$

Aus A2) folgt also $P[A] = P[A \cap B] + P[A \setminus B]$, d.h.

$$P[A \setminus B] = P[A] - P[A \cap B],$$

ebenso

$$P[B \setminus A] = P[B] - P[A \cap B],$$

und die disjunkte Zerlegung von $A \cup B$ gibt dann dank A2)

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

Das liefert die Behauptung. ⌟

Die Interpretation oder Festlegung von Wahrscheinlichkeiten ist eine eher philosophische Frage und hat entsprechend schon zu vielen Diskussionen geführt. Grundsätzlich unterscheidet man dabei

a) die *frequentistische Interpretation*:

Man betrachtet Wiederholungen eines Zufallsexperimentes unter identischen Bedingungen und bestimmt für ein Ereignis $A \in \mathcal{F}$

$$f_n(A) := \frac{1}{n} \# \left(\{ i \in \{1, \ldots, n\} : A \text{ tritt beim Experiment } i \text{ ein} \} \right),$$

also die *relative Häufigkeit* (Frequenz) des Eintretens von A bei n Wiederholungen. Dabei steht $\#(A)$ für die Anzahl der Elemente von A. Dann interpretiert man $P[A]$ als $\lim_{n\to\infty} f_n(A)$, also als die asymptotische relative Häufigkeit bei unendlicher Wiederholung des Experiments. Man sieht einfach, dass $f_n(\,\cdot\,)$ für jedes n die Axiome A0) – A2) erfüllt; für den Limes, sofern er überhaupt existiert, geht das aber in der Regel schief. Siehe dazu [Williams, 1.3, S.25 und 4.3, S.115].

b) die subjektive oder *Bayes'sche Interpretation*:

Hier ist $P[A]$ ein Mass für den Glauben daran, dass A eintreten wird; z.B. hat man dann $P[\text{Kopf bei Münzwurf}] = \frac{1}{2}$ oder $P[\text{es wird morgen regnen}] = \frac{1}{3}$.

Welche der obigen Interpretationen einem besser zusagt, ist letzten Endes eine Frage der persönlichen Einstellung. Wir diskutieren das hier nicht weiter. Bei der Wahl eines konkreten Modells werden wir aber ab und zu auf diesen Punkt zurückkommen.

1.2 Diskrete Wahrscheinlichkeitsräume

In vielen interessanten Fällen, gerade im Zusammenhang mit Anwendungen aus der Informatik, ist $\Omega = \{\omega_1, \omega_2, \ldots, \omega_N\}$ endlich oder $\Omega = \{\omega_1, \omega_2, \ldots\}$ abzählbar und $\mathcal{F} = 2^\Omega$, d.h. jede Teilmenge von Ω ist ein (beobachtbares) Ereignis. In dieser Situation ist die Beschreibung eines Wahrscheinlichkeitsmasses P äquivalent dazu, die Wahrscheinlichkeiten $p_i = P[\{\omega_i\}]$, $i = 1, \ldots, N$ bzw. $i \in \mathbb{N}$, aller Elementarereignisse anzugeben. Für eine beliebige Menge $A \subseteq \Omega$ ist dann nämlich $A \in \mathcal{F}$ und

$$P[A] = P\Bigg[\bigcup_{i \text{ mit } \omega_i \in A} \{\omega_i\}\Bigg] = \sum_{i \text{ mit } \omega_i \in A} P[\{\omega_i\}] = \sum_{i \text{ mit } \omega_i \in A} p_i$$

ebenfalls festgelegt.

Ist $\Omega = \{\omega_1, \ldots, \omega_N\}$ endlich mit $|\Omega| = N$ und $\mathcal{F} = 2^\Omega$, und sind $\omega_1, \ldots, \omega_N$ alle gleich wahrscheinlich, also $p_1 = p_2 = \cdots = p_N = \frac{1}{N}$, so heisst Ω ein *Laplace-Raum* und P die

diskrete Gleichverteilung auf Ω. Für beliebige $A \subseteq \Omega$ ist dann

$$P[A] = \frac{\text{Anzahl der Elementarereignisse in } A}{\text{Anzahl der Elementarereignisse in } \Omega} = \frac{|A|}{|\Omega|} = \frac{\#(A)}{\#(\Omega)}.$$

Hier muss man also im Wesentlichen nur zählen können; das ist allerdings in konkreten Situationen nicht immer ganz einfach. Zur Berechnung von $|A|$ und $|\Omega|$ wird oft *Kombinatorik* gebraucht. Eine kurze Übersicht über die wichtigsten Formeln und Ideen dazu findet sich im Kapitel 10.

Bemerkung. Ist Ω *abzählbar* statt endlich, so existiert eine diskrete Gleichverteilung natürlich nicht mehr. ◇

In den nachfolgenden Beispielen wählen wir immer stillschweigend $\mathcal{F} = 2^\Omega$.

Beispiel 3 (zwei Münzwürfe). Beim zweimaligen Münzwurf ist der Grundraum $\Omega = \{KK, KZ, ZK, ZZ\}$, also $|\Omega| = 4$, und damit $p_i = \frac{1}{4}$, $i = 1, \ldots, 4$, für die diskrete Gleichverteilung. Also ist dann beispielsweise

$$P[\text{mindestens einmal Kopf}] = P[\{KK, KZ, ZK\}] = \frac{3}{4}.$$

◇

Beispiel. Wie gross ist die Wahrscheinlichkeit, dass eine zufällig gewählte dreistellige Zahl wiederholte Ziffern enthält?

Hier ist offenbar $|\Omega| = |\{000, \ldots, 999\}| = 10^3$. Betrachten wir das Ereignis

$$A = \{\text{Zahl enthält wiederholte Ziffern}\},$$

so ist

$$P[A] = 1 - P[A^c] = 1 - \frac{|A^c|}{|\Omega|}$$

und

$$\begin{aligned} |A^c| &= \text{Anzahl der dreistelligen Zahlen mit lauter verschiedenen Ziffern} \\ &\quad (= \text{Anzahl der Permutationen von 3 aus 10 Elementen ohne Zurücklegen}) \\ &\quad = 10 \times 9 \times 8 = 720. \end{aligned}$$

Also ist $P[A] = 1 - 0.72 = 0.28$. ◇

Bemerkung. Falls man unter dreistelligen Zahlen nur solche verstehen will, die nicht mit Nullen beginnen, so ist $\Omega = \{100, \ldots, 999\}$. Man sieht völlig analog, dass die Wahrscheinlichkeit für wiederholte Ziffern aber auch hier 0.28 beträgt (was a priori nicht selbstverständlich ist). ◇

Beispiel 4 (Geburtstage). Das *Geburtstagsproblem* ist eines der bekannten Probleme aus der Wahrscheinlichkeitstheorie mit überraschenden Antworten. Seine Formulierung (der klassische Teil ist a)) ist wie folgt:

Ein Raum enthält n Personen.

a) Wie gross ist die Wahrscheinlichkeit, dass mindestens zwei davon am gleichen Tag Geburtstag haben?

b) Wie gross muss n sein, damit diese Wahrscheinlichkeit $> \frac{1}{2}$ ist?

Die Antwort auf b) ist überraschend: es braucht nur gerade 23 Personen! Um das zu sehen, müssen wir zuerst a) beantworten.

Sei t_i der Geburtstag der i-ten Person. Ignorieren wir der Einfachheit halber Schaltjahre, so ist $t_i \in \{1, \ldots, 365\}$, $i = 1, \ldots, n$, und ein Elementarereignis ω ist eine Folge $\omega = (t_1, \ldots, t_n)$ der Länge n. Wir nehmen an, alle diese Folgen seien gleich wahrscheinlich, so dass die Geburtstage über das ganze Jahr gleichverteilt sind. Dann ist $|\Omega| = 365^n$, und mit

$$A = \{\text{mindestens zwei Personen haben am gleichen Tag Geburtstag}\}$$

ist

$$\begin{aligned} |A^c| &= |\{\text{alle Personen haben verschiedene Geburtstage}\}| \\ &= 365 \times 364 \times \cdots \times (365 - n + 1) \\ &= \frac{365\,!}{(365-n)\,!} \\ &= \binom{365}{n} n\,!, \end{aligned}$$

denn um die Elementarereignisse in A^c zu erzeugen, muss man aus den insgesamt 365 Geburtstagen zuerst ohne "Zurücklegen" insgesamt n auswählen und kann diese n anschliessend noch beliebig permutieren. Also ist

$$P[A] = 1 - P[A^c] = 1 - \frac{365\,!}{365^n \times (365-n)\,!} = 1 - \frac{365}{365} \times \cdots \times \frac{365-n+1}{365}.$$

Durch Einsetzen erhält man beispielsweise

$$\begin{aligned} &\text{für } n = 23 : P[A] = 0.5073, \\ &\text{für } n = 41 : P[A] = 0.9032, \\ &\text{für } n = 61 : P[A] = 0.9951, \\ &\text{für } n = 91 : P[A] = 0.999995, \\ &\text{für } n = 130 : P[A] = 1, \text{ bis auf 10 Nachkommastellen.} \end{aligned}$$

Mit analogen Argumenten können wir nun weitere Fragen beantworten.

c) Wie gross ist die Wahrscheinlichkeit, dass heute jemand Geburtstag hat?

d) Wie gross muss n sein, damit diese Wahrscheinlichkeit $> \frac{1}{2}$ ist?

Mit $B = \{\text{eine Person hat heute Geburtstag}\}$ ist

$$|B^c| = |\{\text{niemand hat heute Geburtstag}\}| = 364^n$$

und damit

$$P[B] = 1 - P[B^c] = 1 - \left(\frac{364}{365}\right)^n.$$

Durch Einsetzen erhält man

$$\begin{aligned}
&\text{für } n = 23 : P[B] = 0.0612,\\
&\text{für } n = 100 : P[B] = 0.2399,\\
&\text{für } n = 300 : P[B] = 0.5609,\\
&\text{für } n = 365 : P[B] = 0.6326.
\end{aligned}$$

Das beantwortet c). Für d) braucht man $P[B] > \frac{1}{2}$, also $1 - (\frac{364}{365})^n > \frac{1}{2}$ oder $n \log \frac{364}{365} < \log \frac{1}{2} = -\log 2$, also

$$n > \frac{-\log 2}{\log \frac{364}{365}} = 252.6.$$

Man braucht also mindestens $n = 253$ Personen; dann ist $P[B] = 0.5005$.

e) Wie gross ist die Wahrscheinlichkeit, dass jemand am gleichen Tag wie der ETH-Präsident Geburtstag hat?

Wie in c) bedeutet das, dass der Geburtstag auf ein bestimmtes Datum (statt heute) fallen muss; also ist die Wahrscheinlichkeit dieselbe wie in c), nämlich $1 - (\frac{364}{365})^n$.

f) Wie gross ist die Wahrscheinlichkeit, dass jemand heute oder am gleichen Tag wie der ETH-Präsident Geburtstag hat?

Nennen wir das betreffende Ereignis C, so ist (unter der Annahme, dass der ETH-Präsident nicht gerade heute Geburtstag hat) $|C^c| = 363^n$ und damit

$$P[C] = 1 - \left(\frac{363}{365}\right)^n.$$

Einsetzen gibt

$$\begin{aligned}
&\text{für } n = 100\text{: } P[C] = 0.4227,\\
&\text{für } n = 253\text{: } P[C] = 0.7510,\\
&\text{für } n = 365\text{: } P[C] = 0.8654.
\end{aligned}$$

Die Unterschiede zu c) oder e) sind also erstaunlich gross.

◇

Beispiel. Beim *Lotto* in der Schweiz werden aus 42 Zahlen 6 gezogen. Das Zufallsexperiment besteht hier also aus einer solchen Ziehung. Ein Elementarereignis ist eine Menge $\{z_1, \ldots, z_6\}$ von 6 verschiedenen Zahlen $z_i \in \{1, 2, \ldots, 42\}$, $i = 1, \ldots, 6$; es ist eine Menge und (im Gegensatz zu Beispiel 4) nicht eine Folge, weil die Reihenfolge der Zahlen keine Rolle spielt. Damit ist $|\Omega|$ die Anzahl der sogenannten *Kombinationen* von 6 aus 42 Elementen *ohne Zurücklegen*, also

$$|\Omega| = \binom{42}{6} = 5'245'786.$$

[Zum Vergleich: In Deutschland werden 6 aus 49 Zahlen gezogen; also haben wir dort $|\Omega| = \binom{49}{6} = 13'983'816$.]

Nehmen wir (im Gegensatz zur Realität) einmal an, dass man einzelne Tipps abgeben kann (in der Praxis muss man mindestens zweimal 6 Zahlen ankreuzen). Wie gross ist dann die Wahrscheinlichkeit

a) für 6 Richtige?

b) für einen Vierer, d.h. genau 4 richtige Zahlen?

Wir nennen diese Ereignisse A_6 bzw. A_4. Alle möglichen Ziehungen werden als gleich wahrscheinlich angenommen; also ist

$$P[A_6] = \frac{|A_6|}{|\Omega|} = \frac{1}{|\Omega|} = \frac{1}{5'245'786} = 1.906 \times 10^{-7}.$$

Für A_4 müssen wir aus den 6 gezogenen Zahlen durch Tippen 4 erwischen und aus den restlichen 36 die übrigen 2; also ist

$$|A_4| = \binom{6}{4} \times \binom{36}{2} = 9450$$

und

$$P[A_4] = \frac{9450}{5'245'786} = 1.801 \times 10^{-3}.$$

[Die entsprechenden Ergebnisse für das Lotto in Deutschland sind $P[B_6] = 7.15 \times 10^{-8}$ und $P[B_4] = 9.69 \times 10^{-4}$.] ◇

Bemerkung. Die Wahrscheinlichkeiten für alle möglichen Trefferzahlen finden sich im Kapitel 3 als Beispiel zur hypergeometrischen Verteilung. ◇

1.3 Bedingte Wahrscheinlichkeiten

Definition. Seien A, B Ereignisse und $P[A] > 0$. Die *bedingte Wahrscheinlichkeit von B* unter der Bedingung, dass A eintritt (kurz: *gegeben A*) wird definiert durch

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

Beispiel 2 (ein Würfelwurf). Beim Würfeln mit einem Würfel sei

$$A = \{\text{gerade Augenzahl}\} = \{2, 4, 6\},$$
$$B = \{\text{Augenzahl} > 3\} = \{4, 5, 6\}.$$

Dann ist $A \cap B = \{4, 6\}$ und

$$P[B \,|\, A] = \frac{P[B \cap A]}{P[A]} = \frac{\frac{2}{6}}{\frac{3}{6}} = \frac{2}{3}.$$

Veranschaulichung: Mein Kollege würfelt und sagt mir, dass A eingetreten sei, d.h. dass die gewürfelte Zahl gerade ist. Mit dieser zusätzlichen Information berechne ich $P[B \,|\, A]$, die bedingte Wahrscheinlichkeit, dass auch B eingetreten ist, d.h. dass die Augenzahl auch noch über 3 liegt. ◇

Bemerkungen. 1) In der Regel ist $P[B \,|\, A] \neq P[B]$; das sieht man schon im obigen Beispiel, wo offenbar $P[B] = \frac{1}{2}$ ist. Das bedeutet anschaulich, dass A tatsächlich Information über B liefert.

2) Bei fixierter Bedingung A sind die bedingten Wahrscheinlichkeiten $P[\,\cdot\,|\,A]$ wieder Wahrscheinlichkeiten auf $(\Omega, \mathcal{F})$. Man sieht durch Nachrechnen nämlich sehr leicht, dass $P^*[\,\cdot\,] := P[\,\cdot\,|\,A]$ die Axiome A0) – A2) erfüllt, und damit ist P^* wieder ein Wahrscheinlichkeitsmass.

3) Die bedingte Wahrscheinlichkeit ist nicht symmetrisch in den beiden Argumenten. Insbesondere ist bei fixiertem Ereignis B die Funktion $A \mapsto P[B\,|\,A]$ kein Wahrscheinlichkeitsmass.

4) Ist Ω endlich oder abzählbar mit $\mathcal{F} = 2^\Omega$, so ist P gegeben durch die Gewichte $p_i = P[\{\omega_i\}]$. Das bedingte Wahrscheinlichkeitsmass $P^*[\,\cdot\,] := P[\,\cdot\,|\,A]$ hat dann die Gewichte

$$p_i^* = P^*[\{\omega_i\}] = P[\{\omega_i\}\,|\,A] = \begin{cases} \frac{p_i}{P[A]} & \text{für } \omega_i \in A, \\ 0 & \text{für } \omega_i \in A^c, \end{cases}$$

wie man durch Einsetzen sofort sieht. Anschaulich heisst das also, dass wir alle Gewichte ausserhalb von A auf Null setzen und anschliessend alle Gewichte in A mit einem festen Faktor so skalieren, dass ihre Summe wieder 1 ergibt. ◇

Direkt aus der Definition der bedingten Wahrscheinlichkeit erhält man (mit einer beliebigen Festlegung von $P[B\,|\,A]$ für $P[A] = 0$ sowie der Konvention $x \times 0 = 0$ für alle x) die sogenannte *Multiplikationsregel*: Für beliebige Ereignisse A, B ist

$$P[A \cap B] = P[B\,|\,A]P[A].$$

(Diese ergänzt die schon bekannte Additionsregel

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

Mit Hilfe der Multiplikationsregel kann man oft Wahrscheinlichkeiten auf einfache Weise in mehreren Schritten berechnen.

Beispiel 5 (Urne). Eine Urne enthält 4 Kugeln, nämlich 3 rote und 1 blaue. Wie gross ist die Wahrscheinlichkeit, beim Ziehen von 2 Kugeln (ohne Zurücklegen) 2 rote zu erwischen?

Sei $R_i = \{i\text{-te gezogene Kugel ist rot}\}$ für $i = 1, 2$. Dann ist

$$P[R_1 \cap R_2] = P[R_1]P[R_2\,|\,R_1] = \frac{3}{4} \times \frac{2}{3} = \frac{1}{2}.$$

Diese Überlegung ist wesentlich einfacher als ein Abzählen aller möglichen und günstigen Ziehungen für das Gesamtexperiment. (Zudem müsste man sich strenggenommen noch überlegen, dass das die gleiche Situation beschreibt wie die Gleichverteilung auf der Menge aller Ziehungen.) ◇

Beispiel 4 (Geburtstage). In unserem schon bekannten Raum mit n Personen suchen wir die Wahrscheinlichkeit, dass niemand heute oder morgen Geburtstag hat. Das ist natürlich nur eine Variation (des Komplements) von f) in Kapitel 1.2, wird hier aber etwas anders gelöst.

Mit $A_1 = \{\text{niemand hat heute Geburtstag}\}$ und $A_2 = \{\text{niemand hat morgen Geburtstag}\}$ ist

$$P[A_1] = \left(\frac{364}{365}\right)^n, \qquad P[A_2 \,|\, A_1] = \left(\frac{363}{364}\right)^n$$

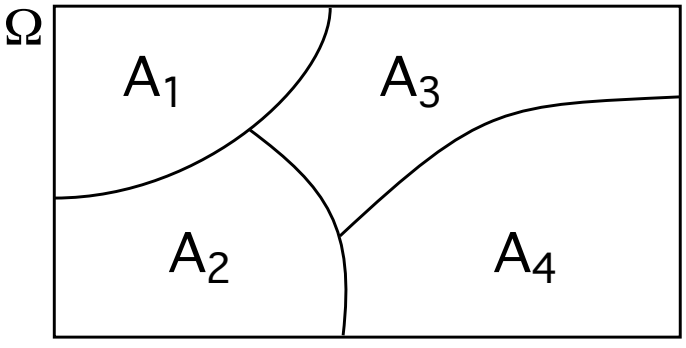
und

$$P[A_1 \cap A_2] = P[A_1]P[A_2 \,|\, A_1] = \left(\frac{363}{365}\right)^n.$$

◇

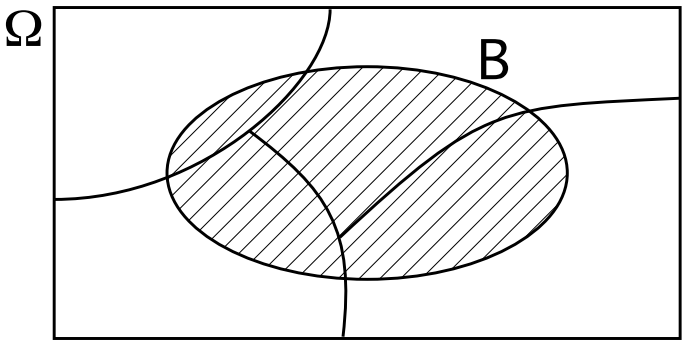
Satz 1.1. (Satz von der totalen Wahrscheinlichkeit) *Sei $A_1, \ldots, A_n$ eine Zerlegung von Ω (in paarweise disjunkte Ereignisse), d.h. $\dot{\bigcup}_{i=1}^n A_i = \Omega$ oder explizit*

$$\bigcup_{i=1}^{n} A_i = \Omega,$$
$$A_i \cap A_k = \emptyset \qquad \text{für } i \neq k.$$



Für beliebige Ereignisse B gilt dann

$$P[B] = \sum_{i=1}^{n} P[B \mid A_i] P[A_i].$$



(Dieses Resultat gilt auch für eine abzählbare Zerlegung $(A_i)_{i\in\mathbb{N}}$; die Summe geht dann über $i = 1$ bis ∞.)

Beweis. Mit elementarer Mengentheorie ist

$$B = B \cap \Omega = B \cap \left(\bigcup_{i=1}^{n} A_i \right) = \bigcup_{i=1}^{n} (B \cap A_i)$$

und

$$(B \cap A_i) \cap (B \cap A_k) = B \cap (A_i \cap A_k) = \emptyset \qquad \text{für } i \neq k.$$

Das bedeutet, dass die $B \cap A_i$ eine disjunkte Zerlegung von B bilden. Also liefern Axiom A2) und die Multiplikationsregel

$$P[B] = P\left[\dot{\bigcup_{i=1}^{n}} (B \cap A_i) \right] = \sum_{i=1}^{n} P[B \cap A_i] = \sum_{i=1}^{n} P[B \mid A_i] P[A_i].$$

q.e.d.

Der Nutzen von Satz 1.1 ist, dass manchmal die Berechnung von $P[A_i]$ und $P[B \mid A_i]$ einfacher ist als eine direkte Berechnung von $P[B]$.

Beispiel 1 (gezinkter Würfel). Eine Urne enthält gleich viele gewöhnliche und gezinkte Würfel; bei den letzteren ist die 6 durch eine 7 ersetzt. Man zieht zufällig einen Würfel und

würfelt damit. Wie gross ist die Wahrscheinlichkeit, dass dabei eine gerade Zahl gewürfelt wird?

Sei $Z = \{\text{gezogener Würfel ist gezinkt}\}$ und $G = \{\text{gewürfelte Zahl ist gerade}\}$. Dann ist $P[Z] = P[Z^c] = \frac{1}{2}$, weil die Urne gleich viele gewöhnliche und gezinkte Würfel enthält, und

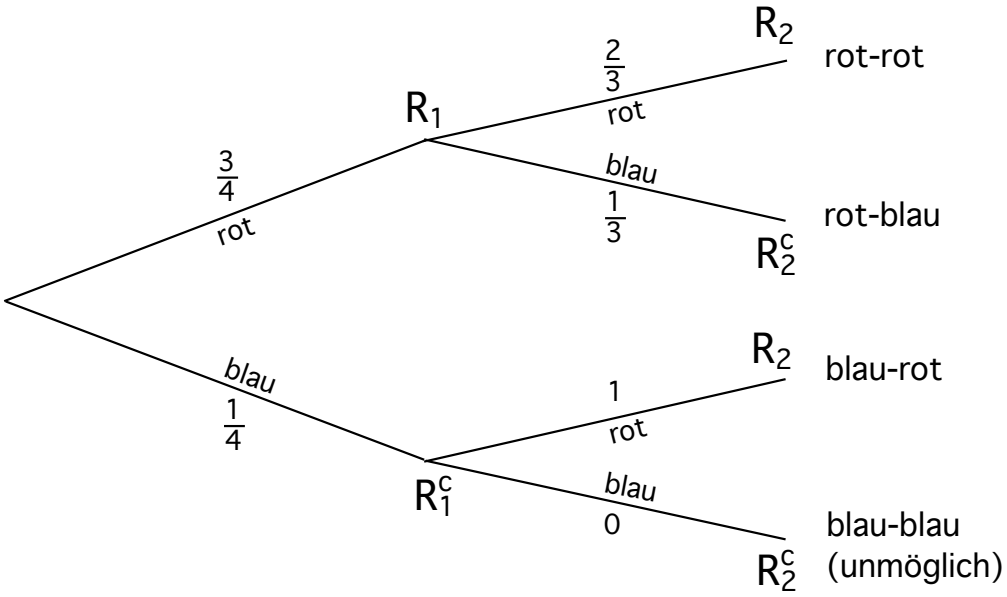
$$P[G \mid Z] = \frac{|\{2,4\}|}{|\{1,2,3,4,5,7\}|} = \frac{1}{3},$$

denn bei einem gezinkten Würfel (den wir ja unter der Bedingung Z betrachten) treten die Zahlen $1, \ldots, 5$ und 7 alle mit gleicher Wahrscheinlichkeit auf. Also ist

$$P[G] = P[Z]P[G \mid Z] + P[Z^c]P[G \mid Z^c] = \frac{1}{2} \times \frac{1}{3} + \frac{1}{2} \times \frac{1}{2} = \frac{1}{6} + \frac{1}{4} = \frac{5}{12}.$$

Insbesondere ist also, wie intuitiv erwartet, die Wahrscheinlichkeit für eine gerade Zahl kleiner als $\frac{1}{2}$. ◇

Überlegungen wie oben oder wie im Satz von der totalen Wahrscheinlichkeit werden oft in einem *Baumdiagramm* dargestellt. Im obigen Beispiel 5 (Urne) mit den 4 Kugeln in der Urne sieht das so aus:



In mehrstufigen Experimenten hat man entsprechend längere Bäume. Diese graphische Darstellung ist wie folgt zu interpretieren. Bei jeder Strecke (an jedem Zweig) steht die *bedingte* Wahrscheinlichkeit, über diesen Weg von einem Knoten zum nächsten zu gelangen (jeweils gegeben, dass man den vorderen Knoten schon erreicht hat). Die Wahrscheinlichkeit für einen gesamten Pfad (= Elementarereignis) von der Wurzel bis zum Baumende ergibt sich nach der Multiplikationsregel durch sukzessives Multiplizieren aller bedingten Wahrscheinlichkeiten entlang des Pfades; und die Wahrscheinlichkeit für ein Ereignis ist nach Axiom A2) die Summe der Wahrscheinlichkeiten aller Pfade (= Elementarereignisse), die zu diesem Ereignis gehören.

Beispiel 5 (Urne). Im obigen Beispiel der Urne mit drei roten und einer blauen Kugel bilden $R_1 = \{\text{erste gezogene Kugel ist rot}\}$ und $R_1^c = \{\text{erste gezogene Kugel ist blau}\}$ offenbar eine Zerlegung von Ω. Für

$$R_2 = \{\text{zweite gezogene Kugel ist rot}\}$$

erhalten wir also

$$P[R_2] = P[R_2 \,|\, R_1]P[R_1] + P[R_2 \,|\, R_1^c]P[R_1^c] = \frac{2}{3} \times \frac{3}{4} + 1 \times \frac{1}{4} = \frac{3}{4}$$

(und damit auch $P[R_2] = P[R_1]$). ◇

Eine häufige Fragestellung betrifft die Berechnung von bedingten Wahrscheinlichkeiten $P[A \,|\, B]$, wenn man umgekehrt bedingte Wahrscheinlichkeiten der Form $P[B \,|\, A]$ kennt. Wir illustrieren das zuerst an zwei Beispielen.

Beispiel 1 (gezinkter Würfel). Eine Urne enthält gleich viele gewöhnliche und gezinkte Würfel; bei den letzteren ist die 6 durch eine 7 ersetzt. Man zieht zufällig einen Würfel und würfelt damit. Dabei beobachten wir eine ungerade Zahl. Wie gross ist dann die (bedingte) Wahrscheinlichkeit, dass der Würfel gezinkt ist?

Sei $Z = \{\text{gezogener Würfel ist gezinkt}\}$ und $G = \{\text{gewürfelte Zahl ist gerade}\}$. Dann haben wir aus den oben gemachten Rechnungen schon $P[Z] = P[Z^c] = \frac{1}{2}$, $P[G \,|\, Z] = \frac{1}{3}$

und $P[G] = \frac{5}{12}$, und das liefert $P[G^c \mid Z] = 1 - P[G \mid Z] = \frac{2}{3}$ und $P[G^c] = 1 - P[G] = \frac{7}{12}$. Gesucht ist

$$P[Z \mid G^c] = \frac{P[Z \cap G^c]}{P[G^c]} = \frac{P[G^c \mid Z]P[Z]}{P[G^c]} = \frac{\frac{2}{3} \times \frac{1}{2}}{\frac{7}{12}} = \frac{12}{21} = \frac{4}{7}.$$

Auch das ist qualitativ einleuchtend: Wir wissen, dass ein gezinkter Würfel mehr ungerade als gerade Zahlen produziert; wenn wir also soeben eine ungerade Zahl beobachtet haben, so glauben wir umgekehrt eher an einen gezinkten als an einen gewöhnlichen Würfel. ◇

Beispiel. Bei einer Krankheitsdiagnose sind die folgenden Angaben bekannt:

- In der gesamten Bevölkerung sind 0.1% krank.
- Von den kranken Personen werden 90% durch die Untersuchung entdeckt.
- Von den gesunden Personen werden 99% durch die Untersuchung als gesund eingestuft.

Nun wird eine Person aus der Bevölkerung herausgegriffen, untersucht und als krank eingestuft. Wie wahrscheinlich ist es, dass das stimmt?

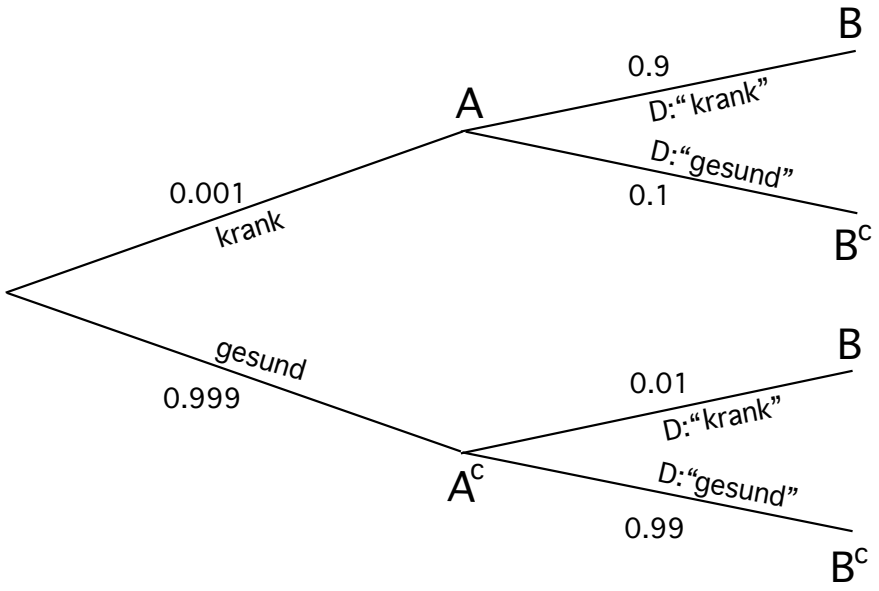
Sei A das Ereignis, dass eine zufällig gewählte Person krank ist, und B das Ereignis, dass die Untersuchung einer zufällig ausgewählten Person die Diagnose "krank" ergibt. Aus den obigen Angaben haben wir dann

$$\begin{aligned} P[A] &= 0.001, \qquad \text{also } P[A^c] = 0.999; \\ P[B \mid A] &= 0.9, \qquad \text{also } P[B^c \mid A] = 0.1; \\ P[B^c \mid A^c] &= 0.99, \qquad \text{also } P[B \mid A^c] = 0.01. \end{aligned}$$

Die Wahrscheinlichkeiten $P[B]$ und $P[B^c]$ kennen wir (noch) nicht.

Gesucht ist nun $P[A \mid B]$, also die *bedingte* Wahrscheinlichkeit, dass unsere Person krank ist, gegeben, dass sie als krank eingestuft wurde.

Graphisch können wir diese Situation wie folgt darstellen; D steht dabei kurz für Diagnose:



Nach der Definition ist $P[A \mid B] = \frac{P[A \cap B]}{P[B]}$. Also erhalten wir nach der Multiplikationsregel

$$P[A \cap B] = P[B \mid A]P[A] = 0.9 \times 0.001 = 0.0009$$

und nach dem Satz von der totalen Wahrscheinlichkeit

$$P[B] = P[B \mid A]P[A] + P[B \mid A^c]P[A^c] = 0.9 \times 0.001 + 0.01 \times 0.999 = 0.01089.$$

Damit ist die gesuchte bedingte Wahrscheinlichkeit gegeben durch

$$P[A \mid B] = \frac{P[A \cap B]}{P[B]} = \frac{0.0009}{0.01089} = 0.0826;$$

also sind nur etwa 8.3% aller als krank diagnostizierten Personen tatsächlich krank! Mit anderen Worten: Eine Diagnose ist bei weitem nicht unfehlbar. ◇

Bemerkung. Verbessert man im obigen Beispiel die Diagnostik von 99% zu 99.9% bzw. zu $P[B \mid A^c] = 0.001$, d.h. zu weniger falschen Krank-Diagnosen bei gesunden Personen, so ergibt sich analog

$$P[A \mid B] = 0.4739,$$

d.h. die Vertrauenswürdigkeit einer Krank-Diagnose steigt massiv — von 8.3% auf 47.4%. Dieses Beispiel illustriert also die Problematik der Fehldiagnosen bei eher seltenen Krankheiten. ◇

In beiden obigen Beispielen haben wir den sogenannten *Satz von Bayes* benutzt:

$$P[A \mid B] = \frac{P[A \cap B]}{P[B]} = \frac{P[B \mid A]P[A]}{P[B \mid A]P[A] + P[B \mid A^c]P[A^c]}.$$

Allgemeiner hat man

Satz 1.2. (Formel von Bayes) *Ist $A_1, \ldots, A_n$ eine Zerlegung von Ω mit $P[A_i] > 0$ für $i = 1, \ldots, n$ und B ein Ereignis mit $P[B] > 0$, so gilt für jedes k*

$$P[A_k \mid B] = \frac{P[B \mid A_k]P[A_k]}{\sum_{i=1}^{n} P[B \mid A_i]P[A_i]}.$$

(Wie Satz 1.1 gilt auch dieses Resultat für eine abzählbare Zerlegung $(A_i)_{i\in\mathbb{N}}$, wobei dann die Summe wieder über $i = 1$ bis ∞ geht.)

Beweis. Wie im Beispiel benutzt man die Definition

$$P[A_k \mid B] = \frac{P[A_k \cap B]}{P[B]}.$$

Im Zähler liefert die Multiplikationsregel

$$P[A_k \cap B] = P[B \mid A_k]P[A_k],$$

im Nenner der Satz von der totalen Wahrscheinlichkeit (Satz 1.1)

$$P[B] = \sum_{i=1}^{n} P[B \mid A_i]P[A_i].$$

q.e.d.

Das letzte Beispiel in diesem Abschnitt soll illustrieren, dass ein Problem oft durchaus auf verschiedene Arten gelöst werden kann. Die Beurteilung, welche der nachfolgend präsentierten Lösungen einfacher, plausibler oder eleganter ist, überlassen wir dem Leser.

Beispiel. Das sogenannte *Ziegenproblem* hat Anfang der 90er Jahre zu heftigen Diskussionen geführt, die teilweise bis in die Literatur hineinreichten. Die amerikanische Journalistin Marilyn vos Savant (mit dem angeblich höchsten IQ der Welt) bekam 1990 von einem Leser für ihre Denksport-Kolumne im "Parade Magazine" die folgende Aufgabe:

Suppose you are on a game show, and you are given the choice of three doors. Behind one door is a car, behind the others, goats. You pick a door, say #1, and the host, who knows what is behind the doors, opens another door, say #3, which has a goat. He says to you, "Do you want to pick door #2?" Is it to your advantage to switch your choice of doors?

Erste Lösung: Wir spezifizieren keinen Wahrscheinlichkeitsraum, sondern nur bedingte Wahrscheinlichkeiten. Sei dazu

$$A := \{\text{der Kandidat hat bei der ersten Wahl die Tür mit dem Auto erwischt}\}$$

und

$$G := \{\text{der Kandidat gewinnt nach Wechseln der Tür}\}.$$

Gesucht ist dann $P[G]$. Nun gilt offenbar $P[G \mid A] = 0$, weil ja der Kandidat durch das Wechseln die Tür mit dem Auto verlässt. Ferner gilt $P[G \mid A^c] = 1$, weil im Fall A^c der Kandidat bei der ersten Wahl eine der zwei Türen mit einer Ziege gewählt hat und der Showmaster danach die andere Tür mit einer Ziege geöffnet hat; durch das Wechseln landet der Kandidat also automatisch bei der Tür mit dem Auto. Schliesslich muss $P[A] = \frac{1}{3}$ sein, weil ja nur eine der drei Türen ein Auto versteckt. Also gibt uns der Satz von der totalen Wahrscheinlichkeit (Satz 1.1)

$$P[G] = P[A]P[G \mid A] + P[A^c]P[G \mid A^c] = \frac{1}{3} \times 0 + \frac{2}{3} \times 1 = \frac{2}{3}.$$

Also lohnt es sich, die Tür zu wechseln.

Zweite Lösung: Die zufälligen Elemente bei diesem Problem sind zum einen, das Auto hinter eine Tür zu stellen, und zum anderen, in Schritt 1 eine Tür auszuwählen. Als Ω kann man also die Menge aller Paare (i, k) mit $i, k \in \{1, 2, 3\}$ wählen und darauf die Gleichverteilung betrachten. Die Entscheidung des Kandidaten nach der Aktion des Showmasters ist nun, ob er die Tür wechselt oder nicht. Dabei gibt es zwei Situationen. Ist $i = k$, so hat der Kandidat schon beim ersten Mal die richtige Tür erwischt und verliert damit beim Wechseln; das passiert in drei Fällen, hat also eine Wahrscheinlichkeit von $\frac{1}{3}$. Ist $i \neq k$, so hat der Kandidat beim ersten Mal eine Tür mit einer Ziege erwischt; das

Auto befindet sich also hinter einer der zwei nach seiner Wahl übrigen Türen, und der Showmaster zeigt ihm durch Öffnen, welche dieser zwei Türen falsch ist. Mit Wechseln gewinnt der Kandidat also; das passiert in sechs Fällen, hat also eine Wahrscheinlichkeit von $\frac{2}{3}$, und damit lohnt es sich, die Tür zu wechseln. ◇

1.4 Unabhängigkeit von Ereignissen

Definition. Zwei Ereignisse A, B heissen *(stochastisch) unabhängig*, falls

$$P[A \cap B] = P[A]P[B].$$

Ist $P[A] = 0$ oder $P[B] = 0$, so sind A und B immer unabhängig. Für $P[A] \neq 0$ gilt

$$A, B \text{ unabhängig} \iff P[B \mid A] = P[B],$$

und symmetrisch gilt für $P[B] \neq 0$

$$A, B \text{ unabhängig} \iff P[A \mid B] = P[A].$$

Anschaulich: Die Tatsache, dass eines der Ereignisse eingetreten ist, hat keinen Einfluss auf die Wahrscheinlichkeit des anderen.

⌜ Sei $P[A] \neq 0$. Dann ist

$$P[B \mid A] = \frac{P[A \cap B]}{P[A]} \text{ gleich } P[B]$$

offenbar äquivalent zu $P[A \cap B] = P[A]P[B]$, also zur Unabhängigkeit. Der Fall $P[B] \neq 0$ geht symmetrisch. ⌟

Beispiel 3 (zwei Münzwürfe). Eine Münze wird zweimal geworfen. Hier ist der Grundraum $\Omega = \{KK, KZ, ZK, ZZ\}$, und wir nehmen an, dass alle Ausgänge gleich wahrscheinlich sind. Für die Ergebnisse

$$A = \{\text{Kopf beim 1. Wurf}\} = \{KK, KZ\},$$
$$B = \{\text{Kopf beim 2. Wurf}\} = \{KK, ZK\}$$

ist

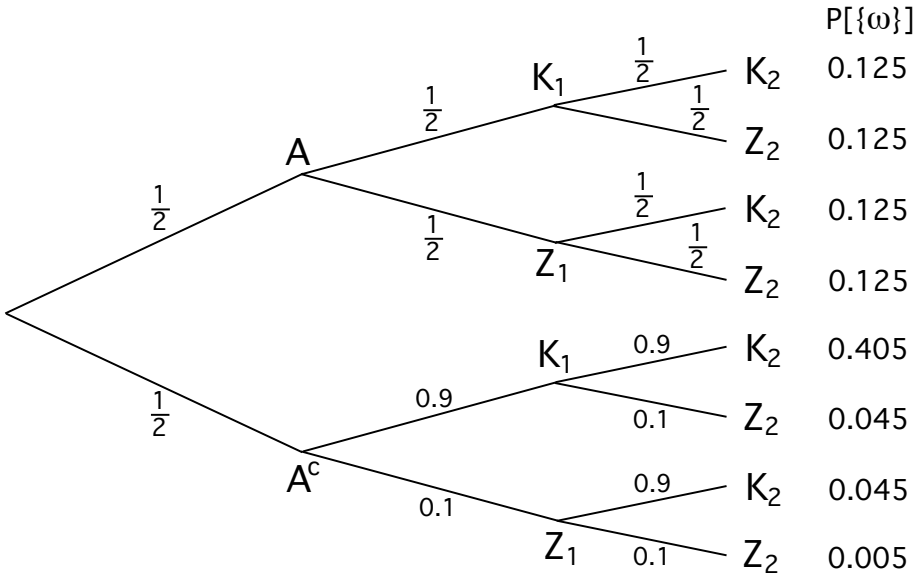
$$P[A] = P[B] = \frac{1}{2},$$
$$P[A \cap B] = P[\{KK\}] = \frac{1}{4},$$

und damit sind A und B (wie erwartet) unabhängig. ◇

Beispiel. In einer Urne befinden sich 2 Sorten von Münzen, gleich viele von jeder Sorte. Die Münzen der Sorte F sind fair: die Wahrscheinlichkeiten für Kopf und Zahl bei einem Wurf sind je $\frac{1}{2}$. Bei Münzen der unfairen Sorte U ist hingegen die Wahrscheinlichkeit für Kopf 0.9, die für Zahl also 0.1. Vom Aussehen her sind die Münzen aber nicht unterscheidbar.

Nun wird zufällig eine Münze gezogen und zweimal geworfen. Für $i = 1, 2$ sei K_i das Ereignis "Kopf beim i-ten Wurf". Sind K_1 und K_2 unabhängig?

Sei $A = \{$Münze von Sorte F wird gezogen$\}$. Dann haben wir folgende Situation:



Schreiben wir k_i und z_i für den Ausgang Kopf bzw. Zahl bei Wurf i, so ist

$$P[K_1] = P[\{Fk_1k_2, Fk_1z_2, Uk_1k_2, Uk_1z_2\}] = 0.125 + 0.125 + 0.405 + 0.045 = 0.7$$

und ebenso $P[K_2] = 0.7$. Hingegen ist

$$P[K_1 \cap K_2] = P[\{Fk_1k_2, Uk_1k_2\}] = 0.125 + 0.405 = 0.53$$

und damit

$$P[K_1 \cap K_2] \neq 0.49 = P[K_1]P[K_2].$$

Also sind hier K_1 und K_2 *abhängig*. Wir haben auch

$$P[K_2 \,|\, K_1] = \frac{P[K_1 \cap K_2]}{P[K_1]} = \frac{0.53}{0.7} = 0.757 > P[K_2].$$

Wegen

$$P[K_1 \,|\, A] = \frac{1}{2} \neq 0.7 = P[K_1],$$
$$P[K_2 \,|\, A] = \frac{1}{2} \neq 0.7 = P[K_2]$$

sind K_1 und A abhängig, und ebenso sind K_2 und A abhängig. Das ist auch die Erklärung für das obige Resultat: Durch ihre gemeinsame Abhängigkeit von A sind auch K_1 und K_2 abhängig.

Anschaulich ist dieses Resultat auch *einleuchtend*. Obwohl wir der Münze ihren Typ nicht ansehen können, wissen wir, dass ihr Typ einen Einfluss auf die Wahrscheinlichkeit von Kopf hat. Gibt der erste Wurf also Kopf, so kann das auf eine Münze vom Typ U hindeuten, und dann erwarten wir bei Wurf 2 auch wieder eher Kopf. Das Ereignis K_1 hat also unsere Einschätzung für die (bedingte) Wahrscheinlichkeit von K_2 geändert. $\diamond$

Bisher haben wir nur Unabhängigkeit von 2 Ereignissen betrachtet. Allgemeiner hat man:

Definition. Die Ereignisse $A_1, \ldots, A_n$ heissen *(stochastisch) unabhängig*, wenn für jede endliche Teilfamilie die Produktformel gilt, d.h. für $m \in \mathbb{N}$ und $\{k_1, \ldots, k_m\} \subseteq \{1, \ldots, n\}$ gilt immer

$$P\left[\bigcap_{i=1}^{m} A_{k_i}\right] = \prod_{i=1}^{m} P[A_{k_i}].$$

Bemerkungen. 1) Statt endlich viele Ereignisse $A_1, \ldots, A_n$ kann man in der obigen Definition auch eine beliebige Familie (abzählbar oder überabzählbar) betrachten; die Produktformel fordert man aber immer nur für alle *endlichen* Teilfamilien.

2) Hat man die Produktformel nur für alle Paare von Ereignissen, so heissen die Ereignisse $A_1, \ldots, A_n$ *paarweise unabhängig*. Aus Unabhängigkeit folgt also immer paarweise Unabhängigkeit; das folgende Beispiel zeigt aber, dass paarweise Unabhängigkeit echt schwächer ist als Unabhängigkeit. ◇

Beispiel. Eine faire Münze wird zweimal geworfen. Wir betrachten die Ereignisse

$$A = \{\text{Kopf bei Wurf 1}\} = \{KK, KZ\},$$
$$B = \{\text{Kopf bei Wurf 2}\} = \{KK, ZK\},$$
$$C = \{\text{beide Würfe gleich}\} = \{KK, ZZ\}.$$

Dann ist offenbar

$$P[A] = P[B] = P[C] = \frac{1}{2}$$

und

$$P[A \cap B] = P[A \cap C] = P[B \cap C] = P[\{KK\}] = \frac{1}{4}.$$

Also sind A, B, C paarweise unabhängig. Wegen

$$A \cap B \cap C = \{KK\}$$

ist aber

$$P[A \cap B \cap C] = \frac{1}{4} \neq \frac{1}{8} = P[A]P[B]P[C],$$

und damit sind die Ereignisse A, B, C nicht unabhängig. ◇

2 Diskrete Zufallsvariablen und Verteilungen

Grundidee: In den meisten Fällen lässt sich ein Zufallsexperiment beschreiben durch eine Abbildung auf einem Grundraum Ω und durch die Wahrscheinlichkeiten, mit denen diese Abbildung Werte in ihrem Wertebereich annimmt. Die Abbildung nennt man *Zufallsvariable*, und die Wahrscheinlichkeiten werden durch die *Verteilung* der Zufallsvariablen beschrieben. Wir wollen das im Folgenden präzisieren.

Um die Darstellung möglichst einfach zu halten, erklären wir zuerst alle Begriffe im diskreten Fall. *In diesem ganzen Kapitel ist also $\Omega \neq \emptyset$ endlich oder abzählbar, $\mathcal{F} = 2^\Omega$ die Potenzmenge von Ω, und das Wahrscheinlichkeitsmass P damit gegeben durch seine Gewichte $p_i = P[\{\omega_i\}]$ für alle i.*

Zur Erinnerung: Eine Menge $\Omega \neq \emptyset$ heisst *abzählbar*, falls es eine bijektive Abbildung $f : \Omega \to \mathbb{N}$ gibt; das bedeutet, dass man Ω via f durchnumerieren (abzählen) kann. Insbesondere ist also eine abzählbare Menge immer unendlich. Beispielsweise sind $\mathbb{N}$, $\mathbb{Z}$ oder $\mathbb{Q}$ (die Menge der rationalen Zahlen) abzählbar, während $\mathbb{R}$ oder $[0, 1]$ überabzählbar sind.

2.1 Grundbegriffe

Definition. Eine (genauer: reellwertige) *diskrete Zufallsvariable (ZV)* auf Ω ist eine Funktion $X : \Omega \to \mathbb{R}$. Mit Ω ist natürlich auch der Wertebereich $\mathcal{W}(X) = \{x_1, x_2, \dots\}$ von X endlich oder abzählbar. Die *Verteilungsfunktion (VF)* von X ist die Abbildung $F_X : \mathbb{R} \to [0, 1]$, die definiert ist durch

$$t \mapsto F_X(t) := P[X \leq t] := P[\{\omega : X(\omega) \leq t\}].$$

Die *Gewichtsfunktion* oder *diskrete Dichte* von X ist die Funktion $p_X : \mathcal{W}(X) \to [0, 1]$,

die durch

$$p_X(x_k) := P[X = x_k] = P[\{\omega : X(\omega) = x_k\}] \qquad \text{für } k = 1, 2, \ldots$$

definiert ist.

Bemerkung. Ist Ω endlich oder abzählbar mit $\mathcal{F} = 2^\Omega$, so ist *jede* Funktion $X : \Omega \to \mathbb{R}$ eine diskrete Zufallsvariable. Für allgemeine Ω und $\mathcal{F}$ braucht die obige Definition von F_X, dass die Menge $\{X \leq t\} = \{\omega : X(\omega) \leq t\}$ für jedes t ein (beobachtbares) Ereignis, also in $\mathcal{F}$ ist, weil P nur auf $\mathcal{F}$ definiert ist. Das bedeutet (per Definition), dass die Funktion X ($\mathcal{F}$-)*messbar* sein soll, und deshalb ist im allgemeinen Fall eine Zufallsvariable definiert als eine messbare Funktion $X : \Omega \to \mathbb{R}$. Die Forderung der Messbarkeit ist mathematisch nötig, bedeutet aber keine starke Einschränkung. ◇

Beispiel 2 (ein Würfelwurf). Beim Würfeln mit einem Würfel ist $\Omega = \{1, \ldots, 6\}$. Die geworfene Augenzahl ist dann eine Zufallsvariable, beschrieben durch die Funktion $X(\omega) = \omega$. ◇

Beispiel 3 (zwei Münzwürfe). Wirft man eine Münze zweimal, so ist der Grundraum $\Omega = \{KK, KZ, ZK, ZZ\}$. Beispiele für Zufallsvariablen sind die Gesamtanzahl der geworfenen Köpfe (X_1) oder die Differenz Anzahl der Köpfe minus Anzahl der Zahlen (X_2):

ω	KK	KZ	ZK	ZZ
$X_1(\omega)$	2	1	1	0
$X_2(\omega)$	2	0	0	-2

◇

Beispiel 6 (Wartezeit). Wirft man eine Münze so lange, bis man Kopf erhält, so ist eine mögliche Wahl für den Grundraum $\Omega = \{K, ZK, ZZK, ZZZK, \ldots\}$. Die Gesamtzahl der Würfe ist dann eine Zufallsvariable:

ω	K	ZK	ZZK	$\dots$
$X(\omega)$	1	2	3	$\dots$

X heisst auch *Wartezeit bis zum ersten Erfolg* (Kopf). $\diamond$

Beispiel. Für jede Teilmenge $A \subseteq \Omega$ ist die *Indikatorfunktion* I_A von A definiert durch

$$I_A(\omega) := \begin{cases} 1 & \text{für } \omega \in A, \\ 0 & \text{für } \omega \in A^c. \end{cases}$$

Ist Ω endlich oder abzählbar und $\mathcal{F} = 2^\Omega$, so ist I_A für jedes $A \subseteq \Omega$ eine diskrete Zufallsvariable. Allgemeiner, d.h. für allgemeines $(\Omega, \mathcal{F})$, ist I_A eine Zufallsvariable genau dann, wenn $A \in \mathcal{F}$ ist.) $\diamond$

Sei nun allgemein X eine diskrete Zufallsvariable mit Wertebereich $\mathcal{W}(X) = \{x_1, x_2, \dots\}$. Wegen

$$\{X \leq t\} = \bigcup_{k \text{ mit } x_k \leq t} \{X = x_k\}$$

ist die Verteilungsfunktion von X gemäss Axiom A2) gegeben durch

$$F_X(t) = P[X \leq t] = \sum_{k \text{ mit } x_k \leq t} p_X(x_k).$$

Insbesondere ist F_X also durch p_X vollständig festgelegt. Mit dem gleichen Argument erhält man auch

$$P[X \in B] = \sum_{x_k \in B} p_X(x_k) \qquad \text{für jede Menge } B \subseteq \mathcal{W}(X).$$

Die Gewichtsfunktion p_X einer diskreten Zufallsvariablen X hat die Eigenschaften $0 \leq p_X(x_k) \leq 1$ für jedes $x_k \in \mathcal{W}(X)$ und

$$\sum_{x_k \in \mathcal{W}(X)} p_X(x_k) = P[X \in \mathcal{W}(X)] = 1.$$

Ist umgekehrt $\mathcal{W} = \{w_1, w_2, \dots\}$ irgendeine nichtleere endliche oder abzählbare Menge und $f : \mathcal{W} \to \mathbb{R}$ eine Funktion mit $0 \leq f(w_k) \leq 1$ für jedes $w_k \in \mathcal{W}$ und

$$\sum_{w_k \in \mathcal{W}} f(w_k) = 1,$$

so kann man einen Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, P)$ und darauf eine diskrete Zufallsvariable X konstruieren, deren Gewichtsfunktion p_X gerade die vorgegebene Funktion f ist. Beispielsweise genügt $\Omega := \mathcal{W}$, $\mathcal{F} := 2^\Omega = 2^\mathcal{W}$ und $X(\omega) := \omega$.

Das stochastische Verhalten einer (reellwertigen) Zufallsvariablen X wird durch ihre *Verteilung* beschrieben; das ist dasjenige Wahrscheinlichkeitsmass μ_X auf $\mathbb{R}$, das durch

$$\mu_X(B) := P[X \in B] := P[\{\omega \in \Omega : X(\omega) \in B\}]$$

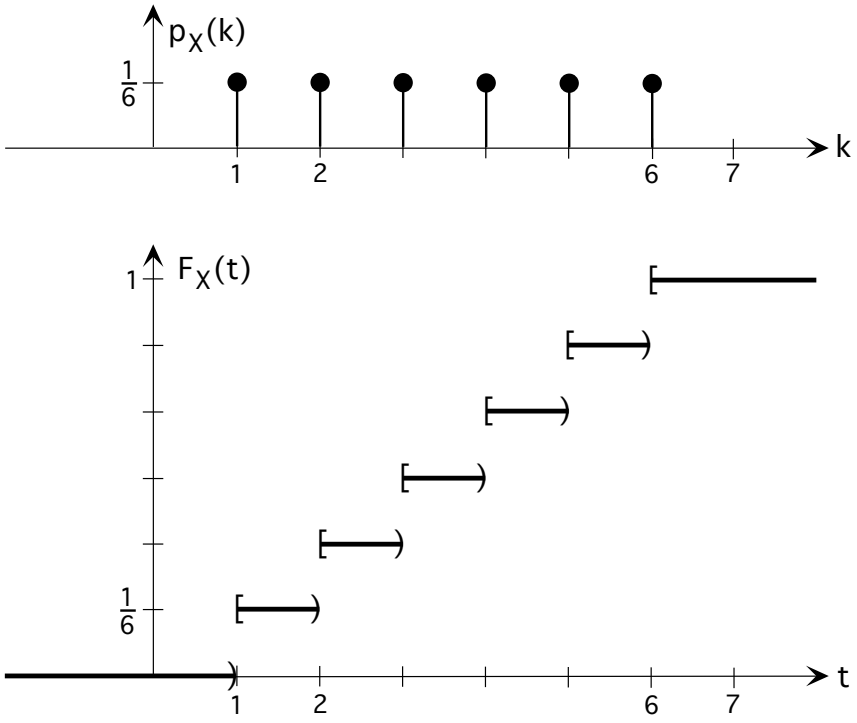
definiert ist. Ist X eine diskrete Zufallsvariable, so heisst μ_X eine *diskrete Verteilung*; dann ist μ_X de facto ein Wahrscheinlichkeitsmass auf dem Wertebereich $\mathcal{W}(X)$ von X, der wie Ω endlich oder abzählbar ist, und damit ist μ_X wie üblich festgelegt durch die Gewichte

$$\mu_X(\{x_k\}) = P[X = x_k] = p_X(x_k) \qquad \text{für alle } x_k \in \mathcal{W}(X).$$

Für diskrete Zufallsvariablen kann man also die Verteilung μ_X und die Gewichtsfunktion p_X direkt miteinander identifizieren; der einzige Unterschied ist, dass μ_X als Argumente *Teilmengen* von $\mathcal{W}(X)$ hat, p_X hingegen *Elemente* von $\mathcal{W}(X)$. Wie oben ist der Zusammenhang zwischen μ_X und p_X gegeben durch

$$\mu_X(B) = P[X \in B] = \sum_{x_k \in B} p_X(x_k) \qquad \text{für } B \subseteq \mathcal{W}(X).$$

Beispiel 2 (ein Würfelwurf). Augenzahl beim Würfeln: Hier ist $\mathcal{W}(X) = \{1, \dots, 6\}$ und $p_X(k) = \frac{1}{6}$ für $k = 1, \dots, 6$. Die Gewichtsfunktion p_X und die Verteilungsfunktion F_X sehen wie folgt aus:

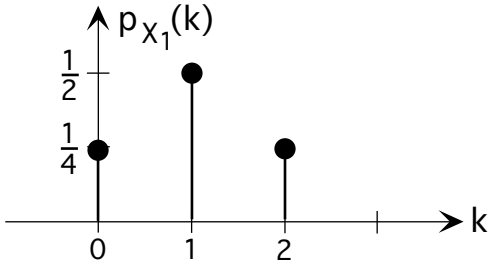


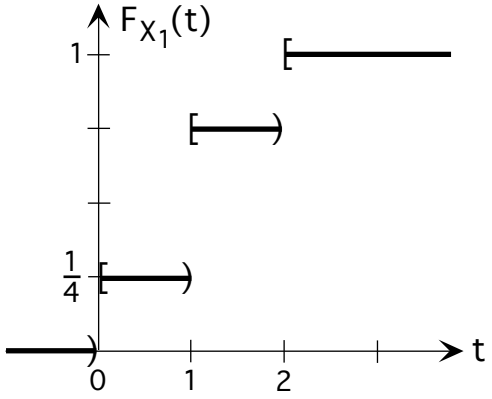
◇

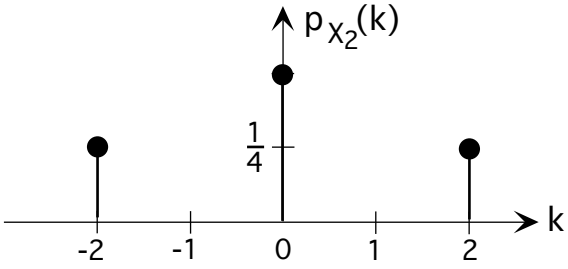
Beispiel 3 (zwei Münzwürfe). Zweimaliger Münzwurf, mit X_1 = Gesamtanzahl der Köpfe und X_2 = Anzahl der Köpfe minus Anzahl der Zahlen:

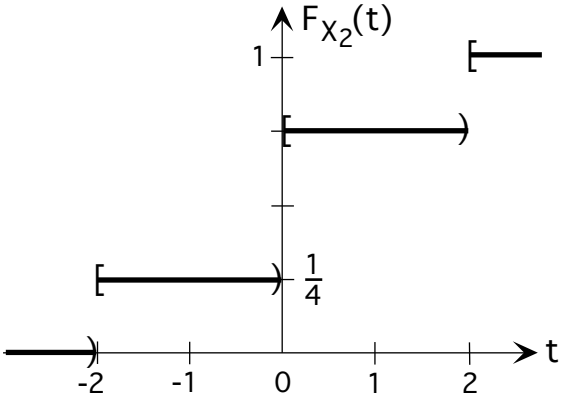
Hier ist $\mathcal{W}(X_1) = \{0, 1, 2\}$ und $p_{X_1}(0) = \frac{1}{4}$, $p_{X_1}(1) = \frac{1}{2}$, $p_{X_1}(2) = \frac{1}{4}$; ferner haben wir $\mathcal{W}(X_2) = \{-2, 0, 2\}$ und $p_{X_2}(-2) = \frac{1}{4}$, $p_{X_2}(0) = \frac{1}{2}$, $p_{X_2}(2) = \frac{1}{4}$.

Die Gewichtsfunktionen p_{X_i} und Verteilungsfunktionen F_{X_i} sehen wie folgt aus:


$p_{X_1}(k)$
$\frac{1}{2}$
$\frac{1}{4}$
k
0
1
2


$F_{X_1}(t)$
1
$\frac{1}{4}$
t
0
1
2


$p_{X_2}(k)$
$\frac{1}{4}$
k
-2
-1
0
1
2


$F_{X_2}(t)$
1
$\frac{1}{4}$
t
-2
-1
0
1
2

◇

Beispiel 6 (Wartezeit). Wartezeit auf den ersten Erfolg (Kopf) bei einer Folge von Münzwürfen: Diese Zufallsvariable hat eine sogenannte *geometrische Verteilung* mit Parameter p, wenn wir annehmen, dass die einzelnen Münzwürfe unabhängig voneinander sind und Kopf jeweils mit Wahrscheinlichkeit $p \in (0,1)$ ergeben. Das bedeutet, dass die Gewichtsfunktion von X gegeben ist durch

$$p_X(k) = P[X = k] = p(1-p)^{k-1} \qquad \text{für } k = 1, 2, \ldots$$

Eine exakte Herleitung werden wir im Kapitel 3 geben. ◇

Ist $X : \Omega \to \mathbb{R}$ eine diskrete Zufallsvariable und $g : \mathbb{R} \to \mathbb{R}$ eine Abbildung, so ist offenbar die Zusammensetzung $Y := g \circ X : \Omega \to \mathbb{R}$ wieder eine diskrete Zufallsvariable; wir schreiben kurz $Y = g(X)$. Beispiele für solche Transformationen sind $Y = X^2$, $Y = e^X$, $Y = (X - a)^n$ für $a \in \mathbb{R}$ und $n \in \mathbb{N}$, $Y = \sin X$, usw.

2.2 Erwartungswerte

Grundidee: Wir möchten für eine Zufallsvariable X gewisse Kennzahlen finden, die in geeigneter Form das durchschnittliche Verhalten von X beschreiben.

Um sowohl die Idee als auch die Definition zu *motivieren*, beginnen wir mit einem

Beispiel. Spieler 1 verspricht Spieler 2, ihm beim Würfelspiel die folgenden Gewinne auszuzahlen:

10 Rappen, falls 1 oder 2 gewürfelt wird,

20 Rp. bei 3 oder 4,

40 Rp. bei 5, und

80 Rp. bei 6.

Wieviel muss Spieler 2 vor jeder Runde an Spieler 1 bezahlen, damit das Spiel fair ist? "Faires Spiel" soll dabei heissen, dass der Einsatz gleich dem "durchschnittlichen Gewinn" sein soll.

Stellen wir uns einmal vor, dass n Runden gespielt werden und dass die Augenzahl i jeweils in n_i Runden auftritt. Der Gesamtgewinn von Spieler 2 ist dann (in Rappen)

$$G(n) = 10n_1 + 10n_2 + 20n_3 + 20n_4 + 40n_5 + 80n_6,$$

und sein durchschnittlicher Gewinn pro Spielrunde ist

$$\frac{1}{n}G(n) = 10\frac{n_1}{n} + 10\frac{n_2}{n} + 20\frac{n_3}{n} + 20\frac{n_4}{n} + 40\frac{n_5}{n} + 80\frac{n_6}{n}.$$

Dieses Ergebnis ist nun aber abhängig von der Anzahl $n = n_1 + \cdots + n_6$ der gespielten Runden, und ausserdem erst ganz am Schluss bekannt. Um das zu vermeiden, ersetzen wir die relativen Häufigkeiten $\frac{n_i}{n}$ durch die entsprechenden Wahrscheinlichkeiten $p_i = \frac{1}{6}$; wir benutzen also die frequentistische Interpretation von Wahrscheinlichkeiten (und gehen davon aus, dass sich der Würfel symmetrisch verhält). Damit erhalten wir den Erwartungswert des Gewinns als "idealisierten durchschnittlichen Gewinn bei unendlich vielen Spielrunden":

$$\text{erwarteter Gewinn} := 10p_1 + 10p_2 + 20p_3 + 20p_4 + 40p_5 + 80p_6 = 180 \times \frac{1}{6} = 30,$$

d.h. der "faire Einsatz" in diesem Sinn ist 30 Rappen. ◇

Definition. Sei X eine diskrete Zufallsvariable mit Gewichtsfunktion p_X. Dann definieren wir den *Erwartungswert* von X als

$$E[X] := \sum_{x_k \in \mathcal{W}(X)} x_k p_X(x_k),$$

sofern die Reihe absolut konvergiert, d.h. falls $\sum_{x_k \in \mathcal{W}(X)} |x_k| p_X(x_k) < \infty$ gilt. Ist die Reihe nicht absolut konvergent, so existiert der Erwartungswert nicht.

Beispiel 7 (Roulette). Ein normales *Roulette*-Rad enthält die Zahlen $0, 1, 2, \ldots, 36$ und 00 (Doppelnull). Nun wettet man 1 CHF darauf, dass eine ungerade Zahl kommt.

Tritt das ein, so bekommt man 2 CHF zurück, sonst nichts; in jedem Fall verliert man aber den Einsatz.

Ist X der Nettogewinn bei einem solchen Spiel, so ist $X = +1$ mit Wahrscheinlichkeit $\frac{18}{38}$ und $X = -1$ mit Wahrscheinlichkeit $\frac{20}{38}$, weil 0 und 00 beide als weder gerade noch ungerade gelten. Der Erwartungswert ist also

$$E[X] = (+1) \times \frac{18}{38} + (-1) \times \frac{20}{38} = -\frac{1}{19}.$$

Dieses Spiel ist (für den Spieler) unfair; auf Dauer verliert man im Schnitt etwas mehr als 5 Rappen pro Spiel. ◇

In der Definition des Erwartungswertes $E[X]$ taucht der zugrundeliegende Raum Ω nicht auf, sondern nur die Verteilung von X, d.h. der Wertebereich $\mathcal{W}(X)$ und die Wahrscheinlichkeiten $p_X(x_k)$, mit denen die Werte $x_k \in \mathcal{W}(X)$ angenommen werden. Wir können $E[X]$ aber auch als eine Summe über Ω schreiben, denn sofern der Erwartungswert existiert, so gilt

$$E[X] = \sum_{\omega_i \in \Omega} X(\omega_i) P[\{\omega_i\}] = \sum_{\omega_i \in \Omega} p_i X(\omega_i). \tag{2.1}$$

⌜ Nach Definition und wegen $\{X = x_k\} = \{\omega_i \in \Omega : X(\omega_i) = x_k\}$ ist

$$E[X] = \sum_{x_k \in \mathcal{W}(X)} x_k P[X = x_k] = \sum_{x_k \in \mathcal{W}(X)} x_k \sum_{\omega_i \in \Omega \text{ mit } X(\omega_i) = x_k} P[\{\omega_i\}] = \sum_{\omega_i \in \Omega} X(\omega_i) P[\{\omega_i\}].$$

Dabei darf man die Reihenfolge der beiden Summen vertauschen, weil der Erwartungswert nach Voraussetzung existiert. ⌟

Nun betrachten wir die Frage, wie man für eine Transformierte $Y = g(X)$ einer diskreten Zufallsvariablen X den Erwartungswert berechnet. Der wesentliche Punkt im folgenden Resultat ist dabei, dass es dafür nicht nötig ist, zuerst die Verteilung von Y zu bestimmen und dann die Definition für $E[Y]$ zu benutzen. Es genügt schon, die Verteilung von X zu kennen.

Satz 2.1. *Sei X eine diskrete Zufallsvariable mit Gewichtsfunktion p_X, und sei $Y = g(X)$ für eine Funktion $g : \mathbb{R} \to \mathbb{R}$. Dann ist*

$$E[Y] = E[g(X)] = \sum_{x_k \in \mathcal{W}(X)} g(x_k) p_X(x_k),$$

sofern die Reihe absolut konvergiert.

Beweis. Sei $y_j \in \mathcal{W}(Y)$ und $A_j := \{x_k \in \mathcal{W}(X) : g(x_k) = y_j\}$. Dann ist offensichtlich $\{Y = y_j\} = \dot{\bigcup}_{x_k \in A_j} \{X = x_k\}$ und damit zum einen

$$p_Y(y_j) = P[Y = y_j] = \sum_{x_k \in A_j} p_X(x_k)$$

und zum anderen

$$\mathcal{W}(X) = \dot{\bigcup}_{y_j \in \mathcal{W}(Y)} \dot{\bigcup}_{x_k \in A_j} \{x_k\}.$$

Wegen $y_j = g(x_k)$ für $x_k \in A_j$ folgt also

$$\begin{aligned} E[Y] &= \sum_{y_j \in \mathcal{W}(Y)} y_j p_Y(y_j) \\ &= \sum_{y_j \in \mathcal{W}(Y)} \sum_{x_k \in A_j} y_j p_X(x_k) \\ &= \sum_{y_j \in \mathcal{W}(Y)} \sum_{x_k \in A_j} g(x_k) p_X(x_k) \\ &= \sum_{x_k \in \mathcal{W}(X)} g(x_k) p_X(x_k). \end{aligned}$$

Die absolute Konvergenz der Reihe braucht man dabei, damit alle Ausdrücke in der obigen Gleichungskette wohldefiniert sind. **q.e.d.**

Erste einfache Eigenschaften des Erwartungswertes gibt das nächste Resultat.

Satz 2.2. *Seien X und Y diskrete Zufallsvariablen, für die jeweils der Erwartungswert existiert. Dann gilt:*

1) Monotonie: *Ist $X \leq Y$ (d.h. $X(\omega) \leq Y(\omega)$ für alle ω), so gilt auch $E[X] \leq E[Y]$.*

2) Linearität: *Für beliebige $a, b \in \mathbb{R}$ gilt $E[aX + b] = aE[X] + b$.*

3) *Falls X nur Werte in $\mathbb{N}_0 = \{0, 1, 2, \dots\}$ annimmt, so gilt*

$$E[X] = \sum_{j=1}^{\infty} P[X \geq j] = \sum_{\ell=0}^{\infty} P[X > \ell].$$

Beweis. 1) Nach der alternativen Darstellung des Erwartungswertes in (2.1) ist

$$E[X] = \sum_{\omega_i \in \Omega} X(\omega_i) P[\{\omega_i\}].$$

Daraus folgt sofort die Behauptung, weil alle $P[\{\omega_i\}] \geq 0$ sind.

2) Durch Einsetzen und Benutzen von $\sum_{\omega_i \in \Omega} P[\{\omega_i\}] = 1$ ergibt sich

$$E[aX + b] = \sum_{\omega_i \in \Omega} \big(aX(\omega_i) + b\big) P[\{\omega_i\}] = a \sum_{\omega_i \in \Omega} X(\omega_i) P[\{\omega_i\}] + b = aE[X] + b.$$

Alternativ erhält man das via Satz 2.1 und mit Benutzung von $\sum_{x_k \in \mathcal{W}(X)} p_X(x_k) = 1$.

3) Wegen $\mathcal{W}(X) = \mathbb{N}_0$ und $\{X \geq j\} = \biguplus_{k=j}^{\infty} \{X = k\}$ ist

$$\begin{aligned} E[X] &= \sum_{k=0}^{\infty} k p_X(k) = \sum_{k=1}^{\infty} k P[X = k] \\ &= \sum_{k=1}^{\infty} \sum_{j=1}^{k} P[X = k] \\ &= \sum_{j=1}^{\infty} \sum_{k=j}^{\infty} P[X = k] \\ &= \sum_{j=1}^{\infty} P[X \geq j] = \sum_{\ell=0}^{\infty} P[X > \ell]. \end{aligned}$$

q.e.d.

Eine schöne Anwendung des letzten Resultats in Satz 2.2 werden wir sehen, wenn wir im Kapitel 3 den Erwartungswert einer geometrischen Verteilung berechnen.

Definition. Sei X eine diskrete Zufallsvariable. Ist $E[X^2] < \infty$, so heisst

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

die *Varianz* von X, und $\sqrt{\mathrm{Var}[X]}$ heisst die *Standardabweichung* von X. Manchmal schreibt man auch $\mathrm{sd}(X) := \sqrt{\mathrm{Var}[X]}$ oder $\sigma(X) := \sqrt{\mathrm{Var}[X]}$.

Sowohl die Varianz als auch die Standardabweichung sind Kennzahlen für die *Streuung* der Verteilung von X. Wegen der Bedingung $E[X^2] < \infty$ existiert automatisch der Erwartungswert von X. Schreiben wir der Kürze halber $m_X := E[X]$, so erhalten wir aus Satz 2.1 mit $g(x) := (x - m_X)^2$ sofort

$$\mathrm{Var}[X] = \sum_{x_k \in \mathcal{W}(X)} (x_k - m_X)^2 p_X(x_k).$$

Daraus sieht man auch, dass die Varianz die durchschnittliche quadratische Abweichung der Zufallsvariablen X von ihrem Erwartungswert $m_X = E[X]$ beschreibt.

Um Beispiele bequemer berechnen zu können, brauchen wir ein kleines Hilfsresultat.

Lemma 2.3. *Sei X eine diskrete Zufallsvariable mit $E[X^2] < \infty$, und sei $Y = aX + b$. Dann gilt:*

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

2) $\mathrm{Var}[Y] = \mathrm{Var}[aX + b] = a^2 \mathrm{Var}[X]$.

Beweis. Der Kürze halber schreiben wir wieder $m_X = E[X]$.

1) Einsetzen gibt dank der Linearität in Satz 2.2

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

2) $E[Y] = aE[X] + b$ nach Satz 2.2 liefert

$$Y - E[Y] = aX + b - (aE[X] + b) = a(X - m_X)$$

und damit

$$\mathrm{Var}[Y] = E[(Y - E[Y])^2] = E[a^2(X - m_X)^2] = a^2\mathrm{Var}[X].$$

q.e.d.

Beispiel 7 (Roulette). In unserem *Roulette*-Spiel mit den Zahlen $0, 1, 2, \ldots, 36$ und 00 setzen wir nun 1 CHF nicht auf Ungerade, sondern auf eine bestimmte Zahl (z.B. 13); falls wir gewinnen, bekommen wir dann 36 CHF ausbezahlt.

Ist Y der entsprechende Nettogewinn, so ist $Y = +35$ mit Wahrscheinlichkeit $\frac{1}{38}$ und $Y = -1$ mit Wahrscheinlichkeit $\frac{37}{38}$, weil man den Einsatz in jedem Fall verliert und nur bei dem Ergebnis (z.B. 13) gewinnt, auf das man gesetzt hat. Der Erwartungswert von Y ist also

$$E[Y] = (+35) \times \frac{1}{38} + (-1) \times \frac{37}{38} = -\frac{1}{19},$$

so dass wir im Schnitt den gleichen (kleinen) Verlust haben wie bei der ersten Strategie, wo man auf Ungerade setzt.

Das Setzen auf eine bestimmte Zahl gibt aber eine viel grössere Streuung, wie wir nun ausrechnen. Wegen

$$E[Y^2] = (+35)^2 \times \frac{1}{38} + (-1)^2 \times \frac{37}{38} = \frac{1262}{38} = \frac{631}{19}$$

ist nämlich

$$\mathrm{Var}[Y] = E[Y^2] - (E[Y])^2 = \frac{631}{19} - \left(\frac{1}{19}\right)^2 = \frac{11988}{361} = 33.21$$

und $\sqrt{\mathrm{Var}[Y]} = 5.763$. Ist hingegen X der Nettogewinn beim Setzen auf Ungerade, so hat X nur die Werte ± 1; also ist $E[X^2] = 1$ und damit

$$\mathrm{Var}[X] = E[X^2] - (E[X])^2 = 1 - \left(\frac{1}{19}\right)^2 = \frac{360}{361} = 0.9973,$$

also $\sqrt{\mathrm{Var}[X]} = 0.9986$.

Qualitativ ist natürlich ohne Rechnung klar, dass X viel weniger stark streut als Y; mit Hilfe von Varianz und Standardabweichung kann man solche Unterschiede aber auch quantitativ systematisch erfassen und messen. ◇

2.3 Gemeinsame Verteilungen, unabhängige Zufallsvariablen

Grundidee: Wir möchten das gemeinsame Verhalten von und die Zusammenhänge zwischen mehreren Zufallsvariablen beschreiben und untersuchen können. Das wird uns auch erlauben, den Begriff der Unabhängigkeit von Ereignissen auf Zufallsvariablen zu verallgemeinern.

Im Folgenden betrachten wir oft n diskrete Zufallsvariablen $X_1, \ldots, X_n$. Ist $n = 2$, so nennen wir die beiden Zufallsvariablen meist nur X und Y.

Definition. Seien $X_1, \ldots, X_n$ (diskrete oder auch beliebige) Zufallsvariablen. Die *gemeinsame Verteilungsfunktion* von $X_1, \ldots, X_n$ ist die Abbildung $F : \mathbb{R}^n \to [0, 1]$, definiert durch

$$(x_1, \ldots, x_n) \mapsto F(x_1, \ldots, x_n) := P[X_1 \leq x_1, \ldots, X_n \leq x_n].$$

Sind $X_1, \ldots, X_n$ diskrete Zufallsvariablen, so definiert man ihre *gemeinsame Gewichtsfunktion* $p : \mathbb{R}^n \to [0, 1]$ durch

$$p(x_1, \ldots, x_n) := P[X_1 = x_1, \ldots, X_n = x_n].$$

Natürlich ist $p(x_1, \ldots, x_n) = 0$ für $(x_1, \ldots, x_n) \notin \mathcal{W}(X_1, \ldots, X_n)$. Aus der gemeinsamen Gewichtsfunktion p bekommt man die gemeinsame Verteilungsfunktion via

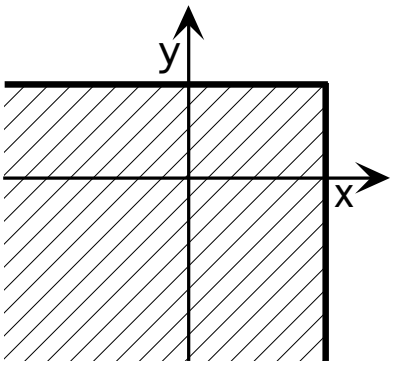
$$\begin{aligned} F(x_1, \ldots, x_n) &= P[X_1 \leq x_1, \ldots, X_n \leq x_n] \\ &= \sum_{y_1 \leq x_1, \ldots, y_n \leq x_n} P[X_1 = y_1, \ldots, X_n = y_n] \\ &= \sum_{y_1 \leq x_1, \ldots, y_n \leq x_n} p(y_1, \ldots, y_n); \end{aligned}$$

das ist völlig analog zum Zusammenhang zwischen F_X und p_X im Abschnitt 2.1.

Beispiel. Für $n = 2$ mit Zufallsvariablen X, Y beschreibt die gemeinsame Verteilungsfunktion gerade die Wahrscheinlichkeiten aller Mengen $\{X \leq x, Y \leq y\}$, d.h. die Wahrscheinlichkeiten dafür, dass der Vektor (X, Y) in ein unendliches Rechteck der Form

$(-\infty, x] \times (-\infty, y]$ fällt, denn

$$F(x, y) = P[X \leq x, Y \leq y].$$



Man kann zeigen, dass dann auch die Wahrscheinlichkeiten $P[(X, Y) \in B]$ für beliebige (genauer: Borel-messbare) Teilmengen $B \subseteq \mathbb{R}^2$ $\big($oder $B \subseteq \mathcal{W}(X, Y)$, falls X und Y diskrete Zufallsvariablen sind$\big)$ eindeutig festgelegt sind. Das bedeutet, dass man das stochastische Verhalten von Zufallsvariablen kennt, sobald man ihre gemeinsame Verteilungsfunktion kennt. $\diamond$

Beispiel 8 (drei Münzwürfe). Eine Münze wird dreimal geworfen. Wir nennen X die Anzahl der Köpfe beim ersten Wurf und Y die Gesamtanzahl der Köpfe. Der zugehörige Grundraum ist

$$\Omega = \{KKK, KKZ, KZK, KZZ, ZKK, ZKZ, ZZK, ZZZ\},$$

und alle Elementarereignisse sind gleich wahrscheinlich.

Die Zufallsvariablen X und Y sind beide diskret mit Wertebereichen $\mathcal{W}(X) = \{0, 1\}$, $\mathcal{W}(Y) = \{0, 1, 2, 3\}$. Ihre gemeinsame Gewichtsfunktion lässt sich am einfachsten in der folgenden Tabelle darstellen:

x \ y	0	1	2	3
0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	0
1	0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$

◇

Beispiel 1 (gezinkter Würfel). Eine Urne enthält gleich viele gewöhnliche und gezinkte Würfel; bei den letzteren ist die 6 durch eine 7 ersetzt. Man zieht zufällig einen Würfel und würfelt damit.

Sei X die Art des gezogenen Würfels, wobei wir die Werte 0 für fair und 1 für gezinkt wählen, und Y die gewürfelte Zahl. Die gemeinsame Gewichtsfunktion von X und Y ist dann in der folgenden Tabelle gegeben:

x \ y	1	2	3	4	5	6	7
0	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	0
1	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	0	$\frac{1}{12}$

Sei ferner Z nur die Angabe, ob die gewürfelte Zahl gerade (wir schreiben dafür 0) oder ungerade (wir schreiben 1) ist. Die gemeinsame Gewichtsfunktion von X und Z ist dann durch die folgende Tabelle gegeben:

x \ z	0	1
0	$\frac{1}{4}$	$\frac{1}{4}$
1	$\frac{1}{6}$	$\frac{1}{3}$

Um die gemeinsame Gewichtsfunktion von X, Y und Z darzustellen, würde man schon eine dreidimensionale Matrix brauchen. $\diamond$

Wie schon erwähnt, beschreibt die gemeinsame Verteilungsfunktion (oder Gewichtsfunktion) von X und Y das gemeinsame Verhalten von X und Y. Daraus sollten wir das Verhalten von X oder Y jeweils für sich allein herleiten können, d.h. die Verteilungsfunktion (oder Gewichtsfunktion) von X bzw. Y. Dazu betrachten wir die sogenannten *Randverteilungen* von X bzw. Y.

Definition. Haben X und Y die gemeinsame Verteilungsfunktion F, so ist die Funktion $F_X : \mathbb{R} \to [0,1]$,

$$x \mapsto F_X(x) := P[X \leq x] = P[X \leq x, Y < \infty] = \lim_{y \to \infty} F(x,y)$$

die Verteilungsfunktion der *Randverteilung* (kurz RV) von X. Analog ist $F_Y : \mathbb{R} \to [0,1]$,

$$y \mapsto F_Y(y) := P[Y \leq y] = P[X < \infty, Y \leq y] = \lim_{x \to \infty} F(x,y)$$

die Verteilungsfunktion der Randverteilung von Y.

Sind X und Y diskrete Zufallsvariablen mit $\mathcal{W}(Y) = \{y_1, y_2, \dots\}$ und gemeinsamer Gewichtsfunktion $p(x,y)$, so ist die *Gewichtsfunktion der Randverteilung* von X gegeben durch $p_X : \mathcal{W}(X) \to [0,1]$,

$$x \mapsto p_X(x) = P[X = x] = \sum_{y_j \in \mathcal{W}(Y)} P[X = x, Y = y_j] = \sum_{y_j \in \mathcal{W}(Y)} p(x, y_j) \qquad \text{für } x \in \mathcal{W}(X).$$

Analog erhalten wir $p_Y(y) = \sum_{x_k \in \mathcal{W}(X)} p(x_k, y)$ für $y \in \mathcal{W}(Y)$. Die Gewichtsfunktion einer Randverteilung erhalten wir also, indem wir die andere(n) Variable(n) in der gemeinsamen Gewichtsfunktion gewissermassen "wegsummieren".

Bemerkung. Betrachtet man statt eines Paares (X,Y) einen Vektor $(X_1, \dots, X_n)$ von diskreten Zufallsvariablen, so definiert man analog die Randverteilung für jeden möglichen "Teilvektor" von $(X_1, \dots, X_n)$; es gibt dann also die n eindimensionalen Randverteilungen

von $X_1, \ldots, X_n$, ferner $\binom{n}{2}$ zweidimensionale Randverteilungen für alle Paare (X_k, X_ℓ) mit $k \neq \ell$, usw. Zu jeder Randverteilung erhält man dann auch die entsprechende Gewichtsfunktion durch "Wegsummieren der überflüssigen Variablen". ◇

Beispiel 8 (drei Münzwürfe). Beim dreimaligen Münzwurf mit X = Kopfanzahl im ersten Wurf und Y = Gesamtanzahl der Köpfe hatten wir die gemeinsame Gewichtsfunktion $p(x, y)$ wie folgt:

x \ y	0	1	2	3
0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	0
1	0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$

Hier ist

$$p_X(1) = P[X = 1] = \sum_{y_j \in \mathcal{W}(Y)} p(1, y_j) = 0 + \frac{1}{8} + \frac{2}{8} + \frac{1}{8} = \frac{1}{2},$$

und ebenso $p_X(0) = \frac{1}{2}$. Das ist natürlich auch das, was wir intuitiv erwarten, denn X nimmt offensichtlich mit gleicher Wahrscheinlichkeit die Werte 0 oder 1 an.

Analog ist z.B.

$$p_Y(1) = P[Y = 1] = \sum_{x_i \in \mathcal{W}(X)} p(x_i, 1) = \frac{2}{8} + \frac{1}{8} = \frac{3}{8},$$

und die anderen Werte von $p_Y(y)$ erhält man analog. Damit ergibt sich die folgende mit den Randverteilungen ergänzte Tabelle:

x \ y	0	1	2	3	$p_X(x)$
0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	0	$\frac{1}{2}$
1	0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	$\frac{1}{2}$
$p_Y(y)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	

Allgemein illustriert dieses Beispiel auch, dass man bei (zweidimensionalen) diskreten Zufallsvariablen die Gewichtsfunktionen der Randverteilungen als Zeilen- bzw. Spaltensummen der gemeinsamen Gewichtsfunktion erhält. ◇

Wie man oben sieht, kann man immer aus der gemeinsamen Verteilungsfunktion die Randverteilungen herleiten; das ist auch intuitiv klar. Umgekehrt ist ebenfalls einleuchtend, dass man aus den Randverteilungen allein in der Regel nicht die gemeinsame Verteilungsfunktion bestimmen kann, weil die Information über den Zusammenhang (die sogenannte *Abhängigkeitsstruktur*) zwischen den einzelnen Zufallsvariablen fehlt. Eine Möglichkeit für einen solchen Zusammenhang ist die Unabhängigkeit, die wir als Nächstes diskutieren.

Definition. Die Zufallsvariablen $X_1, \ldots, X_n$ heissen *unabhängig*, falls gilt

$$F(x_1, \ldots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n),$$

d.h. die gemeinsame Verteilungsfunktion F ist das Produkt der Verteilungsfunktionen F_{X_i} der (eindimensionalen) Randverteilungen.

Eine bequemere Beschreibung als über die Verteilungsfunktionen ist über die Gewichtsfunktionen: Die diskreten Zufallsvariablen $X_1, \ldots, X_n$ sind unabhängig genau dann, wenn gilt

$$p(x_1, \ldots, x_n) = p_{X_1}(x_1) \cdots p_{X_n}(x_n) \qquad \text{für alle } x_1, \ldots, x_n, \tag{2.2}$$

d.h. die gemeinsame Gewichtsfunktion p ist das Produkt der Gewichtsfunktionen p_{X_i} der (eindimensionalen) Randverteilungen. Das ist nicht schwierig zu beweisen, und wir verzichten hier darauf; siehe auch unten das Argument für (2.3).

Manchmal wird Unabhängigkeit von diskreten Zufallsvariablen auch direkt über die obige äquivalente Beschreibung (2.2) definiert. Unsere Definition hat den Vorteil, dass sie auch für allgemeine (nicht unbedingt diskrete) Zufallsvariablen sinnvoll ist.

Natürlich gibt es auch einen Zusammenhang zu unabhängigen Ereignissen, und zwar wie folgt. Die diskreten Zufallsvariablen $X_1, \ldots, X_n$ sind unabhängig genau dann, wenn

für beliebige Teilmengen $B_i \subseteq \mathcal{W}(X_i)$, $i = 1, \ldots, n$, die Ereignisse $A_i := \{X_i \in B_i\}$, $i = 1, \ldots, n$, unabhängig sind. Ebenfalls äquivalent ist, dass für beliebige Teilmengen $B_i \subseteq \mathcal{W}(X_i)$, $i = 1, \ldots, n$, gilt

$$P[X_1 \in B_1, \ldots, X_n \in B_n] = \prod_{i=1}^{n} P[X_i \in B_i]. \tag{2.3}$$

⌜ Aus (2.3) folgt sofort (2.2), wenn wir $B_i := \{x_i\}$, $i = 1, \ldots, n$, wählen. Um umgekehrt von (2.2) zu (2.3) zu kommen, benutzen wir die offensichtliche Beziehung

$$\{X_1 \in B_1, \ldots, X_n \in B_n\} = \bigcup_{x_1 \in B_1, \ldots, x_n \in B_n}^{\cdot} \{X_1 = x_1, \ldots, X_n = x_n\}$$

und analog

$$\{X_i \in B_i\} = \bigcup_{x_i \in B_i}^{\cdot} \{X_i = x_i\}.$$

Das liefert, dank (2.2) im dritten Schritt,

$$\begin{aligned} P[X_1 \in B_1, \ldots, X_n \in B_n] &= \sum_{x_1 \in B_1, \ldots, x_n \in B_n} P[X_1 = x_1, \ldots, X_n = x_n] \\ &= \sum_{x_1 \in B_1, \ldots, x_n \in B_n} p(x_1, \ldots, x_n) \\ &= \sum_{x_1 \in B_1, \ldots, x_n \in B_n} p_{X_1}(x_1) \cdots p_{X_n}(x_n) \\ &= \prod_{i=1}^{n} \sum_{x_i \in B_i} p_{X_i}(x_i) \\ &= \prod_{i=1}^{n} P[X_i \in B_i], \end{aligned}$$

und damit haben wir (2.3). ⌟

Beispiel 8 (drei Münzwürfe). Beim dreimaligen Münzwurf ist $p(1,1) = \frac{1}{8}$ und $p_X(1) = \frac{1}{2}$, $p_Y(1) = \frac{3}{8}$. Also sind X und Y hier nicht unabhängig, denn wir haben $p(1,1) \neq p_X(1)p_Y(1)$. Das ist auch anschaulich klar, wenn man sich überlegt, was X und Y beschreiben: Die Anzahl X der Köpfe beim ersten Wurf hängt ganz offensichtlich mit der Gesamtanzahl Y der geworfenen Köpfe zusammen. ◇

Beispiel. Wir nehmen ein gut gemischtes Spiel mit 36 *Jasskarten* und ziehen zufällig eine davon heraus. Seien X eine Zahl zwischen 1 und 4, welche die gezogene Farbe repräsentiert, und Y der Punktwert der gezogenen Karte (11 für As, 4 für König, 3 für Dame, 2 für Bube, 10 für Zehn), ohne Berücksichtigung von Trümpfen. Wir nehmen an, dass jede einzelne Karte mit gleicher Wahrscheinlichkeit gezogen wird. Dann ist X gleichverteilt auf der Menge $\{1, 2, 3, 4\}$, weil es von jeder Farbe gleich viele Karten gibt, und für jeden möglichen Wert y von Y gilt

$$P[X = x, Y = y] = \frac{1}{36} \times (\text{Anzahl der Karten mit Farbe } x \text{ und Wert } y).$$

Für jeden Wert von x (d.h. für jede Farbe) gibt es aber gleich viele Karten mit Wert y; also ist die obige Anzahl dieselbe für alle x, wir nennen sie einmal $A(y)$, and das gibt zuerst $P[X = x, Y = y] = \frac{1}{36}A(y)$ und daraus

$$P[Y = y] = \sum_{x=1}^{4} P[X = x, Y = y] = \frac{1}{9}A(y).$$

Damit sehen wir wegen

$$P[X = x, Y = y] = \frac{1}{4}P[Y = y] = P[X = x]P[Y = y]$$

also, dass X und Y unabhängig sind. ◇

Eine nützliche Eigenschaft der Unabhängigkeit von Zufallsvariablen ist, dass sie unter individuellen Transformationen erhalten bleibt. Um das genauer zu formulieren, seien $X_1, \ldots, X_n$ diskrete Zufallsvariablen und $Y_i = f_i(X_i)$ für beliebige Abbildungen $f_i : \mathbb{R} \to \mathbb{R}$, $i = 1, \ldots, n$. Wir wissen schon, dass $Y_1, \ldots, Y_n$ dann auch wieder diskrete Zufallsvariablen sind. Falls nun $X_1, \ldots, X_n$ unabhängig sind, so sind auch $Y_1, \ldots, Y_n$ unabhängig, denn für $y_i \in \mathcal{W}(Y_i)$ ist $B_i := \{x \in \mathcal{W}(X_i) : f_i(x) = y_i\} \subseteq \mathcal{W}(X_i)$, und damit folgt aus (2.3)

$$P[Y_1 = y_1, \ldots, Y_n = y_n] = P[X_1 \in B_1, \ldots, X_n \in B_n] = \prod_{i=1}^{n} P[X_i \in B_i] = \prod_{i=1}^{n} P[Y_i = y_i].$$

2.4 Funktionen von mehreren Zufallsvariablen

Seien $X_1, \ldots, X_n$ diskrete Zufallsvariablen und $Y = g(X_1, \ldots, X_n)$ für eine Funktion $g : \mathbb{R}^n \to \mathbb{R}$. Dann ist Y wieder eine diskrete Zufallsvariable, und wir können nach ihrer Verteilungsfunktion usw. fragen. Die wichtigsten Beispiele für g sind Summen und Produkte, und wir untersuchen diese hier etwas genauer.

Zuerst zeigen wir, dass der Erwartungswert sich immer linear verhält.

Satz 2.4. *Seien $X_1, \ldots, X_n$ diskrete Zufallsvariablen mit endlichen Erwartungswerten $E[X_1], \ldots, E[X_n]$. Sei $Y = a + \sum_{\ell=1}^n b_\ell X_\ell$ mit Konstanten $a, b_1, \ldots, b_n$. Dann gilt*

$$E[Y] = a + \sum_{\ell=1}^n b_\ell E[X_\ell].$$

Beweis. Weil die Erwartungswerte $E[X_\ell]$, $\ell = 1, \ldots, n$, alle existieren, sind im Folgenden alle Reihen absolut konvergent, und wir dürfen in beliebiger Reihenfolge summieren. Nach der alternativen Darstellung (2.1) für den Erwartungswert ist dann

$$\begin{aligned} E[Y] &= \sum_{\omega_i \in \Omega} Y(\omega_i) P[\{\omega_i\}] \\ &= \sum_{\omega_i \in \Omega} \left(a + \sum_{\ell=1}^n b_\ell X_\ell(\omega_i) \right) P[\{\omega_i\}] \\ &= a + \sum_{\ell=1}^n b_\ell \sum_{\omega_i \in \Omega} X_\ell(\omega_i) P[\{\omega_i\}] \\ &= a + \sum_{\ell=1}^n b_\ell E[X_\ell]. \end{aligned}$$

Das gibt die Behauptung. **q.e.d.**

Beispiel 9 (Permutationen). *Anzahl der Fixpunkte bei zufälligen Permutationen:* Sei Ω die Menge aller Permutationen der Zahlen $\{1, \ldots, n\}$, $\mathcal{F} = 2^\Omega$ die Potenzmenge von Ω und P die diskrete Gleichverteilung auf Ω. Dieser Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, P)$

ist dann ein Modell für *zufällige Permutationen*, d.h. für das zufällige Anordnen (oder Sortieren) von n verschiedenen Objekten.

Für $\omega \in \Omega$ ist

$$X(\omega) := \#\big\{i \in \{1, \ldots, n\} : \omega(i) = i\big\}$$

die Anzahl der *Fixpunkte* der Permutation ω. Diese Zufallsvariable taucht zum Beispiel auf, wenn wir n Personen ihre Hüte wegnehmen und sie dann zufällig gemischt wieder verteilen; dann ist X die Anzahl der Personen, die ihren eigenen Hut zurückbekommen. Was ist der Erwartungswert $E[X]$?

Benutzt man die Definition

$$E[X] = \sum_{x_k \in \mathcal{W}(X)} x_k P[X = x_k],$$

so muss man die Verteilung von X bestimmen, und das ist schwierig. Besser geht es via die Linearität in Satz 2.4. Setzen wir

$$A_i := \{\omega \in \Omega : \omega(i) = i\},$$

so ist nämlich

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

Weil P die diskrete Gleichverteilung ist, gilt offenbar

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

und damit erhalten wir ganz einfach

$$E[X] = \sum_{i=1}^{n} E[I_{A_i}] = \sum_{i=1}^{n} P[A_i] = 1.$$

In der obigen Interpretation mit den Hüten erhält also im Durchschnitt nur eine Person wieder den richtigen Hut — und zwar unabhänging von n. ◇

Wie oben gesehen ist der Erwartungswert einer Summe von Zufallsvariablen also immer die Summe der Erwartungswerte. Die *Varianz einer Summe* ist in der Regel komplizierter,

und wir betrachten das zuerst für $n = 2$ Summanden. Dann ist zunächst durch Ausquadrieren und Benutzen der Linearität

$$E[(X_1 + X_2)^2] = E[X_1^2] + E[X_2^2] + 2E[X_1X_2]$$

und damit, wegen $E[X_1 + X_2] = E[X_1] + E[X_2]$,

$$\begin{aligned}\mathrm{Var}[X_1 + X_2] &= E[(X_1 + X_2)^2] - (E[X_1 + X_2])^2 \\ &= E[X_1^2] + E[X_2^2] + 2E[X_1X_2] - ((E[X_1])^2 + (E[X_2])^2 + 2E[X_1]E[X_2]) \\ &= \mathrm{Var}[X_1] + \mathrm{Var}[X_2] + 2(E[X_1X_2] - E[X_1]E[X_2]).\end{aligned}$$

Die Grösse

$$\mathrm{Cov}(X_1, X_2) := E[X_1X_2] - E[X_1]E[X_2] = E\big[(X_1 - E[X_1])(X_2 - E[X_2])\big]$$

heisst die *Kovarianz* von X_1 und X_2, und damit haben wir allgemein die Formel

$$\mathrm{Var}[X_1 + X_2] = \mathrm{Var}[X_1] + \mathrm{Var}[X_2] + 2\mathrm{Cov}(X_1, X_2).$$

Wir bemerken auch, dass offenbar

$$\mathrm{Cov}(X, X) = \mathrm{Var}[X]$$

gilt.

Die obige Formel für $\mathrm{Var}[X_1 + X_2]$ verallgemeinert man leicht von 2 auf n Zufallsvariablen und erhält so die allgemeine *Summenformel für Varianzen*, nämlich

$$\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathrm{Var}[X_i] + 2\sum_{i<j} \mathrm{Cov}(X_i, X_j).$$

Falls $\mathrm{Cov}(X_1, X_2) = 0$ ist, so heissen die Zufallsvariablen X_1 und X_2 *unkorreliert*; die Zufallsvariablen $X_1, \ldots, X_n$ heissen *paarweise unkorreliert*, wenn alle Paare X_i, X_j mit $i \neq j$ unkorreliert sind. Damit folgt sofort

$$\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathrm{Var}[X_i] \qquad \text{für paarweise unkorrelierte } X_1, \ldots, X_n.$$

Im Gegensatz zum Erwartungswert verhält sich also die Varianz im Allgemeinen nicht additiv, sondern nur unter zusätzlichen Annahmen.

Nun betrachten wir Produkte von Zufallsvariablen.

Satz 2.5. *Seien $X_1, \ldots, X_n$ diskrete Zufallsvariablen mit endlichen Erwartungswerten $E[X_1], \ldots, E[X_n]$. Falls $X_1, \ldots, X_n$ unabhängig sind, so ist*

$$E\left[\prod_{i=1}^{n} X_i\right] = \prod_{i=1}^{n} E[X_i]$$

(und das beinhaltet auch, dass der Erwartungswert links existiert). Insbesondere sind dann $X_1, \ldots, X_n$ paarweise unkorreliert, und es gilt

$$\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathrm{Var}[X_i],$$

sofern $\mathrm{Var}[X_1], \ldots, \mathrm{Var}[X_n]$ alle existieren und endlich sind.

Beweis. Wegen Unabhängigkeit von $X_1, \ldots, X_n$ gilt

$$P[X_1 = x_1, \ldots, X_n = x_n] = \prod_{i=1}^{n} P[X_i = x_i] \qquad \text{für alle } x_i \in \mathcal{W}(X_i),\ i = 1, \ldots, n$$

und damit nach Satz 2.1

$$\begin{aligned} E\left[\prod_{i=1}^{n} X_i\right] &= \sum_{x_1,\ldots,x_n} \left(\prod_{i=1}^{n} x_i\right) P[X_1 = x_1, \ldots, X_n = x_n] \\ &= \sum_{x_1,\ldots,x_n} \left(\prod_{i=1}^{n} x_i P[X_i = x_i]\right) \\ &= \prod_{i=1}^{n} \left(\sum_{x_i \in \mathcal{W}(X_i)} x_i P[X_i = x_i]\right) \\ &= \prod_{i=1}^{n} E[X_i]. \end{aligned}$$

Weil natürlich für $i \neq j$ die Zufallsvariablen X_i, X_j auch unabhängig sind, folgt

$$\mathrm{Cov}(X_i, X_j) = E[X_i X_j] - E[X_i]E[X_j] = 0,$$

und der Rest ist dann klar. **q.e.d.**

Bemerkung. Es gelten die Implikationen

$$\text{unabhängig} \quad \Longrightarrow \quad \text{paarweise unabhängig} \quad \Longrightarrow \quad \text{unkorreliert},$$

und man kann sich durch Gegenbeispiele überlegen, dass beide umgekehrten Implikationen in der Regel falsch sind. ◇

Beispiel 9 (Permutationen). Sei wieder

$$X = \sum_{i=1}^{n} I_{A_i} =: \sum_{i=1}^{n} Y_i$$

mit

$$A_i := \big\{\text{Permutationen } \omega \text{ von } \{1, \ldots, n\} : \omega(i) = i\big\},$$

d.h. X ist die Anzahl der Fixpunkte einer zufälligen Permutation. Dann ist $Y_i = Y_i^2$ und damit

$$\mathrm{Var}[Y_i] = E[Y_i^2] - (E[Y_i])^2 = E[Y_i](1 - E[Y_i]) = \frac{1}{n}\left(1 - \frac{1}{n}\right) = \frac{n-1}{n^2}.$$

Für $i \neq j$ ist ferner

$$E[Y_i Y_j] = P[A_i \cap A_j] = \frac{|A_i \cap A_j|}{|\Omega|} = \frac{(n-2)\,!}{n\,!} = \frac{1}{n(n-1)}$$

und damit für $i \neq j$

$$\mathrm{Cov}(Y_i, Y_j) = E[Y_i Y_j] - E[Y_i]E[Y_j] = \frac{1}{n(n-1)} - \frac{1}{n^2}.$$

Also erhalten wir nach der Summenformel

$$\begin{aligned}
\mathrm{Var}[X] &= \sum_{i=1}^{n} \mathrm{Var}[Y_i] + 2\sum_{i<j} \mathrm{Cov}(Y_i, Y_j) \\
&= n\frac{n-1}{n^2} + 2\frac{n(n-1)}{2}\left(\frac{1}{n(n-1)} - \frac{1}{n^2}\right) \\
&= 1 - \frac{1}{n} + \left(1 - \frac{n-1}{n}\right) \\
&= 1.
\end{aligned}$$

◇

Wir haben schon gesehen, dass man für die Summe von Zufallsvariablen den Erwartungswert immer einfach berechnen kann, während die Varianz schon schwieriger ist. Nun untersuchen wir die Verteilungsfunktion oder besser (im diskreten Fall) die Gewichtsfunktion einer Summe von Zufallsvariablen.

Seien also X und Y diskrete Zufallsvariablen mit gemeinsamer Gewichtsfunktion $p(x, y)$. Dann ist auch ihre Summe $Z = X + Y$ diskret. Schreiben wir

$$\{Z = z\} = \bigcup_{x_k \in \mathcal{W}(X)} \{X = x_k, Y = z - x_k\},$$

so erhalten wir die Gewichtsfunktion von Z als

$$p_Z(z) = P[Z = z] = \sum_{x_k \in \mathcal{W}(X)} P[X = x_k, Y = z - x_k] = \sum_{x_k \in \mathcal{W}(X)} p(x_k, z - x_k)$$

oder via Symmetrie auch als

$$p_Z(z) = P[Z = z] = \sum_{y_j \in \mathcal{W}(Y)} P[X = z - y_j, Y = y_j] = \sum_{y_j \in \mathcal{W}(Y)} p(z - y_j, y_j).$$

Das gilt völlig allgemein.

Sind X und Y *unabhängig*, so ist $p(x, y) = p_X(x)p_Y(y)$ und damit

$$\begin{aligned} p_Z(z) &= \sum_{x_k \in \mathcal{W}(X)} p_X(x_k)p_Y(z - x_k) \\ &= \sum_{y_j \in \mathcal{W}(Y)} p_X(z - y_j)p_Y(y_j). \end{aligned}$$

Bei unabhängigen Summanden X und Y ist also die Gewichtsfunktion p_Z der Summe $Z = X + Y$ gerade die sogenannte *Faltung* der Gewichtsfunktionen p_X und p_Y. Wir schreiben das auch kurz als $p_Z = p_X * p_Y = p_Y * p_X$.

Beispiel 8 (drei Münzwürfe). Beim *dreimaligen Münzwurf* mit X = Kopfanzahl beim ersten Wurf und Y = Gesamtanzahl der Köpfe sind X und Y nicht unabhängig. Die Summe $Z = X+Y$ hat den Wertebereich $\mathcal{W}(Z) = \{0, 1, \ldots, 4\}$, und die Gewichtsfunktion

von Z ergibt sich aus der Tabelle mit der gemeinsamen Gewichtsfunktion als

$$\begin{aligned}
p_Z(0) &= p(0,0) = \frac{1}{8},\\
p_Z(1) &= p(0,1) + p(1,0) = \frac{2}{8} + 0 = \frac{2}{8},\\
p_Z(2) &= p(0,2) + p(1,1) = \frac{1}{8} + \frac{1}{8} = \frac{2}{8},\\
p_Z(3) &= p(0,3) + p(1,2) = 0 + \frac{2}{8} = \frac{2}{8},\\
p_Z(4) &= p(1,3) = \frac{1}{8}.
\end{aligned}$$

Dieses Beispiel illustriert also die allgemeine Formel (ohne Faltung). ◇

Beispiel. Seien $Y_1, \ldots, Y_n$ unabhängig mit Werten 0 und 1 und

$$P[Y_i = 1] = p = 1 - P[Y_i = 0];$$

man nennt die Y_i dann auch *unabhängige 0-1-Experimente mit Erfolgsparameter p*. Das ist ein Modell für die n-fache Wiederholung eines Experiments (z.B. Werfen einer möglicherweise gezinkten Münze) mit binärem Ausgang, und

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

ist die *Anzahl der Erfolge* (Einsen) in den n Versuchen. Man kann dann die obige Faltungsformel benutzen, um induktiv zu zeigen, dass die Gewichtsfunktion von X gegeben ist durch

$$P[X = k] = \binom{n}{k} p^k (1-p)^{n-k} \qquad \text{für } k = 0, 1, \ldots, n.$$

Das ist eine *Binomialverteilung* mit Parametern n und p; siehe Kapitel 3 für mehr Informationen (und eine einfachere Herleitung). ◇

2.5 Bedingte Verteilungen

Grundidee: Wir haben die gemeinsame Verteilung von zwei Zufallsvariablen und möchten Kenntnisse über die eine Zufallsvariable ausnutzen, um genauere Aussagen über die andere zu machen.

Definition. Seien X und Y diskrete Zufallsvariablen mit gemeinsamer Gewichtsfunktion $p(x, y)$. Die *bedingte Gewichtsfunktion* von X, gegeben dass $Y = y$, ist definiert durch

$$p_{X\,|\,Y}(x\,|\,y) := P[X = x\,|\,Y = y] = \frac{P[X = x, Y = y]}{P[Y = y]} = \frac{p(x, y)}{p_Y(y)}$$

für $p_Y(y) > 0$ und 0 sonst. Das legt dann auch die bedingte Verteilung von X gegeben $Y = y$ fest; siehe Abschnitt 2.1.

Bemerkungen. 1) Für jedes y mit $p_Y(y) > 0$ ist $p_{X\,|\,Y}(\,\cdot\,|\,y)$ als Funktion von x eine *Gewichtsfunktion*, denn offenbar ist $p_{X\,|\,Y}(x\,|\,y) \geq 0$ und

$$\sum_{x_k \in \mathcal{W}(X)} p_{X\,|\,Y}(x_k\,|\,y) = \frac{1}{p_Y(y)} \sum_{x_k \in \mathcal{W}(X)} p(x_k, y) = 1.$$

Das ist genau analog dazu, dass für $P[A] > 0$ die bedingte Wahrscheinlichkeit $P[\,\cdot\,|\,A]$ wieder ein Wahrscheinlichkeitsmass ist.

2) Aus der Charakterisierung der Unabhängigkeit via $p(x, y) = p_X(x)p_Y(y)$ sehen wir sofort, dass X und Y genau dann *unabhängig* sind, wenn für alle y mit $p_Y(y) > 0$ gilt

$$p_{X\,|\,Y}(x\,|\,y) = p_X(x) \qquad \text{für alle } x \in \mathcal{W}(X).$$

Eine symmetrische Aussage gilt, wenn man X und Y vertauscht. ◇

Beispiel 8 (drei Münzwürfe). Beim *dreimaligen Münzwurf* mit X = Kopfanzahl beim ersten Wurf und Y = Gesamtanzahl der Köpfe suchen wir die bedingte Verteilung von X, gegeben dass $Y = 1$. Die Tabelle mit gemeinsamer Gewichtsfunktion und den Gewichtsfunktionen der Randverteilungen ist

x \ y	0	1	2	3	$p_X(x)$
0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	0	$\frac{1}{2}$
1	0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	$\frac{1}{2}$
$p_Y(y)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	

und daraus erhalten wir sofort die gewünschte bedingte Gewichtsfunktion als

$$p_{X\,|\,Y}(0\,|\,1) = \frac{p(0,1)}{p_Y(1)} = \frac{\frac{2}{8}}{\frac{3}{8}} = \frac{2}{3},$$
$$p_{X\,|\,Y}(1\,|\,1) = \frac{p(1,1)}{p_Y(1)} = \frac{\frac{1}{8}}{\frac{3}{8}} = \frac{1}{3}.$$

Die bedingte Verteilung von X gegeben $Y = 1$ hat also unterschiedliche Wahrscheinlichkeiten für die möglichen Werte 0 und 1 von X, während X ohne Kenntnis von Y die Werte 0 und 1 mit gleicher Wahrscheinlichkeit annimmt; siehe Abschnitt 2.3. Das bedeutet (nicht überraschend), dass wir die Wahrscheinlichkeiten von Kopf und Zahl für den ersten Wurf anders einschätzen, wenn wir wissen, dass insgesamt genau einmal Kopf aufgetreten ist. Insbesondere erhalten wir genau wie im Abschnitt 2.3 wieder, dass X und Y abhängig sind. ◇

Beispiel. Sei X eine diskrete Zufallsvariable und A ein Ereignis mit $P[A] > 0$. Die *bedingte Verteilung von X gegeben A* ist dann definiert durch ihre Gewichtsfunktion

$$p_{X\,|\,A}(x) := P[X = x\,|\,A] = \frac{P[\{X = x\} \cap A]}{P[A]} \qquad \text{für } x \in \mathcal{W}(X).$$

Setzen wir $Y := I_A$, so ist

$$P[A] = P[Y = 1] = p_Y(1)$$

und

$$P[\{X = x\} \cap A] = P[X = x, Y = 1] = p(x, 1),$$

wenn wir mit p die gemeinsame Gewichtsfunktion von X und Y bezeichnen. Also ist die bedingte Gewichtsfunktion von X gegeben A gerade der Spezialfall, wo wir die bedingte

Gewichtsfunktion von X, gegeben dass $I_A = 1$, betrachten, d.h.

$$p_{X \mid A}(x) = \frac{p(x,1)}{p_{I_A}(1)} = p_{X \mid I_A}(x \mid 1).$$

$\diamond$

3 Wichtige diskrete Verteilungen

Ziel: Dieses Kapitel gibt eine Übersicht über einige der wichtigsten diskreten Verteilungen, mit Herleitungen und Angaben zur Verwendung in Beispielen.

Eine diskrete Verteilung ist nach Definition die Verteilung μ_X einer diskreten Zufallsvariablen X; siehe Abschnitt 2.1. Sie wird im Prinzip vollständig dadurch beschrieben, dass man den Wertebereich $\mathcal{W}(X)$ von X und die Gewichtsfunktion $p_X(x_k) = P[X = x_k]$, $x_k \in \mathcal{W}(X)$, spezifiziert. Für ein besseres Verständnis ist es aber nützlich, zusätzlich eine typische Situation zu beschreiben, in der diese Verteilung oder Zufallsvariable auftritt. In den meisten Fällen werden wir auch noch gleich den Erwartungswert $E[X]$ und die Varianz $\mathrm{Var}[X]$ berechnen oder zumindest angeben.

3.1 Diskrete Gleichverteilung

Die *diskrete Gleichverteilung* auf einer endlichen Menge $\mathcal{W} = \{x_1, \ldots, x_N\}$ gehört zu einer Zufallsvariablen X mit Wertebereich $\mathcal{W}$ und Gewichtsfunktion

$$p_X(x_k) = P[X = x_k] = \frac{1}{N} \qquad \text{für } k = 1, \ldots, N.$$

Das Typische hier ist also, dass alle individuellen Ergebnisse (Elementarereignisse) mit der gleichen Wahrscheinlichkeit auftreten. Beispiele sind das Ergebnis eines Würfelwurfs ($N = 6$, $\mathcal{W} = \{1, \ldots, 6\}$) oder eines Münzwurfs ($N = 2$, $\mathcal{W} = \{K, Z\}$), eine zufällig ausgewählte dreistellige Zahl ($N = 1000$, $\mathcal{W} = \{000, 001, \ldots, 999\}$ oder alternativ auch $N = 900$, $\mathcal{W} = \{100, \ldots, 999\}$), oder zufällige Permutationen von $\{1, \ldots, n\}$ ($N = n!$, $\mathcal{W} = \{\text{alle Permutationen von } \{1, \ldots, n\}\}$).

3.2 Unabhängige 0-1-Experimente

Um die nächsten vier diskreten Verteilungen gut motivieren und beschreiben zu können, führen wir in diesem Zwischenabschnitt einen allgemeinen einheitlichen Rahmen ein, in den sie sich alle schön einbetten lassen. Dazu betrachten wir eine Folge gleichartiger Experimente, die alle nur mit Erfolg oder Misserfolg enden können, und betrachten die Ereignisse

$$A_i = \{\text{Erfolg beim } i\text{-ten Experiment}\}.$$

Wir machen zwei *Annahmen*:

i) Die A_i sind unabhängig.

ii) $P[A_i] = p$ für alle i.

Die Annahme ii) formalisiert die Gleichartigkeit der Experimente; $p \in [0,1]$ heisst *Erfolgsparameter*. Ist z.B. das Experiment ein Würfelwurf und Erfolg das Würfeln einer Fünf, so ist $p = \frac{1}{6}$. Am einfachsten kann man sich das Ganze als eine Folge von (unabhängigen) Münzwürfen vorstellen, bei denen die "gewünschte" Seite mit Wahrscheinlichkeit p oben landet; es gibt aber natürlich beliebig viele andere Interpretationen. Mit $Y_i = I_{A_i}$ bezeichnen wir die *Indikatorfunktion* des Ereignisses A_i; also ist

$$Y_i(\omega) = \begin{cases} 1 & \text{für } \omega \in A_i, \\ 0 & \text{für } \omega \notin A_i, \end{cases}$$

und nimmt nur die Werte 0 oder 1 an, mit $P[Y_i = 1] = P[A_i] = p$. Wir codieren also die Folge der Ergebnisse als binäre Folge, indem wir 1 für Erfolg und 0 für Misserfolg schreiben. Unter den Annahmen i) und ii) bilden die Zufallsvariablen Y_i eine *Folge unabhängiger 0-1-Experimente* mit Erfolgsparameter p.

Ausgehend von einer Folge unabhängiger 0-1-Experimente mit Erfolgsparameter p können wir nun in den folgenden Abschnitten weitere wichtige Beispiele für diskrete Verteilungen einführen.

Bemerkung. Um die Existenz eines Wahrscheinlichkeitsraumes $(\Omega, \mathcal{F}, P)$ zu zeigen, auf dem es eine (unendliche) Folge von unabhängigen 0-1-Experimenten gibt, muss man schon ernsthaft arbeiten. Insbesondere ist dabei Ω überabzählbar. Wir gehen auf diese Aspekte hier aber nicht näher ein. $\diamond$

3.3 Bernoulli-Verteilung

Machen wir ein einziges 0-1-Experiment und nennen wir das Ergebnis X, so hat X eine *Bernoulli-Verteilung* mit Parameter p. Also ist $\mathcal{W}(X) = \{0, 1\}$, und die Gewichtsfunktion ist gegeben durch $p_X(1) = P[X = 1] = p$, $p_X(0) = P[X = 0] = 1 - p$, d.h. wir haben $p_X(x) = p^x(1-p)^{1-x}$ für $x \in \{0, 1\} = \mathcal{W}(X)$. Wir schreiben kurz $X \sim Be(p)$. Anders gesagt ist X die *Anzahl der Erfolge* bei einem *einzelnen 0-1-Experiment* mit Erfolgsparameter p; das kann also nur die Werte 0 oder 1 liefern.

Erwartungswert und Varianz einer Bernoulli-verteilten Zufallsvariablen $X \sim Be(p)$ berechnet man leicht als

$$E[X] = 1 \times P[X = 1] + 0 \times P[X = 0] = p$$

und, dank $X^2 = X$,

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

3.4 Binomialverteilung

Die *Binomialverteilung* mit Parametern n und p beschreibt die *Anzahl der Erfolge* bei n unabhängigen 0-1-Experimenten mit Erfolgsparameter p. In unserem allgemeinen Rahmen hat also die Zufallsvariable

$$X = \sum_{i=1}^{n} I_{A_i} = \sum_{i=1}^{n} Y_i$$

diese Verteilung. Der Wertebereich von X ist offenbar $\mathcal{W}(X) = \{0, 1, 2, \ldots, n\}$, und die Gewichtsfunktion ist

$$p_X(k) = P[X = k] = \binom{n}{k} p^k (1-p)^{n-k} \qquad \text{für } k = 0, 1, \ldots, n.$$

Wir schreiben kurz $X \sim Bin(n, p)$. Im Spezialfall $n = 1$ erhalten wir gerade die Bernoulli-Verteilung, d.h. $Bin(1, p) = Be(p)$.

Wie erhält man die obige Gewichtsfunktion? Wie schon am Ende von Abschnitt 2.4 erwähnt, könnten wir das induktiv mit Hilfe der Faltungsformel beweisen. Einfacher geht es aber mit einem direkten Beweis, und zwar wie folgt. Weil die Y_i unabhängig sind, gilt für jede feste 0-1-Folge $(y_1, \ldots, y_n)$ der Länge n

$$P[Y_1 = y_1, \ldots, Y_n = y_n] = \prod_{i=1}^{n} P[Y_i = y_i] = p^{\sum_{i=1}^n y_i} (1-p)^{n - \sum_{i=1}^n y_i},$$

denn im Produkt haben genau die Faktoren mit $y_i = 1$ den Wert p, und ihre Anzahl ist $\sum_{i=1}^n y_i$; die restlichen $n - \sum_{i=1}^n y_i$ Faktoren haben den Wert $1 - p$. Weiter ist $\{X = k\}$ die disjunkte Vereinigung aller Ereignisse der Form $\{Y_1 = y_1, \ldots, Y_n = y_n\}$ für 0-1-Folgen $(y_1, \ldots, y_n)$ mit $\sum_{i=1}^n y_i = k$, denn $X = k$ heisst gerade, dass k Experimente erfolgreich ausgehen und die anderen $n - k$ nicht. Schliesslich gibt es offenbar $\binom{n}{k}$ 0-1-Folgen der Länge n mit genau k Einsen, weil man diese k Einsen auf n Plätze zu verteilen hat. Also folgt

$$P[X = k] = \sum_{\substack{(y_1,\ldots,y_n) \in \{0,1\}^n \\ \text{mit } \sum_{i=1}^n y_i = k}} P[Y_1 = y_1, \ldots, Y_n = y_n] = \binom{n}{k} p^k (1-p)^{n-k}.$$

Eine binomialverteilte Zufallsvariable X ist also die Summe von n unabhängigen Bernoulli-verteilten Zufallsvariablen $Y_1, \ldots, Y_n$, die alle den gleichen Parameter p haben. Aus der Linearität des Erwartungswertes (Satz 2.4) und (dank der Unabhängigkeit der Y_i) der

Summenformel für Varianzen in Satz 2.5 folgt deshalb für $X \sim Bin(n,p)$ sofort

$$E[X] = \sum_{i=1}^{n} E[Y_i] = np,$$
$$\mathrm{Var}[X] = \sum_{i=1}^{n} \mathrm{Var}[Y_i] = np(1-p).$$

Beispiel 4 (Geburtstage). Wie gross ist die Wahrscheinlichkeit, dass von insgesamt n Personen genau k heute Geburtstag haben?

In unserer obigen Notation ist A_i das Ereignis, dass Person i heute Geburtstag hat. Wir nehmen an, dass die A_i unabhängig sind mit $P[A_i] = \frac{1}{365} =: p$ für alle i. Dann ist

$$X = \sum_{i=1}^{n} I_{A_i}$$

die Anzahl der Personen, die heute Geburtstag haben, und damit gilt $X \sim Bin(n, \frac{1}{365})$. Also ist

$$P[X=k] = \binom{n}{k} \left(\frac{1}{365}\right)^k \left(\frac{364}{365}\right)^{n-k} =: p(k;n).$$

Numerische Beispiele für $p(k;n)$ liefert die folgende Tabelle.

k \ n	23	100	200	300	365
0	0.9388	0.7601	0.5777	0.4391	0.3674
1	0.0593	0.2088	0.3174	0.3619	0.3684
2	0.0018	0.0284	0.0868	0.1486	0.1842
3	3.4×10^{-5}	0.0025	0.0157	0.0406	0.0612
≥ 4	5.0×10^{-9}	9.4×10^{-5}	2.5×10^{-4}	1.6×10^{-3}	3.6×10^{-3}

Die Berechnung in der obigen Tabelle benutzt wegen der Binomialkoeffizienten die Rekursion

$$p(k+1;n) = \frac{p}{1-p}\frac{n-k}{k+1}p(k;n),$$

die man leicht durch Einsetzen nachrechnet. ◇

3.5 Geometrische Verteilung

Betrachten wir nun eine unendliche Folge von unabhängigen 0-1-Experimenten mit Erfolgsparameter p. Sei X die Wartezeit auf den ersten Erfolg, also

$$X = \inf\{i \in \mathbb{N} : A_i \text{ tritt ein}\} = \inf\{i \in \mathbb{N} : Y_i = 1\},$$

d.h. X ist die Nummer des ersten erfolgreichen Experiments bzw. der Index der ersten Eins in unserer zufälligen 0-1-Folge. Dann hat X eine *geometrische Verteilung* mit Parameter p. Der Wertebereich von X ist offenbar $\mathcal{W}(X) = \mathbb{N} = \{1, 2, 3, \dots\}$, und die Gewichtsfunktion ist

$$p_X(k) = P[X = k] = p(1-p)^{k-1} \qquad \text{für } k = 1, 2, 3, \dots$$

Wir schreiben kurz $X \sim Geom(p)$.

⌜ $X = k$ heisst gerade, dass $A_1, \dots, A_{k-1}$ nicht eintreten und dass dann aber A_k eintritt; also ist wegen der Unabhängigkeit der Y_i und $P[Y_i = 1] = p$

$$P[X = k] = P[Y_1 = 0, \dots, Y_{k-1} = 0, Y_k = 1] = \left(\prod_{i=1}^{k-1} P[Y_i = 0]\right) P[Y_k = 1] = (1-p)^{k-1}p,$$

wie behauptet. ⌟

Die obigen $p_X(k)$ bilden tatsächlich eine Gewichtsfunktion, denn für jedes $k \in \mathcal{W}(X)$ gilt $0 \le p_X(k) \le 1$ (das ist klar) und $\sum_{k\in\mathcal{W}(X)} p_X(k) = 1$:

$$\sum_{k=1}^{\infty} p_X(k) = p\sum_{k=1}^{\infty}(1-p)^{k-1} = p\sum_{j=0}^{\infty}(1-p)^j = p\frac{1}{1-(1-p)} = 1$$

nach der Summenformel für die geometrische Reihe. ⌟

Um den Erwartungswert von X auf elegante Weise zu finden, berechnen wir zuerst

$$P[X > \ell] = P[Y_1 = 0, \ldots, Y_\ell = 0] = (1-p)^\ell.$$

Weil X Werte in $\mathbb{N}_0$ hat, folgt aus Teil 3 in Satz 2.2 also

$$E[X] = \sum_{\ell=0}^{\infty} P[X > \ell] = \sum_{\ell=0}^{\infty}(1-p)^\ell = \frac{1}{1-(1-p)} = \frac{1}{p}$$

wieder nach der Summenformel für die geometrische Reihe. Je kleiner also die Wahrscheinlichkeit p für Erfolg bei einem einzelnen Experiment ist, desto grösser ist (wie anschaulich einleuchtend) die durchschnittliche Wartezeit auf den ersten Erfolg.

Eine Rechnung gibt ferner

$$E[X(X-1)] = \sum_{k=1}^{\infty} k(k-1)p(1-p)^{k-1} = \frac{2(1-p)}{p^2},$$

und daraus erhält man

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

Beispiel. Betrachten wir eine Folge von Münzwürfen mit einer fairen Münze, so dass bei jedem Wurf mit gleicher Wahrscheinlichkeit $\frac{1}{2}$ Kopf (K) oder Zahl (Z) auftreten. Wie lange muss man im Schnitt warten, bis erstmals die Sequenz KKK auftaucht?

Nennen wir diese Wartezeit T_{KKK}, so ist eine exakte Berechnung von $E[T_{KKK}]$ zwar möglich, aber nicht ganz einfach. Wir begnügen uns hier deshalb mit einer Abschätzung. Dazu unterteilen wir die Folge der Münzwürfe in *disjunkte Blöcke* von jeweils drei Würfen und setzen

$$A_i := \{\text{der } i\text{-te disjunkte Block ist } KKK\}.$$

Weil wir die Dreierblöcke disjunkt wählen, sind dann die Ereignisse A_i unabhängig mit $P[A_i] = (\frac{1}{2})^3 = \frac{1}{8} = p$. Ist

$$X = \inf\{i \in \mathbb{N} : A_i \text{ tritt ein}\}$$

die Wartezeit auf den ersten Erfolg bei der zugehörigen Folge von unabhängigen 0-1-Experimenten, so gilt ferner

$$T_{KKK} \leq 3X,$$

denn wir sehen uns immer nur Blöcke der Länge 3 an und bemerken es dabei nicht, falls die Sequenz KKK bei einer Überlappung von zweien unserer disjunkten Blöcke auftaucht, und ein Wert $X = k$ bedeutet, dass wir insgesamt $3k$ einzelne Münzwürfe gemacht haben. Also erhalten wir aus der Monotonie in Satz 2.2

$$E[T_{KKK}] \leq E[3X] = 3\frac{1}{p} = 24.$$

Dieselbe obere Schranke erhalten wir (mit dem gleichen Argument) für jede andere gewünschte Dreiersequenz von Symbolen aus Z und K; also ist zum Beispiel auch

$$E[T_{KZK}] \leq 24.$$

Berechnet man allerdings die obigen Erwartungswerte exakt, so erhält man

$$E[T_{KKK}] = 14,$$
$$E[T_{KZK}] = 10.$$

Erstens ist die obige Abschätzung also relativ grob — und zweitens spielt es offensichtlich für die Wartezeit eine Rolle, welches Muster man haben möchte. ◇

Beispiel. Spielt man *Eile-mit-Weile*, so muss man in bestimmten Situationen warten, bis man eine Fünf würfelt. Mit welcher Wahrscheinlichkeit dauert das länger als (oder genau) drei Runden? Wie oft muss man würfeln, damit die Wahrscheinlichkeit für eine Fünf mindestens 90% ist? Welche Anzahl von Würfen ist die wahrscheinlichste Wartedauer?

Die Wartezeit X auf eine Fünf ist geometrisch verteilt mit Parameter $p = \frac{1}{6}$. Also ist

$$P[X > 3] = (1-p)^3 = \left(\frac{5}{6}\right)^3 = 0.5787$$

und

$$P[X = 3] = p(1-p)^2 = \frac{1}{6} \times \left(\frac{5}{6}\right)^2 = 0.1157.$$

Für die zweite Frage suchen wir ein ganzzahliges k mit $P[X \leq k] \geq 90\%$ bzw.

$$0.1 \geq P[X > k] = (1-p)^k,$$

also

$$k \geq \frac{\log 0.1}{\log(1-p)} = \frac{\log \frac{1}{10}}{\log \frac{5}{6}} = 12.63;$$

also muss man mindestens 13 Mal würfeln.

Schliesslich suchen wir für die letzte Frage dasjenige k, das

$$k \mapsto P[X = k] = p(1-p)^{k-1}$$

über $k \in \mathbb{N}$ maximiert. Aber diese Funktion ist strikt fallend (denn der Quotient

$$\frac{P[X = k+1]}{P[X = k]} = 1 - p$$

ist strikt kleiner als 1), und damit ist sie maximal für $k = 1$. Also ist der wahrscheinlichste Wert von X gerade 1. $\diamond$

Beispiel. In Cornflakes-Packungen findet man *Fussballer-Bildchen*, und zwar jeweils eines pro Packung. Wieviele Packungen muss man im Schnitt kaufen, bis man die Serie von n Bildchen vollständig hat?

Dieses Problem ist unter dem Namen *coupon collector problem* bekannt und taucht in sehr vielen Situationen auf. Eine mögliche Anwendung aus der Informatik wäre zum Beispiel die durchschnittliche Anzahl von Versionen, bis man in einem Programm alle Fehler gefunden hat (wobei man dann die Anzahl der Fehler am Anfang als bekannt annimmt).

Sei also in unserem Beispiel X die Anzahl der Packungen, die man kaufen muss, bis man die ganze Serie hat. Wir teilen X zur Berechnung von $E[X]$ wie folgt in eine Summe auf. Zunächst sei $X_1 = 1$ die Anzahl der gekauften Packungen, bis man das erste Bildchen hat. Iterativ sei dann X_k die Anzahl der gekauften Packungen nach dem Finden des $(k-1)$-ten neuen Bildchens, bis man wieder ein neues Bildchen hat. Dann ist

$$X = \sum_{k=1}^{n} X_k.$$

Als Wartezeit auf einen Erfolg ist jedes X_k geometrisch verteilt, und der Erfolgsparameter von X_k ist

$$p_k = \frac{n-(k-1)}{n},$$

weil man ja schon $k-1$ der insgesamt n Bildchen hat. Also ist

$$\begin{aligned} E[X] &= \sum_{k=1}^{n} E[X_k] = \sum_{k=1}^{n} \frac{1}{p_k} \\ &= \sum_{k=1}^{n} \frac{n}{n-k+1} = \frac{n}{n} + \frac{n}{n-1} + \cdots + \frac{n}{1} \\ &= n \sum_{k=1}^{n} \frac{1}{k} \approx n(\log n + \gamma) \qquad \text{für grosse } n, \end{aligned}$$

wobei $\gamma \approx 0.57$ die sogenannte Euler-Konstante ist. $E[X]$ wächst also etwa wie $n \log n$, insbesondere also schneller als n. ◇

3.6 Negativbinomiale Verteilung

Betrachten wir eine unendliche Folge von unabhängigen 0-1-Experimenten mit Erfolgsparameter p, so können wir für $r \in \mathbb{N}$ auch die Wartezeit auf den r-ten Erfolg betrachten. Dann hat X eine *negativbinomiale Verteilung* mit Parametern r und p. Das ist eine Verallgemeinerung der geometrischen Verteilung, die man als Spezialfall für $r = 1$ erhält. In unserem allgemeinen Rahmen lässt sich X schreiben als

$$X = \inf\left\{k \in \mathbb{N} : \sum_{i=1}^{k} I_{A_i} = r\right\} = \inf\left\{k \in \mathbb{N} : \sum_{i=1}^{k} Y_i = r\right\}.$$

Der Wertebereich von X ist offenbar $\mathcal{W}(X) = \{r, r+1, r+2, \dots\}$, und die Gewichtsfunktion ist

$$p_X(k) = P[X = k] = \binom{k-1}{r-1} p^r (1-p)^{k-r} \qquad \text{für } k = r, r+1, \dots$$

Wir schreiben kurz $X \sim NB(r, p)$.

$X = k$ heisst gerade, dass A_k eintritt und dass zudem in den vorherigen $k-1$ Experimenten genau $r-1$ Erfolge eingetreten sind. Nun gilt aber wegen Unabhängigkeit der Y_i und $P[Y_i = 1] = p$ für jede Familie von Indizes $\{i_1, \dots, i_{k-1}\} = \{1, \dots, k-1\}$

$$\begin{aligned}
&P[Y_{i_1} = 1, \dots, Y_{i_{r-1}} = 1, Y_{i_r} = 0, \dots, Y_{i_{k-1}} = 0, Y_k = 1] \\
&= \Big(\prod_{j=1}^{r-1} P[Y_{i_j} = 1] \Big) \Big(\prod_{j=r}^{k-1} P[Y_{i_j} = 0] \Big) P[Y_k = 1] \\
&= p^{r-1} (1-p)^{k-1-(r-1)} p \\
&= p^r (1-p)^{k-r},
\end{aligned}$$

und es gibt genau $\binom{k-1}{r-1}$ solche Möglichkeiten, in $k-1$ Versuchen $r-1$ Erfolge zu haben; man muss nämlich gerade die "erfolgreichen" Indizes $i_1, \dots, i_{r-1}$ aus $\{1, \dots, k-1\}$ auswählen, bzw. die $r-1$ Erfolge auf die insgesamt $k-1$ Experimente verteilen.

Bemerkung. Sind die Zufallsvariablen $X_1, \dots, X_r \sim Geom(p)$ und unabhängig, so ist ihre Summe $X := X_1 + \dots + X_r \sim NB(r, p)$. Das ist anschaulich klar aus der Interpretation als Wartezeiten, muss aber trotzdem bewiesen werden. Ein mathematisch sauberes Argument benutzt beispielsweise die starke Markov-Eigenschaft von Irrfahrten oder einen Beweis via Induktion über die Faltung. Als Folgerung erhält man für $X \sim NB(r, p)$

$$E[X] = \sum_{i=1}^{r} E[X_i] = \frac{r}{p},$$
$$\mathrm{Var}[X] = \sum_{i=1}^{r} \mathrm{Var}[X_i] = \frac{r(1-p)}{p^2}.$$

◇

3.7 Hypergeometrische Verteilung

In einer Urne seien n Gegenstände, davon r vom Typ 1 und $n-r$ vom Typ 2. Man zieht ohne Zurücklegen m der Gegenstände; die Zufallsvariable X beschreibe die Anzahl der Gegenstände vom Typ 1 in dieser Stichprobe vom Umfang m. Dann hat X eine *hypergeometrische Verteilung* mit Parametern $n \in \mathbb{N}$ und $m, r \in \{1, \ldots, n\}$. Der Wertebereich von X ist $\mathcal{W}(X) = \{0, 1, \ldots, \min(m, r)\}$, und X hat die Gewichtsfunktion

$$p_X(k) = \frac{\binom{r}{k}\binom{n-r}{m-k}}{\binom{n}{m}} \qquad \text{für } k \in \mathcal{W}(X).$$

Zur Herleitung der Gewichtsfunktion nehmen wir an, dass alle der insgesamt $\binom{n}{m}$ möglichen Stichproben gleich wahrscheinlich sind. $X = k$ heisst, dass wir aus den r Gegenständen vom Typ 1 gerade k und aus den restlichen $n-r$ die noch fehlenden $m-k$ erwischen müssen; die Anzahl der Stichproben mit $X = k$ ist also $\binom{\binom{r}{k}n-r}{m-k}$. Daraus ergibt sich $p_X(k) = P[X = k]$ wie in der obigen Formel.

Beispiel. Im Schweizer *Lotto*, wo man 6 aus 42 Zahlen richtig tippen soll, ist die Anzahl der richtig getippten Zahlen bei einem einzelnen Tipp hypergeometrisch verteilt mit Parametern $n = 42$, $r = 6$ und $m = 6$. Die Gewichte $p_X(k)$ sind hier explizit durch folgende Tabelle gegeben:

k	0	1	2	3	4	5	6
$p_X(k)$	0.3713	0.4312	0.1684	0.0272	0.0018	4.118×10^{-5}	1.906×10^{-7}

Zur Berechnung verwenden wir

$$p_X(0) = \frac{\binom{6}{0}\binom{36}{6}}{\binom{42}{6}} = \frac{(36\,!)^2}{30\,!42\,!} = \frac{36 \times \cdots \times 31}{42 \times \cdots \times 37}$$

und die Rekursion

$$p_X(k) = p_X(k-1)\frac{(6-k+1)^2}{k(30+k)}.$$

[Zum Vergleich die Werte aus Deutschland, wo man 6 aus 49 tippen muss:

k	0	1	2	3	4	5	6
$p_X(k)$	0.4360	0.4130	0.1324	0.0177	0.0010	1.85×10^{-5}	7.15×10^{-8}

Wie man sieht, sind die Erfolgsaussichten hier natürlich schlechter.]

◇

3.8 Poisson-Verteilung

Die *Poisson-Verteilung* mit Parameter $\lambda \in (0,\infty)$ ist eine Verteilung auf der Menge $\mathbb{N}_0 = \{0,1,2,\dots\}$ mit Gewichtsfunktion

$$p_X(k) = e^{-\lambda}\frac{\lambda^k}{k!} \qquad \text{für } k = 0,1,2,\dots$$

Ist eine Zufallsvariable X Poisson-verteilt mit Parameter λ, so schreiben wir dafür kurz $X \sim \mathcal{P}(\lambda)$.

Im Gegensatz zu den bisherigen Beispielen erhält man die Poisson-Verteilung nicht aus einem konkreten Experiment, sondern durch einen Grenzübergang aus der Binomialverteilung. Diese Herleitung zeigt auch, dass die Poissonverteilung zur Modellierung von *seltenen Ereignissen* geeignet ist. Wir machen zuerst diese Herleitung und erklären anschliessend, wie man sie interpretieren und motivieren kann.

Sei also X_n für jedes n eine Zufallsvariable mit $X_n \sim Bin(n,p_n)$ und $np_n = \lambda$. Lassen wir $n \to \infty$ gehen, so geht $p_n = \frac{\lambda}{n} \to 0$, und X_n beschreibt die Anzahl der Erfolge bei sehr vielen Versuchen mit sehr kleiner Erfolgswahrscheinlichkeit, also bei seltenen Ereignissen.

Dann gilt

$$\begin{aligned} P[X_n = k] &= \binom{n}{k} p_n^k (1-p_n)^{n-k} \\ &= \frac{n\,!}{k\,!(n-k)\,!} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k} \\ &= \frac{\lambda^k}{k\,!} \frac{n(n-1)\cdots(n-k+1)}{n^k} \left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^{-k}. \end{aligned}$$

Für festes k und $n \to \infty$ gilt dann auf der rechten Seite

$$\begin{aligned} \text{Term 2} &\longrightarrow 1, \\ \text{Term 3} &\longrightarrow e^{-\lambda}, \\ \text{Term 4} &\longrightarrow 1, \end{aligned}$$

und damit insgesamt

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

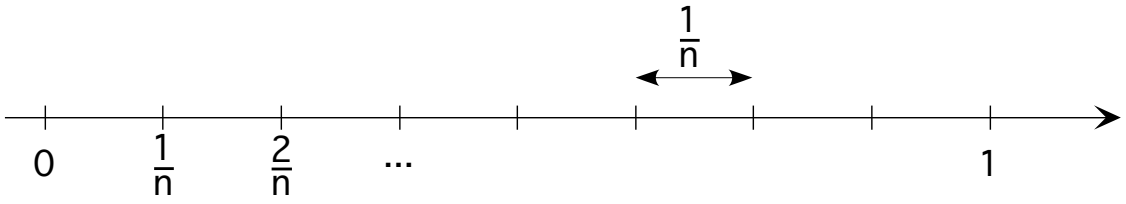
wenn $X \sim \mathcal{P}(\lambda)$. In diesem Sinn ist also die Poisson-Verteilung ein Grenzwert von Binomialverteilungen bei geeigneter Skalierung der Parameter. Etwas allgemeiner gilt das obige Argument auch noch, wenn statt $np_n = \lambda$ nur $np_n \to \lambda$ für $n \to \infty$ gilt.

Neben der Motivation für die Poisson-Verteilung hat das obige Limesargument auch eine praktische *Anwendung*. Für grosse n und kleine p kann man die (sonst eher unhandliche) Binomial-Wahrscheinlichkeit $\binom{n}{k}p^k(1-p)^{n-k}$ *approximativ* durch die Poisson-Wahrscheinlichkeit $e^{-\lambda}\frac{\lambda^k}{k\,!}$ berechnen, wobei $\lambda = np$ ist. Für den dabei entstehenden Fehler gibt es sehr gute theoretische Schranken; siehe z.B. [Krengel, Satz 5.9]. Als *Faustregel* gilt, dass man die Approximation für $np^2 \leq 0.05$ benutzen kann.

Die obige Herleitung erklärt teilweise, warum seltene Ereignisse mit einer Poisson-Verteilung modelliert werden. Beispiele sind:

- die Anzahl der Anrufe bei einer Telefonzentrale in einer bestimmten Periode
- die Anzahl der Grossschäden bei einer Versicherung in einer bestimmten Periode
- die Anzahl der Jobs, die bei einem Server in einer bestimmten Periode eintreffen
- usw.

Um den Zusammenhang mit dem obigen Limesargument herzustellen, betrachten wir als Periode das Zeitintervall $[0, 1]$ und unterteilen das in n Teilintervalle der Länge $\frac{1}{n}$:



Wir machen folgende Annahmen:

- In jedem Teilintervall kann höchstens ein Anruf/Grossschaden/Job eintreffen. (Das kann man zumindest für sehr grosse n vernünftig annehmen.)
- Die Wahrscheinlichkeit einer Ankunft ist dieselbe in jedem Teilintervall; wir nennen sie p_n.
- Die Ankünfte in verschiedenen Teilintervallen sind unabhängig.

Dann bilden die Ereignisse $A_i^{(n)}$ = "Ankunft im Teilintervall i" für jedes n eine (endliche) Familie von unabhängigen 0-1-Experimenten mit Erfolgsparameter p_n, und die Gesamtanzahl der Ankünfte in der n-ten Approximation ist

$$X_n = \sum_{i=1}^{n} I_{A_i^{(n)}} \sim Bin(n, p_n).$$

Für $n \to \infty$ und $np_n \to \lambda$ lässt sich das durch $\mathcal{P}(\lambda)$ approximieren, wie wir in der obigen Herleitung gesehen haben. Wegen $E[X_n] = np_n \approx \lambda$ ist dabei λ etwa die *durchschnittliche Anzahl der Ankünfte pro Zeiteinheit* in der betrachteten Periode. Der Parameter λ hat also auch eine anschauliche Interpretation.

Für eine Poisson-verteilte Zufallsvariable $X \sim \mathcal{P}(\lambda)$ ist

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

Analog berechnet man $E[X(X-1)] = \lambda^2$ und erhält so

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

Beispiel. In der preussischen Kavallerie wurden im 19. Jahrhundert über einen Zeitraum von 20 Jahren bei 10 Korps (Regimenten) die Anzahlen der *Todesfälle durch Hufschlag* gezählt. (Ein Regiment umfasste in der Regel 4000 bis 5000 Mann.) Dabei ergaben sich in diesen 200 Datensätzen die folgenden Daten:

Anzahl der Todesfälle	0	1	2	3	4	≥ 5
Anzahl der Beobachtungen	109	65	22	3	1	0
Relative Häufigkeit	0.545	0.325	0.110	0.015	0.005	0

Von Interesse ist hier die Zufallsvariable X, welche die Anzahl der Todesfälle pro (typisches) Jahr und pro (typisches) Regiment beschreibt. Die Gesamtanzahl der Todesfälle ist

$$65 \times 1 + 22 \times 2 + 3 \times 3 + 1 \times 4 = 122.$$

Die durchschnittliche Anzahl pro Einheit ist also bei den 200 Datensätzen

$$\lambda := \frac{122}{200} = 0.61.$$

Für $X \sim \mathcal{P}(\lambda)$ ergeben sich dann folgende Wahrscheinlichkeiten:

Anzahl der Todesfälle	0	1	2	3	4	≥ 5
Anzahl der Beobachtungen	109	65	22	3	1	0
Relative Häufigkeit	0.545	0.325	0.110	0.015	0.005	0
$p_X(k) = P[X = k]$	0.543	0.331	0.101	0.021	0.0003	4.3×10^{-4}

Wie man sieht, ist die Übereinstimmung der theoretischen Poisson-Wahrscheinlichkeiten mit den beobachteten relativen Häufigkeiten bemerkenswert gut.

Sowohl die Daten als auch die Idee der Anpassung einer Poissonverteilung in diesem Beispiel stammen vom Ökonomen und Statistiker von Bortkiewicz (1898). $\diamond$

4 Allgemeine Zufallsvariablen

In diesem Kapitel führen wir allgemeine (nicht unbedingt diskrete) Zufallsvariablen und Verteilungen ein. Die meisten Resultate sind dabei gleich oder analog wie im diskreten Fall, so dass wir uns relativ kurz fassen können.

4.1 Grundbegriffe

Definition. Sei $(\Omega, \mathcal{F}, P)$ ein Wahrscheinlichkeitsraum, also Ω ein Grundraum, $\mathcal{F} \subseteq 2^\Omega$ eine σ-Algebra von beobachtbaren Ereignissen und P ein Wahrscheinlichkeitsmass auf $\mathcal{F}$. Eine (reellwertige) *Zufallsvariable (ZV)* auf Ω ist eine messbare Funktion $X : \Omega \to \mathbb{R}$; das bedeutet, dass die Menge $\{X \leq t\} = \{\omega : X(\omega) \leq t\}$ für jedes t ein (beobachtbares) Ereignis, also in $\mathcal{F}$, sein muss. Die *Verteilungsfunktion (VF)* von X ist dann die Abbildung $F_X : \mathbb{R} \to [0, 1]$,

$$t \mapsto F_X(t) := P[X \leq t] := P[\{\omega : X(\omega) \leq t\}].$$

Bemerkung. In allen Beispielen in dieser Vorlesung sind die auftretenden Abbildungen X immer messbar. Und natürlich ist jede diskrete Zufallsvariable im Sinne von Kapitel 1.2 auch eine Zufallsvariable im obigen Sinn. ◇

Jede Verteilungsfunktion F_X hat die folgenden *Eigenschaften*:

1) F_X ist *wachsend* und *rechtsstetig*; das bedeutet, dass $F_X(s) \leq F_X(t)$ für $s \leq t$ gilt und $F_X(u) \to F_X(t)$ für $u \to t$ mit $u > t$.

2) $\lim_{t\to-\infty} F_X(t) = 0$, $\lim_{t\to+\infty} F_X(t) = 1$.

Umgekehrt kann man zeigen, dass jede Funktion F mit den Eigenschaften 1) und 2) die Verteilungsfunktion F_X einer Zufallsvariablen X ist.

Das stochastische Verhalten einer Zufallsvariablen X wird durch ihre *Verteilung* beschrieben; das ist dasjenige Wahrscheinlichkeitsmass μ_X auf $\mathbb{R}$, das durch

$$\mu_X(B) := P[X \in B]$$

definiert ist. Man kann zeigen, dass das Mass μ_X festgelegt ist, sobald man die Verteilungsfunktion F_X kennt, und umgekehrt ist natürlich

$$F_X(t) = \mu_X\big((-\infty, t]\big),$$

so dass F_X auch durch μ_X festgelegt ist.

Bei diskreten Zufallsvariablen haben wir meistens mit der Gewichtsfunktion gearbeitet, d.h. mit der Funktion $x \mapsto p_X(x) = P[X = x]$ für $x \in \mathcal{W}(X)$. Das Analogon im allgemeinen Fall ist die sogenannte Dichtefunktion, sofern eine existiert.

Definition. Eine Zufallsvariable X mit Verteilungsfunktion $F_X(t) = P[X \leq t]$ heisst *(absolut)stetig mit Dichte(funktion)* $f_X : \mathbb{R} \to [0, \infty)$, falls gilt

$$F_X(t) = \int_{-\infty}^{t} f_X(s)\,\mathrm{d}s \qquad \text{für alle } t \in \mathbb{R}.$$

Bemerkung. Genaugenommen heisst X stetig, falls F_X nur stetig ist, während bei einer Zufallsvariablen X mit einer Dichte die Verteilungsfunktion F_X sogar (fastüberall) differenzierbar ist. Wir werden in der Regel aber die kürzere Terminologie "stetig mit Dichte" benutzen. ◇

Eine Dichtefunktion f_X hat (analog zu einer Gewichtsfunktion im diskreten Fall) stets folgende *Eigenschaften*:

i) $f_X \geq 0$, und $f_X = 0$ ausserhalb von $\mathcal{W}(X)$.

ii) $\displaystyle\int_{-\infty}^{\infty} f_X(s)\,\mathrm{d}s = 1$; das folgt aus $\lim_{t\to+\infty} F_X(t) = 1$.

In allen praktisch auftretenden Beispielen ist zudem f_X stetig oder zumindest stückweise stetig.

Umgekehrt kann man zu einer gegebenen (messbaren) Funktion $f : \mathbb{R} \to [0, \infty)$ mit

$$\int_{-\infty}^{\infty} f(s)\,\mathrm{d}s = 1$$

eine stetige Zufallsvariable X konstruieren, deren Dichtefunktion f_X gerade f ist; das ist analog zur Situation bei der Gewichtsfunktion p_X im diskreten Fall.

Die Dichtefunktion f_X ist (fast exakt) das *stetige Analogon* zur Gewichtsfunktion p_X einer diskreten Zufallsvariablen. Das sieht man wie folgt. Für $a < b$ ist zunächst

$$P[a < X \le b] = P[X \le b] - P[X \le a] = F_X(b) - F_X(a) = \int_a^b f_X(s)\,\mathrm{d}s.$$

Allgemeiner gilt für (messbare) Mengen $B \subseteq \mathbb{R}$

$$P[X \in B] = \int_B f_X(s)\,\mathrm{d}s;$$

das ist das Analogon zur Beziehung

$$P[X \in B] = \sum_{x_k \in B} p_X(x_k)$$

für diskrete Zufallsvariablen.

Also ist

$$P[t - \varepsilon < X \le t + \varepsilon] = \int_{t-\varepsilon}^{t+\varepsilon} f_X(s)\,\mathrm{d}s.$$

Für $\varepsilon \searrow 0$ folgt daraus

$$P[X = t] = 0 \qquad \text{für jedes } t \in \mathbb{R};$$

in diesem Sinn hat also X keine Gewichtsfunktion wie eine diskrete Zufallsvariable. Ist zudem f_X stetig an der Stelle t, so haben wir aber auch

$$f_X(t) = \lim_{\varepsilon \searrow 0} \frac{1}{\varepsilon} \int_t^{t+\varepsilon} f_X(s)\,\mathrm{d}s = \lim_{\varepsilon \searrow 0} \frac{P[t < X \le t + \varepsilon]}{\varepsilon}$$

und damit

$$P\big[X \in (t, t + \varepsilon]\big] \approx \varepsilon f_X(t) \qquad \text{für } \varepsilon \text{ klein.}$$

Formal schreibt man dafür häufig

$$\text{“}P\big[X \in (t, t+\mathrm{d}t]\big] = f_X(t)\,\mathrm{d}t\text{”}.$$

Analog liefert der Hauptsatz der Infinitesimalrechnung

$$f_X(t) = \frac{\mathrm{d}}{\mathrm{d}t} F_X(t) = F_X'(t)$$

an jeder Stelle t, wo f_X stetig ist. In allen vernünftigen Situationen gilt also

$$\text{Dichtefunktion} = \text{Ableitung der Verteilungsfunktion},$$

so dass Verteilungsfunktion und Dichte auseinander durch Integrieren bzw. Differenzieren entstehen.

Einfache *Merkregel*: Vom diskreten zum stetigen Fall kommt man, indem man die Kombination (Summe, Gewichtsfunktion) systematisch durch (Integral, Dichtefunktion) ersetzt. Wir werden das wiederholt an mehreren Orten antreffen.

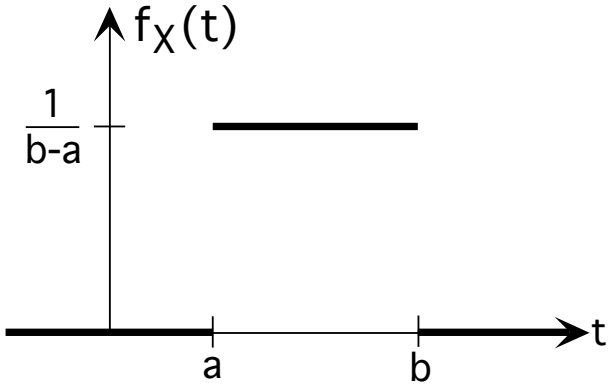
4.2 Wichtige stetige Verteilungen

Ziel: Vorstellung von einigen stetigen Verteilungen mit Dichte, sowie Angaben zur Verwendung in Beispielen.

4.2.1 Gleichverteilung

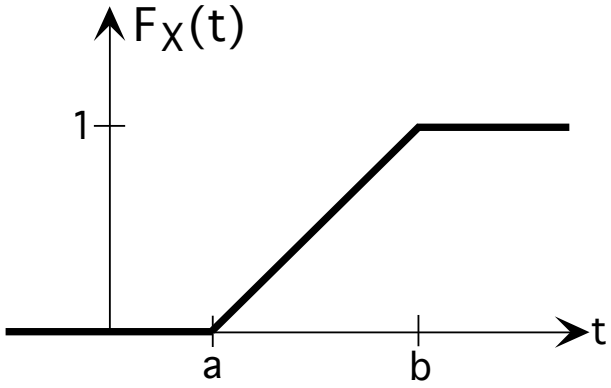
Die *Gleichverteilung* auf einem Intervall $[a, b]$ ist ein Modell für die zufällige Wahl eines Punktes in $[a, b]$. Die zugehörige Zufallsvariable X hat den Wertebereich $\mathcal{W}(X) = [a, b]$, die Dichtefunktion

$$f_X(t) = \begin{cases} \frac{1}{b-a} & \text{für } a \le t \le b, \\ 0 & \text{sonst,} \end{cases}$$



und die Verteilungsfunktion

$$F_X(t) = \begin{cases} 0 & \text{für } t < a, \\ \frac{t-a}{b-a} & \text{für } a \le t \le b, \\ 1 & \text{für } t > b. \end{cases}$$

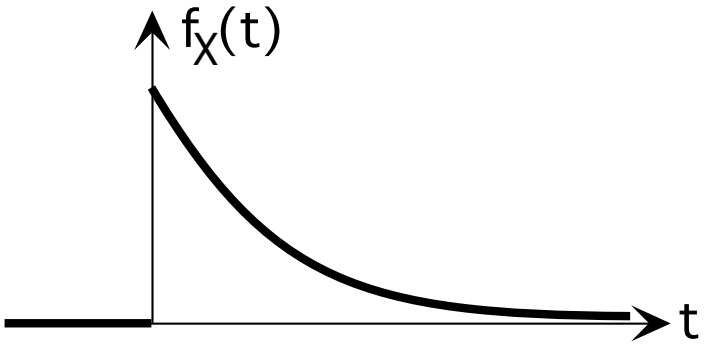


Wir schreiben kurz $X \sim \mathcal{U}(a, b)$; dabei steht $\mathcal{U}$ für uniform. Ein wichtiger Spezialfall ist $\mathcal{U}(0, 1)$, d.h. $a = 0$, $b = 1$, wo die obigen Formeln auch einfacher aussehen.

4.2.2 Exponentialverteilung

Die *Exponentialverteilung* mit Parameter $\lambda > 0$ ist das stetige Analogon der geometrischen Verteilung. Die zugehörige Zufallsvariable X hat $\mathcal{W}(X) = [0, \infty)$, Dichte

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



und Verteilungsfunktion

$$F_X(t) = \int_{-\infty}^{t} f_X(s)\,\mathrm{d}s = \begin{cases} 1 - e^{-\lambda t} & \text{für } t \geq 0, \\ 0 & \text{für } t < 0. \end{cases}$$

Wir schreiben $X \sim Exp(\lambda)$.

Wie die geometrische Verteilung ist die Exponentialverteilung ein Modell für *Wartezeiten* oder *Lebensdauern*; der einzige Unterschied besteht darin, dass nicht nur ganzzahlige, sondern beliebige Werte ≥ 0 angenommen werden können. Auch wie die geometrische Verteilung hat sie die Eigenschaft der *"Gedächtnislosigkeit"* in dem Sinne, dass gilt

$$P[X > t + s \,|\, X > s] = P[X > t],$$

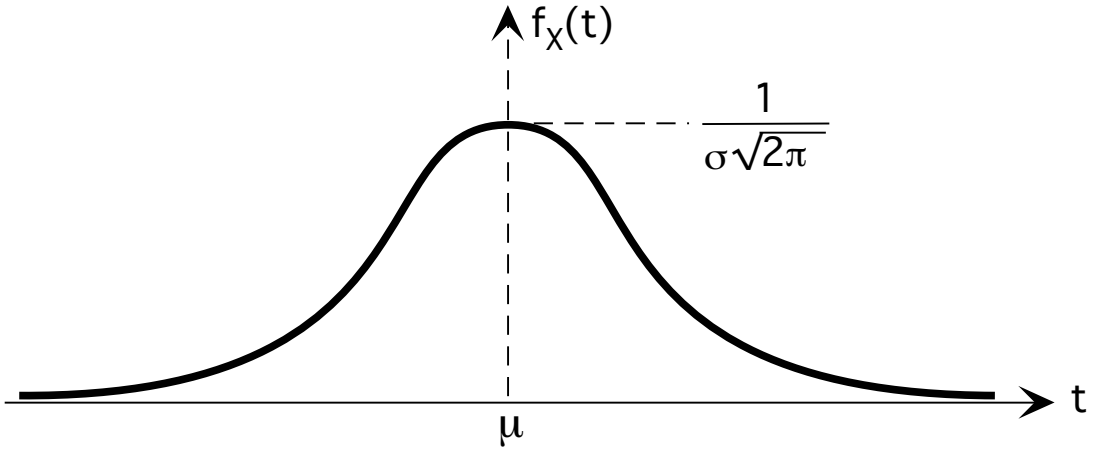
d.h. die Wahrscheinlichkeit, noch eine Zeitdauer t zu überleben, hängt nicht vom schon erreichten Alter s ab. (Bei einer konkreten Anwendung sollte man sich jeweils überlegen, ob das im betrachteten Modell sinnvoll ist.)

$$
\begin{aligned}
P[X > t+s \,|\, X > s] &= \frac{P[X > t+s, X > s]}{P[X > s]} = \frac{P[X > t+s]}{P[X > s]} \\
&= \frac{1 - F_X(t+s)}{1 - F_X(s)} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} = e^{-\lambda t} \\
&= 1 - F_X(t) = P[X > t].
\end{aligned}
$$

4.2.3 Normalverteilung

Die *Normalverteilung*, auch kurz NV oder Gauss-Verteilung genannt, spielt eine absolut zentrale Rolle. Sie hat zwei Parameter $\mu \in \mathbb{R}$, $\sigma^2 > 0$. Die zugehörige Zufallsvariable X hat den Wertebereich $\mathcal{W}(X) = \mathbb{R}$ und die Dichtefunktion

$$
f_X(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} \qquad \text{für } t \in \mathbb{R}.
$$



Die Dichte ist offenbar symmetrisch um μ und hat eine charakteristische glockenförmige Gestalt. Wir schreiben $X \sim \mathcal{N}(\mu, \sigma^2)$. Wir werden später noch sehen, dass μ der Erwartungswert und σ^2 die Varianz von X sind, so dass die Parameter auch eine anschauliche Bedeutung haben.

Ein wichtiger Spezialfall ist die *Standard-Normalverteilung* mit $\mu = 0$, $\sigma^2 = 1$, also $\mathcal{N}(0,1)$. Die zugehörige Dichte wird mit $\varphi(t)$ und die Verteilungsfunktion mit $\Phi(t)$ bezeichnet. Weder für F_X noch für Φ gibt es einen geschlossenen Ausdruck, aber das Integral

$$\Phi(t) = \int_{-\infty}^{t} \varphi(s)\,\mathrm{d}s = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{1}{2}s^2}\,\mathrm{d}s$$

ist *tabelliert*.

Ist $X \sim \mathcal{N}(\mu, \sigma^2)$, so ist $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$, also

$$F_X(t) = P[X \le t] = P\left[\frac{X-\mu}{\sigma} \le \frac{t-\mu}{\sigma}\right] = \Phi\left(\frac{t-\mu}{\sigma}\right);$$

deshalb genügt es, Φ zu tabellieren.

$$P\left[\frac{X-\mu}{\sigma} \le v\right] = \Phi(v) \qquad \text{für } X \sim \mathcal{N}(\mu, \sigma^2),$$

denn die Variablentransformation $y = \frac{s-\mu}{\sigma}$, also $s = \sigma y + \mu$, $\mathrm{d}s = \sigma\,\mathrm{d}y$ liefert

$$\begin{aligned} P\left[\frac{X-\mu}{\sigma} \le v\right] &= P[X \le \mu + \sigma v] \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\mu+\sigma v} e^{-\frac{(s-\mu)^2}{2\sigma^2}}\,\mathrm{d}s \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{v} e^{-\frac{1}{2}y^2} \sigma\,\mathrm{d}y = \int_{-\infty}^{v} \varphi(y)\,\mathrm{d}y = \Phi(v); \end{aligned}$$

also ist $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$.

Es gibt sehr viele Phänomene, die mit einer Normalverteilung modelliert werden; das ist je nach Situation unterschiedlich sinnvoll und sollte deshalb immer individuell betrachtet werden. *Beispiele* sind

- die Streuung von Messwerten um ihren Mittelwert
- die Gewichte oder Grössen von Individuen in einer grossen Bevölkerung
- die Leistungen in IQ-Tests

– die Renditen von Aktien: ist S_t der Endkurs am Tag t, so ist die Rendite (in Prozent) am Tag $t+1$ gegeben als

$$X_{t+1} = 100\frac{S_{t+1} - S_t}{S_t}$$

– usw.

Eine Erklärung für das häufige Auftreten der Normalverteilung ist der *zentrale Grenzwertsatz*; siehe später in Kapitel 5. Er besagt (unter gewissen Annahmen), dass Zufallsvariablen, die sich als Summen oder Mittelwerte von vielen unabhängigen "gleichartigen" Zufallsvariablen darstellen lassen, approximativ normalverteilt sind.

4.3 Erwartungswerte

Ist X eine beliebige reellwertige Zufallsvariable, so kann man X immer durch eine Folge von diskreten Zufallsvariablen approximieren; für $X \geq 0$ kann man zum Beispiel

$$X_n := \sum_{k=1}^{n2^n} (k-1)2^{-n} I_{\{(k-1)2^{-n} \leq X < k2^{-n}\}} + nI_{\{X \geq n\}}$$

für $X_n \nearrow X$ wählen. Den Erwartungswert von $X \geq 0$ erhält man dann als

$$E[X] := \lim_{n\to\infty} E[X_n];$$

dabei ist rechts jeder Erwartungswert wohldefiniert, weil jedes X_n eine diskrete Zufallsvariable ist, aber natürlich muss (und kann) man noch zeigen, dass der Limes existiert und nicht von der approximierenden Folge von diskreten Zufallsvariablen abhängt. Beliebige Zufallsvariablen X zerlegt man als

$$X = X^+ - X^- := \max(X, 0) - \max(-X, 0)$$

mit $X^+ \geq 0$, $X^- \geq 0$ und setzt $E[X] := E[X^+] - E[X^-]$, sofern beide diese Werte endlich sind; andernfalls existiert der Erwartungswert von X nicht (zumindest nicht in $\mathbb{R}$).

Ist X stetig mit Dichte $f_X(x)$, so liefert das

$$E[X] = \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x,$$

sofern das Integral absolut konvergiert, d.h. falls $\int_{-\infty}^{\infty} |x| f_X(x)\,\mathrm{d}x < \infty$. Ist das Integral nicht absolut konvergent, so existiert der Erwartungswert nicht (zumindest nicht in $\mathbb{R}$). Man beachte, dass das wieder völlig analog zum diskreten Fall ist.

Beispiel. *Gleichverteilung* auf $[a,b]$: Für $X \sim \mathcal{U}(a,b)$ ist $\mathcal{W}(X) = [a,b]$ und die Dichte $f_X(x) = \frac{1}{b-a}$ für $a \le x \le b$ und 0 sonst. Also ist

$$E[X] = \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x = \int_a^b x \frac{1}{b-a}\,\mathrm{d}x = \frac{1}{b-a}\frac{1}{2}x^2\Big|_a^b = \frac{1}{2}\frac{b^2-a^2}{b-a} = \frac{a+b}{2}.$$

Das ist gerade der Mittelpunkt oder Schwerpunkt des Intervalls $[a,b]$, und auch das, was man anschaulich erwartet. ◇

Beispiel. *Normalverteilung:* Für $X \sim \mathcal{N}(\mu,\sigma^2)$ haben wir $\mathcal{W}(X) = \mathbb{R}$ und die Dichtefunktion $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{1}{2}(\frac{x-\mu}{\sigma})^2)$. Also ist

$$E[X] = \int_{-\infty}^{\infty} x \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}\,\mathrm{d}x.$$

Um das zu berechnen, betrachten wir zuerst den Spezialfall $\mu = 0$, $\sigma^2 = 1$. Für das entsprechende $Y \sim \mathcal{N}(0,1)$ ist

$$E[Y] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y e^{-\frac{1}{2}y^2}\,\mathrm{d}y = 0,$$

weil der Integrand $g(y) = y e^{-\frac{1}{2}y^2}$ eine ungerade Funktion auf $\mathbb{R}$ ist, d.h. $g(-y) = -g(y)$; deshalb ist

$$\int_{-\infty}^{0} g(y)\,\mathrm{d}y = -\int_0^{\infty} g(y)\,\mathrm{d}y.$$

Für allgemeine $\mu \in \mathbb{R}$, $\sigma > 0$ machen wir die Variablentransformation $z = \frac{x-\mu}{\sigma}$ mit $x = \sigma z + \mu$, $\mathrm{d}x = \sigma\,\mathrm{d}z$ und erhalten so

$$\begin{aligned} E[X] &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (\sigma z + \mu) e^{-\frac{1}{2}z^2} \sigma\,\mathrm{d}z \\ &= \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z e^{-\frac{1}{2}z^2}\,\mathrm{d}z + \mu \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2}\,\mathrm{d}z \\ &= 0 + \mu \int_{-\infty}^{\infty} \varphi(z)\,\mathrm{d}z = \mu \times 1 = \mu, \end{aligned}$$

weil $\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$ als Dichtefunktion (einer $\mathcal{N}(0,1)$-Verteilung) Integral 1 hat.

Alternativ können wir etwas kürzer wie folgt argumentieren. Ist $X \sim \mathcal{N}(\mu, \sigma^2)$, so wissen wir schon, dass $Y := \frac{X-\mu}{\sigma}$ $\mathcal{N}(0,1)$-verteilt ist. Also ist nach obiger Rechnung $E[Y] = 0$, und wegen $X = \sigma Y + \mu$ folgt sofort

$$E[X] = \sigma E[Y] + \mu = \mu.$$

Weil die Dichtefunktion $f_X(x)$ um $x = \mu$ symmetrisch ist, ist μ gerade der "Massenschwerpunkt" der Verteilung. So gesehen hätte man das Ergebnis $E[X] = \mu$ also auch erraten können. ◇

Beispiel. *Cauchy-Verteilung*: Eine Cauchy-verteilte Zufallsvariable X hat den Wertebereich $\mathcal{W}(X) = \mathbb{R}$ und die Dichtefunktion

$$f_X(x) = \frac{1}{\pi} \frac{1}{1+x^2} \qquad \text{für } x \in \mathbb{R}.$$

Die zugehörige Verteilungsfunktion ist

$$F_X(x) = \frac{1}{2} + \frac{1}{\pi} \arctan(x) \qquad \text{für } x \in \mathbb{R}.$$

Sind Y und Z unabhängige $\mathcal{N}(0,1)$-verteilte Zufallsvariablen, so kann man zeigen, dass der Quotient $X := Y/Z$ gerade Cauchy-verteilt ist; siehe [Rice, Abschnitt 3.6.1]. Die Cauchy-Verteilung ist ein Beispiel für eine sogenannte *langschwänzige Verteilung* (heavy-tailed distribution); ihre Dichte geht für $|x| \to \infty$ nur sehr langsam gegen 0 (quadratisch, im Vergleich zum exponentiellen Abfallen bei der Normalverteilung). Anschaulich bedeutet das, dass X auch sehr grosse Werte noch mit substantieller Wahrscheinlichkeit annimmt. Diese Eigenschaft der Cauchy-Verteilung ist oft nützlich, um Gegenbeispiele zu konstruieren.

Für die Cauchy-Verteilung existiert kein Erwartungswert, denn $x f_X(x) = \frac{1}{\pi} \frac{x}{1+x^2}$ ist zwar wie bei der $\mathcal{N}(0,1)$-Verteilung eine ungerade Funktion, im Gegensatz zur Normalvertei-

lung aber nicht absolut integrierbar. Es gilt nämlich wegen Symmetrie

$$\begin{aligned}\int_{-\infty}^{\infty} |x| f_X(x)\,\mathrm{d}x &= 2\int_0^{\infty} \frac{1}{\pi}\frac{x}{1+x^2}\,\mathrm{d}x \\ &= \lim_{b\to\infty}\int_0^b \frac{1}{\pi}\frac{2x}{1+x^2}\,\mathrm{d}x = \lim_{b\to\infty}\frac{1}{\pi}\log(1+x^2)\Big|_0^b \\ &= \lim_{b\to\infty}\frac{1}{\pi}\log(1+b^2) = \infty.\end{aligned}$$

Also ist $E[X]$ hier undefiniert, und genauer gilt sogar $E[X^+] = E[X^-] = \infty$. Weil die Dichte symmetrisch um 0 ist, ist der Median der Verteilung aber 0. ◇

Nun betrachten wir wie im diskreten Fall die Frage, wie man für eine Funktion $Y = g(X)$ einer bekannten Zufallsvariablen X den Erwartungswert berechnet. Analog zum diskreten Fall (siehe Satz 2.1) haben wir

Satz 4.1. *Sei X eine Zufallsvariable und $Y = g(X)$ eine weitere Zufallsvariable. Ist X stetig mit Dichte f_X, so ist*

$$E[Y] = E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\,\mathrm{d}x,$$

sofern das Integral absolut konvergiert.

Alle weiteren allgemeinen Eigenschaften und Resultate für Erwartungswerte gelten im allgemeinen Fall genau gleich wie für diskrete Zufallsvariablen. Der einzige Unterschied taucht jeweils bei der konkreten Berechnung auf.

4.4 Gemeinsame Verteilungen, unabhängige Zufallsvariablen

Auch dieser Abschnitt ist fast völlig parallel zum diskreten Fall.

Definition. Die *gemeinsame Verteilungsfunktion* von n Zufallsvariablen $X_1, \ldots, X_n$ ist die Abbildung $F : \mathbb{R}^n \to [0,1]$,

$$(x_1, \ldots, x_n) \mapsto F(x_1, \ldots, x_n) := P[X_1 \le x_1, \ldots, X_n \le x_n].$$

Falls die gemeinsame Verteilungsfunktion F von $X_1, \ldots, X_n$ sich schreiben lässt als

$$F(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(t_1, \ldots, t_n)\, \mathrm{d}t_n \cdots \mathrm{d}t_1$$

für eine Funktion $f : \mathbb{R}^n \to [0, \infty)$, so heisst $f(x_1, \ldots, x_n)$ die *gemeinsame Dichte* von $X_1, \ldots, X_n$.

Falls $X_1, \ldots, X_n$ eine gemeinsame Dichte f haben, so gilt analog zum diskreten Fall

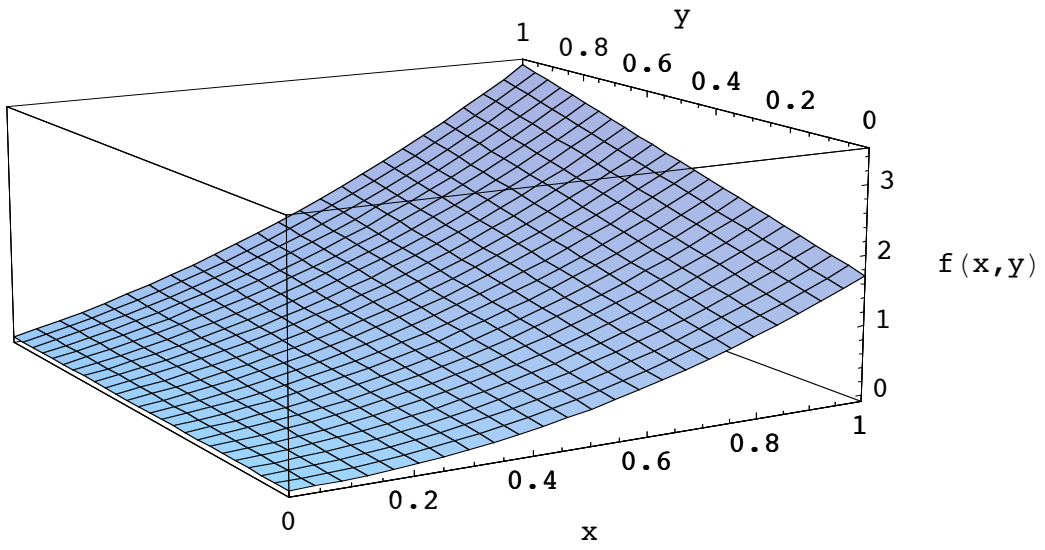
i) $f(x_1, \ldots, x_n) \geq 0$, und $= 0$ ausserhalb von $\mathcal{W}(X_1, \ldots, X_n)$.

ii) $\displaystyle\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_n)\, \mathrm{d}x_n \cdots \mathrm{d}x_1 = 1.$

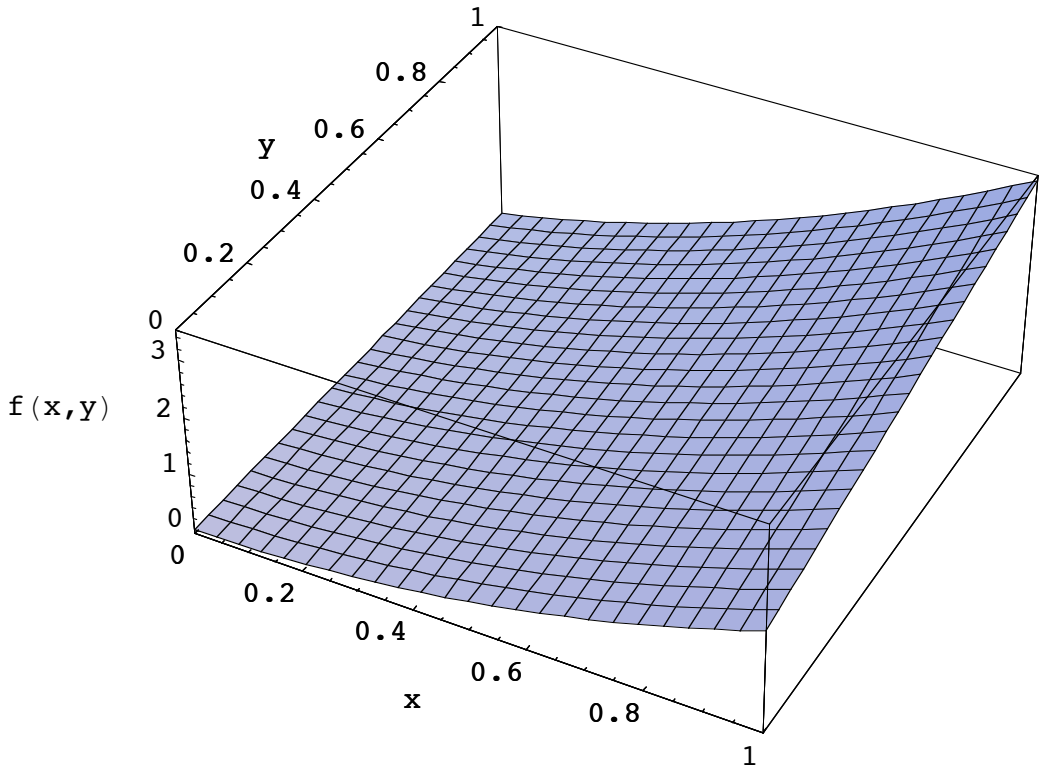
iii) $\displaystyle P[(X_1, \ldots, X_n) \in A] = \int\limits_{(x_1,\ldots,x_n)\in A} \cdots \int f(x_1, \ldots, x_n)\, \mathrm{d}x_n \ldots \mathrm{d}x_1$ für $A \subseteq \mathbb{R}^n$.

Beispiel 10 (zwei stetige Zufallsvariablen). Die gemeinsame Verteilungsfunktion von X und Y habe die Dichte

$$f(x, y) = \begin{cases} \frac{12}{7}(x^2 + xy) & \text{für } 0 \leq x \leq 1,\ 0 \leq y \leq 1, \\ 0 & \text{sonst.} \end{cases}$$



Graphische Darstellung der Dichte $f(x,y) = \frac{12}{7}(x^2 + xy)$ für $0 \leq x \leq 1$ und $0 \leq y \leq 1$.



Graphische Darstellung der Dichte $f(x,y) = \frac{12}{7}(x^2 + xy)$ für $0 \leq x \leq 1$ und $0 \leq y \leq 1$.

Das ist eine Dichtefunktion, denn offenbar ist $f \geq 0$ und

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y\,\mathrm{d}x = \int_0^1\int_0^1 \frac{12}{7}(x^2+xy)\,\mathrm{d}y\,\mathrm{d}x = \int_0^1 \frac{12}{7}\left(x^2y+\frac{1}{2}xy^2\right)\Big|_{y=0}^{y=1}\mathrm{d}x$$
$$= \int_0^1 \frac{12}{7}\left(x^2+\frac{1}{2}x\right)\mathrm{d}x = \frac{12}{7}\left(\frac{1}{3}x^3+\frac{1}{4}x^2\right)\Big|_{x=0}^{x=1} = 1.$$

Die gemeinsame Verteilungsfunktion von X und Y ist

$$F(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(u,v)\,\mathrm{d}v\,\mathrm{d}u = \int_0^x\int_0^y \frac{12}{7}(u^2+uv)\,\mathrm{d}v\,\mathrm{d}u$$
$$= \int_0^x \frac{12}{7}\left(u^2v+\frac{1}{2}uv^2\right)\Big|_{v=0}^{v=y}\mathrm{d}u = \int_0^x \frac{12}{7}\left(u^2y+\frac{1}{2}uy^2\right)\mathrm{d}u$$
$$= \frac{12}{7}\left(\frac{1}{3}u^3y+\frac{1}{4}u^2y^2\right)\Big|_{u=0}^{u=x} = \frac{4}{7}x^3y+\frac{3}{7}x^2y^2 \qquad \text{für } 0\leq x\leq 1,\ 0\leq y\leq 1.$$

Daraus sehen wir sofort, dass $F(0,0)=0$ und $F(1,1)=1$ ist.

Die Wahrscheinlichkeit $P[X \geq Y]$ beispielsweise erhalten wir hier wie folgt. Setzen wir $A := \{(x,y) : 0 \leq y \leq x \leq 1\}$, so ist nach Obigem

$$P[X\geq Y] = P[(X,Y)\in A] = \iint\limits_{(x,y)\in A} f(x,y)\,\mathrm{d}y\,\mathrm{d}x$$
$$= \int_0^1\int_0^x \frac{12}{7}(x^2+xy)\,\mathrm{d}y\,\mathrm{d}x = \int_0^1 \frac{12}{7}\left(x^3+\frac{1}{2}x^3\right)\mathrm{d}x$$
$$= \frac{12}{7}\times\frac{3}{2}\times\frac{1}{4}x^4\Big|_{x=0}^{x=1} = \frac{9}{14}.$$

◇

Völlig analog zum diskreten Fall haben wir weiter

Definition. Haben X und Y die gemeinsame Verteilungsfunktion F, so ist die Funktion $F_X : \mathbb{R} \to [0,1]$,

$$x \mapsto F_X(x) := P[X\leq x] = P[X\leq x, Y<\infty] = \lim_{y\to\infty} F(x,y)$$

die Verteilungsfunktion der *Randverteilung (RV)* von X. Analog ist $F_Y : \mathbb{R} \to [0,1]$,

$$y \mapsto F_Y(y) := P[Y\leq y] = P[X<\infty, Y\leq y] = \lim_{x\to\infty} F(x,y)$$

die Verteilungsfunktion der Randverteilung von Y.

Falls X und Y eine gemeinsame Dichte $f(x,y)$ haben, so haben auch die Randverteilungen von X und Y Dichten $f_X : \mathbb{R} \to [0,\infty)$ bzw. $f_Y : \mathbb{R} \to [0,\infty)$. Analog zur Gewichtsfunktion p_X ist

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y \qquad \text{bzw.} \qquad f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x,$$

d.h. die *Dichtefunktion einer Randverteilung* (kurz: *Randdichte*) entsteht aus der gemeinsamen Dichtefunktion durch "Wegintegrieren" der anderen Variable(n). Man beachte auch hier wieder einmal die schon oft gesehene Ersetzung von (Summe, Gewichtsfunktion) durch (Integral, Dichtefunktion).

$$f_X(x) = \frac{\mathrm{d}}{\mathrm{d}x} F_X(x) = \frac{\mathrm{d}}{\mathrm{d}x} \lim_{y\to\infty} F(x,y) = \frac{\mathrm{d}}{\mathrm{d}x}\left(\int_{-\infty}^{x} \int_{-\infty}^{\infty} f(u,v)\,\mathrm{d}v\,\mathrm{d}u \right) = \int_{-\infty}^{\infty} f(x,v)\,\mathrm{d}v.$$

Beispiel 10 (zwei stetige Zufallsvariablen). Wie oben sei die gemeinsame Dichte von X und Y gegeben durch

$$f(x,y) = \begin{cases} \frac{12}{7}(x^2+xy) & \text{für } 0 \le x \le 1,\ 0 \le y \le 1, \\ 0 & \text{sonst.} \end{cases}$$

Die Randverteilungen von X und Y haben dann die Dichten

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y = \int_0^1 \frac{12}{7}(x^2+xy)\,\mathrm{d}y = \frac{12}{7}\left(x^2 + \frac{1}{2}x\right) \qquad \text{für } 0 \le x \le 1,$$
$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x = \int_0^1 \frac{12}{7}(x^2+xy)\,\mathrm{d}x = \frac{12}{7}\left(\frac{1}{3} + \frac{1}{2}y\right) \qquad \text{für } 0 \le y \le 1.$$

◇

Für die Unabhängigkeit von Zufallsvariablen haben wir wie bisher

Definition. Die Zufallsvariablen $X_1, \ldots, X_n$ heissen *unabhängig*, falls gilt

$$F(x_1, \ldots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n),$$

d.h. die gemeinsame Verteilungsfunktion F ist das Produkt der Verteilungsfunktionen F_{X_i} der Randverteilungen.

Hat man stetige Zufallsvariablen mit Dichten, so ist das (analog zum diskreten Fall) äquivalent zu

$$f(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n) \qquad \text{für alle } x_1, \ldots, x_n,$$

d.h. die gemeinsame Dichtefunktion f ist das Produkt der einzelnen Randdichten f_{X_i}.

Beispiel 10 (zwei stetige Zufallsvariablen). Im obigen Beispiel mit der gemeinsamen Dichte $f(x, y) = \frac{12}{7}(x^2 + xy)$ ist $f_X(x) = \frac{12}{7}(x^2 + \frac{1}{2}x)$, $f_Y(y) = \frac{12}{7}(\frac{1}{3} + \frac{1}{2}y)$. Damit sind auch hier X und Y nicht unabhängig; z.B. ist

$$f\left(\frac{1}{2}, \frac{1}{3}\right) \neq f_X\left(\frac{1}{2}\right) f_Y\left(\frac{1}{3}\right).$$

◇

Beispiel. Wir wählen unabhängig voneinander zufällig zwei Punkte im Einheitsintervall $[0, 1]$. Wie weit liegen sie im Mittel auseinander?

Seien X_1 und X_2 unabhängig und $\sim \mathcal{U}(0, 1)$; wir suchen dann $E[|X_1 - X_2|]$. Die gemeinsame Dichte von X_1 und X_2 ist wegen Unabhängigkeit

$$f(x_1, x_2) = I_{[0,1]}(x_1) I_{[0,1]}(x_2) = \begin{cases} 1 & \text{für } 0 \leq x_1 \leq 1,\ 0 \leq x_2 \leq 1, \\ 0 & \text{sonst.} \end{cases}$$

Also ist

$$\begin{aligned} E[|X_1 - X_2|] &= \int_0^1 \int_0^1 |x_1 - x_2| \,\mathrm{d}x_2 \,\mathrm{d}x_1 \\ &= \int_0^1 \int_0^{x_1} (x_1 - x_2) \,\mathrm{d}x_2 \,\mathrm{d}x_1 + \int_0^1 \int_{x_1}^1 (x_2 - x_1) \,\mathrm{d}x_2 \,\mathrm{d}x_1 \\ &= \int_0^1 \frac{1}{2} x_1^2 \,\mathrm{d}x_1 + \int_0^1 \int_0^{x_2} (x_2 - x_1) \,\mathrm{d}x_1 \,\mathrm{d}x_2 \\ &= \frac{1}{6} + \int_0^1 \frac{1}{2} x_2^2 \,\mathrm{d}x_2 = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}. \end{aligned}$$

◇

4.5 Funktionen und Transformationen von Zufallsvariablen

Analog zum diskreten Fall fragen wir zuerst einmal, wie sich die Summe von Zufallsvariablen verhält.

Seien also X und Y stetige Zufallsvariablen mit gemeinsamer Dichtefunktion $f(x, y)$. Für $Z = X + Y$ suchen wir zuerst die Verteilungsfunktion

$$F_Z(z) = P[Z \leq z] = P[X + Y \leq z].$$

Setzen wir

$$A_z := \{(x, y) \in \mathbb{R}^2 : x + y \leq z\},$$

so ist

$$F_Z(z) = P[(X, Y) \in A_z] = \iint_{A_z} f(x, y) \,\mathrm{d}y \,\mathrm{d}x = \int_{-\infty}^{\infty} \int_{-\infty}^{z-x} f(x, y) \,\mathrm{d}y \,\mathrm{d}x.$$

Mit der Variablentransformation $v = x + y$ wird $y = v - x$, $\mathrm{d}y = \mathrm{d}v$ und

$$F_Z(z) = \int_{-\infty}^{\infty} \int_{-\infty}^{z} f(x, v - x) \,\mathrm{d}v \,\mathrm{d}x = \int_{-\infty}^{z} \int_{-\infty}^{\infty} f(x, v - x) \,\mathrm{d}x \,\mathrm{d}v.$$

Also hat Z auch eine Dichte, und zwar ist

$$f_Z(z) = \frac{\mathrm{d}}{\mathrm{d}z} F_Z(z) = \int_{-\infty}^{\infty} f(x, z - x) \,\mathrm{d}x = \int_{-\infty}^{\infty} f(z - y, y) \,\mathrm{d}y,$$

wieder in genauer Analogie zur Gewichtsfunktion im diskreten Fall. Die letzte Gleichheit folgt dabei wie bisher aus der Symmetrie zwischen X und Y.

Sind zusätzlich X und Y *unabhängig*, so ist $f(x,y) = f_X(x)f_Y(y)$ und damit wieder wie im diskreten Fall

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)\,\mathrm{d}x = \int_{-\infty}^{\infty} f_X(z-y)f_Y(y)\,\mathrm{d}y =: (f_X * f_Y)(z),$$

d.h. f_Z ist hier die *Faltung* von f_X und f_Y.

Beispiel. Seien X und Y unabhängig und beide $\sim Exp(\lambda)$, also $f_X(x) = \lambda e^{-\lambda x}$ für $x \geq 0$ und ebenso $f_Y(y) = \lambda e^{-\lambda y}$ für $y \geq 0$. Dann ist

$$f_Y(z-x) = \begin{cases} \lambda e^{-\lambda(z-x)} & \text{für } z \geq x, \\ 0 & \text{für } z < x, \end{cases}$$

und damit für $Z = X + Y$

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)\,\mathrm{d}x \\ &= \int_0^z \lambda e^{-\lambda x}\lambda e^{-\lambda(z-x)}\,\mathrm{d}x \\ &= \int_0^z \lambda^2 e^{-\lambda z}\,\mathrm{d}x = \lambda^2 z e^{-\lambda z} \qquad \text{für } z \geq 0. \end{aligned}$$

Also hat Z eine Gamma-Verteilung mit Parametern 2 und λ, und wir schreiben das als $Z \sim Ga(2,\lambda)$. ◇

Allgemein ist die *Gamma-Verteilung* eine stetige Verteilung mit der Dichtefunktion

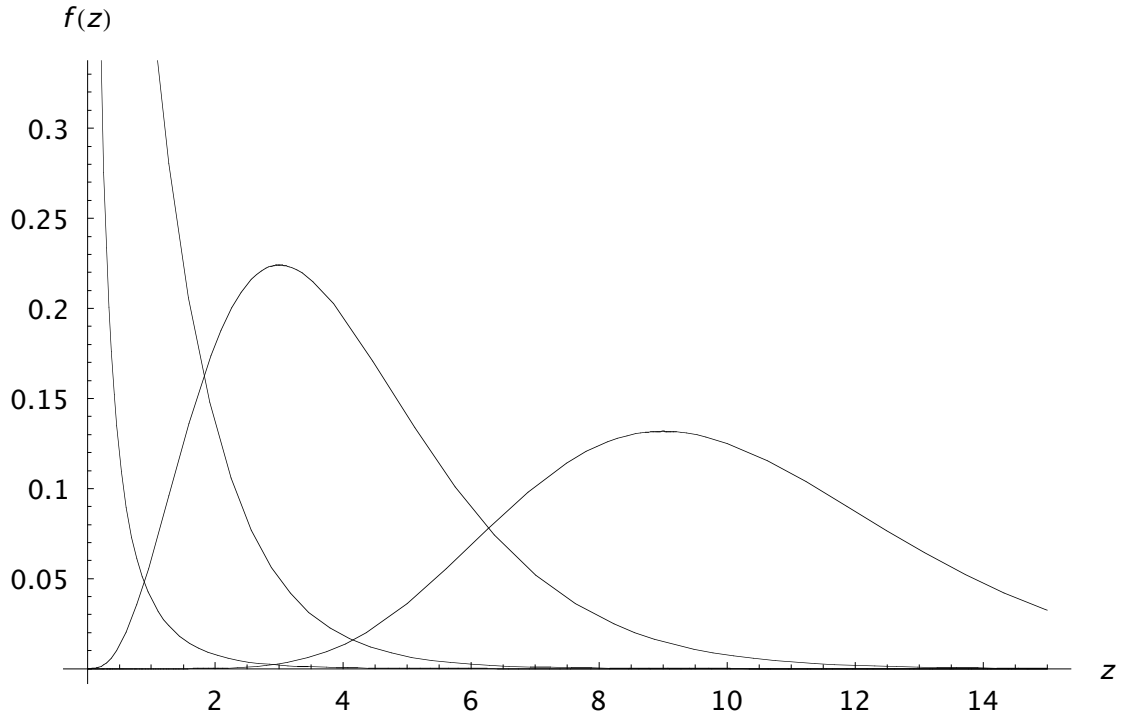
$$f(z) = \frac{1}{\Gamma(\alpha)}\lambda^\alpha z^{\alpha-1}e^{-\lambda z} \qquad \text{für } z \geq 0$$

mit Parametern $\alpha > 0$, $\lambda > 0$. Dabei ist

$$\Gamma(\alpha) := \int_0^\infty u^{\alpha-1}e^{-u}\,\mathrm{d}u \qquad \text{für } \alpha > 0$$

die sogenannte *Gammafunktion*; für $\alpha = n \in \mathbb{N}$ ist $\Gamma(n) = (n-1)!$ die Fakultät von $n-1$. Für eine Zufallsvariable Z mit dieser Verteilung schreiben wir $Z \sim Ga(\alpha,\lambda)$. Für $\alpha = 1$

ist das gerade eine $Exp(\lambda)$-Verteilung, und analog zum obigen Beispiel kann man mit Induktion zeigen, dass eine Summe von n unabhängigen $Exp(\lambda)$-verteilten Zufallsvariablen $Ga(n, \lambda)$-verteilt ist.



Graphische Darstellung der Dichten von Gamma-Verteilungen mit Parametern $\lambda = 1$ und (von links nach rechts) $\alpha = 0.1$, $\alpha = 1$, $\alpha = 4$, $\alpha = 10$.

Sei nun X eine Zufallsvariable mit Verteilungsfunktion F_X und Dichte f_X. Für eine (messbare) Funktion $g : \mathbb{R} \to \mathbb{R}$ betrachten wir die neue Zufallsvariable $Y = g(X)$. Wie sehen dann Verteilungsfunktion F_Y und Dichte f_Y aus, sofern eine Dichte existiert?

Der allgemeine Ansatz zur Lösung dieses Problems ist sehr einfach. Wir schreiben

$$F_Y(t) = P[Y \leq t] = P[g(X) \leq t] = \int_{A_g} f_X(s)\,\mathrm{d}s$$

mit der Menge $A_g := \{s \in \mathbb{R} : g(s) \leq t\}$. Dann versuchen wir die rechte Seite auszurechnen, indem wir die genauere Struktur der Transformation g ausnutzen; siehe Beispiele unten. Die Dichte f_Y von Y erhält man dann, sofern sie existiert, als Ableitung von F_Y.

Für *affine Transformationen*, $g(x) = ax + b$ mit $a > 0$, $b \in \mathbb{R}$, geht das sehr einfach. Aus $Y = g(X) = aX + b$ folgt

$$F_Y(t) = P[Y \le t] = P[aX + b \le t] = P\left[X \le \frac{t-b}{a}\right] = F_X\left(\frac{t-b}{a}\right)$$

und damit nach der Kettenregel

$$f_Y(t) = \frac{\mathrm{d}}{\mathrm{d}t} F_Y(t) = \frac{1}{a} f_X\left(\frac{t-b}{a}\right).$$

Analog bis auf ein Vorzeichen geht das auch mit $a < 0$.

Beispiel. Ist $X \sim \mathcal{N}(\mu, \sigma^2)$ und $Y = aX + b$, so ist

$$f_Y(t) = \frac{1}{a}\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}\Big(\frac{t-b}{a} - \mu\Big)^2\right) = \frac{1}{a\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2a^2\sigma^2}\big(t - (b + \mu a)\big)^2\right).$$

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

Im Spezialfall $a = \frac{1}{\sigma}$, $b = -\frac{\mu}{\sigma}$ ist $Y = \frac{X-\mu}{\sigma}$, und wir erhalten wieder, dass diese Zufallsvariable $\mathcal{N}(0,1)$-verteilt ist. Diese Transformation heisst deshalb auch *Standardisierung* einer Normalverteilung. ◇

Für *nichtlineare Transformationen* betrachten wir zwei Beispiele.

Beispiel. Für $g(x) = x^2$ ist $Y = X^2$ und

$$F_Y(t) = P[Y \le t] = P[X^2 \le t] = P[-\sqrt{t} \le X \le \sqrt{t}] = F_X(\sqrt{t}) - F_X(-\sqrt{t}),$$

falls X stetig ist. Ist zum Beispiel $X \sim \mathcal{N}(0,1)$, so hat $Y = X^2$ die Verteilungsfunktion

$$F_Y(t) = \Phi(\sqrt{t}) - \Phi(-\sqrt{t})$$

und damit nach der Kettenregel wegen Symmetrie von φ die Dichtefunktion

$$f_Y(t) = \frac{1}{2\sqrt{t}}\varphi(\sqrt{t}) - \frac{1}{-2\sqrt{t}}\varphi(-\sqrt{t}) = \frac{1}{\sqrt{t}}\varphi(\sqrt{t}) = \frac{1}{\sqrt{2\pi}} t^{-\frac{1}{2}} e^{-\frac{1}{2}t} \qquad \text{für } t \ge 0.$$

Das ist die Dichte einer sogenannten χ^2_1-Verteilung, einer *Chiquadrat-Verteilung* mit einem Freiheitsgrad. Wir schreiben $Y \sim \chi^2_1$. ◇

Beispiel. Für $g(x) = \frac{1}{x}$ ist $Y = \frac{1}{X}$ und

$$F_Y(t) = P[Y \leq t] = P\left[\frac{1}{X} \leq t\right] = P\left[X \geq \frac{1}{t}\right] = 1 - P\left[X < \frac{1}{t}\right] = 1 - F_X\left(\frac{1}{t}\right),$$

falls X stetig ist, also

$$f_Y(t) = \frac{1}{t^2} f_X\left(\frac{1}{t}\right).$$

Ist zum Beispiel $X \sim \mathcal{U}(0,1)$, so ist $f_X(t) = 1$ für $0 \leq t \leq 1$ und $F_X(t) = t$ für $0 \leq t \leq 1$. Die Zufallsvariable $Y = \frac{1}{X}$ hat dann Wertebereich $\mathcal{W}(Y) = [1, \infty)$ und

$$F_Y(t) = 1 - F_X\left(\frac{1}{t}\right) = 1 - \frac{1}{t} \qquad \text{für } t \geq 1,$$
$$f_Y(t) = \frac{1}{t^2} f_X\left(\frac{1}{t}\right) = \frac{1}{t^2} \qquad \text{für } t \geq 1.$$

Das ist eine sogenannte *Pareto(1)-Verteilung*. ◇

Als eine Anwendung der Transformation von Zufallsvariablen beweisen wir den folgenden

Satz 4.2. *Sei F eine stetige und streng monoton wachsende Verteilungsfunktion, mit Umkehrfunktion F^{-1}. Ist $X \sim \mathcal{U}(0,1)$ und $Y = F^{-1}(X)$, so hat Y gerade die Verteilungsfunktion F.*

Beweis. Direktes Einsetzen gibt

$$F_Y(t) = P[Y \leq t] = P[F^{-1}(X) \leq t] = P[X \leq F(t)] = F(t),$$

weil X nach Annahme $\mathcal{U}(0,1)$-verteilt ist. **q.e.d.**

Der obige Satz erlaubt die explizite Konstruktion einer Zufallsvariablen Y mit einer gewünschten Verteilungsfunktion F, wenn man eine $\mathcal{U}(0,1)$-verteilte Zufallsvariable X hat. Das liefert einen *Simulationsalgorithmus* nach der *Inversionsmethode*, und zwar wie folgt. In der Regel hat man bei einem Computer einen *"Zufallszahlengenerator"*, d.h. einen deterministischen Algorithmus, der eine Folge $(x_1, x_2, \dots)$ von Zahlen in $[0,1]$ produziert,

die sich in einem gewissen Sinn verhält wie die Realisierung einer Folge von unabhängigen $\mathcal{U}(0,1)$-verteilten Zufallsvariablen. Man sagt deshalb auch, dass der Zufallsgenerator eine $\mathcal{U}(0,1)$-Verteilung simuliert. In diesem Sinne simuliert also nach obigem Satz F^{-1}(Zufallsgenerator) die Verteilung F.

Beispiel. Um eine Exponentialverteilung mit Parameter λ zu simulieren, nehmen wir die zugehörige Verteilungsfunktion $F(t) = 1 - e^{-\lambda t}$ für $t \geq 0$. Ihre Inverse F^{-1} erhalten wir via

$$y = F(t) = 1 - e^{-\lambda t} = 1 - e^{-\lambda F^{-1}(y)}$$

als

$$F^{-1}(y) = t = -\frac{\log(1-y)}{\lambda}.$$

Mit $X \sim \mathcal{U}(0,1)$ ist also

$$Y := F^{-1}(X) = -\frac{\log(1-X)}{\lambda} \sim Exp(\lambda).$$

Weil mit X auch $1 - X \sim \mathcal{U}(0,1)$ ist, gilt auch

$$Y' := -\frac{\log X}{\lambda} \sim Exp(\lambda).$$

$\diamond$

5 Ungleichungen und Grenzwertsätze

Grundidee: In vielen Situationen taucht die *Summe von vielen gleichartigen Zufallsvariablen* auf. Wir möchten wissen, wie sich diese Summe etwa verhält, und untersuchen deshalb ihre *Asymptotik*, wenn die Anzahl der Summanden gegen unendlich geht.

Als *Motivation* betrachten wir das folgende

Beispiel. Der Gewinn bei einem *Glücksspiel* wird durch die Zufallsvariable X beschrieben. Bei wiederholtem Spielen erhalten wir also (unabhängige) Zufallsvariablen $X_1, \ldots, X_n$. Der *Gesamtgewinn* nach n Runden ist

$$S_n = \sum_{i=1}^{n} X_i,$$

und der *durchschnittliche Gewinn pro Runde* ist

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

In Abschnitt 2.2 haben wir den Erwartungswert $E[X]$ als idealisierten Durchschnittsgewinn bei unendlich vielen Wiederholungen motiviert. Wir vermuten deshalb, dass die Zufallsvariablen $\overline{X}_n = \frac{1}{n} S_n$ für $n \to \infty$ in einem geeigneten Sinne gegen $E[X]$ konvergieren sollten. Das *Gesetz der grossen Zahlen* gibt dafür eine präzise Formulierung.

Falls wir z.B. die Wahrscheinlichkeit suchen, dass unser Gesamtgewinn zwischen zwei Schranken liegt, also die Wahrscheinlichkeit $P[a < S_n \leq b]$, so brauchen wir dafür die Verteilungsfunktion von S_n. Der *zentrale Grenzwertsatz* gibt dafür eine Approximation für grosses n (genauer: eine Asymptotik für $n \to \infty$). $\diamond$

Grundsätzlich enthält dieses Kapitel also zwei verschiedene Arten von Grenzwertsätzen. Im einen Fall geht es um die Asymptotik von Zufallsvariablen, im anderen um die von Verteilungen oder Verteilungsfunktionen. Für die Zufallsvariablen $\overline{X}_n$ haben wir das Gesetz der grossen Zahlen, für die Verteilungen der S_n den zentralen Grenzwertsatz.

Als Vorbereitung betrachten wir zuerst einige nützliche Ungleichungen. Ohne spezielle Erwähnung gelten alle Resultate in diesem Kapitel für beliebige (diskrete oder allgemeine) Zufallsvariablen.

5.1 Ungleichungen

Wir beginnen mit einem sehr einfachen und doch sehr nützlichen Resultat.

Proposition 5.1. (Markov-Ungleichung) *Sei X eine Zufallsvariable und ferner $g : \mathcal{W}(X) \to [0, \infty)$ eine wachsende Funktion. Für jedes $c \in \mathbb{R}$ mit $g(c) > 0$ gilt dann*

$$P[X \geq c] \leq \frac{E[g(X)]}{g(c)}.$$

Beweis. Weil g wachsend und ≥ 0 ist, gilt

$$g(c)I_{\{X \geq c\}} = g(c)I_{\{g(X) \geq g(c)\}} \leq g(X).$$

Aus der Monotonie des Erwartungswertes folgt also

$$E[g(X)] \geq g(c)E\left[I_{\{X \geq c\}}\right] = g(c)P[X \geq c]$$

und damit die Behauptung. **q.e.d.**

Eine oft benutzte Anwendung ist

Korollar 5.2. (Chebyshev-Ungleichung) *Sei Y eine Zufallsvariable mit endlicher Varianz. Für jedes $b > 0$ gilt dann*

$$P\big[|Y - E[Y]| \geq b\big] \leq \frac{\mathrm{Var}[Y]}{b^2}.$$

Beweis. Wählt man in Proposition 5.1 $X := |Y - E[Y]|$ und $g : [0, \infty) \to [0, \infty)$, $x \mapsto g(x) := x^2$, so ist

$$E[g(X)] = E\left[(Y - E[Y])^2\right] = \mathrm{Var}[Y],$$

und damit folgt die Behauptung aus Proposition 5.1. Alternativ kann man das auch direkt beweisen; das Argument geht dann völlig analog zum Beweis von Proposition 5.1. **q.e.d.**

Die obige Ungleichung illustriert die Bedeutung der Varianz als Streuungsmass: Je kleiner die Varianz ist, desto eher (mit umso grösserer Wahrscheinlichkeit) liegen die Werte von X nahe beim Erwartungswert $E[X]$.

Beispiel. Wir werfen eine faire Münze n Mal und bestimmen die Anzahl S_n der Versuche, in denen Kopf auftritt. Dann ist $S_n \sim Bin(n, p)$ mit $p = \frac{1}{2}$ und damit

$$E[S_n] = np = \frac{n}{2}, \qquad \mathrm{Var}[S_n] = np(1-p) = \frac{n}{4}.$$

Die Wahrscheinlichkeit dafür, dass man mehr als $(1+\delta)\frac{n}{2}$ Mal Kopf beobachtet, ist dann nach der Chebyshev-Ungleichung

$$P[S_n > (1+\delta)E[S_n]] \leq P\left[|S_n - E[S_n]| > \delta\frac{n}{2}\right] \leq \frac{4\mathrm{Var}[S_n]}{n^2\delta^2} = \frac{1}{n\delta^2}.$$

Für $\delta = 10\%$ gibt das zum Beispiel für $n = 1000$

$$P[S_{1000} > 550] \leq \frac{1}{1000 \times 0.01} = 0.1;$$

für $n = 10'000$ erhält man

$$P[S_{10'000} > 5500] \leq \frac{1}{10'000 \times 0.01} = 0.01.$$

Insbesondere sieht man, dass die Wahrscheinlichkeit für eine Abweichung vom Erwartungswert für wachsendes n immer kleiner wird, und wir werden im nächsten Abschnitt sehen, dass dahinter ein allgemeiner Satz steht. In Abschnitt 5.4 werden wir zudem auch wesentlich genauere Abschätzungen für solche Wahrscheinlichkeiten angeben. $\diamond$

5.2 Das Gesetz der grossen Zahlen

Satz 5.3. (schwaches Gesetz der grossen Zahlen) *Sei $X_1, X_2, \ldots$ eine Folge von unabhängigen Zufallsvariablen, die alle den gleichen Erwartungswert $E[X_i] = \mu$ und die gleiche Varianz $\mathrm{Var}[X_i] = \sigma^2$ haben. Sei*

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

Dann konvergiert $\overline{X}_n$ für $n \to \infty$ in Wahrscheinlichkeit/stochastisch gegen $\mu = E[X_i]$, d.h.

$$P[|\overline{X}_n - \mu| > \varepsilon] \xrightarrow[n\to\infty]{} 0 \qquad \text{für jedes } \varepsilon > 0.$$

Anschaulich bedeutet die obige Aussage, dass mit beliebig grosser Wahrscheinlichkeit $1-\delta$ der Wert von $\overline{X}_n$ für $n \geq n_0$ beliebig nahe (bis auf ε) bei μ liegt. Dabei hängt n_0 natürlich von ε und δ ab, aber nicht von ω.

Bemerkung. Wir werden im Beweis sehen, dass es auch schon genügt, wenn die X_i statt unabhängig nur paarweise unkorreliert sind, d.h. $\mathrm{Cov}(X_i, X_k) = 0$ für $i \neq k$. ⋄

Beispiel. Wir simulieren auf einem Computer wie in Abschnitt 4.5 Zufallsvariablen $X_1, \ldots, X_N$ und plotten $n \mapsto \overline{X}_n(\omega)$ für die entsprechende Realisierung ω. Das machen wir

a) für $X_i \sim Exp(\lambda)$: Hier ist $F_X(t) = 1 - e^{-\lambda t}$ für $t > 0$, also $F_X^{-1}(u) = -\frac{1}{\lambda} \log(1-u)$, und damit simulieren wir die X_i als

$$X_i = -\frac{1}{\lambda} \log(1 - U_i),$$

wobei die $U_i \in (0,1)$ mit einem Zufallszahlengenerator erzeugt werden.

b) für X_i Cauchy-verteilt: Hier haben wir $F_X(t) = \frac{1}{2} + \frac{1}{\pi} \arctan t$ für $t \in \mathbb{R}$, also $F_X^{-1}(u) = \tan((u - \frac{1}{2})\pi)$, so dass wir mit

$$X_i = \tan\left(\left(U_i - \frac{1}{2}\right)\pi\right)$$

simulieren.

Wenn wir dieses Experiment durchführen, stellen wir fest:

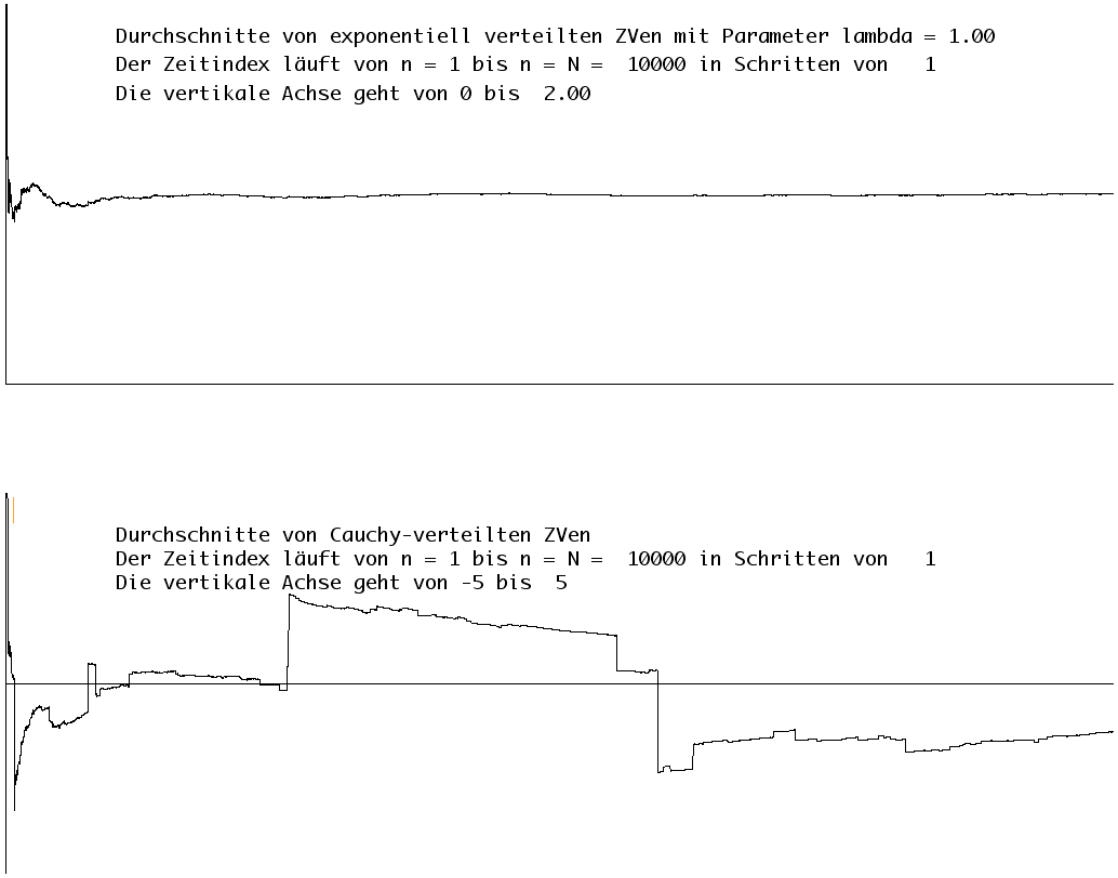


Illustration zum Gesetz der grossen Zahlen

a) $n \mapsto \overline{X}_n(\omega)$ stabilisiert sich für grosse n beim Wert $\frac{1}{\lambda} = E[X_i]$. Das illustriert also das GGZ. (Genauer illustriert das das starke GGZ; siehe die Diskussion am Ende des nächsten Beispiels.)

b) $n \mapsto \overline{X}_n(\omega)$ bleibt auch für grosse n wild oszillierend; das schwache GGZ funktioniert also hier scheinbar nicht. Die Erklärung dafür ist, dass $E[X_i]$ hier nicht existiert, so dass die Voraussetzungen von Satz 5.3 nicht erfüllt sind. In der Tat kann man zeigen, dass $\overline{X}_n$ als Mittel von unabhängigen Cauchy-verteilten Zufallsvariablen

wieder Cauchy-verteilt ist, und damit kann $\overline{X}_n$ natürlich nicht gegen eine Konstante konvergieren. $\diamond$

Nun kommen wir zum Beweis von Satz 5.3. In Anbetracht der Allgemeinheit des Resultates geht das erstaunlich einfach:

Beweis von Satz 5.3: Wir wollen die Chebyshev-Ungleichung auf die Zufallsvariable $Y := \overline{X}_n$ anwenden und berechnen deshalb

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

wobei wir die Linearität des Erwartungswertes ausnutzen. Weil die X_i unabhängig sind (und hier würde nach der Summenformel für Varianzen auch schon paarweise Unkorreliertheit genügen), gilt

$$\mathrm{Var}[\overline{X}_n] = \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}[X_i] = \frac{1}{n}\sigma^2.$$

Also liefert die Chebyshev-Ungleichung für jedes $\varepsilon > 0$

$$P[|\overline{X}_n - \mu| > \varepsilon] \leq \frac{\mathrm{Var}[\overline{X}_n]}{\varepsilon^2} = \frac{\frac{1}{n}\sigma^2}{\varepsilon^2} = \frac{1}{n}\frac{\sigma^2}{\varepsilon^2} \xrightarrow[n\to\infty]{} 0.$$

q.e.d.

Als eine erste Anwendung betrachten wir

Beispiel. *Monte Carlo-Integration:* Nehmen wir an, wir wollen für eine gegebene Funktion $h : [0,1]^d \to \mathbb{R}$ das Integral

$$I := \int_{[0,1]^d} h(\underline{x})\,\mathrm{d}\underline{x} = \int_0^1 \cdots \int_0^1 h(x_1, \ldots, x_d)\,\mathrm{d}x_1 \cdots \mathrm{d}x_d$$

berechnen. Das kann für unangenehme Funktionen h oder grosse Dimension d selbst numerisch schwierig sein. Der Einfachheit halber betrachten wir den Fall $d = 1$. Die Idee ist nun, I als einen Erwartungswert aufzufassen; ist nämlich U gleichverteilt auf $[0,1]$, so ist

$$E[h(U)] = \int_{-\infty}^{\infty} h(x) f_U(x)\,\mathrm{d}x = \int_0^1 h(x)\,\mathrm{d}x = I.$$

Haben wir also eine Folge von Zufallsvariablen $U_1, U_2, \ldots$, die unabhängig und alle $\mathcal{U}(0,1)$-verteilt sind, so liefert das schwache GGZ

$$\overline{h(U_n)} = \frac{1}{n}\sum_{i=1}^{n} h(U_i) \xrightarrow[n\to\infty]{} E[h(U_1)] = I \qquad \text{in Wahrscheinlichkeit,}$$

so dass wir eine Approximation von I erhalten, die wir sehr einfach mit einem Zufallszahlengenerator für die U_i berechnen können.

Dieses Beispiel zeigt bei genauerem Hinsehen allerdings auch, warum man eine stärkere Aussage als das schwache GGZ haben möchte. Konvergenz in Wahrscheinlichkeit bedeutet gemäss der Definition nämlich, dass für grosse n der Wert $\overline{h(U_n)}$ mit grosser Wahrscheinlichkeit nahe bei I liegt. Bei unserer Simulation betrachten wir aber eine feste Realisierung ω, und wir wissen nicht, ob diese in der "guten" Approximationsmenge liegt oder vielleicht im Komplement, das zwar kleine, aber durchaus positive Wahrscheinlichkeit haben kann.

$\diamond$

Mit der obigen Motivation im Hinterkopf beweist man (mit deutlich mehr Aufwand)

Satz 5.4. (starkes Gesetz der grossen Zahlen) *Sei $X_1, X_2, \ldots$ eine Folge von unabhängigen Zufallsvariablen, die alle dieselbe Verteilung haben, und ihr Erwartungswert $\mu = E[X_i]$ sei endlich. Für*

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

gilt dann

$$\overline{X}_n \xrightarrow[n\to\infty]{} \mu \qquad \textit{P-fastsicher (P-f.s.),}$$

d.h.

$$P\Big[\Big\{\omega \in \Omega : \overline{X}_n(\omega) \xrightarrow[n\to\infty]{} \mu\Big\}\Big] = 1.$$

[Einen Beweis unter zusätzlichen Annahmen findet man z.B. bei [Williams, S.113f].]

Expliziter bedeutet die Aussage in Satz 5.4, dass wir ein $\varepsilon > 0$ vorgeben können, und dann ist für jedes ω ausserhalb einer Menge mit Wahrscheinlichkeit 0 die Abweichung von $\overline{X}_n(\omega)$ von μ kleiner als ε für $n \geq n_0$; dabei hängt nun n_0 sowohl von ε als auch von ω ab.

Im obigen Beispiel mit der Monte Carlo-Integration bedeutet Satz 5.4, dass unsere Simulation für $\overline{h(U_n)}$ nicht nur mit grosser Wahrscheinlichkeit, sondern mit Wahrscheinlichkeit 1 nahe bei I liegt. Es kann uns also im Prinzip zwar immer noch passieren, dass wir eine schlechte Realisierung erwischen, aber das passiert mit Wahrscheinlichkeit 0.

Terminologie: Man hat es oft mit Zufallsvariablen zu tun, die unabhängig sind und alle dieselbe Verteilung haben. Der Kürze halber nennt man solche Zufallsvariablen *i.i.d.* (independent identically distributed). Die deutsche Terminologie ist u.i.v. (unabhängig und identisch verteilt), wird aber seltener benutzt.

5.3 Der zentrale Grenzwertsatz

Idee: Wie schon erwähnt fragen wir nach der Asymptotik für die Verteilung einer Summe von vielen "gleichartigen" Zufallsvariablen; das formalisieren wir durch die Bedingung, dass die Summanden i.i.d. sind.

Satz 5.5. (zentraler Grenzwertsatz, ZGS) *Sei $X_1, X_2, \ldots$ eine Folge von i.i.d. Zufallsvariablen mit $E[X_i] = \mu$ und $\mathrm{Var}[X_i] = \sigma^2$. Für die Summe*

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

gilt dann

$$\lim_{n\to\infty} P\left[\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x\right] = \Phi(x) \qquad \text{für alle } x \in \mathbb{R},$$

wobei Φ die Verteilungsfunktion der $\mathcal{N}(0,1)$-Verteilung ist.

[Einen Beweis unter zusätzlichen Annahmen findet man z.B. in [Rice, Abschnitt 5.3].]

Bemerkungen. 1) Wie im Beweis des schwachen GGZ kann man nachrechnen, dass S_n Erwartungswert $E[S_n] = n\mu$ und Varianz $\mathrm{Var}[S_n] = n\sigma^2$ hat. Also hat die Grösse

$$S_n^* = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{S_n - E[S_n]}{\sqrt{\mathrm{Var}[S_n]}},$$

die in Satz 5.5 auftritt, Erwartungswert 0 und Varianz 1; sie heisst deshalb auch die *Standardisierung* von S_n.

2) Für *praktische Anwendungen* schreibt man die Aussage des ZGS meistens in der Form

$$P[S_n^* \leq x] \approx \Phi(x) \qquad \text{für } n \text{ gross}$$

oder

$$S_n^* \overset{\text{approx.}}{\sim} \mathcal{N}(0,1) \qquad \text{für } n \text{ gross,}$$

wobei das Symbol $\overset{\text{approx.}}{\sim}$ für "ist approximativ verteilt gemäss" steht. Zurückübersetzt für S_n oder $\overline{X}_n = \frac{1}{n} S_n$ heisst das dann

$$S_n \overset{\text{approx.}}{\sim} \mathcal{N}(n\mu, n\sigma^2)$$

bzw.

$$\overline{X}_n \overset{\text{approx.}}{\sim} \mathcal{N}\left(\mu, \frac{1}{n}\sigma^2\right),$$

und das wird benutzt, um Wahrscheinlichkeiten für diese Grössen (d.h. für S_n bzw. $\overline{X}_n$) approximativ als Wahrscheinlichkeiten für eine entsprechende Normalverteilung zu berechnen. ◇

Beispiel. Seien $X_1, \ldots, X_{12}$ i.i.d. $\sim \mathcal{U}(-\frac{1}{2}, \frac{1}{2})$ und $S_{12} = \sum_{i=1}^{12} X_i$. Wegen $E[X_i] = 0$, $\text{Var}[X_i] = \frac{1}{12}$ haben wir dann $E[S_{12}] = 0$, $\text{Var}[S_{12}] = 1$, und der ZGS sagt uns, dass

$$S_{12} \overset{\text{approx.}}{\sim} \mathcal{N}(0,1)$$

sein sollte. Obwohl hier $n = 12$ noch gar nicht gross ist, zeigen Simulationen eine erstaunlich gute Übereinstimmung des Histogramms von S_{12} mit der Dichte φ der Standard-Normalverteilung $\mathcal{N}(0,1)$.

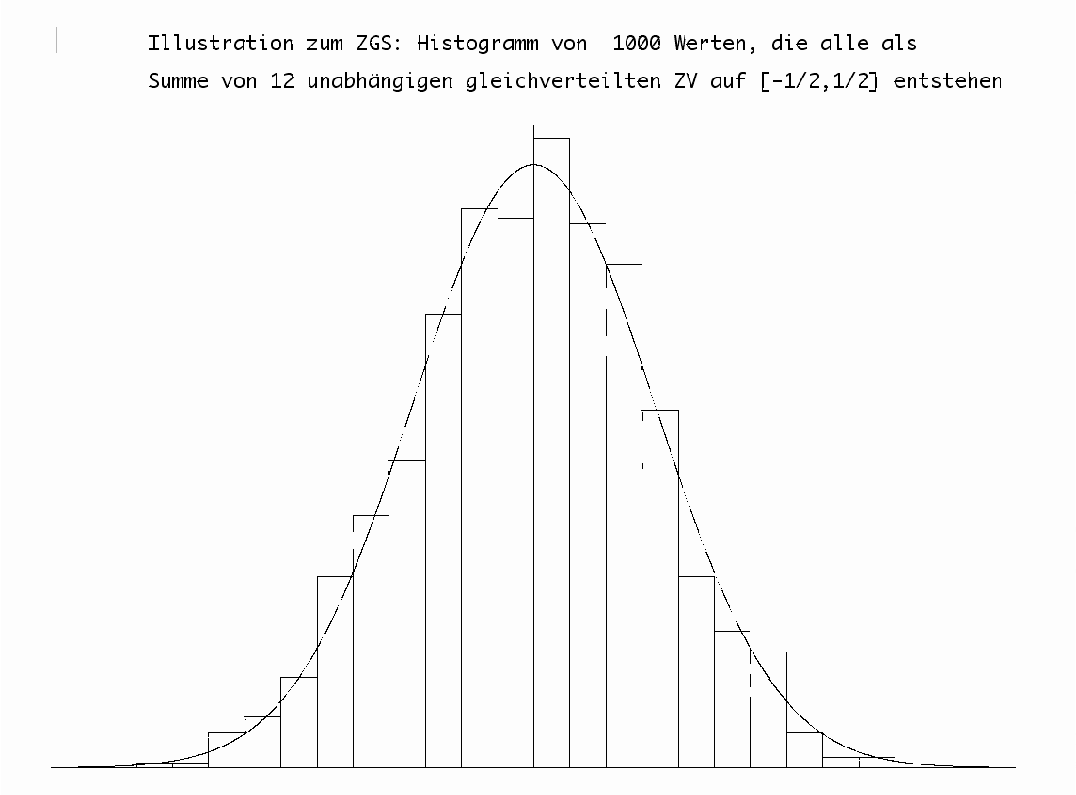


Illustration zum zentralen Grenzwertsatz

◇

Eine sehr häufige Anwendung des ZGS ist

Beispiel. *Normalapproximation für die Binomialverteilung:* In einem Experiment wird eine Münze 100 Mal geworfen, und man erhält 60 Mal Kopf. Ist die Münze fair?

Sei $X_i = I_{\{\text{Kopf bei Wurf } i\}}$ für $i = 1, \ldots, n = 100$. Wir nehmen an, dass die Zufallsvariablen X_i i.i.d. $\sim Be(\frac{1}{2})$ sind, d.h. wir gehen zunächst davon aus, dass die Münze fair ist. Anders gesagt: Wir nehmen an, dass die Münzwürfe unabhängige 0-1-Experimente mit Erfolgsparameter $p = \frac{1}{2}$ sind. Die Anzahl der Würfe mit Kopf ist dann

$$S_{100} = \sum_{i=1}^{100} X_i \sim Bin\left(100, \frac{1}{2}\right),$$

und wir interessieren uns für die Wahrscheinlichkeit des beobachteten Ausgangs, also für

$$P[S_{100} \geq 60] = \sum_{k=60}^{100} P[S_{100} = k] = \sum_{k=60}^{100} \binom{100}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{100-k} = 2^{-100} \sum_{k=60}^{100} \binom{100}{k}.$$

Weil das offensichtlich schwierig zu berechnen ist, benutzen wir den ZGS für eine approximative Berechnung.

Wegen $E[X_i] = \frac{1}{2}$, $\mathrm{Var}[X_i] = \frac{1}{4}$ ist nach dem ZGS

$$S_{100} \overset{\text{approx.}}{\sim} \mathcal{N}\left(100 \times \frac{1}{2}, 100 \times \frac{1}{4}\right) = \mathcal{N}(50, 25).$$

Also ist

$$P[S_{100} \geq 60] = P\left[\frac{S_{100} - 50}{\sqrt{25}} \geq \frac{60 - 50}{5}\right] = 1 - P[S^*_{100} < 2] \approx 1 - \Phi(2) = 0.0228.$$

Für eine faire Münze ist also die Wahrscheinlichkeit für das Ereignis $\{S_{100} \geq 60\}$ sehr klein; dass wir dieses Ereignis nun aber beobachtet haben, stellt also die Fairness der betrachteten Münze in Frage. ◇

Eine etwas genauere Approximation für die Binomialverteilung erhält man, wenn man zusätzlich noch die sogenannte *Kontinuitätskorrektur* benutzt: Für $S_n \sim Bin(n, p)$ ist $S_n \overset{\text{approx.}}{\sim} \mathcal{N}(np, np(1-p))$ und

$$\begin{aligned} P[a < S_n \leq b] &= P\left[\frac{a - np}{\sqrt{np(1-p)}} < S^*_n \leq \frac{b - np}{\sqrt{np(1-p)}}\right] \\ &\approx \Phi\left(\frac{b + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right). \end{aligned}$$

Die Korrektur um $+\frac{1}{2}$ erklärt sich dadurch, dass die Approximation der Binomial-Wahrscheinlichkeiten durch die Fläche unter der Normalverteilungsdichte besser wird, wenn man die Stäbe aus dem Histogramm für die Binomialverteilung jeweils zentriert über die relevanten Werte von a bis b setzt.

Im obigen Beispiel wäre also etwas genauer

$$P[S_{100} \geq 60] = P[S_{100} > 59] \approx 1 - \Phi\left(\frac{59.5 - 50}{\sqrt{25}}\right) = 1 - \Phi(1.9) = 0.0287,$$

was immer noch sehr klein ist.

Im Vergleich zur Chebyshev-Ungleichung erhalten wir im obigen Beispiel so ferner

$$P[S_{1000} > 550] \approx 1 - \Phi\left(\frac{550.5 - 500}{\sqrt{250}}\right) = 1 - \Phi(3.19) = 0.0007$$

und

$$P[S_{10'000} > 5500] \approx 1 - \Phi\left(\frac{5500.5 - 5000}{\sqrt{2500}}\right) = 1 - \Phi(10.01) \approx 0.$$

Das sind nur Approximationen, aber sie zeigen doch ein erstes Mal klar, dass die Abschätzungen aus der Chebyshev-Ungleichung sehr grob sind.

5.4 Grosse Abweichungen und Chernoff-Schranken

In diesem Abschnitt geben wir zuerst ein allgemeines Resultat über die Abschätzung von Restwahrscheinlichkeiten an und zeigen dann, wie man daraus sehr scharfe Schranken für Binomial-Wahrscheinlichkeiten erhält. Die allgemeine Technik stammt aus dem Gebiet der sogenannten *grossen Abweichungen*, und die Resultate im Spezialfall von unabhängigen 0-1–Experimenten sind als *Chernoff-Schranken* bekannt.

Seien also zuerst $X_1, \ldots, X_n$ i.i.d. Zufallsvariablen und $S_n = \sum_{i=1}^n X_i$. Wir wollen die Wahrscheinlichkeit dafür abschätzen, dass S_n sehr gross wird, indem wir obere Schranken für $P[S_n \geq b]$ angeben, wobei typisch b gross ist. Analog kann man natürlich auch obere Schranken für $P[S_n \leq -a]$ mit $a > 0$ gross angeben.

Die *momenterzeugende Funktion* einer Zufallsvariablen X ist

$$M_X(t) := E[e^{tX}] \qquad \text{für } t \in \mathbb{R};$$

das ist immer wohldefiniert in $[0, \infty]$, kann aber $+\infty$ werden.

Satz 5.6. *Seien $X_1, \ldots, X_n$ i.i.d. Zufallsvariablen, für welche die momenterzeugende Funktion $M_X(t)$ für alle $t \in \mathbb{R}$ endlich ist. Für jedes $b \in \mathbb{R}$ gilt dann*

$$P[S_n \geq b] \leq \exp\left(\inf_{t \in \mathbb{R}} \big(n \log M_X(t) - tb \big) \right).$$

Beweis. Die Idee ist sehr einfach: Wir benutzen die Markov-Ungleichung aus Proposition 5.1 mit der Funktion $g(x) = e^{tx}$ und optimieren die so erhaltene Ungleichung über den freien Parameter t.
Sei also $g(x) = e^{tx}$. Diese Funktion ist wachsend, und weil die X_i unabhängig und identisch verteilt sind, gilt

$$E[g(S_n)] = E[e^{tS_n}] = E\left[\exp\left(t \sum_{i=1}^{n} X_i \right) \right] = E\left[\prod_{i=1}^{n} e^{tX_i} \right] = \prod_{i=1}^{n} E[e^{tX_i}] = \big(M_X(t) \big)^n.$$

Also liefert Proposition 5.1

$$P[S_n \geq b] \leq \frac{E[g(S_n)]}{g(b)} = \frac{(M_X(t))^n}{e^{tb}} = \exp\big(n \log M_X(t) - tb \big).$$

Das gilt für jedes $t \in \mathbb{R}$, und die linke Seite hängt nicht von t ab. Also gilt die Abschätzung auch noch, wenn wir rechts über t minimieren, und das liefert die Behauptung. **q.e.d.**

Man beachte, dass die Schranke in Satz 5.6 exponentiell in b und in der Anzahl n der Zufallsvariablen ist; man kann also erwarten, dass das sehr kleine Werte und damit eine sehr gute Abschätzung gibt.

Seien nun die X_i zunächst i.i.d. $\sim Be(p)$, so dass $S_n \sim Bin(n, p)$ ist. Dann ist

$$M_X(t) = E[e^{tX}] = pe^t + (1 - p),$$

und wir müssen

$$t \mapsto n \log M_X(t) - tb = n \log\big(pe^t + (1 - p) \big) - tb$$

minimieren. Weil das aber schlecht zum Rechnen ist, gehen wir etwas anders vor. Wir schätzen zuerst $M_X(t)$ (und damit auch $E[g(S_n)]$) nach oben ab und optimieren dann

diese etwas schlechtere Schranke über t. Dabei formulieren wir das Resultat gleich etwas allgemeiner, indem wir unabhängige 0-1-Experimente mit *variablem* Erfolgsparameter betrachten.

Satz 5.7. *Seien $X_1, \ldots, X_n$ unabhängig mit $X_i \sim Be(p_i)$ und $S_n = \sum_{i=1}^n X_i$. Sei ferner*

$$\mu_n := E[S_n] = \sum_{i=1}^n p_i$$

und $\delta > 0$. Dann gilt

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

Beweis. Wie in Satz 5.6 liefert die Markov-Ungleichung

$$P[S_n \geq b] \leq \frac{1}{e^{tb}} E[e^{tS_n}] = \frac{1}{e^{tb}} \prod_{i=1}^n E[e^{tX_i}] = \frac{1}{e^{tb}} \prod_{i=1}^n M_{X_i}(t).$$

Nun ist aber wegen $1 + z \leq e^z$

$$M_{X_i}(t) = p_i(e^t - 1) + 1 \leq \exp\big(p_i(e^t - 1)\big)$$

und damit

$$\prod_{i=1}^n M_{X_i}(t) \leq \exp\left(\sum_{i=1}^n p_i(e^t - 1) \right) = \exp\big(\mu_n(e^t - 1)\big).$$

Für $b = (1+\delta)\mu_n$ erhalten wir also

$$P[S_n \geq (1+\delta)\mu_n] \leq \exp\big(-(1+\delta)\mu_n t + \mu_n(e^t - 1)\big).$$

Die rechte Seite wird minimal in t für

$$0 = -(1+\delta)\mu_n + \mu_n e^t,$$

also

$$\mu_n(e^t - 1) = \mu_n \delta \qquad \text{und} \qquad t = \log(1+\delta),$$

der minimale Wert ist dann

$$\exp\big(-(1+\delta)\mu_n \log(1+\delta) + \mu_n \delta\big) = \left(\frac{e^\delta}{(1+\delta)^{1+\delta}} \right)^{\mu_n},$$

und das gibt die Behauptung. **q.e.d.**

Beispiel. Wie im Beispiel in Abschnitt 5.1 werfen wir eine faire Münze n Mal und bestimmen die Anzahl S_n der Versuche, in denen Kopf auftritt. Das entspricht der Situation in Satz 5.7 mit allen $p_i = p = \frac{1}{2}$. Die Wahrscheinlichkeit dafür, dass man mehr als $(1+\delta)\frac{n}{2}$ Mal Kopf beobachtet, ist dann nach der Chebyshev-Ungleichung

$$P\big[S_n > (1+\delta)E[S_n]\big] \le \frac{1}{n\delta^2}.$$

Im Vergleich dazu gibt uns Satz 5.7 nun wegen $\mu_n = np = \frac{n}{2}$ die Schranke

$$P\big[S_n > (1+\delta)E[S_n]\big] \le \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^{n/2}.$$

Für $\delta = 10\%$ gibt das zum Beispiel für $n = 1000$

$$P[S_{1000} > 550] \le \left(\frac{e^{0.1}}{1.1^{1.1}}\right)^{500} = 0.0889;$$

die Chebyshev-Ungleichung gab uns hier eine Schranke von 0.1. Ebenso erhält man für $n = 10'000$ die Abschätzung

$$P[S_{10'000} > 5500] \le \left(\frac{e^{0.1}}{1.1^{1.1}}\right)^{5000} = 3.073 \times 10^{-11},$$

was im Vergleich zum Ergebnis 0.01 aus der Chebyshev-Ungleichung nun sehr viel schärfer ist. Für noch grössere n wird dieser Unterschied immer markanter. ◇

Beispiel. Bei unserem obigen Experiment, wo in 100 Münzwürfen 60 Mal Kopf erzielt wurde, können wir die Wahrscheinlichkeit $P[S_{100} \ge 60]$ für dieses Ereignis ebenfalls nach oben abschätzen. In der Notation von Satz 5.7 ist hier $n = 100$, $p_i = \frac{1}{2}$ für alle i und damit $\mu_n = 50$ und $\delta = 0.2$. Die Chebyshev-Ungleichung gibt hier

$$P[S_{100} \ge 60] \le \frac{1}{100 \times (0.2)^2} = 0.25;$$

nach Satz 5.7 ist

$$P[S_{100} \ge 60] \le \left(\frac{e^{0.2}}{(1.2)^{1.2}}\right)^{50} = 0.3909.$$

Beide Abschätzungen sind noch weit vom approximativen Wert entfernt, und hier ist sogar die Chebyshev-Ungleichung noch besser, weil n relativ klein ist. ◇

Teil II: Statistik

Ziel: Einführung einiger Grundideen und Methoden aus der (induktiven) Statistik.

6 Statistische Grundideen

Ausgangssituation: Man hat beobachtete Daten und will daraus Rückschlüsse ziehen auf den zugrundeliegenden Mechanismus, der diese Daten generiert hat.

Ein häufiger erster Schritt ist eine *graphische Aufbereitung* der Daten; das ist oft nützlich, um einen ersten Eindruck und erste Ideen zu bekommen. Die entsprechenden Methoden gehören zur *deskriptiven* oder *beschreibenden Statistik*; diese Aspekte werden hier aber nicht behandelt.

Wir befassen uns im Folgenden mit der *induktiven Statistik*. Die Grundidee dabei ist relativ einfach. Man fasst die Daten $x_1, \ldots, x_n$ auf als Realisierungen/realisierte Werte $X_1(\omega), \ldots, X_n(\omega)$ von Zufallsvariablen $X_1, \ldots, X_n$, und sucht dann (unter geeigneten Zusatzannahmen) Aussagen über die Verteilung von $X_1, \ldots, X_n$.

Wichtig: Man muss immer sauber unterscheiden zwischen den *Daten* $x_1, \ldots, x_n$ (bezeichnet mit kleinen Buchstaben; $x_1, \ldots, x_n$ sind also in der Regel *Zahlen*) und dem generierenden *Mechanismus* $X_1, \ldots, X_n$ (bezeichnet mit grossen Buchstaben; $X_1, \ldots, X_n$ sind *Zufallsvariablen*, also *Funktionen* auf einem Ω). Das wird leider nicht immer eingehalten, obwohl schon der amerikanische Philosoph und Psychologe William James (1842–1910) im 19. Jahrhundert auf diesen Punkt hinwies: “We must be careful not to confuse data with the abstractions we use to analyze them.”

Terminologie: Die Gesamtheit der Beobachtungen $x_1, \ldots, x_n$ oder Zufallsvariablen $X_1, \ldots, X_n$ nennt man oft eine *Stichprobe*; die Anzahl n heisst dann der *Stichprobenumfang*.

Ausgangspunkt unserer Betrachtungen ist in der Regel ein Datensatz $x_1, \ldots, x_n$ aus einer Stichprobe $X_1, \ldots, X_n$, für die wir ein Modell suchen. Dieses ist beschreibbar durch einen (möglicherweise hochdimensionalen) *Parameter* $\vartheta \in \Theta$, und um Begriffe und Notationen sauber definieren und benutzen zu können, muss man genauer spezifizieren, in welcher Art wahrscheinlichkeitstheoretische Aussagen vom Parameter ϑ abhängen. Dazu betrachtet man simultan eine ganze *Familie von Wahrscheinlichkeitsräumen*; man hat typisch einen festen Grundraum $(\Omega, \mathcal{F})$ und für jeden Parameter ϑ aus dem *Parameterraum* Θ ein Wahrscheinlichkeitsmass P_ϑ auf $(\Omega, \mathcal{F})$, also einen Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, P_\vartheta)$ für jedes $\vartheta \in \Theta$. Anschaulich kann man sich vorstellen, dass jemand ("die Natur") einen Parameter $\vartheta \in \Theta$ und damit einen konkreten stochastischen Mechanismus P_ϑ wählt. Als Statistiker weiss man aber nicht, welches ϑ aus Θ gewählt wurde; man betrachtet also die Daten $x_1, \ldots, x_n$ als Ergebnisse von Zufallsvariablen $X_1, \ldots, X_n$ unter dem Mechanismus P_ϑ und versucht, daraus Rückschlüsse über ϑ zu ziehen.

Beispiel. Betrachten wir die obigen Ideen einmal für die Situation eines Münzwurfmodells. Wir haben eine Münze, von der wir nicht recht wissen, wie sie sich beim Werfen verhält. Wir nehmen an, dass die Münze bei jedem Wurf Kopf oder Zahl mit Wahrscheinlichkeit p bzw. $1-p$ produziert; wir kennen aber p nicht und betrachten es deshalb als einen unbekannten Parameter. Hier ist also $\Theta = [0,1]$, der Parameter $\vartheta = p$ entspricht der Erfolgswahrscheinlichkeit für Kopf bei einem einzelnen Münzwurf, und das Wahrscheinlichkeitsmass $P_\vartheta = P_p$ beschreibt das Münzwurfmodell mit dem Erfolgsparameter p. Wir nehmen auch noch an, dass die einzelnen Würfe jeweils (d.h. in jedem Modell P_ϑ) unabhängig sind.

Nun werfen wir unsere Münze n Mal, schreiben jeweils 0 für Zahl, 1 für Kopf und erhalten so Daten $x_1, \ldots, x_n$ aus einer Stichprobe $X_1, \ldots, X_n$. Unser Modell ist

$$\text{die } X_i \text{ sind unter } P_p \text{ i.i.d.} \sim Be(p);$$

der Parameter ist also wie erwähnt $\vartheta = p \in [0,1] = \Theta$, und das Modell P_p beschreibt unabhängige Münzwürfe mit Erfolgsparameter p für "Kopf". Unser Ziel ist es, aus den Daten Rückschlüsse über p zu ziehen. (Zum Beispiel möchten wir vermutlich wissen, ob die Münze wohl fair ist oder nicht.) $\diamond$

In vielen Fällen ist der Parameterraum Θ eine Teilmenge von $\mathbb{R}^m$; wenn man zu gegebenen Daten ein passendes Modell finden und darüber gewisse statistische Aussagen machen möchte, so spricht man dann von einer *parametrischen statistischen Analyse*. Allgemein gehören dazu die folgenden Etappen:

1) Beschreibende Statistik der Daten: In diesem Schritt versucht man mit graphischen Methoden, aufgrund der Daten eine erste Idee für die Wahl einer geeigneten Modellierung zu finden. (Diesen Schritt werden wir hier nicht weiter erklären.)

2) Wahl eines (parametrischen) Modells: Hier spezifiziert man die Parametermenge Θ und die Familie $(P_\vartheta)_{\vartheta\in\Theta}$ von Modellen, mit denen man arbeiten will.

3) Schätzung der Parameter: Aufgrund der Daten will man ein möglichst gut passendes Modell wählen. Dazu benutzt man einen *Schätzer*; die zugehörige *Schätzfunktion* ist eine Abbildung, die gegebenen Daten $x_1, \ldots, x_n$ einen Parameter $\vartheta \in \Theta$ zuordnet.

4) Kritische Modellüberprüfung (Anpassungstest): Hier fragt man, ob die Daten zu dem gewählten Parameter ϑ bzw. Modell P_ϑ gut passen; das macht man mit einem geeigneten *statistischen Test*.

5) Aussagen über Zuverlässigkeit der Schätzungen: Statt eines einzigen Parameterwertes kann man auch versuchen, einen Bereich in Θ so zu spezifizieren, dass die zugehörigen Modelle P_ϑ gut zu den Daten passen; man spricht dann von einem *Konfidenzbereich*.

7 Schätzer

Ziel: Überblick über grundlegende Ideen und Methoden zur Schätzung von Parametern.

7.1 Grundbegriffe

Sei $X_1, \ldots, X_n$ eine Stichprobe, für die wir ein Modell suchen. Wir haben also einen Parameterraum Θ und für jedes $\vartheta \in \Theta$ einen Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, P_\vartheta)$. Meistens ist $\Theta \subseteq \mathbb{R}^m$, und wir suchen dann für die Parameter $\vartheta_1, \ldots, \vartheta_m$ *Schätzer* $T_1, \ldots, T_m$ aufgrund unserer Stichprobe. Solche Schätzer sind Zufallsvariablen der Form $T_j = t_j(X_1, \ldots, X_n)$, wobei wir die Schätzfunktionen $t_j : \mathbb{R}^n \to \mathbb{R}$ noch geeignet wählen/finden müssen. Einsetzen von Daten $x_i = X_i(\omega)$, $i = 1, \ldots, n$, liefert dann *Schätzwerte* $T_j(\omega) = t_j(x_1, \ldots, x_n)$ für ϑ_j, $j = 1, \ldots, m$. Der Kürze halber schreiben wir oft auch $T = (T_1, \ldots, T_m)$ und $\vartheta = (\vartheta_1, \ldots, \vartheta_m)$.

Es ist wichtig, die Konzepte "Schätzer" und "Schätzwert" sauber auseinanderzuhalten. Ein Schätzwert ist eine *Zahl* $T_j(\omega) = t_j(X_1(\omega), \ldots, X_n(\omega)) = t_j(x_1, \ldots, x_n)$; wie die Daten x_i ist das die Realisation $T_j(\omega)$ (im von uns betrachteten konkreten Experiment ω) der Zufallsvariablen T_j. Ein Schätzer T ist dagegen eine *Zufallsvariable*; er hat also eine Verteilung (unter P_ϑ, für jedes $\vartheta \in \Theta$), und wir können z.B. nach der Wahrscheinlichkeit (im Modell P_ϑ, für jedes $\vartheta \in \Theta$) fragen, mit der er nahe beim wahren (aber uns unbekannten) Parameter ϑ liegt. Also ist der Schätzer die Funktion oder Vorschrift, die uns die *Berechnungsmethode* angibt, und der Schätzwert ist das *Ergebnis* einer konkreten Berechnung.

Beispiel 11 (tea tasting lady). Eine englische Lady behauptet, bei Tee mit Milch anhand des Geschmacks unterscheiden zu können, ob zuerst die Milch oder zuerst der Tee in die Tasse eingegossen worden ist. Wie kann man überprüfen, ob das stimmen kann?

Um zunächst einmal Daten zu bekommen, stellen wir der Lady an n Tagen die Aufgabe, zwei Tassen (je eine vom Typ 1 und Typ 2) zu klassifizieren; sie soll also angeben, in welche der Tassen zuerst Milch eingegossen worden ist. Wir notieren die Ergebnisse $x_1, \ldots, x_n \in \{0, 1\}$ (falsch bzw. richtig klassifiziert) und fassen wie üblich diese Daten als Realisationen von Zufallsvariablen $X_1, \ldots, X_n$ auf. Dann ist $S_n = \sum_{i=1}^n X_i$ die (zufällige) Anzahl der korrekt klassifizierten Tassenpaare, und $s_n = \sum_{i=1}^n x_i$ ist die beobachtete Anzahl von Erfolgen.

Als Modelle nehmen wir nun an, dass die X_i unter P_ϑ i.i.d. $\sim Be(\vartheta)$ mit $\vartheta \in \Theta = [0, 1]$ sind. Dann ist natürlich $S_n \sim Bin(n, \vartheta)$ unter P_ϑ, d.h. im Modell P_ϑ, das zu ϑ gehört, ist die Anzahl S_n der Erfolge binomialverteilt mit Parametern n und ϑ.

Weil wir den Parameter ϑ nicht kennen, liegt es nahe, zuerst einmal dafür einen Schätzer zu suchen. Eine erste Möglichkeit wäre, einfach das letzte Ergebnis zu nehmen; unser erster Schätzer $\hat{T}$ für ϑ wäre also $\hat{T} = X_n$. Obwohl das absurd aussieht, werden wir sehen, dass dieser Schätzer durchaus die eine oder andere vernünftige Eigenschaft hat.

Ein zweiter naheliegender Schätzer wäre die durchschnittliche Anzahl der Erfolge der Lady bei ihren n Versuchen; unser zweiter Schätzer wäre also $T = \overline{X}_n = \frac{1}{n} S_n$.

Für gegebene Daten $x_1, \ldots, x_n$ gibt uns das dann zwei Schätzwerte $\hat{t}(x_1, \ldots, x_n) = x_n$ und $t(x_1, \ldots, x_n) = \overline{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$, die wir konkret berechnen könnten. $\diamond$

Wie die X_i sind die Schätzer T_j Zufallsvariablen, deren Verteilung (unter P_ϑ) von den unbekannten Parametern $\vartheta_1, \ldots, \vartheta_m$ abhängt. Nur selten kann man diese Verteilung explizit bestimmen; siehe später die Resultate für die Normalverteilung. Bevor wir zumindest eine systematische Methode zur *Bestimmung* von Schätzern betrachten, führen wir einige allgemeine wünschenswerte *Eigenschaften* auf.

Definition. Ein Schätzer T heisst *erwartungstreu* für ϑ, falls gilt $E_\vartheta[T] = \vartheta$; im Mittel (über alle denkbaren Realisationen ω) schätzt T also richtig. Allgemein heisst $E_\vartheta[T] - \vartheta$ der *Bias* (oder erwartete Schätzfehler) von T; erwartungstreu (auf Englisch "*unbiased*") bedeutet also, dass der Bias Null ist. Der mittlere quadratische Schätzfehler ("*mean squa-*

red error", *MSE*) ist definiert als

$$\mathrm{MSE}_\vartheta[T] := E_\vartheta[(T-\vartheta)^2].$$

Eine Folge von Schätzern $T^{(n)}$, $n \in \mathbb{N}$, heisst *konsistent* für ϑ, falls $T^{(n)}$ für $n \to \infty$ in P_ϑ-Wahrscheinlichkeit gegen ϑ konvergiert, d.h. für jedes $\vartheta \in \Theta$ gilt

$$\lim_{n\to\infty} P_\vartheta[|T^{(n)} - \vartheta| > \varepsilon] = 0 \qquad \text{für jedes } \varepsilon > 0.$$

(Das setzt anschaulich voraus, dass man beliebig viele Daten haben könnte.)

Bemerkung. Nach Lemma 2.3 kann man den MSE zerlegen als

$$\mathrm{MSE}_\vartheta[T] = E_\vartheta[(T-\vartheta)^2] = \mathrm{Var}_\vartheta[T] + (E_\vartheta[T] - \vartheta)^2,$$

also in die Summe aus der Varianz des Schätzers T und dem Quadrat des Bias. Für erwartungstreue Schätzer sind Varianz und MSE dasselbe. ◇

Beispiel 11 (tea tasting lady). Beide oben angegebenen Schätzer $\hat{T} = X_n$ und $T = \overline{X}_n$ sind erwartungstreu. Unter P_ϑ ist ja $\hat{T} = X_n \sim Be(\vartheta)$, also

$$E_\vartheta[\hat{T}] = E_\vartheta[X_n] = \vartheta.$$

Obwohl aber $\hat{T}$ erwartungstreu ist, wird er kaum je das richtige Ergebnis für den Parameter ϑ liefern; für jede konkrete Realisierung ω hat ja $\hat{T}(\omega) = X_n(\omega)$ den Wert 0 oder 1, und nur im theoretischen Mittel über alle ω erhalten wir ϑ. Und natürlich gibt das nie eine konsistente Folge von Schätzern, denn die Folge $\hat{T}^{(n)} = X_n$, $n \in \mathbb{N}$, oszilliert immer zwischen den Werten 0 und 1, konvergiert also nicht gegen ϑ (in irgendeinem Sinn).

Unser zweiter Schätzer für ϑ ist $T = \overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$; er ist auch erwartungstreu, denn

$$E_\vartheta[T] = \frac{1}{n}\sum_{i=1}^n E_\vartheta[X_i] = \vartheta \qquad \text{für alle } \vartheta,$$

weil alle $X_i \sim Be(\vartheta)$ unter P_ϑ sind. Nach dem schwachen Gesetz der grossen Zahlen (Satz 5.3) ist zudem die Folge der Schätzer $T^{(n)}$, $n \in \mathbb{N}$, konsistent für ϑ.

Nicht nur asymptotisch ist $T^{(n)}$ besser als $\hat{T}^{(n)}$; wegen

$$\mathrm{Var}_\vartheta[\hat{T}^{(n)}] = \mathrm{Var}_\vartheta[X_n] = \vartheta(1-\vartheta)$$

und (wegen Unabhängigkeit unter jedem P_ϑ)

$$\mathrm{Var}_\vartheta[T^{(n)}] = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}_\vartheta[X_i] = \frac{1}{n}\vartheta(1-\vartheta)$$

hat $T^{(n)}$ auch kleinere Varianz, ist also im Mittel genauer, als $\hat{T}^{(n)}$. Das ist einleuchtend, weil $T^{(n)}$ die Information in der Stichprobe $X_1, \ldots, X_n$ viel besser ausnutzt. $\diamond$

7.2 Die Maximum-Likelihood-Methode (ML-Methode)

In diesem Abschnitt stellen wir eine Methode vor, um systematisch Schätzer zu bestimmen. Diese Methode liefert in sehr vielen Situationen Ergebnisse, die sowohl plausibel sind als auch gute Eigenschaften haben.

Ausgangspunkt im Folgenden ist immer eine von zwei Situationen, je nachdem, ob wir es mit diskreten oder mit stetigen Zufallsvariablen zu tun haben. In jedem Modell P_ϑ sind $X_1, \ldots, X_n$ entweder diskret mit gemeinsamer Gewichtsfunktion $p(x_1, \ldots, x_n; \vartheta)$ oder stetig mit gemeinsamer Dichtefunktion $f(x_1, \ldots, x_n; \vartheta)$. Meistens sind sogar die X_i unter P_ϑ i.i.d. mit individueller Gewichtsfunktion $p_X(x;\vartheta)$ bzw. Dichtefunktion $f_X(x;\vartheta)$; dann ist also die gemeinsame Gewichtsfunktion

$$p(x_1, \ldots, x_n; \vartheta) = \prod_{i=1}^{n} p_X(x_i; \vartheta)$$

bzw. die gemeinsame Dichtefunktion

$$f(x_1, \ldots, x_n; \vartheta) = \prod_{i=1}^{n} f_X(x_i; \vartheta).$$

Anschaulich ist

$$p(x_1, \ldots, x_n; \vartheta) = P_\vartheta[X_1 = x_1, \ldots, X_n = x_n]$$

gerade die Wahrscheinlichkeit im Modell P_ϑ, dass unsere Stichprobe $X_1, \ldots, X_n$ die Werte $x_1, \ldots, x_n$ liefert, und $f(x_1, \ldots, x_n; \vartheta)$ ist das übliche stetige Analogon.

Definition. Die *Likelihood-Funktion* ist

$$L(x_1, \ldots, x_n; \vartheta) := \begin{cases} p(x_1, \ldots, x_n; \vartheta) & \text{im diskreten Fall,} \\ f(x_1, \ldots, x_n; \vartheta) & \text{im stetigen Fall.} \end{cases}$$

Die Funktion $\log L(x_1, \ldots, x_n; \vartheta)$ heisst *log-Likelihood-Funktion*. Sie hat gegenüber der Likelihood-Funktion den Vorteil, dass sie im i.i.d.-Fall durch eine Summe (statt ein Produkt) gegeben und damit zum Rechnen oft wesentlich einfacher ist.

Nach diesen allgemeinen Vorbereitungen wenden wir uns nun der erwähnten Methode zur Bestimmung von Schätzern zu. Für eine Stichprobe $X_1, \ldots, X_n$ gibt uns die Likelihoodfunktion $L(x_1, \ldots, x_n; \vartheta)$ zumindest im diskreten Fall die Wahrscheinlichkeit im Modell P_ϑ, dass unsere Stichprobe gerade die Werte $x_1, \ldots, x_n$ liefert. Um eine möglichst gute Anpassung des Modells an die Daten zu erreichen, wollen wir diese Wahrscheinlichkeit möglichst gross machen, indem wir den Parameter geschickt wählen. Der *Maximum-Likelihood-Schätzer (ML-Schätzer)* T für ϑ wird also dadurch definiert, dass er

$$\vartheta \mapsto L(X_1, \ldots, X_n; \vartheta) \qquad \text{als Funktion von } \vartheta$$

maximiert.

Meistens sind $X_1, \ldots, X_n$ i.i.d. unter P_ϑ; die Likelihood-Funktion L ist dann ein Produkt, und es ist bequemer, statt L die log-Likelihood-Funktion $\log L$ zu maximieren, weil diese eine Summe ist. Statt zu maximieren sucht man ferner meistens nur Nullstellen der Ableitung (nach ϑ).

Bemerkung. In den Rechnungen arbeitet man oft mit $L(x_1, \ldots, x_n; \vartheta)$, insbesondere beim Maximieren über ϑ. Das optimale ϑ^* ist dann eine Funktion $t(x_1, \ldots, x_n)$ von $x_1, \ldots, x_n$. Damit der resultierende Schätzer T von der Stichprobe $X_1, \ldots, X_n$ abhängt, muss dann aber $x_1, \ldots, x_n$ durch $X_1, \ldots, X_n$ ersetzt werden. $\diamond$

Beispiel. *Bernoulli-Verteilung:* Im Modell P_p seien

$$X_1, \ldots, X_n \text{ i.i.d. } \sim Be(p).$$

Hier ist $\vartheta = p$, und wir wollen also für eine unbekannte Münze den Erfolgsparameter schätzen. Diese Fragestellung haben wir schon im letzten Abschnitt im Beispiel mit der tea tasting lady angetroffen.

Die Gewichtsfunktion einer $Be(\vartheta)$-Verteilung ist

$$p_X(x;\vartheta) = P_\vartheta[X = x] = \vartheta^x(1-\vartheta)^{1-x} \qquad \text{für } x \in \{0,1\}.$$

Weil die X_i i.i.d. sind, ist also

$$L(x_1, \ldots, x_n; \vartheta) = \prod_{i=1}^n p_X(x_i;\vartheta) = \vartheta^{\sum_{i=1}^n x_i}(1-\vartheta)^{n-\sum_{i=1}^n x_i}$$

und

$$\log L(x_1, \ldots, x_n; \vartheta) = \sum_{i=1}^n x_i \log \vartheta + \left(n - \sum_{i=1}^n x_i\right) \log(1-\vartheta).$$

Wir wollen das über ϑ maximieren und setzen dazu die entsprechende Ableitung Null. Die Ableitung nach ϑ ist

$$\frac{\partial}{\partial\vartheta} \log L(x_1, \ldots, x_n; \vartheta) = \frac{1}{\vartheta}\sum_{i=1}^n x_i - \frac{1}{1-\vartheta}\left(n - \sum_{i=1}^n x_i\right),$$

und das ist 0 für

$$(1-\vartheta)\sum_{i=1}^n x_i = \vartheta\left(n - \sum_{i=1}^n x_i\right),$$

d.h. für $\vartheta = \frac{1}{n}\sum_{i=1}^n x_i$.

Der ML-Schätzer für ϑ bzw. p ist hier also

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

◇

Beispiel. *Normalverteilung:* Im Modell P_ϑ seien

$$X_1, \ldots, X_n \text{ i.i.d. } \sim \mathcal{N}(\mu, \sigma^2).$$

Hier ist die Dimension des unbekannten Parameters $m = 2$, wir haben $\vartheta = (\mu, \sigma^2) = (\mu, v)$, und wir wollen μ und $\sigma^2 = v$ schätzen.

Die Dichtefunktion von X_i unter P_ϑ ist

$$f_X(x;\vartheta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) = \frac{1}{\sqrt{2\pi v}} e^{-\frac{(x-\mu)^2}{2v}}.$$

Weil die X_i i.i.d. sind, ist also

$$L(x_1, \ldots, x_n; \vartheta) = \prod_{i=1}^{n} f_X(x_i;\vartheta)$$

und

$$\log L(x_1, \ldots, x_n; \vartheta) = \sum_{i=1}^{n} \log f_X(x_i;\vartheta) = -n\frac{1}{2}(\log 2\pi + \log v) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2v}.$$

Die Ableitungen nach μ und v sind

$$\begin{aligned}
\frac{\partial}{\partial\mu} \log L(x_1, \ldots, x_n; \vartheta) &= 2\sum_{i=1}^{n} \frac{x_i - \mu}{2v}, \\
\frac{\partial}{\partial v} \log L(x_1, \ldots, x_n; \vartheta) &= -\frac{n}{2}\frac{1}{v} + \frac{1}{2v^2}\sum_{i=1}^{n}(x_i-\mu)^2,
\end{aligned}$$

und sie werden beide gleichzeitig 0 für

$$\begin{aligned}
0 = \sum_{i=1}^{n}(x_i-\mu) = \sum_{i=1}^{n} x_i - n\mu, \qquad &\text{d.h. für } \mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}_n, \\
v = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x}_n)^2.&
\end{aligned}$$

Der ML-Schätzer für $\vartheta = (\mu, \sigma^2)$ ist also

$$\begin{aligned}
T_1 &= \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_n, \\
T_2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\overline{X}_n)^2,
\end{aligned}$$

wobei die zweite Gleichheit durch Ausquadrieren folgt. ◇

Der Schätzer $T = (T_1, T_2)$ im obigen Beispiel ist ganz allgemein auch der sogenannte Momentenschätzer für

$$(E_\vartheta[X], \mathrm{Var}_\vartheta[X])$$

in jedem Modell P_ϑ, wo $X_1, \ldots, X_n$ i.i.d. sind. Dieser Schätzer hat aber den allgemeinen Nachteil, dass er nicht erwartungstreu für $(E_\vartheta[X], \mathrm{Var}_\vartheta[X])$ ist. Zwar ist für jedes ϑ

$$E_\vartheta[T_1] = E_\vartheta[\overline{X}_n] = \frac{1}{n}\sum_{i=1}^{n} E_\vartheta[X_i] = E_\vartheta[X];$$

aber

$$E_\vartheta[(\overline{X}_n)^2] = \frac{1}{n^2}\sum_{i,k=1}^{n} E_\vartheta[X_i X_k] = \frac{1}{n^2}\Bigg(\sum_{i=1}^{n} E_\vartheta[X_i^2] + \sum_{i\neq k} E_\vartheta[X_i X_k]\Bigg),$$

und wegen Unabhängigkeit ist für $i \neq k$

$$E_\vartheta[X_i X_k] = E_\vartheta[X_i] E_\vartheta[X_k] = (E_\vartheta[X])^2.$$

Also ist

$$\begin{aligned} E_\vartheta[T_2] &= \frac{1}{n}\sum_{i=1}^{n} E_\vartheta[X_i^2] - \left(\frac{1}{n}E_\vartheta[X^2] + \frac{n^2-n}{n^2}(E_\vartheta[X])^2\right) \\ &= \left(1 - \frac{1}{n}\right)\left(E_\vartheta[X^2] - (E_\vartheta[X])^2\right) \\ &= \frac{n-1}{n}\mathrm{Var}_\vartheta[X]. \end{aligned}$$

Um einen erwartungstreuen Schätzer T' für $(E_\vartheta[X], \mathrm{Var}_\vartheta[X])$ zu haben, benutzt man deshalb meistens

$$\begin{aligned} T_1' &= T_1 = \overline{X}_n \\ T_2' &= \frac{n}{n-1}T_2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = \frac{1}{n-1}\sum_{i=1}^{n} X_i^2 - \frac{n}{n-1}(\overline{X}_n)^2. \end{aligned}$$

Für T_2' benutzt man oft auch die Notation

$$S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

und nennt S^2 die *empirische Stichprobenvarianz*.

7.3 Verteilungsaussagen

In vielen Situationen ist es nützlich oder nötig, die Verteilung (unter P_ϑ, für jedes $\vartheta \in \Theta$) eines Schätzers zu kennen. Exakte allgemeine Aussagen gibt es dazu nicht viele; für die Normalverteilung folgt das weiter unten in Satz 7.1.

Einen allgemeinen *approximativen Zugang* liefert der zentrale Grenzwertsatz. Oft ist ein Schätzer T eine Funktion einer Summe $\sum_{i=1}^n Y_i$, wobei die Y_i im Modell P_ϑ i.i.d. sind; das einfachste Beispiel ist $T = \overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$. Nach dem zentralen Grenzwertsatz ist dann für grosse n

$$\sum_{i=1}^n Y_i \quad \text{approximativ normalverteilt unter } P_\vartheta$$

mit Parametern $\mu = nE_\vartheta[Y_i]$ und $\sigma^2 = n\mathrm{Var}_\vartheta[Y_i]$. Das kann man benutzen, um für die Verteilung von T approximative Aussagen zu bekommen und damit zumindest approximativ gewisse Fragen zu beantworten.

Für normalverteilte Stichproben hat man exakte Aussagen; wir werden das auch bei der Diskussion von Tests später noch ausgiebig benutzen.

Satz 7.1. *Seien $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$. Dann gilt:*

1) *$\overline{X}_n$ ist normalverteilt gemäss $\sim \mathcal{N}(\mu, \frac{1}{n}\sigma^2)$, und damit gilt $\dfrac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$.*

2) *$\dfrac{n-1}{\sigma^2}S^2 = \dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^n (X_i - \overline{X}_n)^2$ ist χ^2-verteilt mit $n-1$ Freiheitsgraden.*

3) *$\overline{X}_n$ und S^2 sind unabhängig.*

4) *Der Quotient*

$$\frac{\overline{X}_n - \mu}{S/\sqrt{n}} = \frac{\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}}{S/\sigma} = \frac{\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{1}{n-1}\frac{n-1}{\sigma^2}S^2}}$$

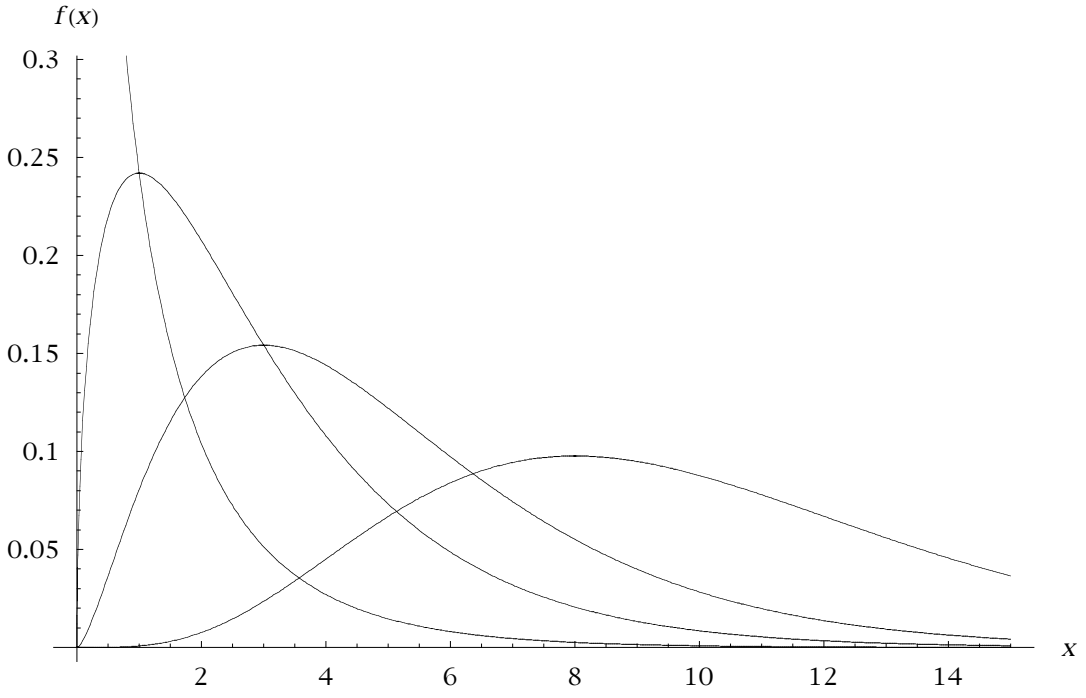
ist t-verteilt mit $n-1$ Freiheitsgraden.

Beweis. Siehe [Rice, Abschnitt 6.3].

Die χ^2-*Verteilung* mit n Freiheitsgraden gehört zu einer stetigen Zufallsvariablen Y mit Dichtefunktion

$$f_Y(y) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} y^{\frac{n}{2}-1} e^{-\frac{1}{2}y} \qquad \text{für } y \geq 0;$$

das ist der Spezialfall einer $Ga(\alpha, \lambda)$-Verteilung mit $\alpha = \frac{n}{2}$ und $\lambda = \frac{1}{2}$. Die χ^2-Verteilung entsteht wie folgt: Sind die Zufallsvariablen $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{N}(0,1)$, so ist die Summe $Y := \sum_{i=1}^n X_i^2 \sim \chi_n^2$. Für $n = 2$ ergibt das eine Exponentialverteilung mit Parameter $\frac{1}{2}$.



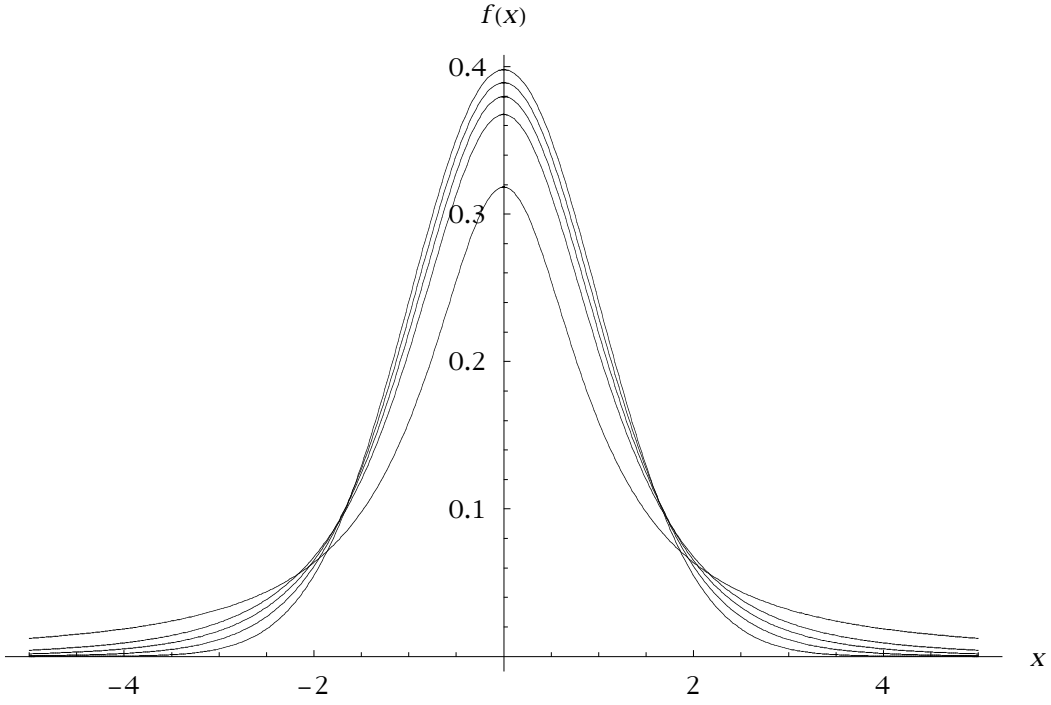
Graphische Darstellung der Dichten von χ^2-Verteilungen
mit Anzahl der Freiheitsgrade (von oben nach unten) $n = 1, 3, 5, 10$.

Die t-*Verteilung* mit n Freiheitsgraden gehört zu einer stetigen Zufallsvariablen Z mit Dichtefunktion

$$f_Z(z) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} \left(1 + \frac{z^2}{n}\right)^{-\frac{n+1}{2}} \qquad \text{für } z \in \mathbb{R}.$$

Für $n = 1$ ist das eine Cauchy-Verteilung, und für $n \to \infty$ erhält man eine $\mathcal{N}(0,1)$-Verteilung. Wie die $\mathcal{N}(0,1)$-Verteilung ist die t-Verteilung symmetrisch um 0; sie ist aber langschwänziger, und zwar umso mehr, je kleiner n ist. Die t-Verteilung entsteht wie folgt: Sind X und Y unabhängig mit $X \sim \mathcal{N}(0,1)$ und $Y \sim \chi^2_n$, so ist der Quotient

$$Z := \frac{X}{\sqrt{\frac{1}{n}Y}} \qquad t\text{-verteilt mit } n \text{ Freiheitsgraden.}$$



Graphische Darstellung der Dichten von t-Verteilungen
mit Anzahl der Freiheitsgrade (von unten nach oben) $n = 1, 3, 5, 10, 100$.

Angesichts der obigen Bemerkungen ist klar, wie die Aussage 4) von Satz 7.1 zustande kommt. Aussage 1) ist schon bekannt, und Aussage 2) ist einigermassen plausibel (auch wenn sie nicht direkt klar ist). Der wichtige (und überraschendste) Teil ist also die Unabhängigkeit in Aussage 3).

8 Tests

Ziel: Überblick über grundlegende Ideen und Methoden sowie einige Beispiele zum Testen von Hypothesen.

8.1 Grundbegriffe

Ausgangspunkt ist wie im letzten Abschnitt eine Stichprobe $X_1, \ldots, X_n$. Wir betrachten wieder eine Familie von Wahrscheinlichkeiten P_ϑ mit $\vartheta \in \Theta$, die unsere möglichen Modelle beschreiben. Wie bisher kann ϑ ein- oder mehrdimensional sein. Wir haben schon eine Vermutung, wo in Θ der richtige (aber unbekannte) Parameter ϑ liegen könnte, und wollen diese mit Hilfe der Daten überprüfen ("testen"). Das Grundproblem ist also, eine Entscheidung zwischen zwei konkurrierenden Modellklassen zu treffen — der *Hypothese* $\Theta_0 \subseteq \Theta$ und der *Alternative* $\Theta_A \subseteq \Theta$, wobei $\Theta_0 \cap \Theta_A = \emptyset$ ist. Meist schreibt man das als

$$\text{Hypothese } H_0 : \vartheta \in \Theta_0,$$
$$\text{Alternative } H_A : \vartheta \in \Theta_A.$$

Ist keine explizite Alternative spezifiziert, so hat man $\Theta_A = \Theta_0^c = \Theta \setminus \Theta_0$. Hypothese und/oder Alternative heissen *einfach*, falls Θ_0 bzw. Θ_A aus einem einzelnen Wert, ϑ_0 bzw. ϑ_A, bestehen, also z.B. $\Theta_0 = \{\vartheta_0\}$ ist; sonst heissen sie *zusammengesetzt*. Expliziter formuliert ist also die Hypothese

$$H_0 : \text{"der wahre (aber unbekannte) Parameter } \vartheta \text{ liegt in der Menge } \Theta_0\text{"}$$

und die Alternative

$$H_A : \text{"der wahre Parameter liegt in } \Theta_A\text{"}.$$

Wir illustrieren das an einem Beispiel.

Beispiel 11 (tea tasting lady). Eine englische Lady behauptet, bei Tee mit Milch anhand des Geschmacks unterscheiden zu können, ob zuerst die Milch oder zuerst der Tee in die Tasse eingegossen worden ist. Wie kann man überprüfen, ob das stimmen kann?

Wie im letzten Abschnitt stellen wir der Lady an n Tagen die Aufgabe, zwei Tassen (je eine vom Typ 1 und Typ 2) zu klassifizieren; sie soll also angeben, in welche der Tassen zuerst Milch eingegossen worden ist. Wir notieren die Ergebnisse $x_1, \ldots, x_n \in \{0, 1\}$ (falsch bzw. richtig klassifiziert) und fassen wie üblich diese Daten als Realisationen von Zufallsvariablen $X_1, \ldots, X_n$ auf. Dann ist $S_n = \sum_{i=1}^n X_i$ die Anzahl der korrekt klassifizierten Tassenpaare.

Als Modelle nehmen wir wieder an, dass die X_i unter P_ϑ i.i.d. $\sim Be(\vartheta)$ mit $\vartheta \in \Theta = [0, 1]$ sind. Dann ist natürlich $S_n \sim Bin(n, \vartheta)$ unter P_ϑ, d.h. im Modell P_ϑ, das zu ϑ gehört, ist die Anzahl S_n der Erfolge binomialverteilt mit Parametern n und ϑ.

Als Skeptiker zweifeln wir an den Fähigkeiten der Lady; wir wählen deshalb als (einfache) Hypothese $H_0 : \vartheta = \frac{1}{2}$, d.h. $\Theta_0 = \{\frac{1}{2}\}$ ("zufälliges Raten — das kann jeder"). Die (zusammengesetzte) Alternative, dass die Lady besondere Fähigkeiten hat, ist dann

$$H_A : \vartheta > \frac{1}{2}, \qquad \text{d.h. } \Theta_A = \left(\frac{1}{2}, 1\right].$$

Um weiterzukommen, müssen wir nun die Entscheidungsfindung anhand der Daten formalisieren. Auf das Beispiel selbst kommen wir später zurück. ◇

Ein *Test* ist allgemein eine Entscheidungsregel, die zu gegebenen Daten $x_1, \ldots, x_n$ einen von zwei Werten liefert; dabei nimmt man als diese zwei Werte 0 und 1, und man interpretiert den *Wert* 1 als die Entscheidung, die Hypothese H_0 abzulehnen.

Konkreter sieht die Entscheidungsregel meist folgendermassen aus. Man hat eine (messbare) Abbildung $t : \mathbb{R}^n \to \mathbb{R}$, $(x_1, \ldots, x_n) \mapsto t(x_1, \ldots, x_n)$, und einen *kritischen Bereich* oder *Verwerfungsbereich* $K \subseteq \mathbb{R}$. Die Zufallsvariable $T = t(X_1, \ldots, X_n)$ heisst dann *Teststatistik*. Die Entscheidungsregel ist $I_{\{t(x_1,\ldots,x_n)\in K\}}$, d.h. man verwirft die Hypothese genau dann, wenn der realisierte Wert $t(x_1, \ldots, x_n) = t(X_1(\omega), \ldots, X_n(\omega)) = T(\omega)$ im Verwerfungsbereich K liegt.

Beachte: Die Entscheidung des Tests hängt via $T(\omega)$ von der Realisierung ω ab. Weil T eine Zufallsvariable ist, ist die Menge

$$\tilde{K} = \{T \in K\} = \{\omega : T(\omega) \in K\}$$

eine (messbare) Teilmenge von Ω, also ein Ereignis, und wir können ihre Wahrscheinlichkeit $P_\vartheta[\tilde{K}] = P_\vartheta[T \in K]$ in jedem Modell P_ϑ betrachten.

Die Entscheidung bei einem Test kann auf zwei verschiedene Arten falsch herauskommen:

1) Bei einem *Fehler 1. Art* wird die Hypothese zu Unrecht abgelehnt, d.h. obwohl sie richtig ist. Das passiert für $\vartheta \in \Theta_0$ und $T \in K$; deshalb heisst $P_\vartheta[T \in K]$ für $\vartheta \in \Theta_0$ die Wahrscheinlichkeit für einen Fehler 1. Art.

2) Bei einem *Fehler 2. Art* wird die Hypothese zu Unrecht nicht verworfen, d.h. man akzeptiert die Hypothese (lehnt sie nicht ab), obwohl sie falsch ist. Das passiert für $\vartheta \in \Theta_A$ und $T \notin K$, und deshalb heisst $P_\vartheta[T \notin K] = 1 - P_\vartheta[T \in K]$ für $\vartheta \in \Theta_A$ die Wahrscheinlichkeit für einen Fehler 2. Art.

Beispiel 11 (tea tasting lady). Ein Fehler 1. Art besteht hier darin, die Hypothese des zufälligen Ratens abzulehnen, obwohl sie richtig ist. Anders gesagt glaubt man hier bei einem Fehler 1. Art an verborgene Fähigkeiten der Lady, obwohl ihre Ergebnisse durch zufälliges Raten entstehen (könnten). Bei einem Fehler 2. Art dagegen glaubt man nicht an die Fähigkeiten der Lady, obwohl diese durchaus vorhanden sind. ◇

Grundsätzlich möchte man einen Test immer so konstruieren (d.h. T und K wählen), dass die obigen Fehler-Wahrscheinlichkeiten möglichst klein werden. Also möchte man die Funktion $\vartheta \mapsto P_\vartheta[T \in K]$ auf Θ_0 möglichst klein und auf Θ_A möglichst gross haben. In der Regel ist aber $\vartheta \mapsto P_\vartheta[T \in K]$ stetig und $\Theta = \Theta_0 \cup \Theta_A$, so dass Θ_0 und Θ_A direkt nebeneinander liegen. Weil ein gleichzeitiges Minimieren auf Θ_0 und Maximieren auf Θ_A deshalb in der Regel grundsätzlich nicht möglich ist, hat sich das folgende zweistufige Vorgehen durchgesetzt:

a) Man wählt zuerst ein *Signifikanzniveau* $\alpha \in (0,1)$ und sorgt zunächst für

$$\sup_{\vartheta \in \Theta_0} P_\vartheta[T \in K] \leq \alpha,$$

d.h. man kontrolliert die Wahrscheinlichkeit für einen Fehler 1. Art durch α.

b) Anschliessend versucht man, die *Macht* des Tests, die Funktion

$$\beta : \Theta_A \to [0,1], \qquad \vartheta \mapsto \beta(\vartheta) := P_\vartheta[T \in K],$$

möglichst gross zu bekommen. Äquivalent formuliert heisst das, dass man die Grösse $1-\beta(\vartheta) = P_\vartheta[T \notin K]$ für $\vartheta \in \Theta_A$, also die Wahrscheinlichkeit für einen Fehler 2. Art, möglichst klein bekommen will.

Das obige asymmetrische Vorgehen macht es schwieriger, die Hypothese zu verwerfen als sie beizubehalten. Ein seriöser Test wird deshalb als Hypothese immer die Negation der eigentlich gewünschten Aussage benutzen. Gelingt es dann nämlich, das trotz der erschwerten Bedingungen zu verwerfen, so kann man viel eher zuversichtlich sein, tatsächlich einen Effekt gefunden zu haben.

Aus der asymmetrischen Behandlung von H_0 und H_A folgt auch, dass die Entscheidung bei einem Test davon abhängt, was man als Hypothese und was als Alternative wählt. Es kann also passieren, dass die gleiche inhaltliche Frage zu unterschiedlichen Entscheidungen führt, wenn man bei ihrem Test Hypothese und Alternative vertauscht. Wir werden das später mit einem Beispiel explizit illustrieren.

Wichtig: Die Entscheidung bei einem Test ist nie ein Beweis, sondern immer nur eine *Interpretation* der Übereinstimmung zwischen Daten und vermutetem Modell. Ist $T(\omega) \in K$, so wird man die Hypothese ablehnen und wegen der Daten nicht mehr glauben, dass $\vartheta \in \Theta_0$ ist. Das kann (muss aber nicht) zur Konsequenz haben, dass man eher glaubt, dass $\vartheta \in \Theta_A$ ist, so dass man die Alternative für plausibler hält. Ist $T(\omega) \notin K$, so wird man die Hypothese nicht verwerfen und sich im Glauben bestärkt fühlen, dass $\vartheta \in \Theta_0$

ist. *Wo aber ϑ tatsächlich liegt, weiss man genauso wenig wie vorher — ein Test liefert keinen Beweis!*

Wie sieht das nun konkreter in unserem Beispiel aus?

Beispiel 11 (tea tasting lady). Unter P_ϑ sind die Zufallsvariablen $X_1, \ldots, X_n$ wieder i.i.d. $\sim Be(\vartheta)$ sowie $S_n = \sum_{i=1}^n X_i \sim Bin(n, \vartheta)$. Hypothese und Alternative sind hier $H_0 : \vartheta = \frac{1}{2}$ und $H_A : \vartheta > \frac{1}{2}$.

Weil man für $\vartheta > \frac{1}{2}$ eher Einsen bei den X_i erwartet als für $\vartheta = \frac{1}{2}$, deuten viele Einsen bzw. ein grosser Wert von S_n eher auf H_A als H_0 hin. Ein plausibler Test könnte also die Teststatistik $T = S_n$ und einen kritischen Bereich der Form $K = (c, \infty)$ nehmen, d.h. man verwirft die Hypothese (des zufälligen Ratens), wenn die Lady viele Erfolge erzielt.

Als Hypothese haben wir hier $\vartheta = \frac{1}{2}$, d.h. "keine besonderen Fähigkeiten"; wir möchten zwar an diese Fähigkeiten eigentlich gerne glauben, aber natürlich nur, wenn sie wirklich von den Daten überzeugend gestützt werden. Wie oben erklärt machen wir es der Lady also bewusst schwer, um bei einem positiven Ergebnis zuversichtlich an ihre Fähigkeiten glauben zu können.

Um den kritischen Wert c zu einem Signifikanzniveau α zu bestimmen, brauchen wir die Wahrscheinlichkeiten $P_\vartheta[T \in K] = P_\vartheta[S_n > c]$ für $\vartheta = \frac{1}{2}$; für die Machtfunktion brauchen wir auch $\beta(\vartheta) = P_\vartheta[T \in K] = P_\vartheta[S_n > c]$ für $\vartheta > \frac{1}{2}$.

Allgemein formuliert bedeutet das, dass wir die Verteilung der Teststatistik T unter jedem P_ϑ (d.h. in jedem Modell) brauchen, um solche Wahrscheinlichkeiten ausrechnen zu können. Das ist in der Regel nicht möglich; um aber zumindest das Signifikanzniveau einhalten zu können, brauchen wir wenigstens die Verteilung von T unter der Hypothese H_0, d.h. in jedem Modell P_ϑ mit $\vartheta \in \Theta_0$.

Sei nun $n = 10$, d.h. wir lassen die Lady 10 Tage lang probieren. Die folgende Tabelle gibt dann die Binomial-Wahrscheinlichkeiten $P_\vartheta[S_{10} > k]$ für verschiedene ϑ und k.

ϑ	$k=7$	$k=8$	$k=9$	$k=10$
0.7	0.3828	0.1493	0.0282	0
0.6	0.1673	0.0464	0.0060	0
0.5	0.0547	0.0107	0.0010	0

Damit wir ein Signifikanzniveau von α erhalten, muss $P_{\frac{1}{2}}[S_{10} > c] \leq \alpha$ sein. Wählen wir $c = 7$, so ist das Niveau $\alpha = 0.0547$, also rund 5%. Auf diesem Niveau sind wir also bereit, bei 8 oder mehr Erfolgen der Lady unsere skeptische Hypothese zu verwerfen und an ihre Fähigkeiten zu glauben.

Die Macht des Tests erhalten wir auch aus der Tabelle; z.B. ist für das gewählte $c = 7$

$$\beta(0.6) = P_{0.6}[S_{10} > 7] = 0.1673$$

oder $\beta(0.7) = 0.3828$. Wir sehen also, dass

$$1 - \beta(\vartheta) = 1 - P_\vartheta[S_{10} > 7] = P_\vartheta[S_{10} \leq 7]$$

für $\vartheta \in \Theta_A$ recht gross wird; der Test hat eine beträchtliche Wahrscheinlichkeit für einen Fehler 2. Art, d.h. für einen Unglauben an tatsächlich vorhandene Fähigkeiten. Das ist die Kehrseite unseres allgemeinen skeptischen Ansatzes; bei nur schwachen Indikationen (p grösser als $\frac{1}{2}$, aber nicht sehr viel grösser) kann es durchaus vorkommen, dass wir diese Fähigkeiten zu Unrecht nicht bemerken. ◇

8.2 Konstruktion von Tests

Im obigen Beispiel haben wir die Wahl der Teststatistik $T = S_n$ und des kritischen Bereichs $K = (c, \infty)$ mit Plausibilitätsargumenten motiviert. Wir wollen nun kurz einen systematischen Ansatz erklären, der in vielen Situationen zu einem optimalen Test führt. Die Idee dazu geht auf Neyman und Pearson zurück.

Sei $L(x_1, \ldots, x_n; \vartheta)$ die Likelihood-Funktion; für diskrete X_i ist das also im Modell P_ϑ die Wahrscheinlichkeit $P_\vartheta[X_1 = x_1, \ldots, X_n = x_n]$, die beobachteten Werte $x_1, \ldots, x_n$ zu erhalten. Für $\vartheta_0 \in \Theta_0$ und $\vartheta_A \in \Theta_A$ betrachten wir den *Likelihood-Quotienten*

$$R(x_1, \ldots, x_n; \vartheta_0, \vartheta_A) := \frac{L(x_1, \ldots, x_n; \vartheta_A)}{L(x_1, \ldots, x_n; \vartheta_0)}.$$

Ist dieser Quotient gross, so ist der Zähler wesentlich grösser als der Nenner. Das bedeutet also, dass die Beobachtungen $x_1, \ldots, x_n$ als Resultate im Modell P_{ϑ_A} deutlich wahrscheinlicher sind als im Modell P_{ϑ_0}; die Daten sprechen dann also gegen ϑ_0 im Vergleich zu ϑ_A. Es liegt deshalb nahe, als Teststatistik $T := R(X_1, \ldots, X_n; \vartheta_0, \vartheta_A)$ und als kritischen Bereich $K := (c, \infty)$ zu wählen, wenn man ϑ_0 gegen ϑ_A testen will; man verwirft dann also die Hypothese H_0, wenn der Quotient R gross wird.

Sind Hypothese und Alternative beide einfach, so ist dieser Test optimal, wie das folgende Resultat zeigt.

Satz 9.1. (Neyman–Pearson-Lemma): *Sei $\Theta_0 = \{\vartheta_0\}$ und $\Theta_A = \{\vartheta_A\}$. Wie oben sei $T := R(X_1, \ldots, X_n; \vartheta_0, \vartheta_A)$ und $K := (c, \infty)$ sowie $\alpha^* := P_{\vartheta_0}[T \in K] = P_{\vartheta_0}[T > c]$. Der Likelihood-Quotienten-Test mit Teststatistik T und kritischem Bereich K ist dann im folgenden Sinn optimal: Jeder andere Test mit Signifikanzniveau $\alpha \leq \alpha^*$ hat kleinere Macht bzw. eine grössere Wahrscheinlichkeit für einen Fehler 2. Art.*
(Etwas formaler: Ist (T', K') ein anderer Test mit $P_{\vartheta_0}[T' \in K'] \leq \alpha^* = P_{\vartheta_0}[T \in K]$, so gilt auch $P_{\vartheta_A}[T' \in K'] \leq P_{\vartheta_A}[T \in K]$.)

Beweis. Siehe [Krengel, Satz 6.2].

Die obige Situation mit einfacher Hypothese und Alternative ist so speziell, dass sie in der Praxis kaum je auftritt. Die Grundidee für den Test lässt sich aber verallgemeinern und liefert in gewissen (weniger restriktiven) Situationen immer noch gute oder optimale Tests, so dass man das Vorgehen mit gutem Gewissen als einen systematischen Ansatz empfehlen kann. Wie wir in Beispielen sehen werden, sind die resultierenden Tests oft auch intuitiv sehr einleuchtend.

Etwas genauer betrachtet man bei zusammengesetzten Hypothesen und Alternativen den sogenannten *verallgemeinerten Likelihood-Quotienten*

$$R(x_1,\dots,x_n) := \frac{\sup_{\vartheta\in\Theta_A} L(x_1,\dots,x_n;\vartheta)}{\sup_{\vartheta\in\Theta_0} L(x_1,\dots,x_n;\vartheta)}$$

oder auch

$$\tilde{R}(x_1,\dots,x_n) := \frac{\sup_{\vartheta\in\Theta_A\cup\Theta_0} L(x_1,\dots,x_n;\vartheta)}{\sup_{\vartheta\in\Theta_0} L(x_1,\dots,x_n;\vartheta)}$$

und wählt als Teststatistik $T_0 := R(X_1,\dots,X_n)$ bzw. $\tilde{T} := \tilde{R}(X_1,\dots,X_n)$ mit kritischem Bereich $K_0 := (c_0,\infty)$. Durch Umformen erhält man daraus oft einen äquivalenten, aber einfacheren Test (T,K) von einer leicht anderen Form; siehe Beispiele. Die Konstante c_0 bzw. den Bereich K muss man dabei noch so wählen, dass der Test ein in der Regel a priori gewähltes Signifikanzniveau einhält.

Beispiel 11 (tea tasting lady). Im Modell P_ϑ sind $X_1,\dots,X_n$ i.i.d. $\sim Be(\vartheta)$; die Gewichtsfunktion eines X_i unter P_ϑ ist also $p_X(x_i;\vartheta) = \vartheta^{x_i}(1-\vartheta)^{1-x_i}$, und damit wird die Likelihood-Funktion

$$L(x_1,\dots,x_n;\vartheta) = \prod_{i=1}^n p_X(x_i;\vartheta) = \vartheta^{\sum_{i=1}^n x_i}(1-\vartheta)^{n-\sum_{i=1}^n x_i}.$$

Der Likelihood-Quotient ist also

$$\begin{aligned} R(x_1,\dots,x_n;\vartheta_0,\vartheta_A) &= \frac{L(x_1,\dots,x_n;\vartheta_A)}{L(x_1,\dots,x_n;\vartheta_0)} \\ &= \left(\frac{\vartheta_A}{\vartheta_0}\right)^{\sum_{i=1}^n x_i} \left(\frac{1-\vartheta_A}{1-\vartheta_0}\right)^{n-\sum_{i=1}^n x_i} \\ &= \left(\frac{\vartheta_A(1-\vartheta_0)}{\vartheta_0(1-\vartheta_A)}\right)^{\sum_{i=1}^n x_i} \left(\frac{1-\vartheta_A}{1-\vartheta_0}\right)^{n}. \end{aligned}$$

Nun ist ja $\vartheta_0 = \frac{1}{2}$ und $\vartheta_A > \frac{1}{2}$, also $\vartheta_0 < \vartheta_A$. Damit ist

$$\frac{\vartheta_A(1-\vartheta_0)}{\vartheta_0(1-\vartheta_A)} = \frac{\vartheta_A - \vartheta_0\vartheta_A}{\vartheta_0 - \vartheta_0\vartheta_A} > 1,$$

und damit ist $R(x_1,\dots,x_n;\vartheta_0,\vartheta_A)$ genau dann gross, wenn der Exponent $\sum_{i=1}^n x_i$ gross ist. Statt des komplizierten Quotienten wählen wir als Teststatistik also

$$T := \sum_{i=1}^n X_i = S_n,$$

und der kritische Bereich "Quotient gross" hat die äquivalente Form "Summe (= Exponent) gross", also

$$K := (c, \infty).$$

Also liefert hier der Neyman–Pearson-Ansatz genau das Testverfahren, das wir oben schon aufgrund von Plausibilitätsargumenten benutzt haben. ◇

Beispiel. Seien $X_1, \ldots, X_n$ unter P_ϑ i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$ mit *bekannter Varianz* σ^2; der Parameter ist hier also $\vartheta = \mu \in \mathbb{R}$. Die Dichtefunktion von X_i ist

$$f_X(x_i; \vartheta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_i - \vartheta)^2}{2\sigma^2}\right),$$

die Likelihood-Funktion ist

$$L(x_1, \ldots, x_n; \vartheta) = \prod_{i=1}^n f_X(x_i; \vartheta) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \vartheta)^2\right),$$

und der Likelihood-Quotient wird

$$\begin{aligned} R(x_1, \ldots, x_n; \vartheta_0, \vartheta_A) &= \frac{L(x_1, \ldots, x_n; \vartheta_A)}{L(x_1, \ldots, x_n; \vartheta_0)} \\ &= \exp\left(-\frac{1}{2\sigma^2}\Big(\sum_{i=1}^n (x_i - \vartheta_A)^2 - \sum_{i=1}^n (x_i - \vartheta_0)^2\Big)\right) \\ &= \text{const.}(\sigma, \vartheta_0, \vartheta_A) \exp\left(\frac{1}{\sigma^2}(\vartheta_A - \vartheta_0)\sum_{i=1}^n x_i\right). \end{aligned}$$

Betrachten wir nun die Hypothese $H_0 : \vartheta = \vartheta_0$ und die Alternative $H_A : \vartheta = \vartheta_A$. Der obige Quotient wird tendenziell gross, falls der Exponent $(\vartheta_A - \vartheta_0)\sum_{i=1}^n x_i$ gross ist. Was das für $\sum_{i=1}^n x_i$ bedeutet, hängt vom Vorzeichen von $\vartheta_A - \vartheta_0$ ab. In jedem Fall wählen wir als Teststatistik

$$T' := \sum_{i=1}^n X_i.$$

Ist $\vartheta_A > \vartheta_0$, so ist $\vartheta_A - \vartheta_0 > 0$, und der Exponent wird dann gross, wenn T' gross ist; hier wählen wir den kritischen Bereich also von der Form $K'_> := (c'_>, \infty)$, d.h. wir lehnen H_0 ab, wenn T' gross ist.

Ist $\vartheta_A < \vartheta_0$, so ist $\vartheta_A - \vartheta_0 < 0$ und der Exponent gross für T' klein (d.h. negativ). Hier ist also der kritische Bereich von der Form $K'_< := (-\infty, c'_<)$.

In beiden Fällen müssen wir den kritischen Bereich, d.h. hier konkret die Konstanten $c'_>$ bzw. $c'_<$, noch so festlegen, dass der Test ein gewähltes Signifikanzniveau α einhält. Wir wollen also $P_{\vartheta_0}[T' \in K'] \leq \alpha$ erreichen, und um diese Wahrscheinlichkeit zu berechnen, brauchen wir die Verteilung der Teststatistik T' unter P_{ϑ_0}, d.h. unter der Hypothese H_0.

Im vorliegenden Fall ist das einfach. Unter jedem P_ϑ sind die X_i i.i.d. $\sim \mathcal{N}(\vartheta, \sigma^2)$; also ist die Summe

$$T' = \sum_{i=1}^{n} X_i \sim \mathcal{N}(n\vartheta, n\sigma^2) \qquad \text{unter } P_\vartheta.$$

Äquivalent ist

$$T = \frac{\overline{X}_n - \vartheta}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1) \qquad \text{unter } P_\vartheta,$$

und wir können T statt T' als Teststatistik benutzen.

Man beachte, dass T im Modell P_ϑ mit $\vartheta \in \Theta_0$, d.h. mit $\vartheta = \vartheta_0$, also unter der Hypothese H_0, tatsächlich berechenbar ist: die Varianz σ^2 ist nach Annahme bekannt, und der zu testende Erwartungswert ϑ_0 ist natürlich auch bekannt. $\diamond$

Das obige Beispiel zeigt den letzten Schritt, den wir allgemein noch machen müssen. Um den kritischen Bereich K passend zum gewünschten Niveau α festlegen zu können, brauchen wir die Verteilung der Teststatistik T unter der Hypothese H_0, d.h. in jedem Modell P_ϑ mit $\vartheta \in \Theta_0$.

8.3 Beispiele

In diesem Abschnitt illustrieren wir die obigen Überlegungen durch einige Beispiele. Dabei verzichten wir weitgehend auf Herleitungen und präsentieren nur die Ergebnisse mehr oder weniger in der Form von "Kochrezepten".

Beispiel. *Normalverteilung, Test für Erwartungswert* bei *bekannter Varianz:* Dieser Test ist unter dem Namen *z-Test* bekannt. Hier sind $X_1, \dots, X_n$ i.i.d. $\sim \mathcal{N}(\vartheta, \sigma^2)$ unter P_ϑ mit bekannter Varianz σ^2, und wir wollen die Hypothese $H_0 : \vartheta = \vartheta_0$ testen. Mögliche Alternativen H_A sind $\vartheta > \vartheta_0$ oder $\vartheta < \vartheta_0$ (*einseitig*), oder $\vartheta \neq \vartheta_0$ (*zweiseitig*). Welche der Alternativen sinnvoll ist, hängt von der konkreten Fragestellung ab.

Die Teststatistik hier ist in jedem Fall

$$T := \frac{\overline{X}_n - \vartheta_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1) \qquad \text{unter } P_{\vartheta_0}.$$

Der kritische Bereich K ist von der Form $(c_>, \infty)$ für den einseitigen Test gegen die Alternative $H_A : \vartheta > \vartheta_0$, bzw. $(-\infty, c_<)$, bzw. $(-\infty, -c_\neq) \cup (c_\neq, +\infty)$. Im zweiseitigen Fall verwirft man H_0 also zugunsten der Alternative $H_A : \vartheta \neq \vartheta_0$, falls $|T| > c_\neq$ ist.

Die Konstanten $c_>$, $c_<$, $c_\neq$ bestimmt man zum gewählten Niveau mit Hilfe der Verteilung von T unter P_{ϑ_0}. Zum Beispiel liefert die Bedingung

$$\alpha = P_{\vartheta_0}[T \in K_>] = P_{\vartheta_0}[T > c_>] = 1 - P_{\vartheta_0}[T \leq c_>] = 1 - \Phi(c_>),$$

dass $c_> = \Phi^{-1}(1-\alpha) =: z_{1-\alpha}$ das sogenannte $(1-\alpha)$-*Quantil* der $\mathcal{N}(0,1)$-Verteilung sein muss; für $\vartheta > \vartheta_0$ verwirft man also H_0, falls

$$\overline{X}_n > \vartheta_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$$

ist. Analog ist $c_< = z_\alpha = -z_{1-\alpha}$ und $c_\neq = z_{1-\frac{\alpha}{2}}$, wobei wir die Symmetrie der $\mathcal{N}(0,1)$-Verteilung ausnutzen; es gilt nämlich

$$\alpha = P_{\vartheta_0}[T \in K_\neq] = P_{\vartheta_0}[T < -c_\neq] + P_{\vartheta_0}[T > c_\neq] = \Phi(-c_\neq) + 1 - \Phi(c_\neq) = 2\big(1 - \Phi(c_\neq)\big).$$

◇

Beispiel 12 (Strausseneier). Die Australier Mr. Smith und Dr. Thurston streiten sich über das Durchschnittsgewicht von *Strausseneiern*. Beide sind damit einverstanden, das Gewicht approximativ als normalverteilt aufzufassen; Mr. Smith behauptet aber, das mittlere Gewicht sei 1100g, während Dr. Thurston darauf besteht, dass die Eier schwerer seien, und zwar im Schnitt 1200g. Um ihren Streit beilegen zu können, reisen die beiden nach

Afrika, um in der Savanne Strausseneier zu suchen. Weil diese aber meistens gut versteckt sind, finden sie nur acht, und zwar mit folgenden Gewichten (in g): 1090, 1150, 1170, 1080, 1210, 1230, 1180, 1140.

Dr. Thurston schlägt nun vor, Mr. Smiths Behauptung als Hypothese $\mu = \mu_0 = 1100$ gegen seine Alternative $\mu > 1100$ (oder auch $\mu = 1200$) auf dem 5%-Niveau zu testen. Die Varianz σ^2 ist beiden bekannt; sie beträgt (in g) $\sigma = 55$. Also berechnet Dr. Thurston

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i = 1156.25$$

und sucht in der Tabelle $z_{1-\alpha} = z_{0.95} = 1.645$. Damit ist

$$T_{\mathrm{Th}}(\omega) = \frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} = \sqrt{8} \times \frac{1156.25 - 1100}{55} = 2.89.$$

Wegen $T_{\mathrm{Th}}(\omega) > z_{0.95}$ wird also die Hypothese $\mu = 1100$ auf dem 5%-Niveau verworfen.

Mr. Smith kommt sich bei diesem Vorgehen benachteiligt vor und macht deshalb den Gegenvorschlag, doch besser Dr. Thurstons Behauptung als Hypothese $\mu = \mu_1 = 1200$ gegen seine Alternative $\mu < 1200$ (oder auch $\mu = 1100$) zu testen. Er berechnet deshalb

$$T_{\mathrm{Sm}}(\omega) = \frac{\overline{x}_n - \mu_1}{\sigma/\sqrt{n}} = \sqrt{8} \times \frac{1156.25 - 1200}{55} = -2.25.$$

Wegen $z_\alpha = z_{0.05} = -z_{0.95} = -1.645$ ist $T_{\mathrm{Sm}}(\omega) < z_{0.05}$; also wird auch die Hypothese $\mu = 1200$ auf dem 5%-Niveau verworfen.

Dieses Beispiel illustriert sehr schön die Bedeutung der Wahl von Hypothese und Alternative und auch ihre asymmetrische Behandlung. Mit dem ersten Test würde man Dr. Thurston Recht geben, mit dem zweiten hingegen Mr. Smith — und das bei völlig identischen Daten. ◇

Beispiel. *Normalverteilung, Test für Erwartungswert* bei *unbekannter Varianz:* Dieser Test ist unter dem Namen *t-Test* bekannt. Hier sind $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$ unter $P_{\overrightarrow{\vartheta}}$, wobei $\overrightarrow{\vartheta} = (\mu, \sigma^2)$ und insbesondere die Varianz σ^2 unbekannt ist. Wir wollen wieder die Hypothese $\mu = \mu_0$ testen.

Genaugenommen ist das eine zusammengesetzte Hypothese, weil der Parameter $\overrightarrow{\vartheta}$ aus den zwei Komponenten μ und σ^2 besteht. Explizit wäre also

$$\Theta_0 = \{\mu_0\} \times (0, \infty) = \{\overrightarrow{\vartheta} = (\mu, \sigma^2) : \mu = \mu_0\}.$$

Die Teststatistik ist hier

$$T := \frac{\overline{X}_n - \mu_0}{S/\sqrt{n}} \sim t_{n-1} \qquad \text{unter } P_{\vartheta_0};$$

wir ersetzen also die unbekannte Varianz durch den Schätzer

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$$

für σ^2 und nutzen die Verteilungsaussagen aus Satz 7.1 aus.

Der kritische Bereich hat (je nach Alternative) eine der drei Formen aus dem letzten Beispiel; die kritischen Werte hier sind $c_> = t_{n-1,1-\alpha}$, bzw. $c_< = t_{n-1,\alpha} = -t_{n-1,1-\alpha}$, bzw. $c_{\neq} = t_{n-1,1-\frac{\alpha}{2}}$. Hier bezeichnen wir mit $t_{m,\gamma}$ das sogenannte γ-Quantil einer t_m-Verteilung, d.h. denjenigen Wert $t_{m,\gamma}$, für den gilt $P[X \leq t_{m,\gamma}] = \gamma$ für X t-verteilt mit m Freiheitsgraden, d.h. $X \sim t_m$. Diese Werte findet man in Tabellen. $\diamond$

Beispiel 12 (Strausseneier). Mr. Smith und Dr. Thurston fragen sich, ob sie bei ihrem ersten Versuch vielleicht eine falsche Information über die Varianz von Strausseneiern benutzt haben. Sie beschliessen deshalb, ihre Tests nochmals ohne die Annahme einer bekannten Varianz durchzuführen, und kommen damit zu einem t-Test.

Dr. Thurston beharrt immer noch darauf, die Hypothese $\mu = 1100$ gegen die Alternative $\mu > 1100$ auf dem 5%-Niveau zu testen. Weil die Varianz σ^2 nun aber unbekannt ist, berechnet er

$$s^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\overline{x}_n)^2\right) = 2798.21, \quad \text{also } s = 52.90; \qquad t_{7,0.95} = 1.895.$$

Damit erhält er als Wert für die Teststatistik

$$\tilde{T}_{\text{Th}}(\omega) = \frac{\overline{x}_n - \mu_0}{s/\sqrt{n}} = \sqrt{8} \times \frac{1156.25 - 1100}{52.90} = 3.008.$$

Wegen $\tilde{T}_{\mathrm{Th}}(\omega) > t_{7,0.95}$ wird also die Hypothese $\mu = 1100$ auf dem 5%-Niveau wieder verworfen.

Nicht überraschend ist Mr. Smith mit diesem Vorgehen immer noch nicht einverstanden und will lieber Dr. Thurstons Behauptung als Hypothese $\mu = \mu_1 = 1200$ gegen seine Alternative $\mu < 1200$ (oder auch $\mu = 1100$) testen. Er berechnet deshalb

$$\tilde{T}_{\mathrm{Sm}}(\omega) = \frac{\overline{x}_n - \mu_1}{s/\sqrt{n}} = \sqrt{8} \times \frac{1156.25 - 1200}{52.90} = -2.339.$$

Wegen $t_{n-1,\alpha} = t_{7,0.05} = -t_{7,0.95} = -1.895$ ist $T_{\mathrm{Sm}}(\omega) < -t_{7,0.95}$; also wird auch die Hypothese $\mu = 1200$ auf dem 5%-Niveau verworfen — und damit sind die beiden wieder gleich weit wie vorher. ◇

Die obigen zwei Tests heissen auch *Einstichproben-Tests*, weil man nur Daten aus einer Stichprobe hat. Bei *Zweistichproben-Tests* geht man aus von Zufallsvariablen $X_1, \ldots, X_n$ und $Y_1, \ldots, Y_m$, die *unter P_ϑ alle unabhängig* sind; zudem sind die X_i und die Y_j unter P_ϑ jeweils für sich betrachtet i.i.d.

Beispiel. *Gepaarter Zweistichproben-Test* bei *Normalverteilung:* Hier sind $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{N}(\mu_X, \sigma^2)$ und $Y_1, \ldots, Y_n$ i.i.d. $\sim \mathcal{N}(\mu_Y, \sigma^2)$ unter P_ϑ; insbesondere ist $m = n$ und die Varianz σ^2 bei beiden Stichproben dieselbe. Eine solche Situation tritt auf, wenn z.B. eine Gruppe von Personen zwei verschiedene Dinge ausprobiert, so dass man eine natürliche Paarbildung zwischen den X_i und Y_i hat.

In dieser Situation kann man Tests über den Vergleich von μ_X und μ_Y auf den Fall nur einer Stichprobe zurückführen; die Differenzen $Z_i := X_i - Y_i$ sind nämlich unter P_ϑ i.i.d. $\sim \mathcal{N}(\mu_X - \mu_Y, 2\sigma^2)$. Damit kann man die bisherigen Tests in leicht angepasster Form benutzen, sowohl für bekannte wie für unbekannte Varianz σ^2. Die resultierenden Tests heissen dann nicht überraschend *gepaarter Zweistichproben-z-Test* (bei bekanntem σ^2) bzw. *gepaarter Zweistichproben-t-Test* (bei unbekanntem σ^2). ◇

Bemerkung. Wir haben oben angenommen, dass X_i und Y_i unabhängig sind. Allgemeiner kann man annehmen, dass die Paare (X_i, Y_i), $i = 1, \ldots, n$, unter P_ϑ unabhängig sind

mit einer zweidimensionalen Normalverteilung mit Erwartungswerten μ_X, μ_Y, bekannten gleichen Varianzen σ^2 und bekannter Korrelation $\varrho \in (-1, +1)$. (Der Fall $\varrho = 0$ entspricht Unabhängigkeit.) Dann sind die $Z_i = X_i - Y_i$ unter P_ϑ i.i.d. $\sim \mathcal{N}(\mu_X - \mu_Y, 2(1-\varrho)\sigma^2)$, und man kann wie oben die bisherigen Tests benutzen. $\diamond$

Beispiel. *Ungepaarter Zweistichproben-Test* bei *Normalverteilung:* Hier sind unter P_ϑ $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{N}(\mu_X, \sigma^2)$ und $Y_1, \ldots, Y_m$ i.i.d. $\sim \mathcal{N}(\mu_Y, \sigma^2)$, wobei die Varianz in beiden Fällen dieselbe ist, aber $m \neq n$ sein kann. Will man einen Vergleich über μ_X und μ_Y hier testen, so kann man nicht mehr paarweise Differenzen bilden. Diesen Test muss man auch benutzen, falls zufällig $m = n$ ist, aber die Daten nicht natürlich gepaart sind. Wir nehmen immer noch an, dass $X_1, \ldots, X_n$ und $Y_1, \ldots, Y_m$ unabhängig sind.

a) Ist σ^2 *bekannt*, so ist die Teststatistik

$$T := \frac{(\overline{X}_n - \overline{Y}_m) - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim \mathcal{N}(0,1)$$

unter jedem P_ϑ. Dabei ist σ nach Annahme bekannt, und $\mu_X - \mu_Y$ muss sich aus der gewünschten Hypothese H_0 als bekannt ergeben. Die kritischen Werte für den Verwerfungsbereich sind wie oben geeignete Quantile der $\mathcal{N}(0,1)$-Verteilung, je nach Alternative. Das ist der *ungepaarte Zweistichproben-z-Test*.

b) Ist σ^2 *unbekannt*, so brauchen wir zuerst die beiden empirischen Varianzen

$$S_X^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2,$$
$$S_Y^2 := \frac{1}{m-1}\sum_{j=1}^{m}(Y_j - \overline{Y}_m)^2.$$

Mit

$$S^2 := \frac{1}{m+n-2}\big((n-1)S_X^2 + (m-1)S_Y^2\big)$$

ist dann die Teststatistik

$$T := \frac{(\overline{X}_n - \overline{Y}_m) - (\mu_X - \mu_Y)}{S\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}$$

unter jedem P_ϑ. Der Rest geht dann analog wie oben. Dieser Test heisst *ungepaarter Zweistichproben-t-Test*. $\diamond$

Die meisten bisherigen Beispiele für Tests gehen von der Annahme normalverteilter Stichproben aus; diese Situation ist sehr angenehm, weil man dann die Verteilung der Teststatistik einfach in expliziter Form hat. Diese Tests sind sehr gut, falls man tatsächlich Normalverteilungen hat; ist das aber nicht der Fall, so verlieren sie sehr schnell einen grossen Teil ihrer Macht. Deshalb ist es gut, auch alternative Tests zu kennen, die von weniger spezifischen Annahmen ausgehen.

8.4 Zusammenfassung zu Tests

Zum Abschluss dieses Abschnitts fassen wir das allgemeine Vorgehen bei Tests noch einmal kurz zusammen. Man hat die folgenden 5 Schritte:

1) Wahl des Modells.

2) Formulierung von Hypothese und Alternative.

3) Bestimmung der Teststatistik T und der Form des kritischen Bereichs K; das kann aus einer Herleitung via verallgemeinerten LQ-Test oder direkt aus einem Statistik-Buch stammen.

4) Festlegung des Niveaus α liefert (die Grenze für) den kritischen Bereich K; dazu braucht man die Verteilung von T unter P_ϑ für alle $\vartheta \in \Theta_0$ (exakt oder approximativ).

5) Berechnen der Teststatistik $T(\omega)$ aus den Daten; ist $T(\omega) \in K$, so wird die Hypothese abgelehnt, andernfalls wird die Hypothese nicht verworfen.

5′) Berechnen von Teststatistik $T(\omega)$ und entsprechendem realisiertem p-Wert(ω) aus den Daten; ist letzterer $\leq \alpha$, so wird die Hypothese abgelehnt, andernfalls nicht.

9 Konfidenzbereiche

Grundidee: Wie in Abschnitt 7 suchen wir aus einer Familie $(P_\vartheta)_{\vartheta\in\Theta}$ von Modellen eines, das zu unseren Daten $x_1,\dots,x_n$ passt. Ein Schätzer für ϑ gibt uns dabei einen einzelnen zufälligen möglichen Parameterwert. Weil es schwierig ist, mit diesem einen Wert den richtigen (aber unbekannten) Parameter zu treffen, suchen wir nun stattdessen eine *(zufällige) Teilmenge des Parameterbereichs*, die hoffentlich den wahren Parameter enthält.

Etwas formaler ist ein (realisierter) *Konfidenzbereich* für ϑ zu Daten $x_1,\dots,x_n$ eine Menge $C(x_1,\dots,x_n)\subseteq\Theta$; in den meisten Fällen ist das ein Intervall, dessen Endpunkte von $x_1,\dots,x_n$ abhängen. Ersetzen wir die Daten $x_1,\dots,x_n$ durch die sie generierenden Zufallsvariablen, so ist $\tilde{C}:=C(X_1,\dots,X_n)$ also eine zufällige Teilmenge von Θ, mit Realisierung $\tilde{C}(\omega)=C(X_1(\omega),\dots,X_n(\omega))$ bei einem festen ω. Ein solches C heisst ein Konfidenzbereich zum *Niveau* $1-\alpha$, falls gilt

$$P_\vartheta[C(X_1,\dots,X_n)\ni\vartheta]\geq 1-\alpha \qquad \text{für alle } \vartheta\in\Theta,$$

d.h. in jedem Modell erwischt man den Parameter mit grosser Wahrscheinlichkeit.

Beispiel 12 (Strausseneier). Die Australier Mr. Smith und Dr. Thurston streiten sich noch immer über das Gewicht von Strausseneiern. Sie haben von ihrer Afrikareise $n=8$ Eier mitgebracht, deren Gewichte (in g) 1090, 1150, 1170, 1080, 1210, 1230, 1180, 1140 betragen. Diese Daten fassen sie auf als Realisationen von Zufallsvariablen $X_1,\dots,X_n$, die alle unter P_ϑ i.i.d. $\sim\mathcal{N}(\mu,\sigma^2)$ sind. Gesucht sind nun Konfidenzbereiche für die unbekannten Parameter μ und σ^2. Im Gegensatz zum Beispiel im letzten Abschnitt wird hier σ^2 auch als unbekannt angenommen.

Die offensichtlichen *Schätzer* für μ und σ^2 sind das Stichprobenmittel $\overline{X}_n=\frac{1}{n}\sum_{i=1}^n X_i$ und die Stichprobenvarianz $S^2=\frac{1}{n-1}\sum_{i=1}^n(X_i-\overline{X}_n)^2$. Es liegt nahe, als Konfidenzbereich jeweils ein Intervall um diese Schätzer herum anzusetzen, und wir machen das zuerst für μ.

Machen wir den Ansatz

$$C(X_1, \dots, X_n) = [\overline{X}_n - \cdots, \overline{X}_n + \cdots],$$

so wollen wir erreichen, dass gilt

$$1 - \alpha \leq P_\vartheta[C(X_1, \dots, X_n) \ni \vartheta] = P_\vartheta\big[[\overline{X}_n - \cdots, \overline{X}_n + \cdots] \ni \mu\big] = P_\vartheta\big[|\overline{X}_n - \mu| \leq \cdots\big].$$

Nach Satz 7.1 ist für jedes $\vartheta \in \Theta$

$$\frac{\overline{X}_n - \mu}{S/\sqrt{n}} \sim t_{n-1} \qquad \text{unter } P_\vartheta;$$

also wollen wir

$$1 - \alpha \leq P_\vartheta\left[\left|\frac{\overline{X}_n - \mu}{S/\sqrt{n}}\right| \leq \frac{\cdots}{S/\sqrt{n}}\right].$$

Um ein möglichst kurzes Intervall zu erhalten, erfüllen wir diese Bedingung mit Gleichheit, und dann brauchen wir gerade

$$\frac{\cdots}{S/\sqrt{n}} = t_{n-1,1-\frac{\alpha}{2}}.$$

Also erhalten wir als Konfidenzintervall für μ zum Niveau $1 - \alpha$

$$C(X_1, \dots, X_n) = \left[\overline{X}_n - t_{n-1,1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}, \overline{X}_n + t_{n-1,1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right].$$

Die konkreten Realisierungen in unserem Beispiel für $1 - \alpha = 99\%$ sind

$$\overline{x}_n = 1156.25, \qquad s = 52.90, \qquad t_{7,0.995} = 3.499$$

und damit

$$C(x_1, \dots, x_n) = [1090.81, 1221.69].$$

Wir sehen, dass sowohl 1100 als auch 1200 in diesem realisierten Intervall liegen. Aufgrund der vorliegenden Daten sind also beide Behauptungen (diejenige von Dr. Thurston und die von Mr. Smith) plausibel.

Um ein Konfidenzintervall für σ^2 zu konstruieren, benutzen wir die ebenfalls aus Satz 7.1 bekannte Tatsache, dass

$$\frac{1}{\sigma^2}(n-1)S^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \sim \chi^2_{n-1} \qquad \text{unter } P_\vartheta.$$

Also ist, mit der Notation $\chi^2_{m,\gamma}$ für das γ-Quantil einer χ^2_m-Verteilung,

$$1-\alpha = P_\vartheta\left[\chi^2_{n-1,\frac{\alpha}{2}} \le \frac{1}{\sigma^2}(n-1)S^2 \le \chi^2_{n-1,1-\frac{\alpha}{2}}\right] = P_\vartheta\left[\frac{(n-1)S^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right],$$

und das Konfidenzintervall für σ^2 zum Niveau $1-\alpha$ wird

$$C(X_1,\dots,X_n) = \left[\frac{(n-1)S^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}, \frac{(n-1)S^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right].$$

Die konkreten Realisierungen in unserem Beispiel für $1-\alpha = 95\%$ sind

$$s^2 = 2798.21, \qquad \chi^2_{7,0.025} = 1.69, \qquad \chi^2_{7,0.975} = 16.01$$

und damit

$$C(x_1,\dots,x_n) = [1223.45, 11'590.23].$$

Übersetzt für $\sigma = \sqrt{\sigma^2}$ erhalten wir als realisiertes Konfidenzintervall $[34.98, 107.66]$. ◇

Bemerkung. Im obigen Beispiel haben wir exakte Konfidenzintervalle erhalten, weil wir genügend genaue Verteilungsaussagen zur Verfügung haben. In allgemeinen Situationen kann man oft nur approximative Konfidenzintervalle mit Hilfe des zentralen Grenzwertsatzes bekommen; siehe Abschnitt 7.3. ◇

Beispiel 11 (tea tasting lady). Nehmen wir nochmals an, dass die Lady in $n = 10$ Versuchen insgesamt 6 Tassenpaare richtig klassifiziert hat. Wie können wir dann einen Konfidenzbereich für ihre Erfolgswahrscheinlichkeit bekommen?

Allgemein ist in jedem Modell P_ϑ die Anzahl S_n der Erfolge $Bin(n,\vartheta)$-verteilt, und wir suchen einen Konfidenzbereich für den unbekannten Parameter ϑ. Wir wollen den zentralen Grenzwertsatz benutzen, um einen approximativen Konfidenzbereich zu bekommen. Nach dem ZGS ist

$$S_n^* := \frac{S_n - n\vartheta}{\sqrt{n\vartheta(1-\vartheta)}} \overset{\text{approx.}}{\sim} \mathcal{N}(0,1).$$

Also gilt

$$P\left[\left|\frac{S_n - n\vartheta}{\sqrt{n\vartheta(1-\vartheta)}}\right| \le z_{1-\frac{\alpha}{2}}\right] = P[|S_n^*| \le z_{1-\frac{\alpha}{2}}] \approx 1-\alpha,$$

und wir können versuchen, diese Ungleichung nach ϑ aufzulösen. Wir wollen also

$$|S_n - n\vartheta| \leq z_{1-\frac{\alpha}{2}} \sqrt{n\vartheta(1-\vartheta)} \qquad \text{oder} \qquad (S_n - n\vartheta)^2 \leq z^2_{1-\frac{\alpha}{2}} n\vartheta(1-\vartheta),$$

aber das wird eher kompliziert.

Alternativ kann man wie folgt vorgehen:

Methode 1: Wir gehen davon aus, dass $\vartheta(1-\vartheta) \approx \frac{1}{4}$ ist und setzen das ein. Dann wollen wir also

$$|S_n - n\vartheta| \leq z_{1-\frac{\alpha}{2}} \sqrt{\frac{n}{4}},$$

und das approximative Konfidenzintervall für ϑ ergibt sich als

$$\left[\overline{S}_n - \frac{z_{1-\frac{\alpha}{2}}}{2\sqrt{n}}, \overline{S}_n + \frac{z_{1-\frac{\alpha}{2}}}{2\sqrt{n}}\right].$$

Methode 2: Wir benutzen den ZGS, um zuerst

$$\overline{S}_n \overset{\text{approx.}}{\sim} \mathcal{N}\left(\vartheta, \frac{\vartheta(1-\vartheta)}{n}\right)$$

zu erhalten. Im approximativen Konfidenzintervall für ϑ ersetzen wir dann ϑ durch seinen Schätzer $\overline{S}_n$ und erhalten so das "doppelt approximative" Konfidenzintervall

$$\left[\overline{S}_n - \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}} \sqrt{\overline{S}_n(1-\overline{S}_n)}, \overline{S}_n + \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}} \sqrt{\overline{S}_n(1-\overline{S}_n)}\right].$$

Für $1-\alpha = 95\%$ ist $z_{1-\frac{\alpha}{2}} = 1.96$. Für $n = 10$ und $s_n = 6$ ergeben sich dann als Schätzwert für ϑ der Wert 0.6 und die realisierten Intervalle $[0.290, 0.910]$ mit Methode 1 und $[0.296, 0.904]$ mit Methode 2. Das zeigt, dass man mit so wenigen Daten (nicht überraschend) keine präzisen Aussagen erwarten kann.

Hätte man stattdessen $n = 100$ Versuche mit $s_n = 60$ Erfolgen, so wäre der Schätzwert für ϑ unverändert 0.6. Die realisierten approximativen Konfidenzintervalle sind aber wesentlich enger — wir erhalten $[0.502, 0.698]$ mit Methode 1 und $[0.504, 0.696]$ mit Methode 2.

$\diamond$

Zwischen Konfidenzbereichen und Tests besteht ganz allgemein ein grundlegender Zusammenhang; die beiden Konzepte sind dual zueinander und lassen sich gegenseitig ineinander überführen. Wir erklären hier diese Beziehung und verzichten anschliessend auf eine weitere Vertiefung von Konfidenzbereichen.

Bemerkung. Ob man lieber Tests oder Konfidenzbereiche ausführlich behandelt, ist Geschmackssache. [Williams] plädiert beispielsweise sehr stark für Konfidenzbereiche. $\diamond$

Sei zuerst $C(X_1, \ldots, X_n)$ ein Konfidenzbereich für ϑ zum Niveau $1-\alpha$. Um die Hypothese $H_0 : \vartheta = \vartheta_0$ zu testen, definieren wir einen Test durch

$$I_{\{\vartheta_0 \notin C(X_1,\ldots,X_n)\}},$$

d.h. wir lehnen H_0 genau dann ab, wenn ϑ_0 nicht in $C(X_1, \ldots, X_n)$ liegt. Dann ist wegen $\Theta_0 = \{\vartheta_0\}$ für jedes $\vartheta \in \Theta_0$

$$P_\vartheta[\vartheta_0 \notin C(X_1, \ldots, X_n)] = 1 - P_{\vartheta_0}[C(X_1, \ldots, X_n) \ni \vartheta_0] \leq \alpha,$$

so dass der Test gerade α als Signifikanzniveau hat. Aus einem Konfidenzbereich für ϑ erhalten wir also eine ganze Familie von Tests, nämlich einen für jede einfache Hypothese $\Theta_0 = \{\vartheta_0\}$ mit einem $\vartheta_0 \in \Theta$.

Sei nun umgekehrt für jede einfache Hypothese $\Theta_0 = \{\vartheta_0\}$ ein Test zum Niveau α gegeben; für jedes ϑ_0 haben wir also einen kritischen Bereich K_{ϑ_0}, so dass $H_0 : \vartheta = \vartheta_0$ genau dann abgelehnt wird, wenn $(X_1, \ldots, X_n) \in K_{\vartheta_0}$ ist. Zudem gilt wegen Niveau α

$$P_{\vartheta_0}[(X_1, \ldots, X_n) \in K_{\vartheta_0}] \leq \alpha \qquad \text{für jedes } \vartheta_0 \in \Theta.$$

Nun definieren wir eine Teilmenge $C(X_1, \ldots, X_n)$ von Θ durch die Bedingung

$$\vartheta \in C(X_1, \ldots, X_n) \quad :\Longleftrightarrow \quad (X_1, \ldots, X_n) \notin K_\vartheta.$$

Dann ist das ein Konfidenzbereich zum Niveau $1 - \alpha$, denn für jedes $\vartheta \in \Theta$ gilt

$$P_\vartheta[C(X_1, \ldots, X_n) \ni \vartheta] = P_\vartheta[(X_1, \ldots, X_n) \notin K_\vartheta] = 1 - P_\vartheta[(X_1, \ldots, X_n) \in K_\vartheta] \geq 1 - \alpha.$$

10 Kombinatorik kurz und knapp

Diese Notizen geben einen kurzen Überblick über die (wenigen) Resultate aus der Kombinatorik, die man sich wirklich merken sollte. Es gibt wesentlich mehr Formeln; aber es ist einfacher und effizienter, sich die Formeln inklusive ihrer Herleitungen von hier zu merken und sich andere Resultate bei Bedarf direkt zu überlegen.

Betrachten wir also n verschiedene Objekte.

1) Auf wie viele Arten kann man diese n Objekte (z.B. nebeneinander) anordnen?

Diese Anzahl heisst die Anzahl der *Permutationen (ohne Wiederholung)* von n Elementen und ist

$$n\,! = n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1.$$

Das Argument für die Herleitung geht so: Das erste Objekt in der Anordnung kann man beliebig aus den insgesamt n wählen; es gibt dafür also n Möglichkeiten. Das zweite Objekt in der Anordnung kann man aus den noch verbleibenden $n-1$ Objekten wählen; es gibt also dafür noch $n-1$ Möglichkeiten, und jede davon kann man mit jeder der n Möglichkeiten für das erste Objekt kreuzen, so dass insgesamt $n\times(n-1)$ Möglichkeiten für die ersten zwei Plätze entstehen. Für das dritte Objekt hat man noch $n-2$ zur Auswahl, usw.; das letzte (n-te) Objekt kann nur noch auf 1 Art gewählt werden, weil nur noch eines da ist, und auf diese Art erhält man die Formel.

2) Auf wie viele Arten kann man k aus den n Objekten auswählen (mit $k \leq n$ und ohne Zurücklegen)?

Diese Anzahl heisst die Anzahl der *Kombinationen (ohne Wiederholung)* und ist gegeben durch

$$\binom{n}{k} = \frac{n\,!}{k\,!(n-k)\,!}.$$

Diese Formel wird sehr oft gebraucht. Ihre Herleitung geht wie folgt. Für die Auswahl des ersten Objektes stehen uns noch alle n zur Verfügung. Das zweite Objekt müssen wir

dann aus den restlichen $n-1$ auswählen, das dritte aus den verbleibenden $n-2$, usw., bis wir das k-te Objekt aus den restlichen $n-(k-1) = n-k+1$ Objekten auswählen müssen. Also können wir auf diesem Weg $n \times (n-1) \times \cdots \times (n-k+1) = \frac{n!}{(n-k)!}$ Sequenzen der Länge k bilden. Nun interessieren wir uns aber nicht für die Reihenfolge, in der die Objekte gezogen worden sind; also identifizieren wir alle Sequenzen, die durch reines Vertauschen der Reihenfolge auseinander entstehen bzw. durch reines Vertauschen ineinander überführt werden können. Nach 1) können die gezogenen k Objekte auf $k!$ Arten angeordnet werden; jede Menge von k Objekten erzeugt also die $(k!)$-fache Anzahl von Anordnungen, die wir miteinander identifizieren, und damit erhalten wir die gesuchte Anzahl, indem wir den obigen Wert noch durch $k!$ dividieren. Das liefert gerade das Ergebnis $\binom{n}{k}$.

Nun betrachten wir die n Objekte als Symbole, die wir beliebig oft benutzen dürfen; wir können die Objekte also wiederholt (mit Zurücklegen) ziehen.

3) Wie viele Sequenzen der Länge m kann man mit den n Symbolen bilden?

Diese Anzahl heisst die Anzahl der *Variationen (mit Wiederholung)* und ist gegeben durch

$$n^m.$$

Das ist ganz einfach herzuleiten. Für jeden der m Plätze in der Sequenz hat man jedes der n Symbole zur Verfügung (da man ja wiederholen darf), also jeweils n Möglichkeiten für jeden Platz, die man mit jeder Möglichkeit für die anderen Plätze kreuzen kann. Also gibt es $n \times n \times \cdots \times n = n^m$ Möglichkeiten, eine Sequenz zu bilden.

Literatur: [Bronstein et al., Abschnitt 16.1 über Kombinatorik].

11 Literatur

[Bronstein et al.] I. N. Bronstein, K. A. Semendjajew, G. Musiol, H. Mühlig, "Taschenbuch der Mathematik", 4. Auflage, Harri Deutsch (1999)

[Krengel] U. Krengel, "Einführung in die Wahrscheinlichkeitstheorie und Statistik", 8. Auflage, Vieweg (2005)

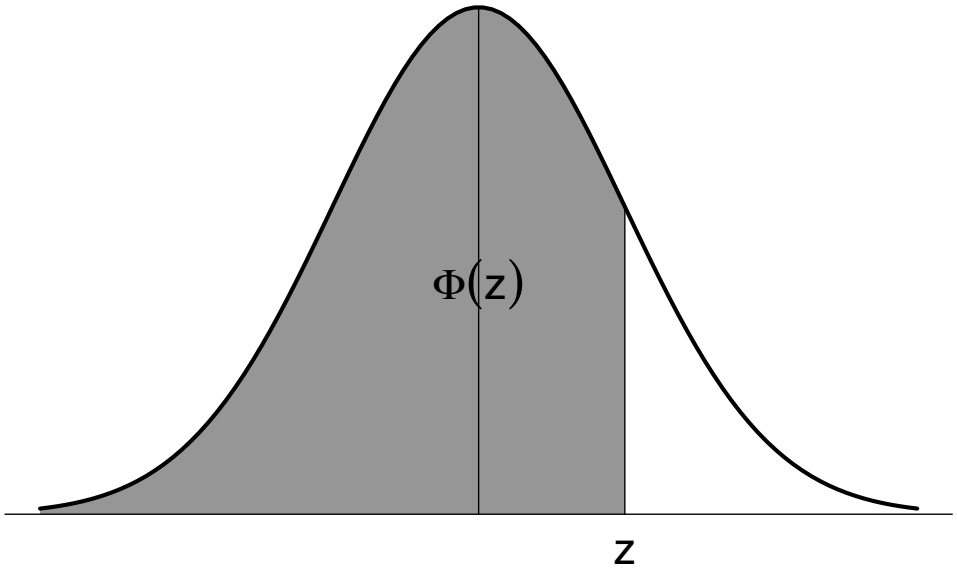
[Lehn, Wegmann] J. Lehn, H. Wegmann, "Einführung in die Statistik", 4. Auflage, Teubner (2004)

[Rice] J. A. Rice, "Mathematical Statistics and Data Analysis", second edition, Duxbury Press (1995)

[Stahel] W. A. Stahel, "Statistische Datenanalyse. Eine Einführung für Naturwissenschaftler", 2. Auflage, Vieweg (1999)

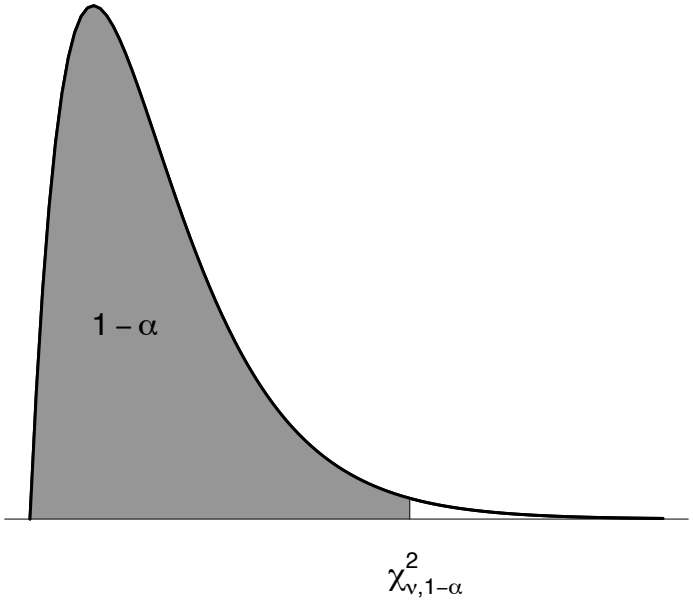
[Williams] D. Williams, "Weighing the Odds. A Course in Probability and Statistics", Cambridge University Press (2001)

12 Tabellen



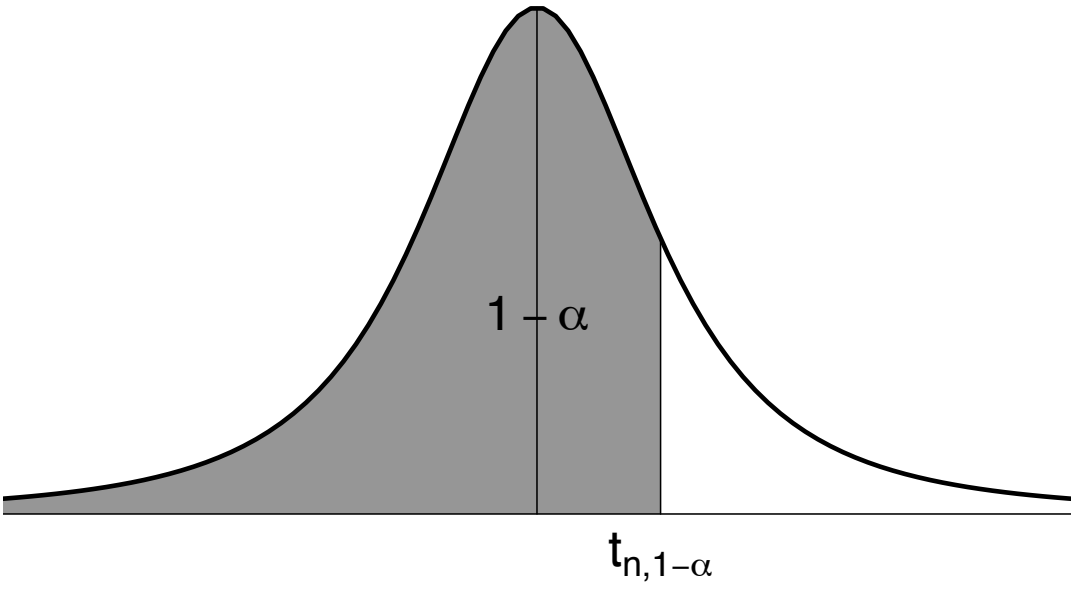
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767

Tabelle der Standard-Normalverteilungsfunktion $\Phi(z) = P[Z \leq z]$ mit $Z \sim \mathcal{N}(0,1)$



	$p = 0.90$	$p = 0.95$	$p = 0.975$	$p = 0.999$	$p = 0.9995$
$\nu = 1$	2.7055	3.8415	5.0239	10.8276	12.1157
$\nu = 2$	4.6052	5.9915	7.3778	13.8155	15.2018
$\nu = 3$	6.2514	7.8147	9.3484	16.2662	17.7300
$\nu = 4$	7.7794	9.4877	11.1433	18.4668	19.9974
$\nu = 5$	9.2364	11.0705	12.8325	20.5150	22.1053
$\nu = 6$	10.6446	12.5916	14.4494	22.4577	24.1028
$\nu = 7$	12.0170	14.0671	16.0128	24.3219	26.0178
$\nu = 8$	13.3616	15.5073	17.5345	26.1245	27.8680
$\nu = 9$	14.6837	16.9190	19.0228	27.8772	29.6658
$\nu = 10$	15.9872	18.3070	20.4832	29.5883	31.4198
$\nu = 11$	17.2750	19.6751	21.9200	31.2641	33.1366
$\nu = 12$	18.5493	21.0261	23.3367	32.9095	34.8213

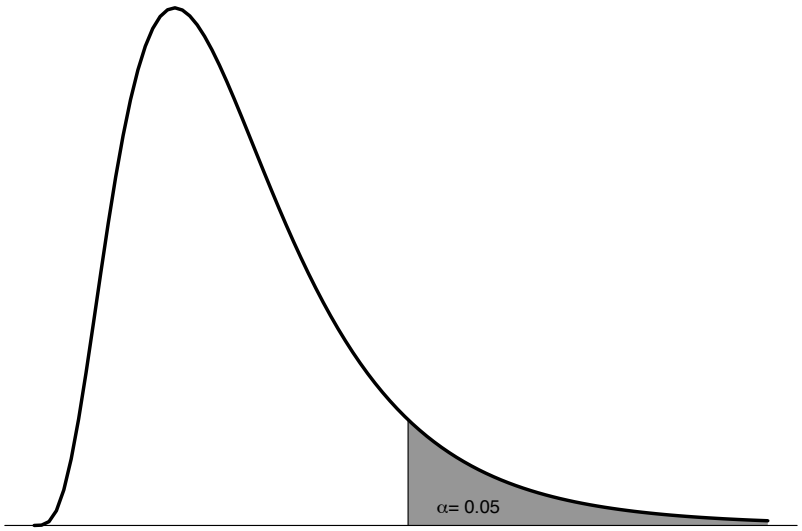
Ausgewählte Quantile $\chi^2_{\nu,1-\alpha}$ der Chiquadrat-Verteilung; in der Tabelle ist $p = 1 - \alpha$.


1 – α
t n,1–α

df	$t_{0.60}$	$t_{0.70}$	$t_{0.80}$	$t_{0.90}$	$t_{0.95}$	$t_{0.975}$	$t_{0.99}$	$t_{0.995}$
1	0.325	0.727	1.376	3.078	6.314	12.706	31.821	63.657
2	0.289	0.617	1.061	1.886	2.920	4.303	6.965	9.925
3	0.277	0.584	0.978	1.638	2.353	3.182	4.541	5.841
4	0.271	0.569	0.941	1.533	2.132	2.776	3.747	4.604
5	0.267	0.559	0.920	1.476	2.015	2.571	3.365	4.032
6	0.265	0.553	0.906	1.440	1.943	2.447	3.143	3.707
7	0.263	0.549	0.896	1.415	1.895	2.365	2.998	3.499
8	0.262	0.546	0.889	1.397	1.860	2.306	2.896	3.355
9	0.261	0.543	0.883	1.383	1.833	2.262	2.821	3.250
10	0.260	0.542	0.879	1.372	1.812	2.228	2.764	3.169
11	0.260	0.540	0.876	1.363	1.796	2.201	2.718	3.106
12	0.259	0.539	0.873	1.356	1.782	2.179	2.681	3.055
13	0.259	0.538	0.870	1.350	1.771	2.160	2.650	3.012
14	0.258	0.537	0.868	1.345	1.761	2.145	2.624	2.977
15	0.258	0.536	0.866	1.341	1.753	2.131	2.602	2.947
16	0.258	0.535	0.865	1.337	1.746	2.120	2.583	2.921
17	0.257	0.534	0.863	1.333	1.740	2.110	2.567	2.898
18	0.257	0.534	0.862	1.330	1.734	2.101	2.552	2.878
19	0.257	0.533	0.861	1.328	1.729	2.093	2.539	2.861
20	0.257	0.533	0.860	1.325	1.725	2.086	2.528	2.845
21	0.257	0.532	0.859	1.323	1.721	2.080	2.518	2.831
22	0.256	0.532	0.858	1.321	1.717	2.074	2.508	2.819
23	0.256	0.532	0.858	1.319	1.714	2.069	2.500	2.807
24	0.256	0.531	0.857	1.318	1.711	2.064	2.492	2.797
25	0.256	0.531	0.856	1.316	1.708	2.060	2.485	2.787
26	0.256	0.531	0.856	1.315	1.706	2.056	2.479	2.779
27	0.256	0.531	0.855	1.314	1.703	2.052	2.473	2.771
28	0.256	0.530	0.855	1.313	1.701	2.048	2.467	2.763
29	0.256	0.530	0.854	1.311	1.699	2.045	2.462	2.756
30	0.256	0.530	0.854	1.310	1.697	2.042	2.457	2.750
31	0.255	0.530	0.853	1.309	1.696	2.040	2.452	2.744
32	0.255	0.530	0.853	1.309	1.694	2.037	2.449	2.738
33	0.255	0.530	0.853	1.308	1.693	2.035	2.445	2.733
34	0.255	0.529	0.852	1.307	1.691	2.032	2.441	2.728
35	0.255	0.529	0.852	1.306	1.690	2.030	2.438	2.724
40	0.255	0.529	0.851	1.303	1.684	2.021	2.423	2.704
60	0.254	0.527	0.848	1.296	1.671	2.000	2.390	2.660
120	0.254	0.526	0.845	1.289	1.658	1.980	2.358	2.617
∞	0.253	0.524	0.842	1.282	1.645	1.960	2.326	2.576

Ausgewählte Quantile $t_{n,1-\alpha}$ der t-Verteilung; in der Tabelle ist $n = df$.

Für $df = \infty$ erhält man die Quantile $z_{1-\alpha}$ der Standard-Normalverteilung.



ν_2 \ ν_1 =	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	242.98	243.91	244.69	245.36	245.95
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.40	19.41	19.42	19.42	19.43
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74	8.73	8.71	8.70
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91	5.89	5.87	5.86
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70	4.68	4.66	4.64	4.62
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.98	3.96	3.94
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57	3.55	3.53	3.51
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28	3.26	3.24	3.22
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07	3.05	3.03	3.01
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91	2.89	2.86	2.85
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79	2.76	2.74	2.72
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69	2.66	2.64	2.62
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60	2.58	2.55	2.53
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53	2.51	2.48	2.46
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48	2.45	2.42	2.40
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46	2.42	2.40	2.37	2.35
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.41	2.38	2.35	2.33	2.31
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.31	2.29	2.27
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.34	2.31	2.28	2.26	2.23
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.31	2.28	2.25	2.22	2.20
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25	2.22	2.20	2.18
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.26	2.23	2.20	2.17	2.15

Kurztabelle der kritischen Werte zum 5%-Niveau der F-Verteilungen mit ν_1 Freiheitsgraden im Zähler und ν_2 Freiheitsgraden im Nenner. In der Tabelle ist $\nu_1 = n_1$, $\nu_2 = n_2$ und $1 - \alpha = 0.95$.

$n\backslash k$	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1	1														
2	1	2	1													
3	1	3	3	1												
4	1	4	6	4	1											
5	1	5	10	10	5	1										
6	1	6	15	20	15	6	1									
7	1	7	21	35	35	21	7	1								
8	1	8	28	56	70	56	28	8	1							
9	1	9	36	84	126	126	84	36	9	1						
10	1	10	45	120	210	252	210	120	45	10	1					
11	1	11	55	165	330	462	462	330	165	55	11	1				
12	1	12	66	220	495	792	924	792	495	220	66	12	1			
13	1	13	78	286	715	1287	1716	1716	1287	715	286	78	13	1		
14	1	14	91	364	1001	2002	3003	3432	3003	2002	1001	364	91	14	1	
15	1	15	105	455	1365	3003	5005	6435	6435	5005	3003	1365	455	105	15	1

Tabelle der Binomialkoeffizienten $\binom{n}{k}$ bis $n = 15$

Spezielle unbestimmte Integrale (die Integrationskonstanten sind weggelassen)**:**

$\int k dx = kx, \quad k$ konstant

$\int x^s dx = \frac{1}{s+1} x^{s+1}, \quad s \neq -1$

$\int \frac{1}{x} dx = \log|x|, \quad x \neq 0$

$\int (ax+b)^s dx = \frac{1}{a(s+1)} (ax+b)^{s+1}, \quad s \neq -1$

$\int \frac{1}{ax+b} dx = \frac{1}{a} \log|ax+b|$

$\int \frac{x}{ax+b} dx = \frac{x}{a} - \frac{b}{a^2} \log(ax+b)$

$\int \frac{x}{(a+bx)^2} dx = \frac{b}{a^2 \cdot (ax+b)} + \frac{1}{a^2} \log(ax+b)$

$\int \frac{x}{(ax+b)^n} dx = \frac{1}{a^2} \left(\frac{-1}{(n-2) \cdot (ax+b)^{n-2}} + \frac{b}{(n-1) \cdot (ax+b)^{n-1}} \right), \quad n \neq 1, \, n \neq 2$

$\int \frac{1}{x^2+a^2} dx = \frac{1}{a} \arctan \frac{x}{a}$

$\int \frac{1}{x^2-a^2} dx = \frac{1}{2a} \log \left| \frac{x-a}{x+a} \right|$

$\int e^{cx} dx = \frac{1}{c} e^{cx}$

$\int a^{cx} dx = \frac{1}{c \cdot \log a} a^{cx}$

$\int x e^{cx} dx = \frac{e^{cx}}{c^2} (cx-1)$

$\int x^2 e^{cx} dx = e^{cx} \left(\frac{x^2}{c} - \frac{2x}{c^2} + \frac{2}{c^3} \right)$

$\int \log|x| dx = x(\log|x| - 1)$

$\int \log_a|x| dx = x(\log_a|x| - \log_a e)$

$\int x^s \log x dx = \frac{x^{s+1}}{s+1} \left(\log x - \frac{1}{s+1} \right), \quad s \neq -1, \; x > 0$

$\int \frac{1}{x} \log x dx = \frac{1}{2} (\log x)^2, \quad x > 0$

$\int \sin x dx = -\cos x$

$\int \cos x dx = \sin x$

$\int \tan x dx = -\log|\cos x|$

$\int \cot x dx = \log|\sin x|$

$\int \sin(ax+b) dx = -\frac{1}{a} \cos(ax+b)$

$\int \cos(ax+b) dx = \frac{1}{a} \sin(ax+b)$

$\int \frac{1}{\sin x} dx = \log \left| \tan \frac{x}{2} \right|$

$\frac{1}{\cos x} dx = \log \left| \tan \left(\frac{x}{2} + \frac{\pi}{4} \right) \right|$

$\int \sin^2 x dx = \frac{1}{2} (x - \sin x \cos x)$

$\int \cos^2 x dx = \frac{1}{2} (x + \sin x \cos x)$

$\int \tan^2 x dx = \tan x - x$

$\int \cot^2 x dx = -\cot x - x$

Anmerkung: Mit $\log$ **ist der natürliche Logarithmus gemeint!**

13 Index