R. Grübel

Universität Hannover
Institut für Mathematische Stochastik

STOCHASTIK I

Sommersemester 2006

Dieses Skript enthält (in geringfügigem Umfang) Material, das in der Vorlesung selbst nicht besprochen wurde; in 'besonders schweren Fällen' ist der entsprechende Passus mit einem '*' gekennzeichnet. Außerdem fehlen natürlich (in größerem Umfang) Illustrationen, Beispiele und Erläuterungen, die in der Vorlesung ad hoc gegeben wurden.

1. Grundbegriffe

Stochastik, ein moderner Sammelbegriff für die Gebiete Wahrscheinlichkeitstheorie und mathematische Statistik, ist die

Mathematik des Zufalls.

Typische Situationen, bei denen der Zufall in der einen oder anderen Form eine Rolle spielt, finden wir

- bei Glücksspielen (Würfelwurf, Kartenmischen),
- in der Physik (statistische Mechanik, Quantenmechanik),
- in den Ingenieurwissenschaften (Signalverarbeitung),
- in den Wirtschaftswissenschaften (Modellierung von Aktienkursen),
- in der Medizin (Vergleich von Medikamenten),
- im Operations Research (Bediennungssysteme), sowie
- in der Informatik (Analyse von Algorithmen, randomisierte Verfahren).

In diesem ersten Abschnitt geht es um einige fundamentale Grundbegriffe, die im gesamten Verlauf der Vorlesung routinemäßig verwendet werden.

1.1 Ein mathematisches Modell für Zufallsexperimente. Bei Zufallsexperimenten ist das Ergebnis nicht durch die Randbedingungen des Experiments festgelegt. Der *Ergebnisraum* Ω ist eine Menge, die die möglichen Ergebnisse (Resultate) des Experiments enthält, *Ereignisse* werden durch Teilmengen von Ω beschrieben. Aussagen über das Ergebnis werden dabei in Teilmengen des Ergebnisraumes übersetzt: eine Aussage wird zu der Menge aller $\omega \in \Omega$, für die diese Aussage richtig ist.

Beispiel 1.1 Beim Wurf eines Würfels ist $\Omega := \{1,2,3,4,5,6\}$ eine geeignete Ergebnismenge; das Ereignis 'eine gerade Zahl erscheint' wird repräsentiert durch (ist) $A = \{2,4,6\}$. Wirft man einen Würfel zweimal, so bietet sich

$$\Omega_2 := \{(i,j) : i,j \in \Omega\} \quad \Big(= \Omega \times \Omega = \Omega^2\Big)$$

an, wobei das Paar (i,j) dafür steht, dass i im ersten und j im zweiten Wurf erscheint. Wirft man zwei Würfel gleichzeitig (und kann man diese nicht unterscheiden), so liegt

$$\tilde{\Omega}_2 := \{(i,j) \in \Omega_2 . i \leq j\}$$

nahe (die Einzelergebnisse sind aufsteigend geordnet). Das Ereignis 'Augensumme 8' wird zu $A = \{(2,6),(3,5),(4,4),(5,3),(6,2)\}$ bei Ergebnisraum Ω_2 und zu $\tilde{A} = \{(2,6),(3,5),(4,4)\}$ bei Ergebnisraum $\tilde{\Omega}_2$. $\lhd$

Ein Ereignis A mit exakt einem Element, also $\mathrm{A} = \{\omega\}$ mit einem $\omega \in \Omega$, nennt man ein *Elementarereignis*. Ergebnisse sind also Elemente von Ω, Ereignisse Teilmengen von Ω. Kombinationen von Ereignissen können durch mengentheoretische Operationen beschrieben werden:

$$\begin{aligned} A \cap B: &\quad A \text{ und } B \text{ treten beide ein,} \\ A \cup B: &\quad A \text{ oder } B \text{ (oder beide) tritt (treten) ein,} \\ A^{\mathrm{c}}: &\quad A \text{ tritt nicht ein.} \end{aligned}$$

Beim Würfelwurf wird beispielsweise das Ereignis 'es erscheint keine gerade Zahl' beschrieben durch $\{2,4,5\}^{\mathrm{c}} = \{1,3,5\}$.

Beispiel 1.2 (Kombinationen von mehr als zwei Ereignissen)
(a) 'Genau eines der Ereignisse A, B, C tritt ein' wird beschrieben durch

$$A \cap B^{\mathrm{c}} \cap C^{\mathrm{c}} + A^{\mathrm{c}} \cap B \cap C^{\mathrm{c}} + A^{\mathrm{c}} \cap B^{\mathrm{c}} \cap C.$$

Hierbei steht $A + B$ für $A \cup B$ bei disjunkten Mengen A, B.

(b) Es sei $A_1, A_2, A_3, \ldots$ eine Folge von Ereignissen. Dann wird das Ereignis 'unendlich viele der A_i's treten ein' repräsentiert durch den *Limes superior* der Mengenfolge,

$$\limsup_{n\to\infty} A_n \;:=\; \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m.$$

Klar: $\cup_{m=n}^{\infty} A_m$ steht für 'mindestens eines der Ereignisse mit Index $\geq n$ tritt ein', und es gilt

$$\begin{aligned} \omega \in \limsup_{n\to\infty} A_n \quad &\Longleftrightarrow \quad \forall n \in \mathbb{N}\ \exists m \geq n:\ \omega \in A_m \\ &\Longleftrightarrow \quad \#\{n \in \mathbb{N} : \omega \in A_n\} = \infty. \end{aligned}$$

◁

Die Menge der Ereignisse (eine Menge von Mengen!) in einem Zufallsexperiment bildet ein Mengensystem $\mathcal{A}$ über Ω, also eine Teilmenge der Potenzmenge $\mathcal{P}(\Omega)$ von Ω. Bei endlichem oder abzählbar unendlichem Ergebnisraum können wir problemlos $\mathcal{A} = \mathcal{P}(\Omega)$ voraussetzen (jede Zusammenfassung von Ergebnissen ist ein Ereignis), bei überabzählbarem Ω geht dies in vielen wichtigen Fällen nicht (wir werden dies später präzisieren). Die obigen Beispiele für Kombinationen von Ereignissen führen auf gewisse Mindestvoraussetzungen an das System $\mathcal{A}$ und damit zur folgenden Definition.

Definition 1.3 $\mathcal{A} \subset \mathcal{P}(\Omega)$ heißt eine *σ-Algebra* über Ω, wenn gilt:

(i) $\Omega \in \mathcal{A}$, (ii) $A \in \mathcal{A} \implies A^{\mathrm{c}} \in \mathcal{A}$,

(iii) $A_1, A_2, \ldots \in \mathcal{A} \implies \bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$.

In Worten: Ein Mengensystem über Ω ist eine σ-Algebra, wenn es die Grundmenge (also den Ergebnisraum) enthält und stabil ist gegenüber den Operationen 'Komplement' und 'abzählbare Vereinigung'.

Was ist nun 'Wahrscheinlichkeit'? Strenggenommen ist dies keine mathematische Frage (analog zu: Was ist eine Gerade?, was ist eine Menge?) Als mathematischer Gegenstand ist Wahrscheinlichkeit eine Funktion, die Ereignissen Zahlen zwischen 0 und 1 zuordnet und dabei gewissen Axiomen genügt. Diese Axiome (Forderungen) werden durch den umgangssprachlichen Wahrscheinlichkeitsbegriff motiviert. Zur Erläuterung betrachten wir die Aussage 'das Ereignis A hat Wahrscheinlichkeit p' (z.B.: 'beim Wurf eines fairen Würfels erscheint mit Wahrscheinlichkeit $1/2$ eine gerade Zahl'). Es gibt zwei hauptsächliche Interpretationen:

(F) Die 'Häufigkeitsauffassung', deren Anhänger auch Frequentisten genannt werden. Es sei $N_n(A)$ die Häufigkeit des Auftretens von A bei n Wiederholungen des Zufallsexperiments; $\frac{1}{n}N_n(A)$ ist die *relative Häufigkeit* von A. Bei großem n würde man erwarten, dass die relative Häufigkeit von A in der Nähe von p liegt (ungefähr die Hälfte der Würfelwürfe sollte eine gerade Zahl liefern).

(S) Die 'Glaubens- oder Plausibilitätsauffassung', deren Anhänger man gelegentlich als Subjektivisten bezeichnet. Der Wert p gibt auf einer Skala von 0 bis 1 die 'Stärke meines Glaubens' an das Eintreten von A wieder. Dies kann über Wetten formalisiert werden und ist im Gegensatz zu (a) auch bei nichtwiederholbaren Experimenten anwendbar (aber eben subjektiv).

Diese Auffassungen sind natürlich nicht disjunkt. Für relative Häufigkeiten gelten die Regeln

$$\frac{1}{n}\,N_n(\Omega) = 1, \quad \frac{1}{n}\,N_n(A) \geq 0 \quad \text{für alle } A \in \mathcal{A},$$

sowie für alle paarweise disjunkten $A_1, \ldots, A_k \in \mathcal{A}$

$$\frac{1}{n}\,N_n(A_1 + \ldots + A_k) \;=\; \frac{1}{n}\,N_n(A_n) + \ldots + \frac{1}{n}\,N_n(A_k).$$

Insgesamt motiviert dies das folgende mathematische Modell für Zufallsexperimente:

Definition 1.4 (Die Kolmogorov-Axiome) Ein *Wahrscheinlichkeitsraum* ist ein Tripel $(\Omega, \mathcal{A}, P)$, bestehend aus einer nichtleeren Menge Ω (dem Ergebnisraum), einer σ-Algebra $\mathcal{A}$ über Ω (dem Ereignissystem), und einer Abbildung $P : \mathcal{A} \to \mathbb{R}$ mit den Eigenschaften

(i) $P(\Omega) = 1$, (ii) $P(A) \geq 0$ für alle $A \in \mathcal{A}$,

(iii) $P\Big(\sum_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} P(A_i)$ für alle paarweise disjunkten $A_1, A_2, \ldots \in \mathcal{A}$.

Eine Abbildung mit diesen Eigenschaften heißt *Wahrscheinlichkeitsmaß*, Eigenschaft (iii) nennt man die *σ-Additivität*.

Beispiel 1.5 (a) Ist Ω eine endliche und nicht-leere Menge, so wird durch

$$P(A) := \frac{\#A}{\#\Omega} \qquad \text{für alle } A \subset \Omega$$

ein Wahrscheinlichkeitsmaß auf $(\Omega, \mathcal{P}(\Omega))$ definiert. Man nennt $(\Omega, \mathcal{A}, P)$ mit $\mathcal{A} = \mathcal{P}(\Omega)$ das *Laplace-Experiment* über Ω. Solche Modelle werden häufig durch Symmetrieüberlegungen nahegelegt. Beim Wurf eines fairen (d.h. symmetrischen) Würfels ergibt sich damit als Wahrscheinlichkeit dafür, dass eine gerade Zahl geworfen wird,

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

(Anzahl der günstigen Fälle dividiert durch die Anzahl der möglichen Fälle, eine vielleicht schon aus dem Schulunterricht bekannte Regel). Ob für ein vorgegebenes Zufallsexperiment ein Laplace-Experiment über einer bestimmten Menge das korrekte Modell ist, ist keine (rein) mathematische Frage. Bei den beiden Ergebnisräumen zum zweimaligen Würfelwurf und zum gleichzeitigen Wurf zweier Würfel würde man unterschiedliche Wahrscheinlichkeiten für die Augensumme 8 bekommen. 'Außermathematische' Überlegungen zeigen, dass Würfel (wie allgemein makroskopische Objekte) unterscheidbar sind und somit 5/36 die richtige Antwort ist; bei der Elementarteilchenphysik können durchaus andere Modelle korrekt sein (in dem Sinne, dass sie die physikalische Realität richtig wiedergeben).

(b) Ein deterministisches Experiment, bei dem nur ein einziges Ergebnis ω_0 möglich ist, kann als degeneriertes Zufallsexperiment $(\Omega, \mathcal{A}, \delta_{\omega_0})$ betrachtet werden. Hierbei ist Ω irgendeine Menge, die ω_0 enthält, $\mathcal{A}$ eine σ-Algebra über Ω und δ_{ω_0} das *Dirac-Maß* oder auch *Einpunktmaß* in ω_0:

$$\delta_{\omega_0}(A) = \begin{cases} 1, & \omega_0 \in A, \\ 0, & \omega_0 \notin A. \end{cases}$$

Man macht sich leicht klar, dass δ_{ω_0} ein Wahrscheinlichkeitsmaß ist. ◁

Im folgenden Satz sind einige erste Folgerungen aus den Axiomen zusammengefasst.

SATZ 1.6 *Es sei* $(\Omega, \mathcal{A}, P)$ *ein Wahrscheinlichkeitsraum. Dann gilt:*

(a) $P(\emptyset) = 0$, $P(A) \leq 1$ *für alle* $A \in \mathcal{A}$,

(b) $P(A^c) = 1 - P(A)$ *für alle* $A \in \mathcal{A}$,

(c) (endliche Additivität) $P(A_1 \cup \ldots \cup A_k) = P(A_1) + \ldots + P(A_k)$ *für alle paarweise disjunkten* $A_1, \ldots, A_k \in \mathcal{A}$,

(d) (Monotonie) $A \subset B \Rightarrow P(A) \leq P(B)$ *für alle* $A, B \in \mathcal{A}$,

(e) (Boolesche Ungleichung) $P(A_1 \cup \ldots \cup A_k) \leq P(A_1) + \ldots + P(A_k)$ *für alle (nicht notwendigerweise disjunkten)* $A_1, \ldots, A_k \in \mathcal{A}$,

(f) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ *für alle* $A, B \in \mathcal{A}$,

(g) (Formel von Poincaré, auch: Einschluss-Ausschluss-Formel oder Siebformel)

$$P(A_1 \cup \ldots \cup A_k) = \sum_{\emptyset \neq H \subset \{1,\ldots,k\}} (-1)^{\#H-1} P\Big(\bigcap_{i \in H} A_i\Big).$$

BEWEIS: Der Nachweis, dass die beteiligten Mengenkombinationen nicht aus der σ-Algebra herausführen, ist Gegenstand einer Übungsaufgabe; beispielsweise gilt $\emptyset \in \mathcal{A}$ wegen $\Omega \in \mathcal{A}$ und $\emptyset = \Omega^c$.

(a) Verwendet man die σ-Additivität von P mit $A_1 = A_2 = \ldots = \emptyset$, so folgt $P(\emptyset) = P(\emptyset) + P(\emptyset) + \ldots$, also $P(\emptyset) = 0$. Die Aussage $P(A) \leq 1$ folgt aus $P(\Omega) = 1$ und der Monotonie (Teil (d)).

(c) Setze $A_{k+1} = A_{k+2} = \ldots = \emptyset$, verwende die σ-Additivität und $P(\emptyset) = 0$.

(b) $A \cup A^c = \Omega$, $A \cap A^c = \emptyset$; verwende nun die endliche Additivität.

(d) Es gilt $B = A + B \cap A^c$, also $P(B) = P(A) + P(B \cap A^c) \geq P(A)$, da $P(B \cap A^c) \geq 0$.

(e) Im Falle $k = 2$ folgt die Aussage aus Teil (f) und $P(A \cap B) \geq 0$. Angenommen, die Aussage ist für ein $k \geq 2$ richtig. Dann folgt

$$P\big((A_1 \cup \ldots \cup A_k) \cup A_{k+1}\big) \;\leq\; P(A_1 \cup \ldots \cup A_k) + P(A_{k+1}),$$

denn für zwei Ereignisse gilt die Formel, also

$$P\big((A_1 \cup \ldots \cup A_k) \cup A_{k+1}\big) \;\leq\; \big(P(A_1) + \ldots + P(A_k)\big) \;+\; P(A_{k+1}),$$

d.h. die Aussage gilt dann auch für $k+1$. Vollständige Induktion liefert nun die gewünschte Aussage.

(f) $A = A \cap B + A \cap B^c$, also ergibt der bereits bewiesene Teil (c) $P(A \cap B^c) = P(A) - P(A \cap B)$. Weiter gilt $A \cup B = B + A \cap B^c$, also

$$P(A \cup B) \;=\; P(B) + P(A \cap B^c) \;=\; P(B) + P(A) - P(A \cap B).$$

(g) Im Falle $k = 2$ erhält man (f). Induktionsschritt: Übungsaufgabe. □

Warum wird in den Kolmogorov-Axiomen die σ-Additivität anstelle beispielsweise der (schwächeren) endlichen Additivität gefordert? Man sieht leicht, dass letztere bereits aus

$$P(A \cup B) = P(A) + P(B) \qquad \text{für alle disjunkten } A, B \in \mathcal{A}$$

folgt. Das folgende Resultat zeigt, dass man σ-Additivität als Stetigkeitseigenschaft interpretieren kann. Wir nennen eine Folge $(A_n)_{n\in\mathbb{N}}$ von Teilmengen von Ω *isoton*, wenn $A_n \subset A_{n+1}$ für alle $n \in \mathbb{N}$ gilt, *antiton* im Falle $A_n \supset A_{n+1}$ für alle $n \in \mathbb{N}$. Wir schreiben beispielsweise $A_n \downarrow A$, wenn $(A_n)_{n\in\mathbb{N}}$ eine antitone Mengenfolge ist mit der Eigenschaft $\bigcap_{n=1}^{\infty} A_n = A$.

SATZ 1.7 *Es seien $\Omega \neq \emptyset$, $\mathcal{A}$ eine σ-Algebra auf Ω und $P : \mathcal{A} \to \mathbb{R}$ eine Abbildung mit den Eigenschaften*
(i) $P(\Omega) = 1$, (ii) $P(A) \geq 0$ *für alle* $A \in \mathcal{A}$,
(iii) $P(A \cup B) = P(A) + P(B)$ *für alle* $A, B \in \mathcal{A}$ *mit* $A \cap B = \emptyset$.
Dann sind äquivalent:

(a) *P ist σ-additiv (also ein Wahrscheinlichkeitsmaß),*

(b) *P ist stetig von unten, d.h. für jede isotone Folge $A_1, A_2, \ldots$ von Ereignissen gilt*

$$\lim_{n\to\infty} P(A_n) = P\Big(\bigcup_{n=1}^{\infty} A_n\Big),$$

(c) *P ist stetig von oben, d.h. für jede antitone Folge $A_1, A_2, \ldots$ von Ereignissen gilt*

$$\lim_{n\to\infty} P(A_n) = P\Big(\bigcap_{n=1}^{\infty} A_n\Big),$$

(d) *P ist stetig in $\emptyset$, d.h. für jede Folge $(A_n)_{n\in\mathbb{N}}$ von Ereignissen mit der Eigenschaft $A_n \downarrow \emptyset$ gilt*

$$\lim_{n\to\infty} P(A_n) = 0.$$

BEWEIS: $(a) \Rightarrow (b)$. Es sei $B_1 := A_1, B_n := A_n \cap A_{n-1}^{\mathrm{c}}$ für alle $n > 1$. Klar: $B_n \in \mathcal{A}$ für alle $n \in \mathbb{N}$, $(B_n)_{n\in\mathbb{N}}$ paarweise disjunkt, $A_n = B_1 + \ldots + B_n$ für alle $n \in \mathbb{N}$, $\bigcup_{n=1}^{\infty} A_n = \sum_{n=1}^{\infty} B_n$. Die σ-Additivität von P liefert

$$\begin{aligned} P\Big(\bigcup_{n=1}^{\infty} A_n\Big) &= P\Big(\sum_{n=1}^{\infty} B_n\Big) = \sum_{n=1}^{\infty} P(B_n) \\ &= \lim_{n\to\infty} \sum_{m=1}^{n} P(B_m) = \lim_{n\to\infty} P\Big(\sum_{m=1}^{n} B_m\Big) \\ &= \lim_{n\to\infty} P(A_n). \end{aligned}$$

(b) ⇒ (c): Über Komplementbildung: Ist $A_n \downarrow$, so ist $A_n^{\mathrm{c}} \uparrow$ und man erhält

$$\begin{aligned} P\Big(\bigcap_{n=1}^{\infty} A_n\Big) &= 1 - P\Big(\bigcup_{n=1}^{\infty} A_n^{\mathrm{c}}\Big) \\ &= 1 - \lim_{n\to\infty} P(A_n^{\mathrm{c}}) \\ &= 1 - \lim_{n\to\infty} \big(1 - P(A_n)\big) \\ &= \lim_{n\to\infty} P(A_n). \end{aligned}$$

(c) ⇒ (d): Trivial.
(d) ⇒ (a): Sind $A_1, A_2, \ldots$ disjunkt, so gilt $B_n \downarrow \emptyset$ für $B_n := \sum_{k=n+1}^{\infty} A_k$, also folgt unter Verwendung der endlichen Additivität

$$\begin{aligned} P\Big(\bigcup_{n=1}^{\infty} A_n\Big) &= P\Big(\sum_{k=1}^{n} A_k + B_n\Big) \\ &= \sum_{k=1}^{n} P(A_k) + P(B_n). \end{aligned}$$

Wegen $P(B_n) \to 0$ konvergiert die Reihe und ist gleich $P\big(\bigcup_{k=1}^{\infty} A_k\big)$. □

Wir werden später noch einmal auf die verschiedenen Varianten der Additivität zurückkommen und bemerken hier nur, dass als Ersatz für die σ-Additivität die endliche Additivität zu schwach für eine befriedigende mathematische Theorie ist.

1.2 Bedingte Wahrscheinlichkeiten und Unabhängigkeit. Es seien A und B Ereignisse in einem Zufallsexperiment, das durch einen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ beschrieben wird. Was ist die Wahrscheinlichkeit von B unter der Bedingung, dass A eintritt? Bei n Wiederholungen tritt A $N_n(A)$-mal ein, unter diesen ist $N_n(A \cap B)$ die (absolute) Häufigkeit von B. Für die relative Häufigkeit von B unter den Experimenten, die A liefern, gilt

$$\frac{N_n(A \cap B)}{N_n(A)} = \frac{\frac{1}{n} N_n(A \cap B)}{\frac{1}{n} N_n(A)}.$$

Durch den frequentistischen Wahrscheinlichkeitsbegriff wird somit die folgende Definition motiviert.

DEFINITION 1.8 Es sei A ein Ereignis mit $P(A) > 0$. Die *bedingte Wahrscheinlichkeit* eines Ereignisses B unter A wird definiert durch

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

Man sieht leicht, dass dann $B \mapsto P(B|A)$ ein Wahrscheinlichkeitsmaß ist, d.h. $(\Omega, \mathcal{A}, P(\,\cdot\,|A))$ ist ein Wahrscheinlichkeitsraum. Er repräsentiert das gegenüber $(\Omega, \mathcal{A}, P)$ dahingehend veränderte Experiment, dass das Eintreten von A bekannt ist.

SATZ 1.9 (a) (Die Multiplikationsregel) *Es seien $A_1, \ldots, A_n$ Ereignisse mit $P(A_1 \cap \ldots \cap A_n) > 0$. Dann gilt*

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

(b) (Das Gesetz von der totalen Wahrscheinlichkeit) *Es sei $A_1, \ldots, A_n$ eine Ereignispartition von Ω, d.h.*

$$A_1, \ldots, A_n \in \mathcal{A}, \quad \bigcup_{i=1}^{n} A_i = \Omega, \quad A_i \cap A_j = \emptyset \text{ für } i \neq j.$$

Dann gilt für alle $B \in \mathcal{A}$

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

(wir lassen hierbei $P(A_i) = 0$ zu und setzen dann $P(B|A_i)P(A_i) = 0$).

(c) (Die Formel von Bayes) *Es seien $A_1, \ldots, A_n$, B wie in (b) und es gelte $P(B) > 0$. Dann folgt*

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{k=1}^{n} P(B|A_k)P(A_k)}.$$

BEWEIS: Verwende $B = \sum_{i=1}^{n} B \cap A_i$ und die Additivität von P bei (b). Alles andere folgt unmittelbar aus den Definitionen. □

Beispiel 1.10 Ein bestimmter medizinischer Test ist zu 95% effektiv beim Erkennen einer bestimmten Krankheit, liefert allerdings bei 1% der gesunden Personen einen 'falschen Alarm'. Angenommen, 0.5% der Bevölkerung leiden unter dieser Krankheit — mit welcher Wahrscheinlichkeit hat jemand die Krankheit, wenn der Test dies behauptet? Wir schreiben A für das Ereignis, dass die getestete Person die Krankheit hat, B für das Ereignis, dass der Test das Vorliegen der Krankheit anzeigt, und übersetzen die obigen Annahmen in

$$P(A) = 0.005, \quad P(B|A) = 0.95, \quad P(B|A^c) = 0.01.$$

Mit der Bayes-Formel ergibt sich dann

$$\begin{aligned} P(A|B) &= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} \\ &= \frac{0.95 \cdot 0.005}{0.95 \cdot 0.005 + 0.01 \cdot 0.995} \approx 0.323, \end{aligned}$$

ein zumindest auf den ersten Blick überraschend hoher Wert. Man beachte, dass der Übersetzung von Prozentzahlen in Wahrscheinlichkeiten bestimmte Annahmen über die Auswahl der Testperson etc. zugrundeliegen.

Es ist hier möglicherweise hilfreich (in dem Sinne, dass dieses Resultat dann weniger paradox wirkt — die mathematische Herleitung bleibt von solchen Verständnishilfen unberührt), wenn man mit einer hypothetischen Population arbeitet: Besteht diese aus 100 000 Personen, so müssten aufgrund der obigen Prozentzahlen 500 Personen krank, 99 500 gesund sein; unter den Kranken würden 475 vom Test als krank deklariert, von den Gesunden 995. Wählt man nun unter den insgesamt $475 + 995$ Personen mit 'positivem' Testresultat eine Person zufällig aus, so erhält man mit Wahrscheinlichkeit $475/(475 + 995) \approx 0.323$ eine kranke Person. ◁

Beispiel 1.10 zeigt auch, dass es nicht immer nötig bzw. sinnvoll ist, einen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ explizit anzugeben.

Einer der zentralen Begriffe der Stochastik ist der der (stochastischen) Unabhängigkeit. Die mathematische Definition soll das intuitive Konzept wiedergeben: B wird von A nicht beeinflusst, wenn sich die Wahrscheinlichkeit von B nicht durch die Information ändert, dass A eingetreten ist. Dies führt auf die Forderung $P(B|A) = P(B)$. Langweilige Fallunterscheidungen (ist $P(A)$ grösser als 0?) werden vermieden durch

Definition 1.11 Zwei Ereignisse A und B heißen *stochastisch unabhängig*, wenn $P(A \cap B) = P(A)P(B)$ gilt.

Bei mehr als zwei Ereignissen ist Vorsicht angesagt:

DEFINITION 1.12 Eine Familie $\{A_i : i \in I\}$ von Ereignissen heißt *paarweise unabhängig*, wenn gilt:

$$P(A_i \cap A_j) = P(A_i)P(A_j) \qquad \text{für alle } i, j \in I \text{ mit } i \neq j;$$

sie heißt *unabhängig*, wenn gilt:

$$P\Big(\bigcap_{i\in H} A_i\Big) \;=\; \prod_{i\in H} P(A_i) \quad \text{für jede endliche Teilmenge } H \text{ von } I.$$

BEISPIEL 1.13 Wir betrachten das Laplace-Experiment über

$$\Omega = \{(0,0),(0,1),(1,0),(1,1)\} \quad \Big(= \{0,1\}^2\Big).$$

Schreibt man '0' für das Resultat 'Kopf' und '1' für Wappen, so ist dieses Laplace-Experiment beispielsweise ein Modell für den zweimaligen Wurf einer fairen Münze. Es seien

$$\begin{aligned} A_1 &:= \{(0,0),(0,1)\} && \text{('Kopf' im ersten Wurf),}\\ A_2 &:= \{(0,0),(1,0)\} && \text{('Kopf' im zweiten Wurf),}\\ A_3 &:= \{(0,1),(1,0)\} && \text{(Resultate verschieden).} \end{aligned}$$

Man sieht leicht (die Durchschnitte sind jeweils einelementig)

$$P(A_1 \cap A_2) = \frac{1}{4} = \frac{1}{2}\cdot\frac{1}{2} = P(A_1)P(A_2),$$

und erhält analog

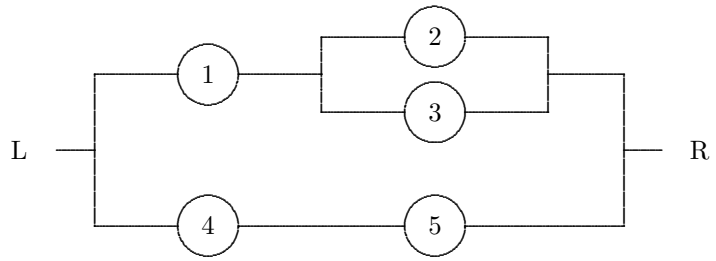
$$P(A_1 \cap A_3) = P(A_1)P(A_3), \quad P(A_2 \cap A_3) = P(A_2)P(A_3).$$

Die Familie $\{A_1, A_2, A_3\}$ ist also paarweise unabhängig. Es gilt jedoch

$$P(A_1 \cap A_2 \cap A_3) \;=\; P(\emptyset) \;=0\; \neq \; P(A_1)P(A_2)P(A_3),$$

die Familie ist also nicht unabhängig. Moral: paarweise Unabhängigkeit impliziert nicht die (volle) Unabhängigkeit. ◁

BEISPIEL 1.14 Eine typische Fragestellung der Angewandten Wahrscheinlichkeitsrechnung bezieht sich auf das Funktionieren von Netzwerken. Wir betrachten einen einfachen Fall, in dem ein System aus fünf wie folgt angeordneten Komponenten besteht:



Wir nehmen an, dass die Komponenten unabhängig voneinander und zwar jeweils mit Wahrscheinlichkeit p funktionieren. Das Gesamtsystem funktioniert, wenn es einen Pfad funktionierender Komponenten vom Eingang zum Ausgang gibt. Mit welcher Wahrscheinlichkeit funktioniert das Gesamtsystem?

Es sei A_i das Ereignis, dass Komponente i funktioniert, B das interessierende Ereignis. Dann gilt $B = B_1 \cup B_2$ mit

$$\begin{aligned} B_1 &:= A_4 \cap A_5 \qquad \text{(unterer Pfad passierbar)} \\ B_2 &:= A_1 \cap (A_2 \cup A_3) \qquad \text{(oberer Pfad passierbar).} \end{aligned}$$

Mit Hilfe der Unabhängigkeit und der Formel $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ aus Satz 1.6 erhalten wir

$$\begin{aligned} P(B_1) &= P(A_4)P(A_5) = p^2, \\ P(B_2) &= P((A_1 \cap A_2) \cup (A_1 \cap A_3)) \\ &= P(A_1 \cap A_2) + P(A_1 \cap A_3) - P(A_1 \cap A_2 \cap A_3) \\ &= 2p^2 - p^3, \\ P(B_1 \cap B_2) &= P(A_4 \cap A_5 \cap A_1 \cap A_2) + P(A_4 \cap A_5 \cap A_1 \cap A_3) \\ &\qquad - P(A_4 \cap A_5 \cap A_1 \cap A_2 \cap A_3) \\ &= 2p^4 - p^5 \end{aligned}$$

(man könnte auch 'B_1, B_2 unabhängig' verwenden — allerdings erfordert dies eine abstrakte Zusatzüberlegung), also insgesamt

$$\begin{aligned} P(B) &= P(B_1) + P(B_2) - P(B_1 \cap B_2) \\ &= p^2 + 2p^2 - p^3 - (2p^4 - p^5) \\ &= p^2(3 - p - 2p^2 + p^3). \end{aligned}$$

Man beachte, dass paarweise Unabhängigkeit hier nicht gereicht hätte. ◁

Beispiel 1.15* ('Simpson's paradox') Das Rechnen mit bedingten Wahrscheinlichkeiten kann gelegentlich in als paradox empfundenen Situationen eine einfache Lösung oder Erklärung liefern; siehe auch das in den Übungen besprochene 'Ziegenproblem'. Ein klassisches Beispiel für das, worum es uns hier geht, liefern die Zulassungszahlen einer amerikanischen Universität aus dem Jahr 1973: Von 1576 männlichen Bewerbern wurden etwa 58% angenommen, von 526 weiblichen Bewerbern nur etwa 46% (aus Zeitgründen betrachten wir nur einen Teil der Daten). Dies wurde damals als Beleg für die Diskriminierung von Frauen angesehen. Die Aufschüsselung nach Fächern sah wie folgt aus:

	Männer		*Frauen*	
Fach	# Bewerber	zugelassen	# Bewerber	zugelassen
1	825	511 (62%)	108	82 (89%)
2	560	352 (63%)	25	17 (68%)
3	191	53 (28%)	393	134 (34%)
Summe	1576	916 (58%)	526	240 (46%)

Berücksichtigt man also den Faktor 'Fach', so ergibt sich ein ganz anderes Bild — offensichtlich bewerben sich Frauen eher in Fächern mit einer höheren Ablehnungsquote.

Was hat dies mit bedingten Wahrscheinlichkeiten zu tun? Wie im Beispiel 1.10 werden Häufigkeiten und Wahrscheinlichkeiten dadurch in Zusammenhang gebracht, dass man die zufällige Auswahl einer Person aus der Grundpopulation der $1576 + 526$ Bewerber, also ein Laplace-Experiment über $\{1, 2, \ldots, 2102\}$ betrachtet. Es seien

S_k : die ausgewählte Person hat sich für Studiengang k beworben,

Z : die ausgewählte Person wird zugelassen,

F, M : die ausgewählte Person ist eine Frau bzw. ein Mann.

Es gilt dann beispielsweise $P(S_1|M) = \frac{825}{1576}$. Die oben eingeführten Rechenregeln liefern

$$P(Z|F) = \sum_{k=1}^{3} P(Z|F \cap S_k) P(S_k|F),$$

$$P(Z|M) = \sum_{k=1}^{3} P(Z|M \cap S_k) P(S_k|M).$$

Man landet also bei dem (ziemlich trivialen) Sachverhalt, dass durchaus

$$P(Z|F \cap S_k) > P(Z|M \cap S_k) \quad \text{für } k = 1, 2, 3$$

und trotzdem $P(Z|F) < P(Z|M)$ gelten kann, da ja die Gewichte verschieden sein können. ◁

2. Laplace-Experimente

Bei Laplace-Experimenten (siehe Beispiel 1.5(a)) haben alle Ergebnisse (korrekt wäre: Elementarereignisse) dieselbe Wahrscheinlichkeit. Zufallsexperimente dieser Art tauchen auf:

- beim Werfen eines symmetrischen Gegenstands (Münze, Würfel, etc.). 'Symmetrisch' heißt dabei, dass alle Seiten mit derselben Wahrscheinlichkeit oben landen.
- beim Mischen von Karten oder allgemeiner beim Herstellen einer zufälligen Reihenfolge. 'Gut gemischt' bzw. 'zufällige Reihenfolge' heißt dabei, dass alle möglichen Anordnungen dieselbe Wahrscheinlichkeit haben.
- beim Entnehmen einer zufälligen Stichprobe aus einer Grundgesamtheit. Zufällige Entnahme einer Stichprobe vom Umfang k aus einer Grundgesamtheit M von n Gegenständen/Personen o.ä. heißt dabei, dass alle Teilmengen vom Umfang k von M mit derselben Wahrscheinlichkeit gezogen werden.

Die Formel 'Anzahl der günstigen, geteilt durch Anzahl der möglichen' Ergebnisse für Wahrscheinlichkeiten in Laplace-Experimenten bedeutet, dass das Bestimmen von Wahrscheinlichkeiten in Laplace-Experimenten letztlich auf das Zählen hinausläuft, wir beschäftigen uns also zunächst mit der 'Kunst des Zählens'. Danach betrachten wir einige konkrete Beispiele und wenden uns schließlich der Frage zu, was 'gleich wahrscheinlich' bei nicht mehr endlichem Ergenisraum bedeuten könnte.

2.1 Etwas Kombinatorik. Es sei wieder $\#A$ der Elemente einer Menge A. In diesem Absatz besprechen wir einige wichtige Formeln für $\#A$ bei bestimmten 'Standardmengen' A. Wir schreiben $A \times B = \big\{(a,b) : a \in A, b \in B\big\}$ für das kartesische Produkt der Mengen A und B und haben einen zugehörigen Potenzbegriff:

$$A^k = \underbrace{A \times \ldots \times A}_{k\text{-mal}} = \big\{(x_1, \ldots, x_k) : x_i \in A \text{ für } i = 1, \ldots, k\big\}.$$

Unser Ausgangspunkt sind die beiden folgenden elementaren Grundregeln:

Regel 1: Gibt es eine bijektive Abbildung von A nach B, so gilt $\#A = \#B$.

Regel 2: Sind A und B disjunkt, so gilt $\#(A \cup B) = \#A + \#B$.

Hat beispielsweise $C \subset A \times B$ die Eigenschaft

$$\#B_x = n \text{ für alle } x \in A \text{ mit } B_x := \{y \in B : (x,y) \in C\},$$

so gilt $\#C = n\,\#A$. Um dies einzusehen, schreibt man die Menge der Paare als disjunkte Vereinigung der Mengen $\{x\} \times B_x$, $x \in A$, verwendet bei den einzelnen Mengen Regel 1 (mit $y \mapsto (x, y)$) und anschließend die auf von zwei auf endlich viele Mengen verallgemeinerte Variante von Regel 2. Als Spezialfall (B_x hängt nicht von x ab) erhält man die Formel $\#(A \times B) = \#A \cdot \#B$.

Wir schreiben abkürzend M_n für $\{1, \ldots, n\}$ (im Folgenden kann anstelle von M_n eine beliebige Menge mit n Elementen stehen). Die obigen Regeln liefern, zusammen mit der anschließenden Diskussion, das folgende Resultat.

Satz 2.1

$$\#M_n^k \;=\; \#\big\{(i_1, \ldots, i_k) :\; 1 \le i_j \le n \;\; \textit{für} \;\; j = 1, \ldots, k\big\} \;=\; n^k.$$

Die Elemente von M_n^k werden gelegentlich *k-Permutationen von M_n mit Wiederholung* genannt. Wir geben zwei typische Anwendungen, bei der Mengen dieses Typs auftauchen:

(i) Einer Menge von n Elementen kann man n^k Stichproben vom Umfang k mit Zurücklegen bei Berücksichtigung der Reihenfolge des Ziehens entnehmen. Das Element $(i_1, \ldots, i_k)$ von M_n^k steht dabei für die Stichprobe, bei der im l-ten Zug das Element i_l erscheint, für $l = 1, \ldots, k$.

(ii) Es gibt n^k Möglichkeiten, k verschiedene Objekte auf n mögliche Plätze zu verteilen, wieder bei Berücksichtigung der Reihenfolge und mit möglicher Mehrfachbelegung. Hierbei steht $(i_1, \ldots, i_k) \in M_n^k$ für die Austeilung, bei der im l-ten Schritt das Objekt mit der Nummer l auf den Platz mit der Nummer i_l gelegt wurde, wieder für $l = 1, \ldots, k$.

Ein recht formaler und möglicherweise weniger anschaulicher Zugang verwendet die Bezeichnung B^A für die Menge der Funktionen $f : A \to B$ und führt auf

$$\#\big(B^A\big) \;=\; (\#B)^{\#A} \quad \text{für endliche Mengen } A, B.$$

Mit $A = \{a_1, \ldots, a_k\}$ und $B = \{b_1, \ldots, b_n\}$ steht dann das k-Tupel $(i_1, \ldots, i_k)$ aus M_n^k für die Funktion $f \in B^A$ mit $f(a_l) = b_{i_l}$ für $l = 1, \ldots, k$.

Was passiert, wenn wir nur injektive Funktionen zulassen?

Satz 2.2 *Für $1 \le k \le n$ gilt*

$$\#\big\{(i_1, \ldots, i_k) \in M_n^k :\; i_l \ne i_j \;\textit{für}\; l \ne j\big\} \;=\; \frac{n!}{(n-k)!}\,.$$

Beweis: Es gibt n Möglichkeiten für i_1, bei gegebenem i_1 bleiben $n-1$ Möglichkeiten für i_2, bei gegebenem (i_1, i_2) bleiben $n-2$ Möglichkeiten für i_3 etc., die gesuchte Anzahl ist also gemäß der oben skizzierten Anwendung der Elementarregeln gleich $n(n-1)(n-2) \cdot \ldots \cdot (n-k+1)$. □

Als wichtigen Spezialfall dieses Satzes erhält man bei $k = n$, dass es genau $n!$ Permutationen einer Menge mit n Elementen gibt. Die Elemente der Menge aus Satz 2.2 werden auch *k-Permutationen von M_n ohne Wiederholung* genannt. Wir haben wieder zwei hauptsächliche Interpretationen:

(i) Einer Menge von n Elementen kann man $\frac{n!}{(n-k)!}$ verschiedene Stichproben vom Umfang k ohne Zurücklegen bei Berücksichtigung der Reihenfolge entnehmen.

(ii) Es gibt $\frac{n!}{(n-k)!}$ verschiedene Möglichkeiten, k Objekte auf n Plätze so zu verteilen, dass keine Mehrfachbesetzungen vorkommen.

Satz 2.3 *Für $1 \leq k \leq n$ gilt*

$$\#\big\{(i_1, \ldots, i_k) \in M_n^k : i_1 < i_2 < \ldots < i_k\big\} = \binom{n}{k}.$$

Beweis: Zu jedem Element dieser Menge gehören genau $k!$ Elemente der Menge aus Satz 2.2, nämlich alle die k-Tupel, die durch Permutation der Koordinaten aus dem geordneten Tupel hervorgehen. □

Man nennt die Elemente der Menge aus Satz 2.3 auch *k-Kombinationen von M_n ohne Wiederholung*. Als wichtigen Spezialfall erhalten wir die Aussage, dass eine Menge mit n Elementen $\binom{n}{k}$ Teilmengen vom Umfang k hat — was wiederum zusammen mit der bekannten Formel für die Mächtigkeit der Potenzmenge einer Menge einen Beweis für $\sum_{k=0}^{n} \binom{n}{k} = 2^n$ liefert. (Wir sehen, dass man Identitäten für Binomialkoeffizienten mit kombinatorischen Überlegungen beweisen kann.)
Wie in den vorangegangenen Fällen haben wir auch hier zwei Standardanwendungen:

(i) Es gibt $\binom{n}{k}$ Möglichkeiten, aus n verschiedenen Objekten k verschiedene herauszugreifen (Stichproben ohne Zurücklegen und ohne Berücksichtigung der Reihenfolge des Ziehens).

(ii) Es gibt $\binom{n}{k}$ verschiedene Möglichkeiten, k Objekte ohne Mehrfachbesetzung auf n Plätze zu verteilen, wenn die Verteilungsreihenfolge nicht berücksichtigt wird.

Satz 2.4 *Für alle $k \in \mathbb{N}$ gilt*

$$\#\big\{(i_1, \ldots, i_k) \in M_n^k : i_1 \leq i_2 \leq \ldots \leq i_k\big\} = \binom{n+k-1}{k}.$$

BEWEIS: Wir definieren eine bijektive Abbildung ϕ von

$$\big\{(i_1,\ldots,i_k)\in M_n^k : i_1\le\ldots\le i_k\big\}$$

nach

$$\big\{(i_1,\ldots,i_k)\in M_{n+k-1}^k : i_1<\ldots<i_k\big\}$$

durch

$$\phi\big((i_1,\ldots,i_k)\big) = (i_1, i_2+1, i_3+2,\ldots,i_k+k-1)$$

und verwenden Regel 1 und Satz 2.3. □

Auch für die Elemente der Menge aus Satz 2.4 gibt es einen Namen, *k-Kombinationen von M_n mit Wiederholung*, sowie zwei klassische Interpretationen:

(i) Einer Menge von n Elementen kann man $\binom{n+k-1}{k}$ verschiedene Stichproben vom Umfang k entnehmen, wenn zurückgelegt wird und die Ziehungsreihenfolge unbeachtet bleibt.

(ii) Es gibt $\binom{n+k-1}{k}$ Möglichkeiten, k Objekte mit möglicher Mehrfachbesetzung auf n Plätze zu verteilen, wenn die Verteilungsreihenfolge nicht berücksichtigt wird.

Aus der zweiten Interpretation ergibt sich als Anwendung, dass man eine natürliche Zahl k auf $\binom{n+k-1}{k}$ Weisen als Summe von n nicht-negativen ganzen Zahlen schreiben kann:

$$\#\big\{(i_1,\ldots,i_n)\in\mathbb{N}_0^n : i_1+\ldots+i_n=k\big\} = \binom{n+k-1}{k}.$$

Hierbei ist i_l die Anzahl der Objekte auf Platz l, ein leeres Fach beispielsweise entspricht einem Summanden 0.

Gibt es auch bei Kombinationen eine formale Definition über Funktionen? Bei den Permutationen sieht man den Zusammenhang zu Funktionen, wenn man $(i_1,\ldots,i_k)$ als Tabelle auffasst: Mit $A=\{a_1,\ldots,a_k\}$ und $B=\{b_1,\ldots,b_n\}$ steht diese dann für die Funktion $f: A\to B$ mit $f(a_l)=b_{i_l}$, $1\le l\le k$. Bei den Kombinationen haben wir nur isotone Tupel zugelassen. Definiert man nun eine Äquivalenzrelation '$\sim$' auf B^A durch

$$f\sim g \quad :\Longleftrightarrow \quad \exists\pi : A\to A,\ \pi \text{ bijektiv},\ f=g\circ\pi,$$

so entsprechen die Kombinationen mit Wiederholung den Äquivalenzklassen in B^A, die ohne Wiederholung den Äquivalenzklassen im Teilraum der injektiven Funktionen. Dies folgt aus zwei einfachen Überlegungen: Zum einen ist Injektivität in dem Sinn mit '$\sim$' verträglich, dass entweder alle Elemente einer Äquivalenzklasse injektiv sind oder keines, zum anderen gibt es bei einer

festgelegten Numerierung der Elemente von A und B stets einen kanonischen Vertreter, nämlich das isotone Element. Satz 2.3 und Satz 2.4 können also auch wie folgt geschrieben werden:

$$\#\big(\{f \in B^A : \ f \text{ injektiv}\}/\sim\big) = \binom{\#B}{\#A}, \quad \#\Big(B^A/\sim\Big) = \binom{\#B+\#A-1}{\#A}.$$

Wir fassen die Formeln aus den Sätzen 2.1-2.4 in der folgenden Tabelle zusammen:

	Wiederholungen: *mit*	*ohne*
Permutationen	n^k	$\dfrac{n!}{(n-k)!}$
Kombinationen	$\binom{n+k-1}{k}$	$\binom{n}{k}$

2.2 Einige typische Probleme.

2.2.1 (Das Geburtstagsproblem) In einem Raum befinden sich n Personen. Mit welcher Wahrscheinlichkeit haben mindestens zwei dieser Personen am gleichen Tag Geburtstag? Wir machen einige vereinfachende Annahmen: Der 29. Februar wird vernachlässigt, ebenso die Möglichkeit von Zwillingen etc., auch saisonale Schwankungen der Geburtenrate werden nicht berücksichtigt. Dann ist ein Laplace-Experiment über

$$\Omega := \big\{(i_1,\dots,i_n) : 1 \le i_1,\dots,i_n \le 365\big\} = \{1,\dots,365\}^n$$

plausibel, wobei $i_j = k$ bedeutet, dass Person j am k-ten Tag des Jahres Geburtstag hat. Es geht um

$$A := \big\{(i_1,\dots,i_n) \in \Omega : \ i_l = i_j \text{ für ein Paar } (l,j) \text{ mit } l \ne j\big\}.$$

Man hat

$$A^{\mathrm{c}} = \big\{(i_1,\dots,i_n) \in \Omega : \ i_l \ne i_j \text{ für } l \ne j\big\}$$

und erhält mit den Formeln aus Abschnitt 2.2

$$P(A) = 1 - \frac{\#A^{\mathrm{c}}}{\#\Omega} = 1 - \frac{365!}{365^n(365-n)!}\,.$$

Dies ist eine (in n) steigende Folge, denn beim Übergang von n zu $n+1$ wird im Nenner ein Faktor $(365-n)$ durch 365 ersetzt. Ab $n = 23$ gilt $P(A) \ge 0.5$, bei $n = 50$ hat man bereits $P(A) \approx 0.97$.

2.2.2 (Ein Bridge-Problem) Beim Kartenspiel Bridge werden 52 Karten an die vier Spieler (Nord, Süd, Ost und West) verteilt. Wir wollen die Wahrscheinlichkeit der Ereignisse

A: einer der Spieler erhält alle vier Asse,

B: jeder der Spieler erhält ein As

bestimmen. Das Mischen der Karten liefert eine zufällige Reihenfolge,

$$\Omega' \;=\; \big\{(\omega_1,\ldots,\omega_{52}) \in \{1,\ldots,52\}^{52} : \omega_i \neq \omega_j \text{ für } i \neq j\big\},$$

Ω' ist also die Menge der Permutationen von $\{1,\ldots,52\}$. Hierbei werden die Karten mit $1,\ldots,52$ durchnumeriert; $\omega_k = j$ bedeutet, dass die k-te Karte im Stapel die Nummer j hat. Alle Elementarereignisse haben dieselbe Wahrscheinlichkeit $\frac{1}{52!}$ (wir können diese Annahme als Definition von 'Karten gut gemischt' betrachten). Die Ereignisse A und B hängen nicht von der Reihenfolge ab, mit der die Karten bei den Spielern ankommen; man kann also auch mit

$$\begin{aligned}\Omega \;:=\; \big\{&(D_1,D_2,D_3,D_4) : D_i \subset \{1,\ldots,52\},\\ &\#D_i = 13 \;\text{ für } i=1,\ldots,4,\; D_i \cap D_j = \emptyset \text{ für } i \neq j\big\}\end{aligned}$$

arbeiten. Hierbei ist D_i die Menge der Karten für Spieler i. Die Austeilreihenfolge definiert eine Abbildung von Ω' in Ω, die jeweils $(13!)^4$ verschiedene Elemente von Ω' auf genau ein Element von Ω abbildet (alle 13! Permutationen der an Spieler 1 ausgegebenen Karten liefern dieselbe Menge D_1 etc.). Betrachten wir also als Resultat des Zufallsexperiments das Vierer-Tupel der 'Hände', so liegt noch stets ein Laplace-Experiment vor, denn es werden jeweils gleich viele Elemente von Ω' zu einem Element von Ω zusammengefasst. Hieraus ergibt sich auch

$$\#\Omega \;=\; \frac{\#\Omega'}{(13!)^4} = \frac{52!}{13!13!13!13!}\,.$$

Man kann dies auch wie folgt einsehen: D_1 ist eine Teilmenge vom Umfang 13 von einer Menge mit 52 Elementen, es gibt also $\binom{52}{13}$ Möglichkeiten für D_1. D_2 ist eine Teilmenge vom Umfang 13 der Menge $\{1,\ldots,52\} - D_1$, die 52-13=39 Elemente hat. Ist also D_1 festgelegt, so bleiben $\binom{39}{13}$ Möglichkeiten für D_2. Für D_3 bleiben $\binom{26}{13}$ Möglichkeiten und der vierte Spieler erhält automatisch die übrigen Karten: Anwendung der Regeln aus Abschnitt 2.2 führt also auf

$$\#\Omega \;=\; \binom{52}{13}\cdot\binom{39}{13}\cdot\binom{26}{13}\cdot 1 \;=\; \frac{52!}{13!13!13!13!}\,.$$

Es sei nun A_i das Ereignis, dass Spieler i alle vier Asse erhält (wir können annehmen, dass diese mit $1,\ldots,4$ durchnumeriert sind). Dann gilt

$$A_1 \;=\; \big\{(D_1,D_2,D_3,D_4) \in \Omega :\, D_1 \supset \{1,2,3,4\}\big\}.$$

Für $D_1 \cap \{1,\ldots,4\}^c$ bleiben $\binom{48}{9}$ Möglichkeiten (9 Karten aus der Menge der 'Nicht-Asse'). Die Anzahl der Möglichkeiten für D_2, D_3 und D_4 bleibt unverändert, also gilt

$$P(A_1) \;=\; \frac{1}{\#\Omega}\binom{48}{9}\binom{39}{13}\binom{26}{13} \;=\; \frac{13\cdot 12\cdot 11\cdot 10}{52\cdot 51\cdot 50\cdot 49}\,.$$

Dieselben Argumente funktionieren bei A_2, A_3, A_4 und führen auf dasselbe Ergebnis. Offensichtlich sind $A_1,\ldots,A_4$ disjunkt und haben Vereinigung A, also ergibt sich

$$P(A) \;=\; P(A_1)+\ldots+P(A_4) \;=\; 4P(A_1) \;\approx\; 0.01056,$$

in ungefähr einem von 100 Spielen wird ein Spieler alle Asse erhalten.

Bei der Behandlung von B kann man ganz analog verfahren. Wir kürzen die Argumentation wie folgt ab: Es gibt 4! Möglichkeiten, die vier Asse so an die vier Spieler zu verteilen, dass jeder genau ein As erhält (4 Möglichkeiten für das Kreuz-As, 3 für das Pik-As etc.). Sind die Asse verteilt, so bleiben

$$\binom{48}{12}\binom{36}{12}\binom{24}{12} \;=\; \frac{48!}{12!12!12!12!}$$

Möglichkeiten für die übrigen Karten. Dies ergibt

$$P(B) \;=\; \frac{\#B}{\#\Omega} \;=\; \frac{4!\,13^4}{52\cdot 51\cdot 50\cdot 49} \;\approx\; 0.1055,$$

in ungefähr einem von 10 Spielen sind also die Asse gleichmässig verteilt.

2.2.3 (Der zerstreute Postbote) Ein Postbote verteilt n Briefe zufällig auf n Briefkästen, einen pro Kasten. Wir nehmen an, dass zu jeder der n Adressen genau einer der n Briefe gehört. Mit welcher Wahrscheinlichkeit erhält keine Person den für sie bestimmten Brief?

Wir numerieren Briefe und Briefkästen so, dass Brief i in Kasten i gehört, $1 \le i \le n$. Die möglichen Austeilungen entsprechen dann den Permutationen von $\{1,\ldots,n\}$. 'Zufällig' soll heißen, dass ein Laplace-Experiment über

$$\Omega_n \;:=\; \big\{(\omega_1,\ldots,\omega_n) : \omega_i \in \{1,\ldots,n\},\, \omega_i \neq \omega_j \text{ für } i \neq j\big\}$$

vorliegt. Sei zunächst

$$A_n \;:=\; \{\omega \in \Omega_n : \omega_i \neq i \text{ für alle } i = 1,\ldots,n\}$$

die Menge der *fixpunktfreien* Permuationen sowie

$$B_{n,i} := \{\omega \in \Omega_n : \omega_i = i\}, \quad 1 \le i \le n.$$

Offensichtlich gilt $A_n^c = \bigcup_{i=1}^n B_{ni}$, also folgt mit der Siebformel (Satz 1.6 (g))

$$\begin{aligned} P_n(A_n) &= 1 - P\Big(\bigcup_{i=1}^n B_{ni}\Big) \\ &= 1 - \sum_{H\subset\{1,\cdots,n\},\, H\neq\emptyset} (-1)^{\#H-1}\, P_n\Big(\bigcap_{i\in H} B_{ni}\Big)\,. \end{aligned}$$

Wir haben

$$\bigcap_{i\in H} B_{ni} = \{\omega \in \Omega_n : \omega_i = i \text{ für alle } i \in H\}\,.$$

Für ein ω aus diesem Durchschnitt sind $\#H$ Positionen festgelegt. Die übrigen $n - \#H$ Positionen können beliebig permutiert werden, also gilt

$$\#\bigcap_{i\in H} B_{ni} = (n - \#H)!\,.$$

Schliesslich ist die Anzahl aller H mit k Elementen gleich $\binom{n}{k}$, also erhalten wir insgesamt

$$\begin{aligned} P_n(A_n) &= 1 - \sum_{H\subset\{1,\ldots,n\},\, H\neq\emptyset} \frac{(-1)^{\#H-1}(n-\#H)!}{n!} \\ &= 1 - \sum_{k=1}^n \binom{n}{k}(-1)^{k-1}\frac{(n-k)!}{n!} \\ &= \sum_{k=0}^n \frac{(-1)^k}{k!}\,. \end{aligned}$$

Aus der Analysis ist $\sum_{k=0}^\infty x^k/k! = e^x$ bekannt. Für große n ist also die Wahrscheinlichkeit dafür, dass kein Brief beim richtigen Empfänger landet, ungefähr $e^{-1} \approx 0.3679$. Wir haben hier ein erstes Grenzwertresultat. Da es im vorliegenden Fall um eine alternierende Reihe geht, können wir darüberhinaus sogar eine Fehlerabschätzung angeben:

$$\left|P_n(A_n) - e^{-1}\right| \le \frac{1}{(n+1)!}\,.$$

Gleichzeitig haben wir eine Aussage bewiesen, die nicht auf Wahrscheinlichkeiten Bezug nimmt: Die Anzahl der fixpunktfreien Permutationen einer Menge von n Elementen ist $n!\sum_{k=0}^n (-1)^k/k!$.

2.3 Unendliche Ergebnisräume. Kann man auch bei unendlichem Ergebnisraum von gleich wahrscheinlichen Resultaten sprechen? Bei abzählbar unendlichem Ω wie beispielsweise $\Omega = \mathbb{N}$ erhält man, wenn $P(\{n\}) = \delta$ für alle $n \in \mathbb{N}$ gilt mit einem festen $\delta > 0$,

$$P\Big(\Big\{1, 2, \ldots, \Big\lceil \frac{2}{\delta} \Big\rceil\Big\}\Big) \;=\; \delta \Big\lceil \frac{2}{\delta} \Big\rceil \;\geq\; 2,$$

was natürlich nicht sein darf (man beachte, dass wir bei diesem Argument nur die endliche Additivität verwendet haben). Im verbleibenden Fall, also bei $P(\{n\}) = 0$ für alle $n \in \mathbb{N}$, hätte man

$$P(\mathbb{N}) \;=\; \sum_{n=1}^{\infty} P(\{n\}) \;=\; 0,$$

was ebenfalls nicht sein darf (bei diesem Argument haben wir die σ-Additivität verwendet). Es gibt in unserem axiomatischen Rahmen also kein Modell für eine zufällige natürliche Zahl, bei dem alle Elementarereignisse $\{n\}$, $n \in \mathbb{N}$, dieselbe Wahrscheinlichkeit haben.

Wir betrachten nun die Situation bei überabzählbarem Ergebnisraum.

2.3.1 (Der rotierende Zeiger) Hält man eine Uhr mit einem Sekundenzeiger zu einem 'zufälligen Zeitpunkt' an und betrachtet den Winkel $\omega \in [0, 2\pi)$ des Sekundenzeigers mit der 12 Uhr-Richtung, so würde man von einem Laplace-Experiment über $\Omega_{60} = \{2\pi k/60 : k = 0, 1, \ldots, 59\}$ ausgehen. Bei einer stets feiner werdenden Zerlegung (oder einem geeigneten Mechanismus mit kontinuierlicher Bewegung) liegt, zumindest als Idealisierung, ein 'Laplace-Experiment' über $\Omega = [0, 1)$ nahe, mit

$$P\big([a, b)\big) \;=\; \frac{b-a}{2\pi} \quad \text{für } 0 \leq a < b < 2\pi.$$

Bei diesem Modell erhält man mit der Stetigkeit von oben von Wahrscheinlichkeitsmaßen (Satz 1.7 (c))

$$P(\{a\}) \;=\; \lim_{n\to\infty} P\Big(\Big[a, a + \frac{1}{n}\Big)\Big) \;=\; 0,$$

alle Elementarerereignisse haben also dann die Wahrscheinlichkeit 0. Im Gegensatz zur Situation im abzählbaren Fall folgt hieraus *nicht* $P(\Omega) = 0$, dazu bräuchte man schon eine Art 'Hyperadditivität'.

2.3.2 (Die Nadel von Buffon) Eine große Fläche wird mit parallelen Linien im Abstand D bedeckt. Eine Nadel der Länge L wird 'in zufälliger Weise' auf diese Fläche geworfen. Mit welcher Wahrscheinlichkeit schneidet die Nadel eine dieser Linien? Wir setzen einfachheitshalber $L \leq D$ voraus. Das Wurfergebnis kann durch ein Paar (x, θ) beschrieben werden, wobei x den Abstand des Nadelzentrums zur nächsten Linie und θ den Winkel zwischen Nadel- und Linienrichtung angibt. Entscheidend ist nun eine Invarianzüberlegung: Drehungen und Verschiebungen sollten keine Rolle spielen, also sollten alle Elemente von

$$\Omega := \big\{(x,\theta) : 0 \leq x \leq D/2,\ 0 \leq \theta < \pi\big\}$$

'dieselbe Wahrscheinlichkeit' haben. Schaut man sich die Formel an, auf die diese Forderung bei endlichem Ergebnisraum führt, so liegt es nahe,

$$P(A) = \frac{\text{Fläche von } A}{\text{Fläche von } \Omega}$$

zu forden.

Bei gegebenem θ schneidet die Nadel genau dann eine der Linien, wenn $x \leq L\sin(\theta)/2$ gilt, das interessierende Ereignis wird also beschrieben durch

$$A = \Big\{(x,\theta) \in \Omega : x \leq \frac{L}{2}\sin(\theta)\Big\}$$

und man erhält

$$P(A) = \Big(\frac{\pi D}{2}\Big)^{-1} \int_0^\pi \frac{L}{2}\,\sin(\theta)\,d\theta = \frac{2L}{\pi D}\,.$$

Schätzt man $P(A)$ durch die beobachtete relative Häufigkeit der Linienüberquerungen beim Wurf einer großen Anzahl von Nadeln, so lässt sich auf diese Weise ein (zufälliger) Näherungswert für π bestimmen. Diese Beobachtung hat allerdings bestenfalls didaktischen Wert als Einstieg in die Monte-Carlo-Methode, da selbst die aus der Numerik als praktisch unbrauchbar bekannte Leibniz-Reihe bessere Resultate liefert.

2.3.3 (Das Paradox von Bertrand) Mit welcher Wahrscheinlichkeit ist die von einer zufälligen Geraden im Einheitskreis gebildete Sekante länger als $\sqrt{3}$, die Seite eines einbeschriebenen gleichseitigen Dreiecks?

Methode 1: Man wählt einen Punkt zufällig und gleichverteilt aus dem Inneren des Kreises und betrachtet die Sehne, die diesen Punkt als Mittelpunkt hat.

In dieser Situation ist die Sekante genau dann länger als die Seite des einbeschriebenen Dreiecks, wenn der Punkt im Inneren des Inkreises des Dreiecks liegt. Dieser hat Radius 1/2, man erhält also die Antwort 1/4.

Methode 2: Man wählt zwei Punkte unabhängig voneinander zufällig und gleichverteilt auf dem Rand des Kreises und verbindet diese.

Betrachtet man den als ersten gewählten Punkt als Eckpunkt eines einbeschriebenen gleichseitigen Dreiecks, so ist das interessierende Ereignis äquivalent dazu, dass der zweite Punkt 'im Schatten' der gegenüberliegenden Seite landet. Dies führt auf die Antwort $1/3$.

Methode 3: Man wählt einen zufälligen Kreisdurchmesser, dann, unabhängig von der ersten Wahl, auf diesem einen zufälligen Punkt (in beiden Fällen gleichverteilt auf dem möglichen Intervall) und betrachtet die Sehne, die diesen Punkt als Mittelpunkt hat.

Die Sekante, die man als Senkrechte zu dem gewählten Durchmesser im Punkt x erhält, ist genau dann länger als $\sqrt{3}$, wenn $x \in (-1/2, 1/2)$ gilt. Diese Argumentation führt auf die Antwort $1/2$.

Welches die richtige Antwort ist, hängt davon ab, wie das Zufallsexperiment ausgeführt wird; Invarianzüberlegungen führen auf die Antwort $1/2$. Man sieht, dass man bei überabzählbarem Ergebnisraum mit dem Konzept 'gleich wahrscheinlich' vorsichtig umgehen muss.

2.3.4 (You can't always get what you want) In den obigen Beispielen mit überabzählbarem Ergebnisraum haben wir uns nicht um den konkreten Definitionsbereich der Wahrscheinlichkeitsmaße gekümmert — aus gutem Grund, wie wir jetzt sehen werden. Bereits im allereinfachsten Beispiel des rotierenden Zeigers aus Absatz 2.3.1 benötigen wir eine Gleichverteilung auf $[0,1)$, also einen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ mit $\Omega = [0,1)$ und

$$P(x + A) \;=\; P(A) \quad \text{für alle } x \in [0,1),\ A \in \mathcal{A}, \tag{$\star$}$$

wobei die Addition modulo 1 zu verstehen ist und $x + A := \{x + y : y \in A\}$.

SATZ 2.5 *Ein Wahrscheinlichkeitsmaß auf $\mathcal{P}([0,1))$ mit der Eigenschaft ($\star$) existiert nicht.*

BEWEIS (unter Verwendung des Auswahlaxioms): Auf $[0,1)$ wird durch

$$x \sim y \quad :\Longleftrightarrow \quad x - y \in \mathbb{Q}$$

eine Äquivalenzrelation definiert. Das Auswahlaxiom erlaubt es, aus jeder der zugehörigen Äquivalenzklassen ein Element auszuwählen; sei A die so erhaltene Menge. Da die Äquivalenzklassen disjunkt sind, enthält A von jeder Äquivalenzklasse genau ein Element. Wir behaupten nun:

(i) $(A+x) \cap (A+y) = \emptyset$ für alle $x, y \in \mathbb{Q} \cap [0,1)$, $x \neq y$,

(ii) $\bigcup_{x \in \mathbb{Q} \cap [0,1)} (x + A) = [0,1)$.

Zu (i): Angenommen, man hat $a + x = b + y$ mit $x, y \in \mathbb{Q} \cap [0,1)$, $x < y$, und $a, b \in A$. Dies führt auf $a \neq b$, wegen $a - b \in \mathbb{Q}$ würde A also im Widerspruch zur Konstruktion zwei Elemente aus einer Äquivalenzklasse enthalten.

Zu (ii): Die Richtung '$\subset$' ist klar, da die Addition modulo 1 geschieht. Ist andererseits $z \in [0,1)$, dann existiert ein $a \in A$ mit $a \sim z$, d.h. $x := a - z \in \mathbb{Q}$ (mit dem 'üblichen' Minus). Ersetzt man ggf. x durch $x + 1$, so erhält man die gewünschte Darstellung von z.

Ist nun P ein Wahrscheinlichkeitsmaß auf $\mathcal{P}([0,1))$ mit der Eigenschaft $(\star)$, so muss P auch der Menge A einen Wert zuordnen. Mit $(\star)$, (ii) und der σ-Additivität von P (deren Anwendbarkeit (i) benötigt) würde dann

$$\sum_{x \in \mathbb{Q} \cap [0,1)} P(A) = 1$$

folgen — dies ist unmöglich. □

Die Potenzmenge ist also zu groß, wir werden uns mit einer kleineren σ-Algebra zufrieden geben müssen. Wir werden dies im übernächsten Abschnitt weiterverfolgen, betrachten aber im folgenden Abschnitt zunächst wieder Wahrscheinlichkeitsräume mit endlichem oder abzählbar unendlichem Ergebnisraum.

Die obigen Betrachtungen werfen auch zusätzliches Licht auf die Additivitätsannahmen bei Wahrscheinlichkeitsmaßen. Bereits in Abschnitt 1 haben wir erwähnt, dass die schwächere Bedingung der endlichen Additivität für eine befriedigende mathematische Theorie nicht reicht. Fordert man dagegen die Additivität für beliebige, also auch überabzählbare Mengenfamilien ('Hyperadditivität'; eine Eigenschaft, die für relative Häufigkeiten gilt), so bleibt nicht genug übrig: Aus $P(\{\omega\}) = 0$ für alle $\omega \in \Omega$ würde $P \equiv 0$ folgen.

3. Diskrete Wahrscheinlichkeitsräume und Zufallsgrößen

3.1 Allgemeines. Wir nennen $(\Omega, \mathcal{A}, P)$ einen *diskreten Wahrscheinlichkeitsraum*, wenn Ω eine endliche oder abzählbar unendliche Menge ist und $\mathcal{A} = \mathcal{P}(\Omega)$ gilt. Aufgrund der σ-Additivität ist P dann durch die zugehörige *Wahrscheinlichkeitsmassenfunktion* (kurz: Massenfunktion) p,

$$p : \Omega \to \mathbb{R}, \quad p(\omega) := P\big(\{\omega\}\big)$$

eindeutig festgelegt:

$$P(A) = \sum_{\omega \in A} p(\omega) \quad \text{für alle } A \in \mathcal{A}.$$

Dies verallgemeinert die im letzten Abschnitt behandelten Laplace-Experimente, bei denen Ω endlich und p eine konstante Funktion ist.

Oft interessiert man sich nicht für das konkrete Ergebnis ω eines Zufallsexperiments, sondern nur für einen hiervon abhängigen Wert $X(\omega)$.

Definition 3.1 Es seien $(\Omega, \mathcal{A}, P)$ eine diskreter Wahrscheinlichkeitsraum und S eine nicht-leere Menge. Dann heißt eine Abbildung $X : \Omega \to S$ eine S-wertige *diskrete Zufallsgrösse*. Im Falle $S = \mathbb{R}$ sprechen wir von *Zufallsvariablen*, bei $S = \mathbb{R}^d$ mit $d > 1$ von *Zufallsvektoren*.

Mit ω ist auch $X(\omega)$ zufällig, triviale Extremfälle ausgenommen. Es wird bei der Behandlung von Zufallsgrößen also nicht darum gehen (können), welchen Wert X annimmt, sondern darum, mit welcher Wahrscheinlichkeit X in einer Teilmenge A von S liegt. Im folgenden sei $X^{-1}(A) := \big\{\omega \in \Omega : X(\omega) \in A\big\}$.

Satz und Definition 3.2 *Es seien $(\Omega, \mathcal{A}, P)$ ein diskreter Wahrscheinlichkeitsraum und $X : \Omega \to S$ eine diskrete Zufallsgrösse. Dann wird durch*

$$P^X : \mathcal{P}(S) \to \mathbb{R}, \quad P^X(A) := P\big(X^{-1}(A)\big) \text{ für alle } A \subset S,$$

ein Wahrscheinlichkeitsmaß auf $(S, \mathcal{P}(S))$ definiert, die Verteilung *von X.*

Beweis: (i) $P^X(S) = P(\{\omega \in \Omega : X(\omega) \in S\}) = P(\Omega) = 1$.

(ii) Sind $A_1, A_2, \ldots \subset S$ paarweise disjunkt, so sind auch die Mengen $X^{-1}(A_1)$, $X^{-1}(A_2), \ldots$ paarweise disjunkt, und mit der σ-Additivität von P folgt

$$\begin{aligned} P^X\Big(\sum_{i=1}^{\infty} A_i\Big) &= P\Big(X^{-1}\Big(\sum_{i=1}^{\infty} A_i\Big)\Big) \\ &= P\Big(\sum_{i=1}^{\infty} X^{-1}(A_i)\Big) \\ &= \sum_{i=1}^{\infty} P\big(X^{-1}(A_i)\big) \;=\; \sum_{i=1}^{\infty} P^X(A_i). \end{aligned}$$

Dies zeigt, dass P^X σ-additiv ist. □

Als alternative Schreibweise für die Verteilung einer Zufallsgröße verwenden wir auch $\mathcal{L}(X)$ (das $\mathcal{L}$ steht für das englische Wort 'law') und schreiben häufig $P(X \in A)$ für $P\big(X^{-1}(A)\big)$.

Beispiel 3.3 Wie oft erscheint 'Kopf' beim fünfmaligen Wurf einer fairen Münze? Das Ausgangsexperiment ist ein Laplace-Experiment über $\Omega = \{0,1\}^5$ (1: Kopf, 0: Wappen). Die Anzahl der 'Kopf'-Würfe ist

$$X(\omega) := \omega_1 + \omega_2 + \ldots + \omega_5, \quad \omega = (\omega_1, \ldots, \omega_5) \in \Omega.$$

Als Bildbereich kommt beispielsweise $S = \{0, 1, \ldots, 5\}$ in Frage. Als Wahrscheinlichkeitsmaß auf einer endlichen Menge wird $\mathcal{L}(X)$ wieder durch die zugehörige Massenfunktion beschrieben, wir benötigen also die Werte

$$P(X = k) \;=\; P\big(\{\omega \in \Omega :\, X(\omega) = k\}\big) \;=\; P\big(X^{-1}(\{k\})\big)$$

für $k = 0, 1, \ldots, 5$. Man erhält

$$\begin{aligned} P\big(\{\omega \in \Omega :\, X(\omega) = k\}\big) &= \frac{\#\{\omega \in \Omega :\, X(\omega) = k\}}{\#\Omega} \\ &= \frac{\#\{(\omega_1, \ldots, \omega_5) \in \{0,1\}^5 : \sum_{i=1}^5 \omega_i = k\}}{2^5} \\ &= \frac{\binom{5}{k}}{32} \quad \text{für } k = 0, 1, \ldots, 5, \end{aligned}$$

denn es gibt $\binom{5}{k}$ Möglichkeiten, die k 1-Werte auf die fünf möglichen Positionen zu verteilen. ◁

Man beachte, dass $\mathcal{L}(X)$ die im Zusammenhang mit X interessierenden Wahrscheinlichkeiten festlegt, keineswegs aber die Zufallsgröße selbst. Bezeichnet beispielsweise Y die Anzahl der 'Wappen'-Würfe in der Situation von Beispiel 3.3, so erhält man $\mathcal{L}(Y) = \mathcal{L}(X)$, obwohl offensichtlich X und Y niemals denselben Wert annehmen.

3.2 Einige wichtige diskrete Verteilungen.

3.2.1 Eine diskrete Zufallsvariable X heißt *binomialverteilt mit Parametern n und p*, kurz: $\mathcal{L}(X) = \mathrm{Bin}(n,p)$ oder $X \sim \mathrm{Bin}(n,p)$, wobei $n \in \mathbb{N}$ und $p \in [0,1]$, wenn

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

gilt. Dies impliziert wegen

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \;=\; \big(p + (1-p)\big)^n = \; 1$$

(binomische Formel), dass die Wahrscheinlichkeit für X-Werte außerhalb von $\{0,1,\ldots,n\}$ gleich 0 ist, also $P(X \in \{0,1,\ldots,n\}) = 1$ gilt.

Die Zufallsvariable X aus Beispiel 3.3 ist $\mathrm{Bin}(5, \frac{1}{2})$-verteilt. In Verallgemeinerung der in diesem Beispiel betrachteten Situation tauchen Binomialverteilungen stets bei Erfolgsanzahlen bei unabhängigen Wiederholungen auf, wenn man 'Erfolg' als das Eintreten eines bestimmten Ereignisses A in einem Einzelexperiment (beispielsweise 'Kopf' beim Münzwurf) interpretiert. Hierbei ist n die Anzahl der Versuchswiederholungen und p die Erfolgswahrscheinlichkeit, d.h. die Wahrscheinlichkeit für das Eintreten von A in einem Einzelexperiment. Zur Begründung bemerken wir, dass jede konkrete Abfolge von A und A^{c}, bei der k-mal A und $(n-k)$-mal A^{c} vorkommt, wegen der vorausgesetzten Unabhängigkeit der Einzelexperimente die Wahrscheinlichkeit $p^k(1-p)^{n-k}$ hat; es gibt $\binom{n}{k}$ Möglichkeiten, die k A-Faktoren auf die n möglichen Positionen zu verteilen.

Im Falle $n = 1$ spricht man auch von *Bernoulli-Verteilungen*; X nimmt dann mit Wahrscheinlichkeit 1 nur die Werte 0 und 1 an.

3.2.2 Die Zufallsvariable X heißt *Poisson-verteilt mit Parameter $\lambda > 0$*, wenn

$$P(X=k) \;=\; e^{-\lambda}\,\frac{\lambda^k}{k!} \quad \text{für alle } k \in \mathbb{N}_0$$

gilt. Diese Verteilung spielt eine wichtige Rolle als Grenzverteilung, sie approximiert beispielsweise Binomialverteilungen $\mathrm{Bin}(n,p)$ bei großem n und kleinem p:

SATZ 3.4 *Ist* $(p_n)_{n\in\mathbb{N}} \subset [0,1]$ *eine Nullfolge mit der Eigenschaft*

$$\lim_{n\to\infty} np_n \;=\; \lambda \;\in (0,\infty),$$

so gilt für alle $k \in \mathbb{N}_0$

$$\lim_{n\to\infty} \binom{n}{k} p_n^k (1-p_n)^{n-k} \;=\; e^{-\lambda}\,\frac{\lambda^k}{k!}\,.$$

BEWEIS: Eine einfache Umformung liefert

$$\binom{n}{k} p_n^k (1-p_n)^{n-k} \;=\; \frac{n(n-1)\cdot\ldots\cdot(n-k+1)}{n^k}\;\frac{(np_n)^k}{k!}\;\frac{\left(1-\frac{np_n}{n}\right)^n}{(1-p_n)^k}\,.$$

Bei festem k ergibt sich mit $n\to\infty$ für den ersten Faktor der Grenzwert 1, für den zweiten $\lambda^k/k!$. Beim Nenner des letzten Faktors erhält man den Limes 1, beim Zähler verwendet man die Monotonie von $x \mapsto (1-x/n)^n$, $x>0$, in Verbindung mit einem Einschachtelungsargument und der bekannten Aussage $\lim_{n\to\infty}(1+x/n)^n = e^x$, um den Grenzwert $e^{-\lambda}$ zu erhalten. □

In Worten besagt dieser Satz, dass bei einer großen Anzahl n von Wiederholungen mit kleiner Erfolgswahrscheinlichkeit p die Zahl X der Erfolge näherungsweise Poisson-verteilt ist mit Parameter $\lambda = np$. Diese Verteilung taucht daher häufig im Zusammenhang mit seltenen Ereignissen auf, beispielsweise bei der Anzahl der Druckfehler pro Seite in einem Buch, der Anzahl emittierter Partikel pro Zeiteinheit bei radioaktivem Material, bei der Anzahl der durch Hufschlag ihres Pferdes ums Leben gekommenen Soldaten eines Kavallerieregiments etc.; Satz 3.4 ist daher auch als das *Gesetz der seltenen Ereignisse* bekannt.

3.2.3 Angenommen, wir werfen einen fairen Würfel solange, bis eine Sechs erscheint. Es sei X die hierfür notwendige Anzahl der Würfe, einschließlich des Wurfes, der die erste Sechs liefert. Offensichtlich gilt $X = n$ (mit $n\in\mathbb{N}$) genau dann, wenn die ersten $n-1$ Versuche keine Sechs ergeben und im n-ten Versuch eine Sechs erscheint. Aufgrund der Unabhängigkeit der Würfe hat dieses Ereignis die Wahrscheinlichkeit

$$\left(1-\frac{1}{6}\right)^{n-1}\frac{1}{6}\,.$$

Wenn allgemeiner X nur Werte aus $\mathbb{N}$ annimmt und

$$P(X=n) = (1-p)^{n-1}p \quad \text{für alle } n\in\mathbb{N}$$

gilt, dann heißt X *geometrisch verteilt mit Parameter* p $(\in (0,1))$.
Diese Verteilung tritt also als Verteilung der Anzahl der Versuche auf, wenn man ein Zufallsexperiment solange wiederholt, bis ein bestimmtes Ereignis, das die Wahrscheinlichkeit p hat, eingetreten ist. Wartet man in Verallgemeinerung hiervon auf das r-te Eintreten des Ereignisses, so erhält man eine Zufallsvariable X, die nur die Werte $r, r+1, \ldots$ annimmt, und für die

$$P(X=n) = \binom{n-1}{r-1}(1-p)^{n-r}p^r \quad \text{für alle } n \in \mathbb{N},\ n \geq r$$

gilt. Man nennt diese Verteilung die *negative Binomalverteilung* mit Parametern r und p, wobei $r \in \mathbb{N}$ und $0 < p < 1$. In der Literatur wird stattdessen häufig auch die Verteilung der Anzahl der Misserfolge bis zum r-ten Versuch (also von $Y = X - r$) so benannt.
Wir haben hier die explizite Angabe des Definitionsbereiches Ω der Zufallsvariablen vermieden. Ergebnisräume der Form $\{0,1\}^{\mathbb{N}}$ (unendlich oft wiederholter Münzwurf) sind überabzählbar, passen also nicht in den gegenwärtigen Rahmen. Alternativ kann man beim Warten auf den ersten Erfolg von der abzählbaren Ergebnismenge $\Omega := \{(0,0,\ldots,0,1) \in \{0,1\}^k : k \in \mathbb{N}\}$ ausgehen.

3.2.4 Eine Urne enthalte N Kugeln, M weiße und $N-M$ schwarze. Dieser Urne werden n Kugeln ohne Zurücklegen entnommen $(n, M \leq N)$, X sei die Anzahl der weißen Kugeln in der 'Stichprobe'. Dann gilt, wobei wie üblich $\binom{i}{j} = 0$ für $j > i$ gesetzt wird,

$$P(X=k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} \quad \text{für } k = 0,\ldots,n,$$

denn es gibt $\binom{M}{k}$ Möglichkeiten für die weißen und $\binom{N-M}{n-k}$ für die schwarzen Kugeln in der Stichprobe und alle $\binom{N}{n}$ möglichen Ziehungen werden als gleich wahrscheinlich vorausgesetzt. Wir nennen diese Verteilung die *hypergeometrische Verteilung* mit Parametern n, N und M, und kürzen dies ab zu $X \sim \mathrm{HypGeo}(N; M, n)$ (bei dieser Reihenfolge darf man die letzten beiden Parameter vertauschen, siehe Übungen). Beispielsweise ist in der in Abschnitt 2.2.2 beschriebenen Situation die Anzahl der Asse, die 'Nord' erhält, hypergeometrisch verteilt mit Parametern 13, 52 und 4. Ein anderes populäres Beispiel: Die Wahrscheinlichkeit für k Richtige beim Zahlenlotto '6 aus 49' ist

$$\frac{\binom{6}{k}\binom{43}{6-k}}{\binom{49}{6}} \quad \text{für } k = 0,\ldots,6,$$

man erhält hypergeometrische Verteilung mit den Parametern 49, 6 und 6.

3.2.5 Es seien $(\Omega, \mathcal{A}, P)$ ein Zufallsexperiment und $A_1, \ldots, A_r$ eine Ereignispartition (siehe Satz 1.9 (b)) von Ω; $p_i := P(A_i)$ für $i = 1, \ldots, r$. Dieses Experiment werde n-mal unabhängig wiederholt, $X = (X_1, \ldots, X_r)$ sei der Zufallsvektor, dessen l-te Komponente zählt, wie oft das Ereignis A_l eingetreten ist. Dann gilt in Verallgemeinerung von 3.2.1

$$P\big(X = (k_1, \ldots, k_r)\big) \;=\; \frac{n!}{k_1! \cdot \ldots \cdot k_r!}\, p_1^{k_1} \cdot \ldots \cdot p_r^{k_r}$$

für alle $k_1, \ldots, k_r \in \mathbb{N}_0$ mit $k_1 + \ldots + k_r = n$. Man nennt diese Verteilung die *Multinomialverteilung* mit Parametern n und $p = (p_1, \ldots, p_r)$; hierbei muss $n \in \mathbb{N}$, $p \in [0,1]^r$ mit $\sum_{i=1}^r p_i = 1$ erfüllt sein.
Zählt man beispielsweise beim n-fachen Wurf eines fairen Würfels, wie häufig die Ergebnisse $1, \ldots, 6$ eingetreten sind, so erhält man die Multinomialverteilung mit Parametern n und $p = \big(\frac{1}{6}, \frac{1}{6}, \ldots, \frac{1}{6}\big)$.

3.3 Erwartungswert und Varianz von Zufallsvariablen. In diesem Unterabschnitt sei stets $(\Omega, \mathcal{A}, P)$ ein diskreter Wahrscheinlichkeitsraum und $X : \Omega \to \mathbb{R}$ (soweit nicht anders erwähnt) eine (diskrete) Zufallsvariable.

Definition 3.5 Der *Erwartungswert von* X, Schreibweise: EX, wird definiert durch

$$EX = \sum_{\omega \in \Omega} X(\omega)\, P\big(\{\omega\}\big),$$

vorausgesetzt, die Summe konvergiert absolut, d.h.

$$\sum_{\omega \in \Omega} |X(\omega)|\, P(\{\omega\}) \;<\; \infty.$$

Ist dies nicht der Fall, so sagen wir, dass der *Erwartungswert von X nicht existiert*.

Der Erwartungswert EX ist also ein mit den jeweiligen Wahrscheinlichkeiten gewogenes Mittel der Werte von X. Das folgende Resultat zeigt, dass man die Summation auf den Bildraum verlagern kann.

Satz 3.6 *Zusätzlich zu $(\Omega, \mathcal{A}, P)$ und X sei $f : \mathbb{R} \to \mathbb{R}$ gegeben, $Y := f(X)$. Dann ist Y eine diskrete Zufallsvariable, und mit p_X, p_Y als zugehörigen Massenfunktionen gilt*

$$EX \;=\; \sum_{x \in \mathbb{R}} x\, p_X(x) \quad \Big(:= \sum_{x \in \mathbb{R}, p_X(x) > 0} x\, p_X(x)\Big),$$

$$EY \;=\; \sum_{y \in \mathbb{R}} y\, p_Y(y) \;=\; \sum_{x \in \mathbb{R}} f(x)\, p_X(x),$$

vorausgesetzt, die beteiligten Summen konvergieren absolut.

BEWEIS: Die Mengen $A_x := \{\omega \in \Omega : X(\omega) = x\}$, $x \in \text{Bild}(X)$, bilden eine Ereignispartition von Ω. Da absolut konvergente Reihen beliebig umgeordnet werden können, erhalten wir

$$\begin{aligned}\sum_{\omega\in\Omega} X(\omega)\,P(\{\omega\}) &= \sum_{x\in\text{Bild}(X)}\ \sum_{\omega\in A_x} X(\omega)\,P(\{\omega\}) \\ &= \sum_{x\in\mathbb{R}} x \sum_{\omega\in A_x} P(\{\omega\}) \quad = \sum_{x\in\mathbb{R}} x\,P(X=x).\end{aligned}$$

Y ist offensichtlich wieder eine reellwertige Abbildung auf Ω, also eine (diskrete) Zufallsvariable. Es gilt

$$\begin{aligned}EY &= \sum_{\omega\in\Omega} Y(\omega)\,P(\{\omega\}) \\ &= \sum_{\omega\in\Omega} f\big(X(\omega)\big)\,P(\{\omega\}) \\ &= \sum_{x\in\text{Bild}(X)}\ \sum_{\omega\in A_x} f(X(\omega))\,P(\{\omega\}) \\ &= \sum_{x\in\mathbb{R}} f(x)\,P(X=x),\end{aligned}$$

denn $f \circ X$ ist auf A_x konstant. □

Wichtige Konsequenz: EX hängt von X nur über die Verteilung von X ab — insbesondere haben Zufallsvariablen mit derselben Verteilung auch denselben Erwartungswert. Für das Verständnis von Erwartungswerten ist vielleicht die folgende Analogie zur Mechanik hilfreich: Platziert man Massen $\pi_1, \pi_2, \pi_3, \ldots$ auf die Punkte $x_1, x_2, x_3, \ldots \in \mathbb{R}$, so ist $\sum x_i p_i$, mit $p_i := \pi_i / \sum_j \pi_j$, der Schwerpunkt des Gesamtgebildes. Beim Würfelwurf hat man die Massen $1/6$ in den Punkten $1, 2, \ldots, 6$ und erhält als Schwerpunkt den Wert 3.5 (dies zeigt übrigens, dass der Erwartungswert nicht unbedingt ein Wert ist, den man erwarten würde).

Betrachtet man allgemeiner eine S-wertige diskrete Zufallsgröße X und eine Abbildung $f : S \to \mathbb{R}$, so erhält man

$$Ef(X) = \sum_{x\in S} f(x)P(X=x),$$

eine in vielen Rechnungen nützliche Formel.

BEISPIEL 3.7 Im Falle $X \sim \mathrm{Bin}(n,p)$ erhalten wir, da das Bild von X aus den Zahlen $0,1,\ldots,n$ besteht,

$$\begin{aligned} EX &= \sum_{k=0}^{n} k\,P(X=k) \\ &= \sum_{k=1}^{n} k\binom{n}{k} p^k\,(1-p)^{n-k} \\ &= np\sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!\,((n-1)-(k-1))!}\,p^{k-1}\,(1-p)^{(n-1)-(k-1)} \\ &= np\sum_{k=0}^{n-1}\binom{n-1}{k}p^k(1-p)^{n-1-k} \quad = \; np\,. \end{aligned}$$

Definiert man Y durch $Y := X(X-1)$, so ergibt sich ganz analog

$$EY \;=\; \sum_{k=2}^{n} k(k-1)\binom{n}{k}\,p^k\,(1-p)^{n-k} \;=\; n(n-1)p^2. \qquad \lhd$$

Der folgende Satz zeigt, dass der Erwartungswertoperator linear und monoton ist.

SATZ 3.8 *Es seien X, Y diskrete Zufallsvariable mit existierendem Erwartungswert und $c \in \mathbb{R}$.*

(a) (Linearität) *Dann existieren auch $E(X+Y)$ sowie $E(cX)$ und es gilt $E(X+Y) = EX + EY$, $E(cX) = cEX$.*

(b) (Monotonie) *Gilt $X \le Y$, also $X(\omega) \le Y(\omega)$ für alle $\omega \in \Omega$, so folgt $EX \le EY$.*

BEWEIS: Die Existenz beispielsweise von $E(X+Y)$ ergibt sich leicht mit der Dreiecksungleichung:

$$\begin{aligned} \sum_{\omega\in\Omega}\big|(X+Y)(\omega)\big|\,P(\{\omega\}) &\le \sum_{\omega\in\Omega}\big(|X(\omega)|+|Y(\omega)|\big)\,P(\{\omega\}) \\ &\le \sum_{\omega\in\Omega}|X(\omega)|\,P(\{\omega\}) \;+\; \sum_{\omega\in\Omega}|Y(\omega)|\,P(\{\omega\}) \\ &< \infty. \end{aligned}$$

Nachdem dies geklärt ist, kann man den Erwartungswert der Summe mit im

wesentlichen denselben Schritten einfach nachrechnen:

$$\begin{aligned} E(X+Y) &= \sum_{\omega\in\Omega}(X+Y)(\omega)\,P(\{\omega\}) \\ &= \sum_{\omega\in\Omega}X(\omega)\,P(\{\omega\}) \;+\; \sum_{\omega\in\Omega}Y(\omega)\,P(\{\omega\}) \\ &= EX \;+\; EY, \end{aligned}$$

die anderen Beweisteile können genauso leicht erbracht werden. □

Mit der Linearität und der Monotonie folgt aus $X \le |X|$, $-X \le |X|$ die wichtige Beziehung

$$|EX| \le E|X|.$$

Der Erwartungswert von X beschreibt die *Lage* der Verteilung von X. Es folgen nun Messzahlen für die *Variabilität* der Verteilung.

Definition 3.9 Das *k-te Moment* einer Zufallsvariablen X ist EX^k, vorausgesetzt, es gilt $\sum_x |x|^k P(X=x) < \infty$ (sonst sagen wir, dass das k-te Moment von X nicht existiert). Existiert das zweite Moment zu X, so nennen wir

$$\operatorname{var}(X) := E(X-EX)^2, \quad \sigma(X) := \big(\operatorname{var}(X)\big)^{1/2}$$

die *Varianz* und die *Standardabweichung* von X.

Die Varianz ist also die mittlere quadratische Abweichung der Zufallsvariablen X von ihrem Mittelwert; durch den Übergang zur Standardabweichung erhält man eine Streuungsmesszahl in den gleichen Dimensionen wie X. Bei der Berechnung dieser Größen sind die folgenden Formeln oft hilfreich.

Lemma 3.10 (a) $\operatorname{var}(X) = EX^2 - (EX)^2$,
(b) $\operatorname{var}(\alpha X) = \alpha^2 \operatorname{var}(X)$ *für alle* $\alpha \in \mathbb{R}$.
(c) *Gilt* $P(X=c)=1$ *für ein* $c\in\mathbb{R}$, *so folgt* $\operatorname{var}(X)=0$.

Beweis: Wir zeigen nur (a), die anderen Teile werden in den Übungen behandelt. Mit den Rechenregeln aus Satz 3.8 erhält man

$$\begin{aligned} \operatorname{var}(X) &= E\big(X^2 - 2(EX)X + (EX)^2\big) \\ &= EX^2 \;-\; 2(EX)EX \;+\; E\big((EX)^2\big) \\ &= EX^2 - (EX)^2, \end{aligned}$$

wobei wir im letzten Schritt Teil (c) verwendet haben. □

Beispiel 3.11 (a) Im Falle $X \sim \mathrm{Bin}(n,p)$ gilt nach Beispiel 3.7

$$EX = np, \quad EX(X-1) = n(n-1)p^2,$$

also

$$EX^2 \;=\; E(X^2 - X) + EX \;=\; EX(X-1) + EX \;=\; n^2p^2 - np^2 + np$$

und damit

$$\mathrm{var}(X) \;=\; EX^2 - (EX)^2 \;=\; n^2p^2 - np^2 + np - n^2p^2 \;=\; np(1-p).$$

(b) Ist X Poisson-verteilt mit Parameter λ (siehe Absatz 3.2.2), so erhält man

$$\begin{aligned} EX \;&=\; \sum_{k=0}^{\infty} k\, e^{-\lambda}\, \frac{\lambda^k}{k!} \\ &=\; \lambda\, e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \\ &=\; \lambda\, e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \quad = \; \lambda\, e^{-\lambda}\, e^{\lambda} \quad = \; \lambda \end{aligned}$$

sowie

$$EX(X-1) \;=\; \sum_{k=2}^{\infty} k(k-1) e^{-\lambda}\, \frac{\lambda^k}{k!} \;=\; \lambda^2,$$

also

$$\mathrm{var}(X) \;=\; EX(X-1) \,+\, EX \,-\, (EX)^2 \;=\; \lambda\,.$$

Bei der Poisson-Verteilung stimmen Erwartungswert und Varianz überein. ◁

Bemerkung und Definition 3.12 Ist M eine beliebige Menge und $A \subset M$, so heißt

$$1_A : M \to \mathbb{R}, \quad x \;\mapsto \begin{cases} 1, & x \in A, \\ 0, & x \notin A, \end{cases}$$

die *Indikatorfunktion* zu A. Man kann $A \mapsto 1_A$ als Einbettung der Potenzmenge von M in den Ring der reellwertigen Funktionen auf M betrachten; so wird beispielsweise aus dem Durchschnitt die Multiplikation. Ist $(\Omega, \mathcal{A}, P)$ ein diskreter Wahrscheinlichkeitsraum und $A \subset \Omega$, so zeigt die Zufallsvariable $X := 1_A$ an, ob das Ereignis A eintritt (Wert 1) oder nicht (Wert 0). Offensichtlich gilt $\mathcal{L}(X) = \mathrm{Bin}(1,p)$ mit $p = P(A)$. Mit dieser Konstruktion sieht man, dass Erwartungswerte Wahrscheinlichkeiten verallgemeinern:

$$E1_A \;=\; 0 \cdot P(1_A = 0) + 1 \cdot P(1_A = 1) \;=\; P(A),$$

d.h. die Wahrscheinlichkeit eines Ereignisses ist gleich dem Erwartungswert der zugehörigen Indikatorfunktion. Mathematisch ergeben sich Erwartungswerte als natürliche Fortsetzung von Wahrscheinlichkeiten, wenn man Ereignisse über ihre Indikatorfunktionen in den Raum der Zufallsvariablen einbettet: Die Additivität des Maßes wird zur Linearität des Erwartungswertes.

3.4 Bedingte Verteilungen und Unabhängigkeit. Sind $X : \Omega \to S_1$ und $Y : \Omega \to S_2$ Zufallsgrößen auf einem diskreten Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$, so ist

$$Z : \Omega \to S_1 \times S_2, \quad \omega \mapsto \big(X(\omega), Y(\omega)\big)$$

eine Zufallsgröße mit Werten in $S_1 \times S_2$. Die Verteilung P^Z von Z nennt man auch die *gemeinsame Verteilung* von X und Y.

Beispiel 3.13 In der Situation von Absatz 2.2.2 (Bridge) sei X die Anzahl der Asse von 'Nord', Y die von 'Süd'. Dann ist $Z := (X, Y)$ eine Zufallsgröße mit Werten in $\{0, \ldots, 4\} \times \{0, \ldots, 4\}$, und die dort eingeführten Techniken führen auf

$$P\big(Z = (k,l)\big) \;=\; \frac{\binom{4}{k}\binom{48}{13-k}\binom{4-k}{l}\binom{35+k}{13-l}\binom{26}{13}}{\frac{52!}{(13!)^4}}.$$

		X					
		0	1	2	3	4	Zeilen-summen:
	0	1150	2600	1950	572	55	6327
	1	2600	4225	2028	286	0	9139
Y	2	1950	2028	468	0	0	4446
	3	572	286	0	0	0	858
	4	55	0	0	0	0	55
Spalten-summen:		6327	9139	4446	858	55	(20825)

Tabelle der mit 20825 multiplizierten Werte

Aus den Werten in der Tabelle ergeben sich wegen

$$P(X = i) \;=\; P(X = i, Y = 0) + P(X = i, Y = 1) + \ldots + P(X = i, Y = 4)$$

für $i = 0, \ldots, 4$ (analog für Y) die *Marginalverteilungen* (oder auch *Randverteilungen*) der Verteilung von Z, also die Verteilungen der Komponenten X und Y von Z. Die gemeinsame Verteilung enthält i.a. mehr Information als die Randverteilungen. Man kann aus der Tabelle die Wahrscheinlichkeit von Ereignissen ablesen, die von X und Y abhängen, beispielsweise

$$\begin{aligned} P(X = Y) &= P(X = 0, Y = 0) + \ldots + P(X = 4, Y = 4) \\ &= \frac{1150 + 4225 + 468 + 0 + 0}{20825} \approx 0.280576\,. \end{aligned}$$ ◁

Die gemeinsame Verteilung erlaubt auch eine Verlagerung der Summation bei der Berechnung von Erwartungswerten von Zufallsvariablen der Form $f(X, Y)$. In der im folgenden Diagramm zusammengefassten Situation

$$\begin{array}{ccccc} & \overset{X}{\nearrow} & S_1 & & \\ \Omega & & \times & \xrightarrow{\quad f \quad} & \mathbb{R} \\ & \underset{Y}{\searrow} & S_2 & & \end{array}$$

erhält man im Stil von Satz 3.6 die für Rechnungen häufig nützliche Formel

$$Ef(X, Y) = \sum_x \sum_y f(x, y)\, P(X = x, Y = y)\,.$$

Analog zum Übergang von Wahrscheinlichkeiten zu bedingten Wahrscheinlichkeiten in Abschnitt 1.2 erhalten wir bei diskreten Zufallsgrößen einen Übergang von Verteilungen zu bedingten Verteilungen und (bei Bildmenge $\mathbb{R}$) von Erwartungswerten zu bedingten Erwartungsweerten.

Satz und Definition 3.14 *Mit $(\Omega, \mathcal{A}, P)$, S_1, S_2, X und Y wie oben gilt für alle $x \in S_1$ mit $P(X = x) > 0$: Durch*

$$A \mapsto P(Y \in A | X = x) \quad \left(= \frac{P\big(\{\omega \in \Omega :\, Y(\omega) \in A \wedge X(\omega) = x\}\big)}{P\big(\{\omega \in \Omega :\, X(\omega) = x\}\big)} \right)$$

wird ein Wahrscheinlichkeitsmaß auf $\big(S_2, \mathcal{P}(S_2)\big)$ definiert, die bedingte Verteilung von Y unter $X = x$*; Schreibweise: $P^{Y|X=x}$ oder $\mathcal{L}(Y|X = x)$.*
Im Falle $S_2 = \mathbb{R}$ und $\sum_y |y| P^{Y|X=x}(\{y\}) < \infty$ nennen wir

$$\begin{aligned} E[Y|X = x] &:= \sum_{y \in \mathbb{R}} y\, P^{Y|X=x}\big(\{y\}\big) \\ &\left(= \frac{1}{P(X = x)} \sum_y y\, P(Y = y, X = x) \right) \end{aligned}$$

den bedingten Erwartungswert von Y unter $X = x$.

Für die Verknüpfung der Abbildungen $X : \Omega \to S_1$ und $x \mapsto P^{Y|X=x}$ bzw. $x \mapsto E[Y|X = x]$ schreiben wir kurz $P^{Y|X}$ oder $\mathcal{L}(Y|X)$ bzw. $E[Y|X]$. Beide Abbildungen sind Zufallsgrössen, die sich als Funktion von X darstellen lassen.

BEWEIS: Klar. □

In der Situation von Beispiel 3.13 ergibt sich beispielsweise als bedingte Erwartung der Anzahl der Asse des Partners, wenn man selbst 2 Asse hat,

$$\begin{aligned} E[Y|X=2] &= 0 \cdot P(Y=0|X=2) + \ldots + 4 \cdot P(Y=4|X=2) \\ &= 0 \cdot \frac{1950}{4446} + 1 \cdot \frac{2028}{4446} + 2 \cdot \frac{468}{4446} + 3 \cdot \frac{0}{4446} + 4 \cdot \frac{0}{4446} \\ &= \frac{2964}{4446} = \frac{2}{3}. \end{aligned}$$

Als Erwartungswert für Y, also ohne die Zusatzinformation $X = 2$, erhält man den Wert 1 — was man übrigens auch begründen kann, ohne zu rechnen. In den Übungen werden einige Eigenschaften bedingter Erwartungswerte behandelt (mit denen man dann auch das obige Ergebnis 2/3 ohne Rechnung erhalten kann), und es wird gezeigt, dass der bedingte Erwartungswert $E[Y|X]$ die Funktion von X ist, die die Zufallsvariable Y in einem gewissen Sinn optimal vorhersagt.

BEISPIEL 3.15 Es sei $(\Omega', \mathcal{A}', P')$ das Modell für ein Zufallsexperiment, in dem ein bestimmtes Ereignis A mit Wahrscheinlichkeit $p > 0$ eintritt. Unser Modell für das n-malige unabhängige Wiederholen des Ausgangsexperiments ist $(\Omega, \mathcal{A}, P)$ mit $\Omega = (\Omega')^n$, $\mathcal{A} = \mathcal{P}(\Omega)$ und

$$P\big(\{(\omega_1, \ldots, \omega_n)\}\big) = P'\big(\{\omega_1\}\big) \cdot \ldots \cdot P'\big(\{\omega_n\}\big).$$

(Man sieht leicht, dass hierdurch in der Tat ein Wahrscheinlichkeitsmaß auf $(\Omega, \mathcal{A})$ definiert wird.) Es sei

$$X : \Omega \to \mathbb{R}, \quad \omega \mapsto \#\{1 \le i \le n : \omega_i \in A\}$$

die Anzahl der Einzelexperimente mit Resultat in A,

$$Y : \Omega \to \mathcal{P}(\{1, \ldots, n\}), \quad \omega \mapsto \{1 \le i \le n : \omega_i \in A\}$$

die Menge der Versuchsnummern, in denen A eintritt. Die gemeinsame Verteilung von X und Y ist offensichtlich auf

$$\big\{(k, B) : k \in \{0, \ldots, n\},\, B \subset \{1, \ldots, n\} \text{ mit } \#B = k\big\}$$

konzentriert, und für jedes Element dieser Menge gilt

$$P(X=k, Y=B) \;=\; \prod_{j\in B} p \prod_{j\notin B}(1-p) \;=\; p^k(1-p)^{n-k}.$$

Aus Abschnitt 3.2.1 ist bereits $P(X=k) = \binom{n}{k}p^k(1-p)^{n-k}$ bekannt, also folgt

$$P^{Y|X=k}(\{B\}) \;=\; \frac{p^k(1-p)^{n-k}}{\binom{n}{k}p^k(1-p)^{n-k}} \;=\; \frac{1}{\binom{n}{k}}\,.$$

Die bedingte Verteilung von Y unter $X=k$ ist also die Gleichverteilung (auch Laplace-Verteilung genannt) auf der Menge der Teilmengen vom Umfang k von $\{1,\ldots,n\}$: Alle möglichen Anordnungen für die 'Erfolge' sind gleich wahrscheinlich. In der Statistik wird es sich als wichtig erweisen, dass in dieser bedingten Verteilung der Parameter p nicht auftaucht — im Gegensatz zur Verteilung von Y selbst, gilt doch beispielsweise $P\big(Y=\{1,\ldots,n\}\big)=p^n$. $\lhd$

Wir dehnen nun den Unabhängigkeitsbegriff auf Zufallsgrößen aus.

Definition 3.16 Für jedes $i\in I$ sei $X_i:\Omega\to S_i$ eine diskrete Zufallsgröße. Die Familie $\{X_i:\, i\in I\}$ heißt *stochastisch unabhängig*, wenn für jede Wahl von $A_i\subset S_i$, $i\in I$, die Ereignisfamilie $\{X_i^{-1}(A_i):\, i\in I\}$ stochastisch unabhängig ist im Sinne von Definition 1.12.

Satz 3.17 *Eine Familie $\{X_i,: i\in I\}$ von diskreten Zufallsgrößen ist genau dann unabhängig, wenn für alle $\{i_1,\ldots,i_n\}\subset I$, $x_{i_1}\in S_{i_1},\ldots,x_{i_n}\in S_{i_n}$ gilt:*

$$P(X_{i_1}=x_{i_1},\ldots,X_{i_n}=x_{i_n}) \;=\; P(X_{i_1}=x_{i_1})\cdot\ldots\cdot P(X_{i_n}=x_{i_n}).$$

Beweis: Für beliebige $A_i\subset S_i$ und $\{i_1,\ldots,i_n\}\subset I$ gilt

$$\begin{aligned}
P\Big(\bigcap_{j=1}^{n} X_{i_j}^{-1}(A_i)\Big) \;&=\; \sum_{x_{i_1}\in A_{i_1},\ldots,x_{i_n}\in A_{i_n}} P(X_{i_n}=x_{i_1},\ldots,X_{i_n}=x_{i_n}) \\
&=\; \sum_{x_{i_1}\in A_{i_1}} P(X_{i_1}=x_{i_1}) \sum_{x_{i_2}\in A_{i_2}} P(X_{i_2}=x_{i_2})\ldots \\
&\qquad\qquad \ldots \sum_{x_{i_n}\in A_{i_n}} P(X_{i_n}=x_{i_n}) \\
&=\; P(X_{i_n}\in A_{i_1})\cdot\ldots\cdot P(X_{i_n}\in A_{i_n}),
\end{aligned}$$

also ist die Bedingung hinreichend. Wählt man Elementarereignisse in Definition 3.16, so folgt auch die Notwendigkeit. $\square$

Bei einer endlichen Familie $X_1, \ldots, X_n$ hat man also Unabhängigkeit genau dann, wenn die gemeinsame Massenfunktion p

$$p(x_1, \ldots, x_n) \;=\; P(X_1 = x_1, \ldots, X_n = x_n),$$

sich als Produkt der marginalen Massenfunktionen p_i, $p_i(x_i) = P(X_i = x_i)$ für $1 \le i \le n$, schreiben lässt, also

$$p(x_1, \ldots, x_n) \;=\; p_1(x_1) \cdot \ldots \cdot p_n(x_n)$$

gilt für alle $x_1 \in S_1, \ldots, x_n \in S_n$. Bei Unabhängigkeit ergibt sich daher die gemeinsame Verteilung aus den Randverteilungen; i.a. ist dies nicht der Fall.

3.5 Reellwertige diskrete Zufallsgrößen. Mit $\mathbb{R}$ als Wertebereich hat man zusätzliche Strukturen und damit spezielle Probleme und Konzepte.

Satz 3.18 (Multiplikationsregel für Erwartungswerte) *Sind X und Y unabhängige Zufallsvariablen mit existierenden Erwartungswerten, so existiert auch der Erwartungswert zu $X \cdot Y$, und es gilt $EXY = EXEY$.*

Beweis: Die Mengen

$$A_{xy} \;:=\; \big\{\omega \in \Omega :\, X(\omega) = x,\, Y(\omega) = y\big\}, \quad x \in \mathrm{Bild}(X),\; y \in \mathrm{Bild}(Y),$$

bilden eine Partition von Ω, also folgt wie im Beweis zu Satz 3.6 (Verlagerung der Summation) unter Ausnutzung der Unabhängigkeit

$$\begin{aligned}
\sum_{\omega\in\Omega} \big|(X\cdot Y)(\omega)\big|\, P\big(\{\omega\}\big) \;&=\; \sum_x \sum_y \sum_{\omega\in A_{x,y}} \big|\big(X\cdot Y\big)(\omega)\big|\, P\big(\{\omega\}\big) \\
&=\; \sum_x \sum_y |xy|\, P(X = x, Y = y) \\
&=\; \sum_x \sum_y |x|\,|y|\, P(X = x)\, P(Y = y) \\
&=\; \Big(\sum_x |x|\, P(X = x)\Big)\Big(\sum_y |y|\, P(Y = y)\Big) \\
&=\; \Big(\sum_{\omega\in\Omega} \big|X(\omega)\big|\, P\big(\{\omega\}\big)\Big)\Big(\sum_{\omega\in\Omega} \big|Y(\omega)\big|\, P\big(\{\omega\}\big)\Big).
\end{aligned}$$

Wegen der vorausgesetzten Existenz der einzelnen Erwartungswerte ist dies endlich, also existiert auch EXY. Wiederholt man nun die Rechnung ohne Betragsstriche, oder verwendet man die Formeln

$$Ef(X,Y) \;=\; \sum_x \sum_y f(x,y)\, P(X = x, Y = y),\; Ef(X) \;=\; \sum_{x\in S} f(x)\, P(X = x),$$

so erhält man $EXY = EXEY$. □

Im allgemeinen folgt die Existenz von EXY nicht aus der von EX, EY. Man hat jedoch:

SATZ 3.19 (Cauchy-Schwarz-Ungleichung) *Existiert zu den Zufallsvariablen X und Y das zweite Moment, so existiert auch EXY und es gilt*

$$(EXY)^2 \leq EX^2 EY^2.$$

BEWEIS: Wegen

$$\big|(X \cdot Y)(\omega)\big| \;=\; \big|X(\omega)\big|\,\big|Y(\omega)\big| \;\leq\; X(\omega)^2 + Y(\omega)^2 \quad \text{für alle } \omega \in \Omega$$

gilt

$$\sum_{\omega\in\Omega} \big|(X \cdot Y)(\omega)\big|\, P\big(\{\omega\}\big) \;\leq\; \sum_{\omega\in\Omega} X(\omega)^2\, P\big(\{\omega\}\big) \;+\; \sum_{\omega\in\Omega} Y(\omega)^2 P\big(\{\omega\}\big),$$

also existiert der Erwartungswert zu XY. Für beliebiges $t \in \mathbb{R}$ existiert dann auch das zweite Moment zu $X + tY$ (Satz 3.8) und ist nicht-negativ:

$$0 \;\leq\; E(X + tY)^2 \;=\; EX^2 + t^2 EY^2 + 2tEXY \quad \text{für alle } t \in \mathbb{R}.$$

Im Falle $EY^2 = 0$ kann die Gerade auf der rechten Seite nur dann oberhalb von 0 bleiben, wenn $EXY = 0$ gilt; in diesem Falle gilt also die behauptete Ungleichung. Im Falle $EY^2 > 0$ erhält man als kleinsten Wert der Parabel auf der rechten Seite

$$\frac{1}{EY^2}\big(EX^2 EY^2 - (EXY)^2\big).$$

Dies ist nur dann nicht-negativ, wenn die behauptete Ungleichung gilt. □

Varianten der Cauchy-Schwarz-Ungleichung tauchen auch in anderen Vorlesungen auf, oft im Zusammenhang mit Begriffen wie Orthogonalität und Projektion. In der folgenden Bemerkung stellen wir die Verbindung her und erhalten gleichzeitig eine geometrische Interpretation bedingter Erwartungswerte; Details sind Gegenstand einer Übungsaufgabe.

BEMERKUNG 3.20 Ist $(\Omega, \mathcal{A}, P)$ ein diskreter Wahrscheinlichkeitsraum mit der Eigenschaft

$$P\big(\{\omega\}\big) > 0 \quad \text{für alle } \omega \in \Omega,$$

so ist

$$H \;:=\; \{X : \Omega \to \mathbb{R} :\; EX^2 < \infty\} \quad \text{mit } \langle X, Y\rangle := EXY$$

ein Hilbert-Raum. Mit $\|X\| := \langle X, X\rangle^{1/2}$ wird die Cauchy-Schwarzsche Ungleichung zu

$$\big|\langle X, Y\rangle\big| \;\le\; \|X\|\,\|Y\|.$$

Ist Z eine Zufallsgröße auf diesem Wahrscheinlichkeitsraum und mit Werten in irgendeiner Menge S, so wird durch

$$H(Z) \;:=\; \big\{X \in H :\; X = \phi(Z) \text{ für ein } \phi : S \to \mathbb{R}\big\}$$

ein Unterraum von H definiert. Die Abbildung

$$H \to H(Z), \quad X \mapsto E[X|Z]$$

ist die Orthogonalprojektion auf diesen Unterraum.

Dies behandelt die allgemeine Situation (im diskreten Fall). Bei endlichen Wahrscheinlichkeitsräumen, beispielsweise bei $\Omega = \{1, \dots, n\}$, $\mathcal{A} = \mathcal{P}(\Omega)$ und $p_i := P(\{i\}) > 0$ für $i = 1, \dots, n$, kann man eine Zufallsvariable X mit dem Vektor

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad x_i := X(i) \text{ für } i = 1, \dots, n,$$

identifizieren und erhält dann den euklidischen Raum $\mathbb{R}^n$ mit dem Skalarprodukt $\langle x, y\rangle = \sum_{i=1}^n p_i x_i y_i$. ◁

Definition 3.21 Es seien X und Y Zufallsvariablen mit endlichem zweiten Moment und den Standardabweichungen σ_X, σ_Y. Dann heißt

$$\operatorname{cov}(X, Y) \;:=\; E(X - EX)(Y - EY)$$

die *Kovarianz* von X und Y. Im Falle $\operatorname{cov}(X, Y) = 0$ nennt man X und Y *unkorreliert*. Ist $\sigma_X \cdot \sigma_Y > 0$, so nennt man

$$\rho(X, Y) \;:=\; \frac{\operatorname{cov}(X, Y)}{\sigma_X \sigma_Y}$$

den *Korrelationskoeffizienten* von X und Y.

Satz 3.22 *Es seien X und Y Zufallsvariablen mit existierendem zweiten Moment. Dann gilt:*

(a) $\operatorname{cov}(X, Y) = EXY - (EX)(EY)$.

(b) *Sind X und Y unabhängig, so sind sie auch unkorreliert.*

(c) *Ist $\rho(X, Y)$ ist definiert, so gilt $-1 \le \rho(X, Y) \le 1$.*

BEWEIS: (a) Mit der Linearität des Erwartungswertoperators (Satz 3.8) folgt

$$\begin{aligned}\operatorname{cov}(X,Y) &= E\big(XY-(EX)Y-X(EY)+(EX)(EY)\big)\\ &= EXY-(EX)(EY)-(EX)(EY)+(EX)(EY)\\ &= EXY-EXEY.\end{aligned}$$

(b) folgt unmittelbar aus (a) und Satz 3.18.

(c) Satz 3.19 liefert

$$\begin{aligned}\operatorname{var}(X)\operatorname{var}(Y)\,\rho(X,Y)^2 &= \big(E(X-EX)(Y-EY)\big)^2\\ &\le E(X-EX)^2\,E(Y-EY)^2\\ &= \operatorname{var}(X)\operatorname{var}(Y).\end{aligned}$$

□

Gemäß Teil (b) des Satzes sind unabhängige Zufallsvariable unkorreliert — die Umkehrung hiervon gilt nicht! Kovarianz und Korrelation können als Maß für die lineare Abhängigkeit von Zufallsvariablen betrachtet werden; auch dies wird in den Übungsaufgaben weiter ausgeführt. Mit Hilfe dieser Begriffe lässt sich auch etwas über die Varianz einer Summe von Zufallsvariablen aussagen. Die zweite Aussage des folgenden Satzes ist auch als *Gleichheit von Bienaymé* bekannt.

SATZ 3.23 *Es seien $X_1,\dots,X_n$ Zufallsvariablen mit existierendem zweiten Moment. Dann gilt*

$$\operatorname{var}(X_1+\ldots+X_n) \;=\; \sum_{i=1}^{n}\operatorname{var}(X_i) \;+\; \sum_{\substack{i,j=1\\ i\neq j}}^{n}\operatorname{cov}(X_i,X_j).$$

Sind die Zufallsvariablen $X_1,\dots,X_n$ darüberhinaus unabhängig, so gilt

$$\operatorname{var}(X_1+\ldots+X_n) \;=\; \operatorname{var}(X_1)+\ldots+\operatorname{var}(X_n).$$

BEWEIS: Unter Verwendung von Satz 3.22 und Lemma 3.10 folgt

$$\begin{aligned}\operatorname{var}\Big(\sum_{i=1}^{n}X_i\Big) &= E\Big(\sum_{i=1}^{n}X_i\Big)^2-\Big(E\sum_{i=1}^{n}X_i\Big)^2\\ &= \sum_{i,j=1}^{n}EX_iX_j-\sum_{i,j=1}^{n}EX_iEX_j\end{aligned}$$

$$= \sum_{i=1}^{n} \big(EX_i^2 - (EX_i)^2\big) \; + \sum_{i \neq j} (EX_iX_j - EX_iEX_j)$$
$$= \sum_{i=1}^{n} \operatorname{var}(X_i) \; + \sum_{i \neq j} \operatorname{cov}(X_i, X_j).$$

Der zweite Teil folgt hieraus sofort mit Satz 3.22 (b). □

BEISPIEL 3.24 (a) In einem Zufallsexperiment sei A ein Ereignis mit der Wahrscheinlichkeit p. Das Experiment werde n-mal unabhängig wiederholt; X_i zeige an, ob das Ereignis in der i-ten Wiederholung eintritt ($X_i = 1$) oder nicht ($X_i = 0$). Dann sind $X_1, \ldots, X_n$ unabhängig mit

$$EX_i \; = \; 0 \cdot P(X_i = 0) + 1 \cdot P(X_i = 1) \; = \; p,$$

$$EX_i^2 = EX_i = p, \quad \operatorname{var}(X_i) = p - p^2 = p(1-p).$$

Somit gilt für $S_n := X_1 + \ldots + X_n$

$$ES_n = \sum_{i=1}^{n} EX_i = np, \quad \operatorname{var}(S_n) = \sum_{i=1}^{n} \operatorname{var}(X_i) = np(1-p).$$

Wegen $S_n \sim \operatorname{Bin}(n,p)$ ist dies ein alternativer Beweis für die Formeln aus Beispiel 3.11 (a).

(b) Es sei X hypergeometrisch verteilt, also

$$P(X = k) \; = \; \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} \quad \text{für } \; k = 0, \ldots, n.$$

Wie in Abschnitt 3.2.4 erklärt, entsteht dies als Verteilung der Anzahl der weißen Kugeln, wenn man einer Urne mit N Kugeln eine Stichprobe vom Umfang n entnimmt; hierbei wird vorausgesetzt, dass M der Kugeln in der Urne weiß sind. Setzt man $X_i = 1$, wenn im i-ten Zug eine weiße Kugel gezogen wird, und $X_i = 0$ sonst, so gilt offensichtlich $X = X_1 + \ldots + X_n$. Im Gegensatz zu der unter (a) betrachteten Situation sind die Summanden nun allerdings nicht mehr unabhängig, wir benötigen also eine Hilfsüberlegung. Hierzu stellen wir uns die Kugeln als mit den Zahlen 1 bis N numeriert vor. Sind $Y_1, \ldots, Y_n$ die (Nummern der) gezogenen Kugeln, so gilt $X_i = \phi(Y_i)$ mit

$$\phi(i) \; := \; \begin{cases} 1, & i\text{-te Kugel weiß}, \\ 0, & \text{sonst}, \end{cases}$$

und mit den in Abschnitt 3 besprochenen Techniken erhält man

$$P(Y_1 = i_1, \dots, Y_n = i_n) \;=\; \frac{(N-n)!}{N!}$$

für alle n-Permutationen $(i_1, \dots, i_n)$ ohne Wiederholung von $\{1, \dots, N\}$. Es sei S_n die Menge der Permutationen von $\{1, \dots, n\}$. Für beliebiges $\pi \in S_n$ und $(i_1, \dots, i_n)$ wie oben ergibt sich

$$\begin{aligned} P\big(Y_{\pi(1)} = i_1, \dots, Y_{\pi(n)} = i_n\big) \;&=\; P\big(Y_1 = i_{\pi^{-1}(1)}, \dots, Y_n = i_{\pi^{-1}(n)}\big) \\ &=\; \frac{(N-n)!}{N!} \\ &=\; P(Y_1 = i_1, \dots, Y_n = i_n), \end{aligned}$$

also gilt $\mathcal{L}\big((Y_1, \dots, Y_n)\big) \;=\; \mathcal{L}\big((Y_{\pi(1)}, \dots, Y_{\pi(n)})\big)$ und damit auch

$$\mathcal{L}\big((X_1, \dots, X_n)\big) \;=\; \mathcal{L}\big((X_{\pi(1)}, \dots, X_{\pi(n)})\big) \qquad \text{für alle } \pi \in S_n$$

(man spricht dann von *vertauschbaren* Zufallsvariablen). Dies impliziert, dass die Verteilung von X_i nicht von i abhängt. Man sieht leicht, dass $X_1 \sim \mathrm{Bin}(1, M/N)$ gilt, erhält also

$$EX \;=\; \sum_{i=1}^{n} EX_i \;=\; n\,EX_1 \;=\; \frac{nM}{N}\,.$$

Bei der Varianz argumentiert man analog und benutzt nun, dass $\mathcal{L}\big((X_i, X_j)\big) = \mathcal{L}\big((X_1, X_2)\big)$ für alle i, j mit $i \neq j$ gilt. Wegen $X_1 + X_2 \sim \mathrm{HypGeo}(2; N, M)$ bedeutet dies

$$EX_1X_2 \;=\; P(X_1 + X_2 = 2) \;=\; \frac{\binom{M}{2}\binom{N-M}{0}}{\binom{N}{2}} \;=\; \frac{M(M-1)}{N(N-1)}\,.$$

Mit Satz 3.23 folgt nun

$$\begin{aligned} \mathrm{var}(X) \;&=\; n\,\mathrm{var}(X_1) \;+\; n(n-1)\,\mathrm{cov}(X_1, X_2) \\ &=\; n\,\frac{M}{N}\Big(1 - \frac{M}{N}\Big) \;+\; n(n-1)\left(\frac{M(M-1)}{N(N-1)} - \frac{M^2}{N^2}\right) \\ &=\; \frac{nM(N-n)(N-M)}{N^2(N-1)}\,. \end{aligned}$$

Beide Formeln kann man natürlich auch im Stil von Beispiel 3.7 'zu Fuß' erhalten. ◁

SATZ UND DEFINITION 3.25 (a) *Es seien P und Q Wahrscheinlichkeitsmaße auf $\mathbb{Z}$ mit Massenfunktionen p und q. Dann ist auch*

$$r : \mathbb{Z} \to \mathbb{R}, \quad r_n := \sum_{k\in\mathbb{Z}} p_k q_{n-k}$$

eine Wahrscheinlichkeitsmassenfunktion. Das zugehörige Wahrscheinlichkeitsmaß R nennen wir die Faltung *von P und Q, Schreibweise: $R = P \star Q$.*

(b) *Sind X und Y unabhängige Zufallsvariablen mit Werten in $\mathbb{Z}$, so ist auch $X + Y$ eine Zufallsvariable mit Werten in $\mathbb{Z}$, und es gilt $P^{X+Y} = P^X \star P^Y$.*

BEWEIS: (a) Offensichtlich hat man $r_n \geq 0$ für alle $n \in \mathbb{Z}$ sowie

$$\begin{aligned}\sum_{n\in\mathbb{Z}} r_n &= \sum_{n\in\mathbb{Z}}\sum_{k\in\mathbb{Z}} p_k q_{n-k} \\ &= \sum_{k\in\mathbb{Z}} p_k \sum_{n\in\mathbb{Z}} q_{n-k} = \sum_{k\in\mathbb{Z}} p_k \cdot 1 = 1,\end{aligned}$$

also definiert r ein Wahrscheinlichkeitsmaß auf $\mathbb{Z}$ (durch $R(A) := \sum_{k\in A} r_k$).

(b) Wir zerlegen nach dem Wert von X:

$$\begin{aligned}P(X+Y=n) &= \sum_{k\in\mathbb{Z}} P(X=k, X+Y=n) \\ &= \sum_{k\in\mathbb{Z}} P(X=k, Y=n-k) \\ &= \sum_{k\in\mathbb{Z}} P(X=k)P(Y=n-k).\end{aligned}$$

Verwende nun Teil (a) mit $p_k = P(X = k)$, $q_k = P(Y = k)$ und $r_k = P(X + Y = k)$. □

BEISPIEL 3.26 Es seien X und Y unabhängige Zufallsvariable; X sei Poisson-verteilt mit Parameter λ und Y sei Poisson-verteilt mit Parameter μ. Dann gilt für alle $n \in \mathbb{N}_0$

$$\begin{aligned}P(X+Y=n) &= \sum_{k\in\mathbb{Z}} P(X=k)P(Y=n-k) \\ &= \sum_{k=0}^{n} e^{-\lambda}\frac{\lambda^k}{k!}\, e^{-\mu}\frac{\mu^{n-k}}{(n-k)!} \\ &= e^{-(\lambda+\mu)}\frac{1}{n!}\sum_{k=0}^{n}\binom{n}{k}\lambda^k\mu^{n-k} \\ &= e^{-(\lambda+\mu)}\frac{(\lambda+\mu)^n}{n!},\end{aligned}$$

$X + Y$ ist also wieder Poisson-verteilt, und zwar mit Parameter $\lambda + \mu$. Die Poisson-Verteilungen bilden eine sog. *Faltungshalbgruppe*.

Was ist die bedingte Verteilung von X unter $X + Y$? Für alle $n \in \mathbb{N}_0, k \in \{0, \dots, n\}$ erhält man

$$\begin{aligned} P(X = k | X + Y = n) &= \frac{P(X = k, X + Y = n)}{P(X + Y = n)} \\ &= \frac{P(X = k)P(Y = n - k)}{P(X + Y = n)} \\ &= \frac{e^{-\lambda}\frac{\lambda^k}{k!}\, e^{-\mu}\frac{\mu^{n-k}}{(n-k)!}}{\frac{e^{-(\lambda+\mu)}(\lambda+\mu)^k}{n!}} \\ &= \binom{n}{k}\Big(\frac{\lambda}{\lambda+\mu}\Big)^k\Big(1 - \frac{\lambda}{\lambda+\mu}\Big)^{n-k}, \end{aligned}$$

also gilt $\mathcal{L}(X|X + Y) = \mathrm{Bin}\big(X + Y, \lambda/(\lambda + \mu)\big)$. Konkret: Angenommen, ein Buch von 100 Seiten hat auf Seite k X_k Druckfehler, wobei $X_1, \dots, X_{100}$ unabhängig und Poisson-verteilt sind mit Parameter $\lambda > 0$ (diese Annahmen sind natürlich bestenfalls näherungsweise erfüllt). Enthält das Buch insgesamt 10 Druckfehler, so ist die bedingte Verteilung der Anzahl der Druckfehler auf der dritten Seite $\mathrm{Bin}\big(10, \frac{1}{100}\big)$. ◁

3.6 Wahrscheinlichkeitserzeugende Funktionen. Ist $(a_n)_{n\in\mathbb{N}_0}$ eine Folge reeller Zahlen, so nennt man bekanntlich die Potenzreihe $\hat{a}(z) := \sum_{n=0}^{\infty} a_n z^n$ die zugehörige *erzeugende Funktion*. Ist die Folge beschränkt, so darf $\hat{a}$ in einer Nullumgebung beliebig oft gliedweise differenziert werden und man kann dann insbesondere die Folge aus ihrer erzeugenden Funktion zurückerhalten:

$$a_n = \frac{1}{n!}\frac{d^n}{dz^n}\hat{a}(z)\Big|_{z=0}.$$

Manche Probleme, insbesondere die Behandlung von Differenzengleichungen, können durch den Übergang zu erzeugenden Funktionen vereinfacht werden.

Beispiel 3.27 (Ein Ruin-Problem) Spieler I besitzt n €, Spieler II $N - n$ €. In jeder Runde gewinnt I von II 1€ mit Wahrscheinlichkeit p und verliert 1€ sonst. Das Spiel wird fortgesetzt, bis einer der Spieler sein gesamtes Geld verloren hat. Mit welcher Wahrscheinlichkeit gewinnt I das Spiel?

Sei $N \in \mathbb{N}$ fest; A_n bezeichne das Ereignis, dass I bei Anfangskapital n gewinnt, B das Ereignis, dass I die erste Runde gewinnt. Das Gesetz von der totalen Wahrscheinlichkeit (Satz 1.9 (b)) liefert

$$P(A_n) = P(A_n|B)P(B) + P(A_n|B^{\mathrm{c}})P(B^{\mathrm{c}}) \quad \text{für } 0 < n < N.$$

Sei $p_n := P(A_n)$. Wir nehmen an, dass die Runden voneinander unabhängig sind und erhalten dann für $(p_0, \dots, p_N)$ die folgende Differenzengleichung zweiter Ordnung mit zwei Randbedingungen:

$$p_n \;=\; p\,p_{n+1} \,+\, (1-p)\,p_{n-1} \quad \text{für } 1 \le n \le N-1, \qquad p_0 = 0,\; p_N = 1. \quad (*)$$

Mit erzeugenden Funktionen lassen sich solche Gleichungen häufig routinemäßig lösen (oft es geht es natürlich auch, wie übrigens auch hier, direkt mit irgendwelchen Tricks oder geschickten Umformungen — die allerdings erst einmal gefunden werden müssen). Sei $r := (1-p)/p$, wir setzen (zunächst) $r \neq 1$ voraus (also $p \neq \frac{1}{2}$). Löst man $(*)$ nach p_{n+1} auf, so erhält man

$$p_{n+1} \;=\; (1+r)p_n \,-\, rp_{n-1}.$$

Multiplikation mit z^{n+1} und Summation über $n \in \mathbb{N}$ liefert unter Beachtung von $p_0 = 0$ für $\hat{p}(z) = \sum_{n=0}^{\infty} p_n z^n$ die Beziehung

$$\hat{p}(z) - p_1 z \;=\; (1+r)z\hat{p}(z) - rz^2\hat{p}(z).$$

Löst man dies nach $\hat{p}(z)$ auf und führt man dann eine Partialbruchzerlegung durch, so ergibt sich

$$\hat{p}(z) \;=\; \frac{p_1 z}{1-(1+r)z+rz^2} \;=\; \frac{p_1}{r-1}\Big(\frac{1}{1-rz} - \frac{1}{1-z}\Big).$$

Erinnert man sich nun an die Formel für die geometrische Reihe, so erhält man hieraus

$$p_n \;=\; \frac{p_1}{r-1}\big(r^n - 1\big).$$

Die übrige Randbedingung $p_N = 1$ führt auf $p_1 = (r-1)/(r^N-1)$, also folgt insgesamt

$$p_n = \frac{r^n - 1}{r^N - 1}, \quad n = 0, \dots, N.$$

Ähnlich erhält man bei $r = 1$ das Resultat $p_n = \frac{n}{N}$, $n = 0, \dots, N$. Konkret: Ich betrete ein Kasino mit 100 € Kapital und setze bei Roulette in jeder Runde einen Euro auf Rot; Rot erscheint mit Wahrscheinlichkeit 18/37 und bringt 2 €. Ich höre auf, wenn ich 100 € gewonnen oder aber alles verloren habe. Dies passt in die obige Situation mit $p = 18/37$, $N = 200$ und $n = 100$. Die zugehörige Erfolgswahrscheinlichkeit ist

$$\frac{(\frac{19}{18})^{100} - 1}{(\frac{19}{18})^{200} - 1} \;\approx\; 0.00447.$$

In dieser Situation ist es offensichtlich geschickter, alles auf einen Schlag auf Rot zu setzen, denn dann ist die Erfolgswahrscheinlichkeit $18/37 \approx 0.4865$. $\lhd$

DEFINITION 3.28 Ist X eine $\mathbb{N}_0$-wertige Zufallsvariable, so heißt

$$\hat{p}_X(z) := \sum_{k=0}^{\infty} P(X=k)\, z^k \quad \Big(= Ez^X \Big)$$

die *wahrscheinlichkeitserzeugende Funktion* zu(r Verteilung von) X.

Wir schreiben $f^{(k)}$ für die k-te Ableitung einer Funktion f.

SATZ 3.29 (a) *Ist X eine $\mathbb{N}_0$-wertige Zufallsvariable mit wahrscheinlichkeitserzeugender Funktion $\hat{p}$, so gilt für alle $k \in \mathbb{N}$: Das* k-te faktorielle Moment $E\big(X(X-1)\cdot\ldots\cdot(X-k+1)\big)$ *existiert genau dann, wenn* $\lim_{z\uparrow 1} \hat{p}^{(k)}(z)$ *existiert, und dann gilt*

$$EX(X-1)\cdot\ldots\cdot(X-k+1) = \lim_{z\uparrow 1} \hat{p}^{(k)}(z).$$

(b) *Sind X und Y unabhängige, $\mathbb{N}_0$-wertige Zufallsvariablen mit wahrscheinlichkeitserzeugenden Funktionen $\hat{p}_X$ und $\hat{p}_Y$, so gilt für die wahrscheinlichkeitserzeugende Funktion $\hat{p}_{X+Y}$ zur Summe $X+Y$:*

$$\hat{p}_{X+Y}(z) = \hat{p}_X(z)\,\hat{p}_Y(z) \quad \text{für alle } z \text{ mit } |z| \le 1.$$

BEWEIS: (a) Innerhalb des Konvergenzradius ist die Vertauschung von Summation und Differentiation erlaubt, d.h. es gilt

$$\hat{p}^{(k)}(z) = \sum_{n=k}^{\infty} n(n-1)\cdot\ldots\cdot(n-k+1)\, P(X=n)\, z^{n-k}.$$

Nach dem aus der Analysis bekannten Satz von Abel gilt für Potenzreihen $\sum_{n=0}^{\infty} a_n z^n$ mit nichtnegativen Koeffizienten

$$\lim_{z\uparrow 1} \sum_{n=0}^{\infty} a_n z^n = \sum_{n=0}^{\infty} a_n,$$

wobei bestimmte Divergenz zugelassen ist (d.h. genau dann kommt auf der einen Seite ∞ heraus, wenn dies auch für die andere Seite gilt). Schließlich gilt nach der letzten Formel in Satz 3.6

$$EX(X-1)\cdot\ldots\cdot(X-k+1) = \sum_{n=0}^{\infty} n(n-1)\cdot\ldots\cdot(n-k+1)\, P(X=n).$$

(b)

$$\begin{aligned} \hat{p}_{X+Y}(z) &= Ez^{X+Y} = Ez^X z^Y \\ &= Ez^X\, Ez^Y = \hat{p}_X(z)\,\hat{p}_Y(z). \end{aligned}$$

Hierbei haben wir verwendet, dass bei festem $|z| \le 1$ mit X und Y auch die Zufallsvariablen z^X und z^Y unabhängig sind (hierzu später mehr) und somit Satz 3.19 angewendet werden kann. □

BEISPIEL 3.30 (a) Ist X Poisson-verteilt mit Parameter $\lambda > 0$, so erhält man

$$\hat{p}_X(z) \;=\; \sum_{n=0}^{\infty} z^n\, e^{-\lambda}\, \frac{\lambda^n}{n!} \;=\; e^{-\lambda} \sum_{n=0}^{\infty} \frac{1}{n!}(\lambda z)^n \;=\; e^{\lambda(z-1)}.$$

Hieraus folgt

$$\hat{p}'_X(z) = \lambda \hat{p}_X(z), \quad \hat{p}''_X(z) = \lambda^2 \hat{p}_X(z),$$

mit Satz 3.29 (a) also

$$EX \;=\; \lim_{z\uparrow 1} \lambda e^{\lambda(z-1)} \;=\; \lambda, \quad EX(X-1) \;=\; \lim_{z\uparrow 1} \lambda^2 e^{\lambda(z-1)} \;=\; \lambda^2,$$

in Übereinstimmung mit Beispiel 3.11 (b). Ist Y eine weitere, von X unabhängige und mit Parameter μ Poisson-verteilte Zufallsvariable, so folgt mit Satz 3.29 (b)

$$\hat{p}_{X+Y}(z) \;=\; \hat{p}_X(z)\, \hat{p}_Y(z) \;=\; e^{\lambda(z-1)} e^{\mu(z-1)} \;=\; e^{(\lambda+\mu)(z-1)}\,.$$

Dies ist die wahrscheinlichkeitserzeugende Funktion zur Poisson-Verteilung mit Parameter $\lambda + \mu$. Da p durch $\hat{p}$ festgelegt ist, muss also die Zufallsvariable $X + Y$ wieder Poisson-verteilt sein, und zwar mit Parameter $\lambda + \mu$. Insgesamt haben wir damit einen alternativen Beweis für einen bereits in Beispiel 3.26 hergeleiteten Sachverhalt.

(b) Die obigen Aussagen lassen sich mit Induktion von zwei auf n Summanden übertragen. Sind beispielsweise $X_1, \ldots, X_n$ unabhängig und identisch verteilt (insbesondere haben sie dann dieselbe wahrscheinlichkeitserzeugende Funktion), so gilt

$$\hat{p}_{X_1+\cdots+X_n}(z) \;=\; p_{X_1}(z)^n.$$

Beim Würfelwurf ergibt sich so für die Augensumme $S = X_1 + \cdots + X_{10}$ von 10 Würfen die wahrscheinlichkeitserzeugende Funktion

$$\hat{p}_S(z) \;=\; \Big(\frac{1}{6}(z + z^2 + \cdots + z^6)\Big)^{10}.$$

Als Wahrscheinlichkeit für die Augensumme 35 erhält man nun mit den Maple-Befehlen

```
p := z -> (sum(z^k,k=1..6)/6)^10;
coeff(p(z),z,35);
```

den Wert

$$\frac{7631}{104976} \;\approx\; 0.0727.$$

3.7 Ungleichungen, das schwache Gesetz der großen Zahlen. Nach den Resultaten aus Abschnitt 3.5 gilt für den Mittelwert $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ von n unabhängigen Zufallsvariablen $X_1, \ldots, X_n$, die alle den Erwartungswert μ und die Varianz σ^2 haben,

$$E\bar{X}_n \;=\; \frac{1}{n}\sum_{i=1}^n \mu \;=\; \mu, \quad \operatorname{var}\bigl(\bar{X}_n\bigr) \;=\; \frac{1}{n^2}\sum_{i=1}^n \sigma^2 \;=\; \frac{\sigma^2}{n}$$

(wir haben hier die Rechenregel $\operatorname{var}(\alpha X) = \alpha^2 \operatorname{var}(X)$ benutzt, die Gegenstand einer Übungsaufgabe ist). Für große n ist also die Verteilung von $\bar{X}_n$ mit kleiner Variabilität um den Mittelwert herum konzentriert. Präzisere Aussagen ermöglichen Ungleichungen vom folgenden Typ.

Satz 3.31 (a) (Die Markovsche Ungleichung)
Es sei $p > 0$ und $E|X|^p < \infty$. Dann gilt

$$P\bigl(|X| \geq \alpha\bigr) \;\leq\; \frac{1}{\alpha^p}\, E|X|^p \quad \textit{für alle } \alpha > 0.$$

(b) (Die Chebyshevsche Ungleichung)
Es sei $EX^2 < \infty$. Dann gilt

$$P\bigl(|X - EX| \geq \alpha\bigr) \;\leq\; \frac{1}{\alpha^2}\operatorname{var}(X) \quad \textit{für alle } \alpha > 0.$$

Beweis: (a) Wir definieren eine neue (diskrete) Zufallsvariable Y durch

$$Y(\omega) \;:=\; \begin{cases} \alpha, & \bigl|X(\omega)\bigr| \geq \alpha, \\ 0 & \bigl|X(\omega)\bigr| < \alpha. \end{cases}$$

Offensichtlich gilt $|Y(\omega)|^p \leq |X(\omega)|^p$ für alle $\omega \in \Omega$, die Monotonieeigenschaft des Erwartungswertes (Satz 3.8) liefert also $E|Y|^p \leq E|X|^p$. Da Y nur die beiden Werte α und 0 annimmt, gilt gemäß Satz 3.6

$$E|Y|^p \;=\; 0^p P\bigl(|X| < \alpha\bigr) \;+\; \alpha^p P\bigl(|X| \geq \alpha\bigr).$$

Insgesamt erhält man also $\alpha^p P(|X| \geq \alpha) \leq E|X|^p$.
(b) Sei $Y = X - EX$. Wir verwenden Teil (a) mit $p = 2$:

$$P\bigl(|X - EX| \geq \alpha\bigr) \;=\; P\bigl(|Y| \geq \alpha\bigr) \;\leq\; \frac{1}{\alpha^2} EY^2 \;=\; \frac{1}{\alpha^2}\operatorname{var}(X).$$

□

Der folgende Satz ist eine einfache Version des schwachen Gesetzes der großen Zahlen.

SATZ 3.32 *Es sei $X_1, X_2, \ldots$ eine Folge von paarweise unkorrelierten Zufallsvariablen mit Erwartungswert μ und Varianz σ^2, $\bar{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$. Dann gilt*

$$P\big(|\bar{X}_n - \mu| \geq \epsilon\big) \;\to\; 0 \quad \textit{mit } n \to \infty \quad \textit{für alle } \epsilon > 0.$$

BEWEIS: Mit Satz 3.23 erhält man $\operatorname{var}(\bar{X}_n) = \sigma^2/n$, also folgt mit Chebyshev (Satz 3.31 (b))

$$P\big(|\bar{X}_n - \mu| \geq \epsilon\big) \;\leq\; \frac{1}{\epsilon^2}\operatorname{var}(\bar{X}_n) \;\to 0$$

mit $n \to \infty$ für jedes feste $\epsilon > 0$. □

Nimmt man also ein festes $\epsilon > 0$ (wie klein auch immer), so geht die Wahrscheinlichkeit dafür, dass der Mittelwert der Beobachtungen vom gemeinsamen Erwartungswert um mehr als ϵ abweicht, mit wachsendem n gegen 0. Ein Spezialfall ist der, bei dem X_i anzeigt, ob im i-ten Experiment ein bestimmtes Ereignis A eingetreten ist. Der obige Satz besagt dann, dass die relative Häufigkeit von A bei n Wiederholungen mit $n \to \infty$ in einem gewissen Sinn gegen die Wahrscheinlichkeit von A konvergiert: Die Wahrscheinlichkeit dafür, dass relative Häufigkeit und Wahrscheinlichkeit um mehr als ϵ ($\epsilon > 0$ fest) voneinander abweichen, wird bei hinreichend großer Anzahl von Versuchswiederholungen beliebig klein. Man kann dieses Resultat als eine (erste) Bestätigung des axiomatischen Aufbaus der Wahrscheinlichkeitstheorie durch die Kolmogorov-Axiome ansehen.

BEISPIEL 3.33 (Eine Anwendung in der Analysis)
Der Approximationssatz von Weierstraß besagt, dass eine stetige reellwertige Funktion auf einem kompakten Intervall $[a, b] \subset \mathbb{R}$ gleichmäßig durch Polynome approximiert werden kann. Wir wollen diesen Satz mit den Mitteln der Stochastik beweisen — sogar konstruktiv! Wir können $[a, b] = [0, 1]$ annehmen. Sei hierzu

$$p_n : [0,1] \to \mathbb{R}, \quad p_n(x) := \sum_{k=0}^{n} f\Big(\frac{k}{n}\Big)\binom{n}{k} x^k (1-x)^{n-k}$$

das n-te *Bernstein-Polynom* zu f. Wir behaupten:

$$\forall \epsilon > 0\; \exists n_0 \in \mathbb{N}\; \forall n \geq n_0\; \forall x \in [0,1] : \; \big|f(x) - p_n(x)\big| \leq \epsilon. \qquad (\star)$$

Sei also $\epsilon > 0$. Da eine stetige Funktion auf einem kompakten Intervall gleichmäßig stetig ist, existiert ein $\delta = \delta(\epsilon) > 0$ mit

$$\forall x, y \in [0,1] : \ |x - y| < \delta \ \Rightarrow \ \big|f(x) - f(y)\big| < \epsilon/2.$$

Außerdem sind stetige Funktionen auf kompakten Intervallen beschränkt, d.h. es gibt ein $K < \infty$ mit $|f(x)| \le K$ für alle $x \in [0,1]$. Nach diesen analytischen Vorbereitungen stellen wir nun wie folgt die Verbindung zur Stochastik her: Wähle $x \in [0,1]$. Wir betrachten den n-fach wiederholten Wurf einer Münze, die mit Wahrscheinlichkeit x das Resultat 1 und sonst 0 liefert. Bezeichnet X_i das Resultat des i-ten Wurfes, so ist $n\bar{X}_n$ die Anzahl der 1-Ergebnisse, also $\mathrm{Bin}(n,x)$-verteilt, und es folgt

$$Ef(\bar{X}_n) \ = \ \sum_{k=0}^{n} f\Big(\frac{k}{n}\Big)\, P\big(n\bar{X}_n = k\big) \ = \ p_n(x).$$

Wie im Beweis zu Satz 3.32 erhalten wir

$$P_n\big(|\bar{X}_n - x| \ge \delta\big) \ \le \ \frac{x(1-x)}{n\delta^2} \ \le \ \frac{1}{4n\delta^2}\,,$$

denn $x(1-x) \le 1/4$. Wähle nun $n_0 \in \mathbb{N}$ so groß, dass die Ungleichung $2K/(4n_0\delta^2) < \epsilon/2$ erfüllt ist. Für alle $n \ge n_0$ gilt dann

$$\begin{aligned}
\big|f(x) - p_n(x)\big| \ &= \ \big|Ef(\bar{X}_n) - f(x)\big| \\
&\le \ E\big|f(\bar{X}_n) - f(x)\big|\, 1_{\{|\bar{X}_n - x| < \delta\}} \\
&\qquad\qquad + \ E\big|f(\bar{X}_n) - f(x)\big|\, 1_{\{|\bar{X}_n - x| \ge \delta\}} \\
&\le \ \frac{\epsilon}{2}\, P\big(|\bar{X}_n - x| < \delta\big) \ + \ 2K\, P\big(|\bar{X}_n - x| \ge \delta\big) \\
&< \ \epsilon.
\end{aligned}$$

Damit ist $(\star)$ bewiesen. $\lhd$

4. Allgemeine Wahrscheinlichkeitsräume

4.1 Mengensysteme. In Abschnitt 2.3.4 haben wir gesehen, dass man bei überabzählbarem Ergebnisraum Ω in der Regel nicht mehr *allen* Teilmengen A von Ω eine Wahrscheinlichkeit zuordnen kann. Der Definitionsbereich von P soll aber häufig zumindest bestimmte Mengen enthalten, beispielsweise die Intervalle im Falle $\Omega = \mathbb{R}$. Wir beschäftigen uns in diesem Unterabschnitt zunächst ganz allgemein mit Mengensystemen.

DEFINITION 4.1 Es sei $\Omega \neq \emptyset$ und $\mathcal{E} \subset \mathcal{P}(\Omega)$. Dann heißt

$$\sigma(\mathcal{E}) := \bigcap_{\mathcal{A} \supset \mathcal{E},\, \mathcal{A}\ \sigma-\text{Algebra}} \mathcal{A}$$

die *von $\mathcal{E}$ erzeugte σ-Algebra*; $\mathcal{E}$ nennt man ein *Erzeugendensystem zu $\mathcal{A}$*.

In dieser Definition haben wir stillschweigend von der (trivialen) Tatsache Gebrauch gemacht, dass der Durchschnitt von beliebig vielen σ-Algebren über derselben Grundmenge wieder eine σ-Algebra ist. Der obige Durchschnitt ist übrigens nicht leer, denn es gilt $\mathcal{E} \subset \mathcal{P}(\Omega)$ und $\mathcal{P}(\Omega)$ ist eine σ-Algebra. Der für uns vorläufig wichtigste Fall ist $\Omega = \mathbb{R}$.

DEFINITION 4.2 Die von den LORA-Intervallen $(a, b]$, $-\infty < a < b < \infty$, erzeugte σ-Algebra heißt die σ-Algebra der *Borel-Mengen von* $\mathbb{R}$; Schreibweisen: $\mathcal{B}$, $\mathcal{B}(\mathbb{R})$ oder $\mathcal{B}_{\mathbb{R}}$.

Eine σ-Algebra $\mathcal{A}$ kann durchaus verschiedene Erzeugendensysteme haben, größere Mengensysteme erzeugen größere σ-Algebren und trivialerweise gilt $\sigma(\mathcal{A}) = \mathcal{A}$. Als 'general abstract nonsense' formuliert: Die Abbildung $\mathcal{E} \mapsto \sigma(\mathcal{E})$ ist isoton und idempotent, aber nicht injektiv.

SATZ 4.3 *Die σ-Algebra $\mathcal{B}(\mathbb{R})$ wird auch erzeugt von den Mengensystemen*

$$\begin{aligned}
\mathcal{E}_1 &:= \big\{[a, b) : -\infty < a < b < \infty\big\} \quad (\textit{den 'LARO-Intervallen'}),\\
\mathcal{E}_2 &:= \big\{(-\infty, a] : -\infty < a < \infty\big\},\\
\mathcal{E}_3 &:= \big\{U \subset \mathbb{R} : U \textit{ offen}\big\}.
\end{aligned}$$

BEWEIS: Es sei $\mathcal{E} := \{(a, b] : -\infty < a < b < \infty\}$ das Erzeugendendsystem aus der Definition von $\mathcal{B}$. Es reicht, jeweils $\mathcal{E}_i \subset \mathcal{B}$ und $\mathcal{E} \subset \sigma(\mathcal{E}_i)$ zu zeigen: Die erste Inklusion impliziert $\sigma(\mathcal{E}_i) \subset \mathcal{B}$, die zweite $\mathcal{B} (= \sigma(\mathcal{E})) \subset \sigma(\mathcal{E}_i)$. Hierbei

können wir die mengenalgebraischen Abgeschlossenheitseigenschaften von σ-Algebren gegenüber endlichen und abzählbar unendlichen Vereinigungen und Durchschnitten sowie Komplementen verwenden. In diesem Sinne ergibt sich $\sigma(\mathcal{E}_1) = \mathcal{B}$ aus

$$[a,b) \;=\; \bigcap_{n=1}^{\infty} \bigcup_{m=1}^{\infty} \Big(a - \frac{1}{n}, b - \frac{1}{m}\Big], \quad (a,b] \;=\; \bigcup_{n=1}^{\infty} \bigcap_{m=1}^{\infty} \Big[a + \frac{1}{n}, b + \frac{1}{m}\Big)$$

und $\sigma(\mathcal{E}_2) = \mathcal{B}$ folgt aus

$$(-\infty, a] \;=\; \bigcup_{n=1}^{\infty} (a-n, a], \quad (a,b] = (-\infty, b] \cap (-\infty, a]^{\mathrm{c}}.$$

Bei $\mathcal{E}_3$ verwenden wir, dass es zu jedem x aus einer offenen Menge U ein x enthaltendes Intervall $(a,b] \subset U$ gibt, von dem wir annehmen können, dass die Endpunkte rationale Zahlen sind:

$$U \;=\; \bigcup_{\{(a,b)\in\mathbb{Q}\times\mathbb{Q}:\,(a,b]\subset U\}} (a,b]\,.$$

Dies zeigt, dass jede offene Menge $U \subset \mathbb{R}$ als abzählbare Vereinigung von LORA-Intervallen dargestellt werden kann, also $\sigma(\mathcal{E}_3) \subset \mathcal{B}$. Die Gegenrichtung folgt aus der Darstellung

$$(a,b] = \bigcap_{n=1}^{\infty} \Big(a, b + \frac{1}{n}\Big)$$

und der bekannten Tatsache, dass offene Intervalle offene Mengen sind. □

Dieser Satz impliziert, dass die Intervalle $[a,b)$, $(-\infty,a]$ Borel-Mengen sind, ebenso wie alle offenen Mengen. Wegen

$$\{a\} \;=\; \bigcap_{n=1}^{\infty} \Big(a - \frac{1}{n}\,,\, a\Big]$$

sind auch alle Einpunktmengen und somit alle abzählbaren Mengen wie beispielsweise $\mathbb{Q}$ Borel-Mengen, damit auch kompakte Intervalle, die irrationalen Zahlen etc.; $\mathcal{B}$ ist für alle praktischen Zwecke reichhaltig genug.

Ist A eine nicht-leere Teilmenge von $\mathbb{R}$, so wird durch

$$\mathcal{B}_A \;=\; \{B \cap A :\, B \in \mathcal{B}\}$$

eine σ-Algebra über A definiert (Übungsaufgabe), die *Spur von $\mathcal{B}$ auf A*; wir nennen $\mathcal{B}_A$ auch das System der Borel-Mengen von A. In der Maßtheorie wird der folgende wichtige Satz bewiesen.

SATZ 4.4 *Es gibt ein Wahrscheinlichkeitsmaß P auf $\big([0,1),\mathcal{B}_{[0,1)}\big)$ mit der Eigenschaft*

$$P\big([a,b)\big) \;=\; b-a \qquad \textit{für alle } a,b \textit{ mit } 0\le a<b<1. \tag{$\star$}$$

BEMERKUNG 4.5 (a) Man kann zeigen, dass ($\star$) auf die Eigenschaft ($\star$) aus Abschnitt 2.3.4 führt; wir werden später sehen, dass (mit $\mathcal{B}_{[0,1)}$ anstelle von $\mathcal{A}$) auch die Gegenrichtung gilt. Satz 4.4 zeigt also, dass durch eine Verkleinerung des Definitionsbereiches, die für praktische Anwendungen bedeutungslos ist, tatsächlich das in Abschnitt 2.3.4 angesprochene Problem gelöst wird.

(b) Man kann P auf $(\mathbb{R},\mathcal{B}_{\mathbb{R}})$ fortsetzen durch

$$P_{\mathbb{R}}(B) \;:=\; P\big(B\cap[0,1)\big) \qquad \text{für alle } B\in\mathcal{B}_{\mathbb{R}}.$$

Umgekehrt erhält man aus einem Wahrscheinlichkeitsmaß P auf $(\mathbb{R},\mathcal{B}_{\mathbb{R}})$ ein Wahrscheinlichkeitsmaß $P_{[0,1)}$ auf $\big([0,1),\mathcal{B}_{[0,1)}\big)$ durch

$$P_{[0,1)}(B) \;:=\; P(B\cap[0,1)),$$

wenn nur $P([0,1)) = 1$ gilt. Das Intervall $[0,1)$ lässt sich hierbei durch ein $A\in\mathcal{B}$ mit $P(A)=1$ ersetzen. In diesem Sinne nennt man das Wahrscheinlichkeitsmaß P aus Satz 4.4 die *Gleichverteilung auf dem Einheitsintervall*, ohne i.a. zu spezifizieren, ob man $[0,1)$, $(0,1]$, $(0,1)$ oder $[0,1]$ meint, denn wegen

$$P\big(\{x\}\big) \;=\; \lim_{n\to\infty} P\Big(\Big[x,x+\frac{1}{n}\Big)\Big) \;=\; \lim_{n\to\infty}\Big(x+\frac{1}{n}-x\Big) \;=\; 0$$

spielen die Randpunkte keine Rolle. Man schreibt für P auch unif$(0,1)$, die 'uniforme' Verteilung; eine weitere Bezeichnung, deren Sinn später klar werden wird, ist *Rechteckverteilung*.

(c) In der Maßtheorie nennt man ein Paar $(\Omega,\mathcal{A})$, $\Omega\neq\emptyset$ und $\mathcal{A}$ eine σ-Algebra über Ω, einen *messbaren Raum*, und eine Abbildung $\mu:\mathcal{A}\to[0,\infty]$ ein *Maß*, wenn

$$\mu(\emptyset)=0, \quad \mu\Big(\sum_{i=1}^{\infty}A_i\Big)=\sum_{i=1}^{\infty}\mu(A_i)$$

für alle paarweise disjunkten $A_1,A_2,\ldots\in\mathcal{A}$ gilt. In diesem Sinne sind Wahrscheinlichkeiten ganz einfach normierte Maße. Die geometrische Variante des Problems aus Abschnitt 2.3.4 lautet: Lässt sich allen Teilmengen von $\mathbb{R}$ (oder allgemeiner $\mathbb{R}^d$) sinnvoll eine Länge (allgemeiner, ein Volumen) zuordnen? Es ist wieder eine Einschränkung des Definitionsbereiches nötig, und man erhält dann: Es gibt ein Mass ℓ (das *Lebesgue-Maß*) auf $(\mathbb{R},\mathcal{B})$ mit

$$\ell\big((a,b]\big) \;=\; b-a \qquad \text{für alle } a<b,\ a,b\in\mathbb{R}.$$

Man kann also unif$(0,1)$ als Einschränkung von ℓ auf das Einheitsintervall auffassen. ◁

Wir müssen uns nun mit dem Problem der Eindeutigkeit auseinandersetzen—ist beispielsweise unif$(0,1)$ durch $(\star)$ eindeutig bestimmt? Hierzu verwenden wir ein auch später sehr nützliches Hilfsmittel.

DEFINITION 4.6 Es sei Ω eine nicht-leere Menge. Dann heißt $\mathcal{D} \subset \mathcal{P}(\Omega)$ ein *Dynkin-System*, wenn gilt

(i) $\Omega \in \mathcal{D}$, (ii) $A \in \mathcal{D} \Rightarrow A^c \in \mathcal{D}$,

(iii) $A_1, A_2, \ldots \in \mathcal{D}$ mit $A_i \cap A_j = \emptyset$ für $i \neq j \quad \Longrightarrow \quad \bigcup_{i=1}^{\infty} A_i \in \mathcal{D}$.

Im Vergleich zu σ-Algebren wird also die Forderung der Abgeschlossenheit gegenüber beliebigen abzählbaren Vereinigungen auf disjunkte Vereinigungen abgeschwächt. Der Durchschnitt von beliebig vielen Dynkin-Systemen ist offensichtlich wieder ein Dynkin-System, wir können also von

$$\delta(\mathcal{E}) := \bigcap_{\mathcal{D} \supset \mathcal{E},\, \mathcal{D} \text{ Dynkin-System}} \mathcal{D}$$

als dem von $\mathcal{E}$ erzeugten Dynkin-System sprechen.

Dynkin-Systeme sind 'fast' σ-Algebren. Um dies präzisieren zu können, benötigen wir den folgenden Begriff: Wir nennen ein Mengensystem $\mathcal{E}$ *durchschnittsstabil* und schreiben kurz $\cap$-stabil, wenn gilt

$$A, B \in \mathcal{E} \quad \Longrightarrow \quad A \cap B \in \mathcal{E}.$$

Der folgende Satz zeigt, dass genau diese Eigenschaft den Schritt vom Dynkin-System zur σ-Algebra ermöglicht.

SATZ 4.7 (a) *Ein $\cap$-stabiles Dynkin-System ist eine σ-Algebra.*

(b) *Ist $\mathcal{E}$ $\cap$-stabil, so gilt $\delta(\mathcal{E}) = \sigma(\mathcal{E})$.*

BEWEIS: (a) Es seien $A_1, A_2, \ldots \in \mathcal{D}$ (nicht notwendigerweise disjunkt!). Wir wollen zeigen, dass $\bigcup_{n=1}^{\infty} A_n \in \mathcal{D}$ gilt und setzen hierzu $B_1 := A_1$,

$$B_n := A_n \cap A_1^c \cap \ldots \cap A_{n-1}^c \quad \big(= A_n \setminus (A_1 \cup \ldots \cup A_{n-1}) \big)$$

für alle $n > 1$. Durchschnittsstabilität und Eigenschaft (ii) liefern $B_n \in \mathcal{D}$ für alle $n \in \mathbb{N}$. Offensichtlich sind die B_n's disjunkt, also gilt nach Eigenschaft (iii) $\bigcup_{n=1}^{\infty} B_n \in \mathcal{D}$. Mit

$$\bigcup_{n=1}^{\infty} B_n = \bigcup_{n=1}^{\infty} A_n$$

folgt nun die gewünschte Aussage (eine ähnliche Konstruktion wurde bereits im Beweis von Satz 1.7 verwendet).

(b) Da jede σ-Algebra ein Dynkin-System ist, folgt $\delta(\mathcal{E}) \subset \sigma(\mathcal{E})$ unmittelbar aus den beteiligten Definitionen. Es sei nun, für jedes $A \in \delta(\mathcal{E})$,

$$\mathcal{D}_A := \big\{B \subset \Omega : B \cap A \in \delta(\mathcal{E})\big\}.$$

Dann ist $\mathcal{D}_A$ ein Dynkin-System: (i) und (iii) sind trivial, (ii) folgt mit

$$B^{\mathrm{c}} \cap A = (A^{\mathrm{c}} + B \cap A + \Omega^{\mathrm{c}} + \Omega^{\mathrm{c}} + \ldots)^{\mathrm{c}}.$$

Da $\mathcal{E}$ $\cap$-stabil ist, gilt $E' \in \mathcal{D}_E$ für alle $E, E' \in \mathcal{E}$, also $\mathcal{E} \subset \mathcal{D}_E$ und damit $\delta(\mathcal{E}) \subset \mathcal{D}_E$ für alle $E \in \mathcal{E}$, denn $\mathcal{D}_E$ ist ja ein Dynkin-System. Dies heißt

$$D \in \delta(\mathcal{E}),\ E \in \mathcal{E} \quad \Longrightarrow \quad D \cap E \in \delta(\mathcal{E}),$$

also $E \in \mathcal{D}_D$ für alle $E \in \mathcal{E}$, $D \in \delta(\mathcal{E})$. Dies wiederum liefert $\mathcal{E} \subset \mathcal{D}_D$, also $\delta(\mathcal{E}) \subset \mathcal{D}_D$ für alle $D \in \delta(\mathcal{E})$ und damit

$$A \in \delta(\mathcal{E}),\ D \in \delta(\mathcal{E}) \quad \Longrightarrow \quad A \cap D \in \delta(\mathcal{E}).$$

Also ist $\delta(\mathcal{E})$ $\cap$-stabil und $\delta(\mathcal{E}) \supset \sigma(\mathcal{E})$ folgt mit Teil (a). □

Satz 4.8 *Es sei $\mathcal{A}$ eine σ-Algebra mit $\cap$-stabilem Erzeuger $\mathcal{E}$. Sind dann P und Q Wahrscheinlichkeitsmaße auf $\mathcal{A}$ mit der Eigenschaft*

$$P(E) = Q(E) \quad \textit{für alle } E \in \mathcal{E},$$

so gilt

$$P(A) = Q(A) \quad \textit{für alle } A \in \mathcal{A}.$$

Beweis: Es sei

$$\mathcal{D} := \big\{A \in \mathcal{A} : P(A) = Q(A)\big\}.$$

Dann gilt $\mathcal{E} \subset \mathcal{D}$ und $\mathcal{D}$ ist, wie man leicht überprüft, ein Dynkin-System. Satz 4.7 (b) liefert nun

$$\mathcal{D} \supset \delta(\mathcal{E}) = \sigma(\mathcal{E}) = \mathcal{A}.$$

□

Stimmen also zwei Wahrscheinlichkeitsmaße auf einem $\cap$-stabilen Erzeuger überein, so sind sie gleich. Die Mengen $[a, b)$, $0 \leq a \leq b < 1$, bilden ein Erzeugendensystem von $\mathcal{B}_{[0,1)}$ (Übungsaufgabe); dieses ist offensichtlich $\cap$-stabil. Insbesondere gibt es also nur ein Wahrscheinlichkeitsmaß auf $\mathcal{B}_{[0,1)}$ mit der Eigenschaft $(\star)$ und wir können von *der* Gleichverteilung auf dem Einheitsintervall sprechen.

4.2 Zufallsgrößen und Verteilungen. Wie im diskreten Fall interessiert man sich auch im allgemeinen Fall oft nicht für das exakte Resultat $\omega \in \Omega$ eines Zufallsexperiments, sondern nur für den Wert $X(\omega)$ einer Funktion X hiervon, und es geht dann um die Wahrscheinlichkeit, dass X in einer bestimmten Menge landet. Da unser Wahrscheinlichkeitsmaß nun u.U. nicht mehr auf der gesamten Potenzmenge des Ergebnisraums definiert ist, ist nicht mehr automatisch gewährleistet, dass $P(X \in A)$ überhaupt 'legal' ist. Wir schreiben weiterhin $X \in A$ oder $X^{-1}(A)$ für $\{\omega \in \Omega : X(\omega) \in \mathcal{A}\}$.

Definition 4.9 Es seien $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum und $(\Omega', \mathcal{A}')$ ein messbarer Raum. Eine Abbildung $X : \Omega \to \Omega'$ heißt *Zufallsgröße* (auf $(\Omega, \mathcal{A}, P)$ und mit Werten in $(\Omega', \mathcal{A}')$), wenn X $(\mathcal{A}, \mathcal{A}')$-*messbar* ist, d.h. wenn gilt:

$$X^{-1}(A') \in \mathcal{A} \quad \text{für alle } A' \in \mathcal{A}'.$$

Für eine Zufallsgröße sind also die Wahrscheinlichkeiten dafür, dass ein Wert in einer messbaren Menge des Bildraums angenommen wird, definiert. Der Begriff Messbarkeit stammt (natürlich) aus der Maßtheorie. Die folgende Analogie zur Topologie ist gelegentlich hilfreich: Auf einer Menge M wird eine Topologie durch das System $\mathcal{U} \subset \mathcal{P}(U)$ der offenen Mengen beschrieben. Eine Abbildung $f : M \to M'$ von einem topologischen Raum $(M,\mathcal{U})$ in einen weiteren topologischen Raum $(M',\mathcal{U}')$ heißt stetig, wenn $f^{-1}(U') \in \mathcal{U}$ gilt für alle $U' \in \mathcal{U}'$. Also: Messbarkeit heißt, dass die Urbilder messbarer Mengen messbar sind, Stetigkeit heißt, dass die Urbilder offener Mengen offen sind. Natürlich ist im Falle $\mathcal{A} = \mathcal{P}(\Omega)$ die Bedingung $X^{-1}(A') \in \mathcal{A}$ sogar für alle $A' \in \mathcal{P}(\Omega')$ erfüllt — dies ist der Grund dafür, dass wir bei diskreten Wahrscheinlichkeitsräumen ohne den Messbarkeitsbegriff ausgekommen sind.

Es ist bekannt, dass Verknüpfungen stetiger Funktionen wieder stetig sind; der folgende Satz enthält den entsprechenden maßtheoretischen Sachverhalt.

Satz 4.10 *Es seien* $(\Omega, \mathcal{A})$, $(\Omega', \mathcal{A}')$, $(\Omega'', \mathcal{A}'')$ *messbare Räume sowie* $X : \Omega \to \Omega'$, $Y : \Omega' \to \Omega''$ $(\mathcal{A}, \mathcal{A}')$- *bzw.* $(\mathcal{A}', \mathcal{A}'')$-*messbare Abbildungen. Dann ist* $Z := Y \circ X$ $(\mathcal{A}, \mathcal{A}'')$-*messbar.*

BEWEIS: Für alle $A'' \in \mathcal{A}''$ gilt

$$\begin{aligned} Z^{-1}(A'') &= \big\{\omega \in \Omega : Y(X(\omega)) \in A''\big\} \\ &= X^{-1}\big(\{\omega' \in \Omega' : Y(\omega') \in A''\}\big) \\ &= X^{-1}\big(Y^{-1}(A'')\big) \in \mathcal{A}, \end{aligned}$$

denn $A' := Y^{-1}(A'') \in \mathcal{A}'$, $X^{-1}(A') \in \mathcal{A}$ gilt aufgrund der vorausgesetzten Messbarkeiten. □

Beim Nachweis der Messbarkeit kann man sich auf Erzeugendensysteme beschränken:

SATZ 4.11 *Es seien $(\Omega, \mathcal{A})$ und $(\Omega', \mathcal{A}')$ messbare Räume und $X : \Omega \to \Omega'$ eine Abbildung. Ist $\mathcal{E}' \subset \mathcal{P}(\Omega')$ ein Erzeugendensystem von $\mathcal{A}'$ und gilt*

$$X^{-1}(E') \in \mathcal{A} \quad \textit{für alle } E' \in \mathcal{E}',$$

so ist X $(\mathcal{A}, \mathcal{A}')$-messbar.

BEWEIS: Es sei $\mathcal{A}_0 = \{A' \subset \Omega' : X^{-1}(A') \in \mathcal{A}\}$. Dann ist $\mathcal{A}_0$ eine σ-Algebra über Ω': $X^{-1}(\Omega') = \Omega \in \mathcal{A}$, also gilt $\Omega' \in \mathcal{A}_0$. Weiter hat man

$$X^{-1}(A^{\mathrm{c}}) = \{\omega \in \Omega : X(\omega) \notin A\} = \big(\{\omega \in \Omega : X(\omega) \in A\}\big)^{\mathrm{c}} = \big(X^{-1}(A)\big)^{\mathrm{c}},$$

also gilt

$$\begin{aligned} A \in \mathcal{A}_0 &\implies X^{-1}(A) \in \mathcal{A} \implies \big(X^{-1}(A)\big)^{\mathrm{c}} \in \mathcal{A} \\ &\implies X^{-1}(A^{\mathrm{c}}) \in \mathcal{A} \implies A^{\mathrm{c}} \in \mathcal{A}_0. \end{aligned}$$

Analog erhält man mit

$$X^{-1}\Big(\bigcup_{n=1}^{\infty} A_n\Big) = \bigcup_{n=1}^{\infty} X^{-1}(A_n)$$

die dritte definierende Eigenschaft einer σ-Algebra. Nach Voraussetzung gilt $\mathcal{E}' \subset \mathcal{A}_0$, also $\mathcal{A}' = \sigma(\mathcal{E}') \subset \mathcal{A}_0$ und damit $X^{-1}(A') \in \mathcal{A}$ für alle $A' \in \mathcal{A}'$. □

Schließlich haben wir die folgende Verallgemeinerung von Satz 3.2.

SATZ UND DEFINITION 4.12 *Ist X eine $(\Omega', \mathcal{A}')$-wertige Zufallsgröße auf $(\Omega, \mathcal{A}, P)$, so wird durch*

$$\mathcal{A}' \ni A' \;\mapsto\; P(X \in A') \quad \Big(= P\big(\{\omega \in \Omega : X(\omega) \in A'\}\big)\Big)$$

ein Wahrscheinlichkeitsmaß auf $(\Omega', \mathcal{A}')$ definiert. Dieses Wahrscheinlichkeitsmaß heißt die Verteilung *von X, Schreibweisen: P^X oder $\mathcal{L}(X)$.*

Bei Beachtung der Messbarkeit ist der Beweis identisch zum Beweis im diskreten Fall. In der Sprache der Maßtheorie ist die Verteilung einer Zufallsgröße das durch die messbare Abbildung auf dem Bildraum induzierte Bildmaß.

BEISPIEL 4.13 Es sei $(\Omega, \mathcal{A}, P) = \big([0,1), \mathcal{B}_{[0,1)}, \mathrm{unif}(0,1)\big)$. Für jedes $x \in \Omega$ werde $T_x : \Omega \to \Omega$ definiert durch

$$T_x(y) \;:=\; \begin{cases} y - x, & \text{wenn } y \geq x, \\ y - x + 1, & \text{wenn } y < x. \end{cases}$$

Für alle $A \in \mathcal{A}$ gilt dann

$$T_x^{-1}(A) \;=\; \{y \in \Omega : y - x \in A \text{ oder } y - x + 1 \in A\} \;=\; x + A \pmod 1,$$

insbesondere also

$$T_x^{-1}\big([0,a)\big) \;=\; \begin{cases} [x, x+a), & \text{wenn } x + a \leq 1, \\ [0, x+a-1) \cup [x, 1), & \text{wenn } x + a > 1 \end{cases} \quad \in \mathcal{A}\,.$$

Mit $\sigma\big(\{[0,a) : 0 < a \leq 1\}\big) = \mathcal{A}$ und Satz 4.11 folgt hieraus die $(\mathcal{A}, \mathcal{A})$-Messbarkeit von T_x. Man sieht auch, dass

$$P\big(T_x^{-1}([0,a))\big) \;=\; a \;=\; P([0,a))$$

für alle $a \in (0,1]$ gilt, mit Satz 4.8 folgt also $P^{T_x} = P$. Dies wiederum liefert

$$P(x + A) = P(A) \quad \text{für alle } A \in \mathcal{A},$$

d.h. das Wahrscheinlichkeitsmaß $\mathrm{unif}(0,1)$ hat die Eigenschaft $(\star)$ (Translationsinvarianz modulo 1). ◁

4.3 Reellwertige Zufallsgrößen. Wie in der in Abschnitt 3 behandelten diskreten Situation verdient der Fall, in dem $\mathbb{R}$ der Wertebereich der Zufallsgrößen ist, besondere Beachtung. Eine reellwertige Zufallsgröße nennen wir auch Zufallsvariable (kurz: ZV). Es sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum; als σ-Algebra auf $\mathbb{R}$ werden wir grundsätzlich die σ-Algebra $\mathcal{B}$ der Borel-Mengen nehmen. Aus Satz 4.3 und Satz 4.11 folgt unmittelbar, dass $X : \Omega \to \mathbb{R}$ genau dann eine Zufallsvariable, also $(\mathcal{A}, \mathcal{B})$-messbar ist, wenn $X^{-1}((-\infty, a]) \in \Omega$ für alle $a \in \mathbb{R}$ erfüllt ist. Den einfachsten Fall solcher Abbildungen liefern die Indikatorfunktionen: Wegen

$$1_A^{-1}\big((-\infty, a]\big) \;=\; \begin{cases} \emptyset, & a < 0, \\ A^{\mathrm{c}}, & 0 \leq a < 1, \\ \Omega, & a \geq 1, \end{cases}$$

ist 1_A genau dann eine Zufallsvariable, wenn $A \in \mathcal{A}$ gilt. Durch den Übergang $A \mapsto 1_A$ werden also die messbaren Mengen in den Raum der messbaren Abbildungen eingebettet.

Häufig werden mit einer Zufallsvariablen X Operationen ausgeführt, im Zusammenhang mit der Streuung ist beispielsweise X^2 interessant. Ist X^2 wieder eine Zufallsvariable?

SATZ 4.14 *Ist $g : \mathbb{R} \to \mathbb{R}$ stetig oder (schwach) monoton steigend oder fallend, so ist g $(\mathcal{B}, \mathcal{B})$-messbar.*

BEWEIS: Ist g stetig, so ist $g^{-1}(U)$ für jede offene Menge offen, also in $\mathcal{B}$. Hieraus folgt die Behauptung mit Satz 4.3 und Satz 4.11. Der Beweis für monotone Funktionen g ist Gegenstand einer Übungsaufgabe. □

Ist X eine Zufallsvariable, so kann X^2 als Verknüpfung der $(\mathcal{A}, \mathcal{B})$-messbaren Abbildung X und der $(\mathcal{B}, \mathcal{B})$-messbaren, weil stetigen, Abbildung $g : \mathbb{R} \to \mathbb{R}$, $g(x) = x^2$, angesehen werden, ist nach Satz 4.10 also $(\mathcal{A}, \mathcal{B})$-messbar und damit wieder eine Zufallsvariable. Wird eine neue Abbildung aus mehreren Zufallsvariablen zusammengesetzt, so lässt sich häufig der folgende Satz anwenden.

SATZ 4.15 (a) *Sind X und Y Zufallsvariablen auf $(\Omega, \mathcal{A}, P)$, so liegen die Mengen $\{X < Y\}$, $\{X \leq Y\}$, $\{X = Y\}$ und $\{X \neq Y\}$ in $\mathcal{A}$ (hierbei steht $\{X < Y\}$ für die Menge $\{\omega \in \Omega : X(\omega) < Y(\omega)\}$ etc.).*

(b) *Sind X, Y Zufallsvariablen auf $(\Omega, \mathcal{A}, P)$ und $\alpha, \beta \in \mathbb{R}$, so sind auch*

$$\alpha X + \beta, \quad X + Y, \quad X \cdot Y, \quad X \wedge Y, \quad X \vee Y$$

Zufallsvariablen. ($a \wedge b := \min\{a, b\}$, $a \vee b := \max\{a, b\}$)

(c) *Ist $(X_n)_{n\in\mathbb{N}}$ eine Folge von Zufallsvariablen auf $(\Omega, \mathcal{A}, P)$, so sind auch*

$$\sup_{n\in\mathbb{N}} X_n, \quad \inf_{n\in\mathbb{N}} X_n, \quad \limsup_{n\to\infty} X_n, \quad \liminf_{n\to\infty} X_n$$

Zufallsvariablen (vorausgesetzt, diese Größen sind $\mathbb{R}$-wertig). Gilt $X_n(\omega) \to X(\omega)$ für alle $\omega \in \Omega$, so ist auch X eine Zufallsvariable.

BEWEIS: (a) Durch $\{X < Y\} = \bigcup_{q\in\mathbb{Q}} \{X < q\} \cap \{Y > q\}$ wird die Menge $\{X < Y\}$ als zugelassene Kombination messbarer Mengen dargestellt. Wegen $\{X \leq Y\} = \{Y < X\}^c$, $\{X = Y\} = \{X \leq Y\} \cap \{X < Y\}^c$, $\{X \neq Y\} = \{X = Y\}^c$ liegen dann auch die anderen Mengen in $\mathcal{A}$.

(b) Die Abbildung $x \to \alpha x + \beta$ ist stetig, also ist $\alpha X + \beta$ als Verknüpfung messbarer Abbildungen messbar (siehe auch das obige Argument für X^2). Weiter erhält man mit dem bereits bewiesenen Teil (a)

$$\{X + Y \leq a\} = \{X \leq a - Y\} \in \mathcal{A} \quad \text{für alle } a \in \mathbb{R},$$

denn $a - Y$ ist ein Zufallsvariable, folglich ist $X + Y$ messbar. Mit

$$X \cdot Y = \frac{1}{4}\big((X + Y)^2 - (X - Y)^2\big)$$

folgt dann auch die Messbarkeit von $X \cdot Y$, mit

$$\{X \vee Y \leq a\} = \{X \leq a\} \cap \{Y \leq a\}, \quad \{X \wedge Y \leq a\} = \{X \leq a\} \cup \{Y \leq a\}$$

die von $X \vee Y$ und $X \wedge Y$ (hierbei haben wir wiederholt verwendet, dass X $(\mathcal{A}, \mathcal{B})$-messbar ist, wenn $\{X \leq a\} \in \mathcal{A}$ gilt für alle $a \in \mathbb{R}$).

(c) Ähnlich wie bei Teil (b) erhält man

$$\Big\{\sup_{n\in\mathbb{N}} X_n \leq a\Big\} = \bigcap_{n=1}^{\infty} \{X_n \leq a\} \in \mathcal{A}.$$

Die Messbarkeit der anderen Abbildungen ergibt sich nun mit

$$\begin{aligned} \inf_{n\in\mathbb{N}} X_n &= -\sup_{n\in\mathbb{N}}(-X_n), \\ \limsup_{n\to\infty} X_n &= \inf_{n\in\mathbb{N}} \sup_{m\geq n} X_m, \\ \liminf_{n\to\infty} X_n &= \sup_{n\in\mathbb{N}} \inf_{m\geq n} X_m. \end{aligned}$$

Konvergiert X_n mit $n \to \infty$ punktweise gegen X, so gilt $X = \limsup_{n\to\infty} X_n$, also ist X eine Zufallsvariable. □

Im Teil (c) lässt sich die Einschränkung auf $\mathbb{R}$-wertige Abbildungen beseitigen, wenn man $\mathbb{R}$ zu $\bar{\mathbb{R}} := \mathbb{R} \cup \{-\infty\} \cup \{+\infty\}$ $\big(= [-\infty, \infty]\big)$ erweitert und auch $\mathcal{B}$ passend ergänzt zu $\mathcal{B}(\bar{\mathbb{R}}) := \sigma\big(\mathcal{B} \cup \{\{-\infty\}, \{\infty\}\}\big)$.

4.4 Verteilungsfunktionen. Die Verteilung einer reellwertigen Zufallsgröße ist ein Wahrscheinlichkeitsmaß auf $(\mathbb{R}, \mathcal{B})$, also eine Abbildung von $\mathcal{B}$ nach $[0, 1]$. Wir wollen nun zeigen, dass sich solche Wahrscheinlichkeitsmaße durch Abbildungen von $\mathbb{R}$ nach $[0, 1]$ beschreiben lassen.

Definition 4.16 Die *Verteilungsfunktion* F zu einem Wahrscheinlichkeitsmaß P auf $(\mathbb{R}, \mathcal{B})$ wird definiert durch

$$F : \mathbb{R} \to \mathbb{R}, \quad F(x) := P\big((-\infty, x]\big) \quad \text{für alle } x \in \mathbb{R}.$$

Ist P die Verteilung einer Zufallsvariablen X, so nennen wir F auch die Verteilungsfunktion zu X.

Da die Mengen $(-\infty, x]$, $x \in \mathbb{R}$, ein $\cap$-stabiles Erzeugendensystem von $\mathcal{B}$ bilden (Satz 4.3), wird P durch das zugehörige F eindeutig festgelegt (Satz 4.8).

SATZ 4.17 *Ist F die Verteilungsfunktion zu einem Wahrscheinlichkeitsmaß P auf $(\mathbb{R}, \mathcal{B})$, so hat F die folgenden Eigenschaften:*

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

(ii) *F ist (schwach) monoton steigend,*

(iii) *F ist stetig von rechts.*

BEWEIS: (ii) folgt unmittelbar aus der Monotonie von P (siehe Satz 1.6 (d)).
(i): Sei $(x_n)_{n\in\mathbb{N}} \subset \mathbb{R}$ mit $\lim_{n\to\infty} x_n = -\infty$ (d.h. $\forall c \in \mathbb{R}\ \exists n_0 \in \mathbb{N}\ \forall n \geq n_0 : x_n \leq c$). Setze $y_n := \sup_{m\geq n} x_m$. Dann gilt $y_n \downarrow -\infty$, also $(-\infty, y_n] \downarrow \emptyset$, und es folgt mit der Stetigkeit von P in $\emptyset$ (Satz 1.7 (d))

$$0 \leq F(x_n) = P\big((-\infty, x_n]\big) \leq P\big((-\infty, y_n]\big) \to 0$$

mit $n \to \infty$. Die andere Aussage erhält man analog mit der Stetigkeit von P von unten (in $\mathbb{R}$, Satz 1.7 (b)).
(iii) Ist $(x_n)_{n\in\mathbb{N}} \subset \mathbb{R}$ mit $x_n \geq x$ für alle $n \in \mathbb{N}$ und $x_n \to x$, so gilt $y_n \downarrow x$ für $y_n := \sup_{m\geq n} x_m$, also

$$\begin{aligned} F(x) &= P\big((-\infty, x]\big) \leq P\big((-\infty, x_n]\big) \\ &= F(x_n) \leq P\big((-\infty, y_n]\big) \to P((-\infty, x]) = F(x), \end{aligned}$$

wobei wir wieder eine Stetigkeitseigenshaft von P verwendet haben. □

Wir wollen nun zeigen, dass die obige Liste vollständig ist, d.h. dass zu jeder Funktion F mit den Eigenschaften (i)-(iii) ein Wahrscheinlichkeitsmaß P existiert, dessen Verteilungsfunktion F ist.

DEFINITION 4.18 Es sei F eine Funktion mit den Eigenschaften (i)-(iii) aus Satz 4.17. Dann definieren wir die *Quantilfunktion* Q zu F durch

$$Q : (0,1) \to \mathbb{R}, \quad Q(y) := \inf\big\{x \in \mathbb{R} : F(x) \geq y\big\}.$$

Wir schreiben auch F^{-1} für die Quantilfunktion zu F.

Ist X eine Zufallsvariable mit Verteilungsfunktion F, so nennt man $F^{-1}(\alpha)$ $(0 < \alpha < 1)$ das *α-Quantil* zu X (bzw. $\mathcal{L}(X)$ oder F); es ist dies der kleinste Wert q_α mit der Eigenschaft, dass der Wert von X mit Mindestwahrscheinlichkeit α nicht größer ist. Nur wenn F stetig und streng monoton wachsend ist, ist F^{-1} die Umkehrfunktion von F im üblichen Sinne.

LEMMA 4.19 $\quad y \le F(x) \iff F^{-1}(y) \le x.$

BEWEIS: '$\Rightarrow$' folgt unmittelbar aus der Definition von F^{-1}. Da außerdem

$$\begin{aligned} F(x) < y &\implies F\Big(x + \frac{1}{n}\Big) < y \quad \text{für ein } n \in \mathbb{N} \quad (\text{denn } F \text{ ist stetig von rechts}) \\ &\implies F^{-1}(y) \ge x + \frac{1}{n} \quad (\text{denn } F \text{ ist schwach monoton steigend}) \\ &\implies F^{-1}(y) > x \end{aligned}$$

gilt, hat man auch die Gegenrichtung. □

SATZ 4.20 *Es sei $F : \mathbb{R} \to \mathbb{R}$ eine Funktion mit den Eigenschaften* (i)-(iii) *aus Satz 4.17. Dann existiert ein Wahrscheinlichkeitsmaß P auf $(\mathbb{R}, \mathcal{B})$ mit Verteilungsfunktion F.*

BEWEIS: Es sei $\Omega = (0,1)$, $\mathcal{A} = \mathcal{B}_{(0,1)}$ und $P_0 = \text{unif}(0,1)$. Wir definieren $X : \Omega \to \mathbb{R}$ durch $X(\omega) := F^{-1}(\omega)$. Dann ist X eine Zufallsvariable (nach einer Übungsaufgabe folgt Messbarkeit von F^{-1} aus der Monotonie von F^{-1}), und Lemma 4.19 liefert für $P := \mathcal{L}(X)$

$$\begin{aligned} P\big((-\infty, x]\big) &= P_0(X \le x) \\ &= P_0\big(\{\omega \in \Omega : F^{-1}(\omega) \le x\}\big) \\ &= P_0\big((0, F(x)]\big) = F(x). \end{aligned}$$

□

Der Übergang von $P : \mathcal{B} \to \mathbb{R}$ zu $F : \mathbb{R} \to \mathbb{R}$, der letzlich durch die spezielle Struktur von $(\mathbb{R}, \mathcal{B})$ ermöglicht wird, bedeutet eine erhebliche Vereinfachung. Satz 4.20 zeigt auch, dass es zu jedem Wahrscheinlichkeitsmaß auf $(\mathbb{R}, \mathcal{B})$ eine Zufallsvariable mit diesem Wahrscheinlichkeitsmaß als Verteilung gibt.

In den Übungen wird gezeigt, dass Verteilungsfunktionen linksseitige Limiten haben, d.h. für alle $x \in \mathbb{R}$ existiert

$$F(x-) := \lim_{y \uparrow x, y < x} F(y),$$

und dass die Wahrscheinlichkeit, mit der X einen Wert x annimmt, durch die Sprunghöhe $F(x) - F(x-)$ von F in x gegeben wird. Insbesondere besteht die Verteilungsfunktion zu einer diskreten Zufallsvariablen nur aus Sprüngen. Ist $f : \mathbb{R} \to \mathbb{R}$ eine Funktion mit $\int_{-\infty}^{\infty} f(x)\,dx = 1$, so wird nach den obigen Resultaten durch

$$P\big((-\infty, x]\big) := \int_{-\infty}^{x} f(y)\,dy \quad \text{für alle } x \in \mathbb{R}$$

ein Wahrscheinlichkeitsmaß auf $(\mathbb{R}, \mathcal{B})$ definiert, das Wahrscheinlichkeitsmaß mit der *Riemann-Dichte* f. Hat die Zufallsvariable X eine solche Verteilung P, so nennen wir f eine *Wahrscheinlichkeitsdichte* von X. Zufallsvariablen mit einer Dichte werden gelegentlich 'stetig' genannt (als Gegensatz zu 'diskret') — dies bezieht sich *nicht* auf X als Abbildung, sondern ist nur als Abkürzung von 'X ist absolutstetig verteilt' zu verstehen. Ist f stetig in x, so ist die zugehörige Verteilungsfunktion F,

$$F(x) \;=\; \int_{-\infty}^{x} f(y)\,dy \qquad \text{für alle } \; x \in \mathbb{R},$$

in x differenzierbar, und es gilt $F'(x) = f(x)$.

Beispiel 4.21 Im Falle $P = \mathrm{unif}(0,1)$ hat man

$$P\big((-\infty, x]\big) \;=\; \int_{-\infty}^{x} f(y)\,dy \quad \text{für alle } x \in \mathbb{R}$$

mit

$$f(y) \;=\; \begin{cases} 1, & 0 < y < 1 \\ 0, & \text{sonst} \end{cases} \qquad \Big(= 1_{(0,1)}(y) \Big).$$

◁

Wahrscheinlichkeitsdichten sind in mancher Hinsicht ein infinitesimales Analogon zu Wahrscheinlichkeitsmassenfunktionen, können aber durchaus Werte größer als 1 annehmen. Ganz allgemein gilt für eine Zufallsvariable X mit Dichte f:

$$P(X \in A) \;=\; \int_A f(x)\,dx,$$

die Wahrscheinlichkeiten ergeben sich also als Fläche unter der Dichtefunktion. Da wir hier nur das Riemann-Integral voraussetzen, macht die rechte Seite nicht für alle Borel-Mengen Sinn — dies wird erst durch den (in der Maßtheorie bzw. der Stochastik II ausgeführten) Übergang zum Lebesgue-Integral erreicht.

4.5 Einige wichtige Verteilungen mit Riemann-Dichten.

4.5.1 Die Funktion

$$f_{a,b} : \mathbb{R} \to \mathbb{R}, \quad f_{a,b}(x) = \begin{cases} 1/(b-a), & a < x < b, \\ 0, & \text{sonst}, \end{cases}$$

hat für alle $a, b \in \mathbb{R}$ mit $a < b$ die Eigenschaften

$$f_{a,b}(x) \geq 0 \;\; \text{für alle } x \in \mathbb{R}, \quad \int_{-\infty}^{\infty} f_{a,b}(x)\,dx = 1,$$

ist also Dichte eines Wahrscheinlichkeitsmaßes auf $(\mathbb{R}, \mathcal{B})$. Wir nennen dieses Wahrscheinlichkeitsmaß die *Gleich-* oder *Rechteckverteilung* auf dem Intervall (a, b) (die Randpunkte spielen keine Rolle) und schreiben hierfür $\mathrm{unif}(a, b)$. Offensichtlich verallgemeinert dies die zu Beginn dieses Abschnitts eingeführte Gleichverteilung auf dem Einheitsintervall. Alle diese Verteilungen gehen durch affine Transformationen auseinander hervor: Hat X die Verteilung $\mathrm{unif}(0, 1)$, so gilt für die Zufallsvariable $Y := a + (b - a)X$

$$\begin{aligned} P(Y \le y) &= P\Big(X \le \frac{y-a}{b-a}\Big) = \frac{y-a}{b-a} \quad \text{für } a < y < b, \\ P(Y \le y) &= 0 \quad \text{für } y \le a, \\ P(Y \le y) &= 1 \quad \text{für } y \ge b, \end{aligned}$$

also insgesamt

$$P(Y \le y) = \int_{-\infty}^{y} f_{ab}(x)\,dx \quad \text{für alle } y \in \mathbb{R},$$

d.h. $Y \sim \mathrm{unif}(a, b)$. (Wir haben Satz 4.15 (b) verwendet.)

Beispiel 4.22 Ein Stab der Länge 1 zerbricht an einer zufälligen Stelle. Wir machen die (einigermaßen unrealistische) Annahme, dass alle Bruchpositionen gleich wahrscheinlich sind und erhalten dann als Modell für dieses Zufallsexperiment den Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ mit $\Omega = (0, 1)$, $\mathcal{A} = \mathcal{B}_{(0,1)}$ und $P = \mathrm{unif}(0, 1)$. Die Länge des kürzeren Stücks ist $X(\omega) = \min\{\omega, 1 - \omega\}$, nach Satz 4.15 ist dies eine Zufallsvariable. Welche Verteilung hat X? Offensichtlich gilt $P(X \le x) = 0$ für $x < 0$ und $P(X \le x) = 1$ für $x \ge 1/2$, und für $x \in (0, 1/2)$ erhält man

$$\begin{aligned} P(X \le x) &= P\big(\{\omega \in (0,1) : \omega \le x \text{ oder } 1 - \omega \le x\}\big) \\ &= P\big((0, x] \cup [1 - x, 1)\big) = 2x. \end{aligned}$$

Dies ist die Verteilungsfunktion zu $\mathrm{unif}(0, 1/2)$, also ist X wieder gleichverteilt, nun auf dem Intervall $(0, 1/2)$. ◁

4.5.2 Die *Gamma-Verteilung* mit Parametern α und λ ($\alpha > 0$, $\lambda > 0$) ist die Verteilung mit der Dichte

$$f_{\alpha,\lambda}(x) = \begin{cases} \frac{1}{\Gamma(\alpha)} x^{\alpha-1} \lambda^\alpha e^{-\lambda x}, & x > 0, \\ 0, & x \le 0, \end{cases}$$

wobei $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\,dx$ die Gamma-Funktion bezeichnet. Wir schreiben hierfür auch $\Gamma(\alpha, \lambda)$ und kurz $X \sim \Gamma(\alpha, \lambda)$, wenn die Zufallsvariable X diese Verteilung hat. Diese Klasse von Wahrscheinlichkeitsmaßen taucht in verschiedenen Zusammenhängen auf. Besonders wichtig ist der Fall $\alpha = 1$, der auf die *Exponentialverteilungen* führt (diese werden in einer Übungsaufgabe näher behandelt).

4.5.3 Die *Normalverteilung* mit Parametern μ und σ^2, kurz $N(\mu, \sigma^2)$, wobei $\mu \in \mathbb{R}$ beliebig und $\sigma^2 > 0$, ist die Verteilung mit der Dichte

$$\phi_{\mu,\sigma^2}(x) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}(x-\mu)^2\Big), \quad x \in \mathbb{R}.$$

Als Graph erhält man die berühmte Gaußsche Glockenkurve; die Parameter μ und σ beschreiben die Lage und Breite von ϕ. Im Falle $\mu = 0$, $\sigma^2 = 1$ spricht man von den Standardparametern, $N(0,1)$ ist die *Standardnormalverteilung*. Offensichtlich gilt

$$\phi_{\mu,\sigma^2}(x) \;=\; \frac{1}{\sigma}\phi_{0,1}\Big(\frac{x-\mu}{\sigma}\Big) \quad \text{für alle } x \in \mathbb{R}.$$

Die Verteilungsfunktion zu $N(0,1)$ ist Φ,

$$\Phi : \mathbb{R} \to [0,1], \quad \Phi(x) := \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}}\, e^{-y^2/2}\, dy.$$

Eine Variante hiervon ist auch als 'Fehlerfunktion' bekannt. Die Funktion Φ ist vertafelt und in gängigen Softwarepaketen enthalten. Die statistischen Anwendungen sind die zugehörige α-Quantile von Bedeutung; für $\alpha = 0.9$, 0.95, 0.975, 0.99, 0.995 erhält man die Werte 1.2816, 1.6449, 1.9600, 2.3263 und 2.5758.

LEMMA 4.23 (a) $\int_{-\infty}^{\infty} \phi_{\mu,\sigma^2}(x)\, dx = 1$ *für alle* $\mu \in \mathbb{R}$, $\sigma^2 > 0$,
(b) $\Phi(x) = 1 - \Phi(-x)$ *für alle* $x \in \mathbb{R}$,
(c) $X \sim N(\mu, \sigma^2)$, $a \neq 0$, $b \in \mathbb{R} \implies Y := aX + b \sim N(a\mu + b, a^2\sigma^2)$.

BEWEIS: (a) Substitution $y = \sigma^{-1}(x - \mu)$ zeigt, dass es reicht, den Fall $\mu = 0$, $\sigma^2 = 1$ zu behandeln. Standardtechniken der Analysis (Transformation auf Polarkoordinaten) ergeben

$$\begin{aligned}\Big(\int_{-\infty}^{\infty} e^{-x^2/2}\, dx\Big)^2 \;&=\; \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)/2}\, dx\, dy \\ &=\; \int_0^{2\pi}\int_0^{\infty} r\, e^{-r^2/2}\, dr\, d\phi \\ &=\; \int_0^{2\pi} \big(-e^{-r^2/2}\big)\Big|_0^{\infty} d\phi \;=\; 2\pi.\end{aligned}$$

(b) folgt mit $\phi(-x) = \phi(x)$.

(c) Im Falle $a > 0$ erhält man mit der Substitution $x' = ax + b$

$$\begin{aligned} P(Y \le y) &= P\Big(X \le \frac{y-b}{a}\Big) \\ &= \int_{-\infty}^{\frac{y-b}{a}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}(x-\mu)^2\Big)\, dx \\ &= \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi\sigma^2 a^2}} \exp\Big(-\frac{1}{2\sigma^2 a^2}\big(x' - (a\mu+b)\big)^2\Big) dx' \,. \end{aligned}$$

Dies zeigt, dass die Verteilungsfunktion zu Y die Verteilungsfunktion zu $N(a\mu + b, a^2\sigma^2)$ ist, also $Y \sim N(a\mu + b, a^2\sigma^2)$ gilt. □

Teil (a) ist ein Nachtrag: ϕ_{μ,σ^2} ist tatsächlich eine Wahrscheinlichkeitsdichte. Wegen (b) und (c) reicht es, die Verteilungsfunktionen zu $N(\mu, \sigma^2)$ für die Standardparameter und Argumente ≥ 0 zu vertafeln; beispielsweise gilt $u_\alpha = -u_{1-\alpha}$ für die Quantile u_α zu $N(0,1)$. In Kombination mit den oben genannten Quantilen ergibt sich als typische Anwendung von Lemma 4.23 (b) und (c) die Aussage, dass

$$P\big(|X - \mu| > 1.96\sigma\big) \approx 0.05$$

gilt, wenn X normalverteilt ist mit Parametern μ und σ^2.

Eines der wichtigsten Resultate der Stochastik, der Zentrale Grenzwertsatz, besagt, dass Normalverteilungen unter bestimmten, recht allgemeinen Bedingungen als Grenzwerte bei (standardisierten) Summen von unabhängigen Zufallsvariablen auftauchen. Dieses Thema wird in der Stochastik II im Detail behandelt; wir begnügen uns hier mit einem wichtigen Spezialfall und verzichten beim Beweis auf die vollständige Ausarbeitung der technischen Details.

Satz 4.24 (de Moivre-Laplace)
Es sei $(X_n)_{n\in\mathbb{N}}$ eine Folge von Zufallsgrößen mit $X_n \sim \mathrm{Bin}(n,p)$ für alle $n \in \mathbb{N}$, mit einem festen p, $0 < p < 1$. Dann gilt für alle $a, b \in \mathbb{R}$ mit $a < b$

$$\lim_{n\to\infty} P\Big(a \le \frac{X_n - np}{\sqrt{np(1-p)}} \le b\Big) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\, dx.$$

Beweisskizze: Wir setzen $\sigma_n^2 := np(1-p)$ und $x_n(k) := \sigma_n^{-1}(k - np)$. Dann gilt

$$P\Big(a \le \frac{X_n - np}{\sqrt{np(1-p)}} \le b\Big) = \sum_{\{k:\, a \le x_n(k) \le b\}} \frac{1}{\sigma_n}\, \phi_n\big(x_n(k)\big) \qquad (\star)$$

mit

$$\phi_n(x) \;:=\; \sigma_n P\Big(\frac{X_n - np}{\sqrt{np(1-p)}} = x\Big),$$

also

$$\phi_n\big(x_n(k)\big) \;=\; \sigma_n P(X_n = k) \;=\; \sqrt{np(1-p)}\,\binom{n}{k} p^k (1-p)^{n-k}.$$

Wegen $x_n(k) - x_n(k-1) = \sigma_n^{-1}$ lässt sich die rechte Seite von $(\star)$ als Riemann-Summe interpretieren, wobei allerdings die Funktion ϕ_n noch von n abhängt. Wir wollen nun zeigen, dass für jede Folge $(k_n)_{n\in\mathbb{N}}$ mit $\lim_{n\to\infty} x_n(k_n) = x$, $x \in [a,b]$,

$$\lim_{n\to\infty} \phi_n\big(x_n(k)\big) \;=\; \phi(x)$$

gilt, wobei $\phi = \phi_{0,1}$ die Dichte zur Standardnormalverteilung bezeichnet. Im Limes wird die erwähnte Summe dann zum Integral von ϕ über $[a,b]$, und dies ist der behauptete Grenzwert.

Es ist etwas angenehmer, mit den Logarithmen zu arbeiten. Die Stirling-Formel wird dann zu

$$\log(n!) \;=\; \Big(n + \frac{1}{2}\Big)\log(n) - n + \frac{1}{2}\log(2\pi) \;+\; o(1),$$

und man erhält, wobei wir k_n zu k abkürzen,

$$\begin{aligned}
\log\big(\phi_n\big(x_n(k)\big)\big) \;&=\; \frac{1}{2}\log(n) + \frac{1}{2}\log(p) + \frac{1}{2}\log(1-p) \\
&\quad + \Big(n + \frac{1}{2}\Big)\log(n) - \;n\; + \frac{1}{2}\log(2\pi) \\
&\quad - \Big(k + \frac{1}{2}\Big)\log(k) + \;k\; - \frac{1}{2}\log(2\pi) \\
&\quad - \Big(n - k + \frac{1}{2}\Big)\log(n-k) + \;(n-k)\; - \frac{1}{2}\log(2\pi) \\
&\quad + \;k\log(p) \;+\; (n-k)\log(1-p) \;+\; o(1) \\
&=\; -\frac{1}{2}\log(2\pi) \;-\; n\cdot\psi\Big(\frac{k}{n}\Big) \;+\; o(1)
\end{aligned}$$

mit

$$\psi(y) \;:=\; y\log\Big(\frac{y}{p}\Big) \;+\; (1-y)\log\Big(\frac{1-y}{1-p}\Big),$$

wobei wir

$$\frac{1}{2}\log(n) - \frac{1}{2}\log(k) + \frac{1}{2}\log(p) \;=\; o(1)$$

etc. benutzt haben. Eine Taylor-Entwicklung von ψ an der Stelle $y = p$ liefert

$$\begin{aligned}\psi(y) &= \psi(p) + \psi'(p)\,(y-p) + \frac{1}{2}\psi''(p)\,(y-p)^2 + o((y-p)^2) \\ &= \frac{1}{2p(1-p)}\,(y-p)^2 + o((y-p)^2).\end{aligned}$$

Mit $y = k/n$ und $k = k_n$ wie oben erhält man

$$n\,\psi\Big(\frac{k}{n}\Big) = \frac{1}{2}\,x^2 + o(1),$$

also ergibt sich der gewünschte Grenzwert. □

Die bekannten Formeln für die Momente von Binomialverteilungen führen auf

$$E\Big(\frac{X_n - np}{\sqrt{np(1-p)}}\Big) = 0, \quad \operatorname{var}\Big(\frac{X_n - np}{\sqrt{np(1-p)}}\Big) = 1,$$

die Zufallsgrößen X_n wurden also durch eine geeignete Verschiebung auf Erwartungswert 0 und durch eine geeignete Skalierung auf Varianz 1 transformiert. Satz 4.24 zeigt, dass auf diese Weise standardisierte Binomialverteilungen durch eine Standardnormalverteilung approximiert werden können. Im Gegensatz zu der Situation beim Gesetz der seltenen Ereignisse (Satz 3.4) geht die Erfolgswahrscheinlichkeit p mit wachsender Zahl n von Wiederholungen nicht gegen 0, sondern bleibt konstant. Der oben erwähnte Zentrale Grenzwertsatz betrachtet Summen von Zufallsvariablen; im hier behandelten Spezialfall sind die einzelnen Summanden die Indikatorfunktionen, die anzeigen, ob in den einzelnen Versuchswiederholungen ein Erfolg eintritt.

Beispiel 4.25 Mit welcher Wahrscheinlichkeit erscheint beim 600-maligen Wurf eines Würfels mindestens 90-mal und höchstens 105-mal eine Sechs? Als tatsächlicher Wert ergibt sich

$$\sum_{k=90}^{105} \binom{600}{k} \Big(\frac{1}{6}\Big)^k \Big(\frac{5}{6}\Big)^{600-k} = 0.60501\ldots,$$

Satz 4.24 führt mit $n = 600$ und $p = 1/6$ auf

$$\begin{aligned}P(90 \le X_{600} \le 105) &= P\Big(\frac{105-100}{\sqrt{500/6}} \le X^\star_{600} \le \frac{90-100}{\sqrt{500/6}}\Big) \\ &\approx \Phi\Big(\frac{5}{\sqrt{500/6}}\Big) - \Phi\Big(\frac{-10}{\sqrt{500/6}}\Big) \\ &= 0.571398\ldots.\end{aligned}$$

(Man kann diese Approximation mit der sog. *Stetigkeitskorrektur* verbessern, bei der beispielsweise $P(X_{600} \le 105) = P(X_{600} \le 105.5)$ ausgenutzt wird.) ◁

4.6 Erwartungswerte. Die 'offizielle' Verallgemeinerung erfordert das allgemeine Lebesgue-Integral, das beispielsweise zu Beginn der Vorlesung Stochastik II besprochen wird. Wir begnügen uns hier mit Andeutungen.
Ist X eine Zufallsvariable mit Dichte f und setzt man für alle $x \in \mathbb{R}$

$$\lceil x \rceil := \min\{k \in \mathbb{Z} : k \geq x\}, \quad \lfloor x \rfloor := \max\{k \in \mathbb{Z} : k \leq x\},$$

so wird durch

$$\underline{X}_n := 2^{-n}\lfloor 2^n X \rfloor, \quad \overline{X}_n := 2^{-n}\lceil 2^n X \rceil$$

eine Familie von diskreten Zufallsvariablen definiert, für die $\underline{X}_n \uparrow X$, $\overline{X}_n \downarrow X$ mit $n \to \infty$ gilt. Bei diesen können wir die bereits vorhandene Definition des Erwartungswertes verwenden:

$$\begin{aligned}
E\underline{X}_n &= \sum_{k\in\mathbb{Z}} k2^{-n}\, P\big(\underline{X}_n = k2^{-n}\big) \\
&= \sum_{k\in\mathbb{Z}} k2^{-n} \int_{k2^{-n}}^{(k+1)2^{-n}} f(x)\,dx \\
&= \sum_{k\in\mathbb{Z}} \int_{k2^{-n}}^{(k+1)2^{-n}} \frac{\lfloor 2^n x \rfloor}{2^n} f(x)\,dx \\
&= \int_{-\infty}^{\infty} \frac{\lfloor 2^n x \rfloor}{2^n} f(x)\,dx \\
&\leq \int_{-\infty}^{\infty} x f(x)\,dx \\
&\leq \int_{-\infty}^{\infty} \frac{\lceil 2^n x \rceil}{2^n} f(x)\,dx \;=\; \ldots \;=\; E\overline{X}_n.
\end{aligned}$$

Wegen $\overline{X}_n - \underline{X}_n \leq 2^{-n}$ gilt

$$E\overline{X}_n - E\underline{X}_n \;=\; E(\overline{X}_n - \underline{X}_n) \;\leq\; 2^{-n},$$

es liegt also nahe, den Erwartungswert von X im Falle $\int |x| f(x)dx < \infty$ durch

$$EX = \int x f(x)\,dx$$

zu definieren. Obwohl dies für praktische Zwecke (Rechnungen) i.a. reicht, ist es doch mathematisch unbefriedigend: Eine nützliche Formel wie

$$Eg(X) = \int g(x) f(x)\,dx,$$

die wir im folgenden häufig verwenden werden, ergibt sich nicht ohne weiteres.

BEISPIEL 4.26 Im Falle $X \sim N(\mu, \sigma^2)$ erhält man

$$\begin{aligned} EX &= \int_{-\infty}^{\infty} x \, \frac{1}{\sqrt{2\pi\sigma^2}} \, e^{-(x-\mu)^2/2} \, dx \\ &= \int_{-\infty}^{\infty} (x-\mu) \frac{1}{\sqrt{2\mu\sigma^2}} \, e^{-(x-\mu)^2/2} \, dx \; + \; \mu \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\mu\sigma^2}} \, e^{-(x-\mu)^2/2} \, dx \\ &= \mu, \end{aligned}$$

denn das erste Integral hat aus Symmetriegründen den Wert 0 und das zweite Integral ist als Integral über eine Wahrscheinlichkeitsdichte gleich 1. ◁

4.7 Unabhängigkeit. Bisher sind uns σ-Algebren nur als 'notwendiges Übel' begegnet; sie spielen aber in der Stochastik eine weitaus wichtigere Rolle, beispielsweise als natürliche Heimat des Unabhängigkeitsbegriffs und als Repräsentanten von Teilinformation.

SATZ UND DEFINITION 4.27 *Es sei X eine Zufallsgröße auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ mit Werten in dem messbaren Raum $(\Omega', \mathcal{A}')$. Dann ist $\{X^{-1}(A) : A \in \mathcal{A}'\}$ eine σ-Algebra. Diese nennt man die* von X erzeugte σ-Algebra*, Schreibweise: $\sigma(X)$.*

BEWEIS: Übungsaufgabe. □

Kennen wir das Resultat ω des Zufallsexperiments, so können wir von jedem Ereignis $A \in \mathcal{A}$ sagen, ob es eingetreten ist oder nicht. Die von X erzeugte σ-Algebra $\sigma(X)$ ist die Menge der Ereignisse, für die wir diese Entscheidung treffen können, wenn uns nur $X(\omega)$ bekannt ist.

Wir haben in Abschnitt 1 der Vorlesung zwei Ereignisse A und B unabhängig genannt, wenn $P(A \cap B) = P(A)P(B)$ gilt, und in Aufgabe 7 (d) gesehen, dass dann auch A^c und B^c unabhängig sind. Es gilt sogar, dass dann zwei beliebige Mengen aus den jeweiligen erzeugten σ-Algebren

$$\sigma(\{A\}) = \{\emptyset, A, A^c, \Omega\}, \quad \sigma(\{B\}) = \{\emptyset, B, B^c, \Omega\}$$

in diesem Sinne unabhängig sind. Dies führt auf:

DEFINITION 4.28 Es sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum, $I \neq \emptyset$.

(a) Eine Familie $\{\mathcal{A}_i :\in I\}$ von Unter-σ-Algebren von $\mathcal{A}$ heißt *stochastisch unabhängig*, wenn für jede endliche Teilmenge $J = \{j_1, \ldots, j_n\}$ von I und alle $A_{j_1} \in \mathcal{A}_{j_1}, \ldots, A_{j_n} \in \mathcal{A}_{j_n}$ gilt:

$$P\Big(\bigcap_{j \in J} A_j\Big) \; = \; \prod_{j \in J} P(A_j). \qquad (*)$$

(b) Ist für jedes $i \in I$ X_i eine Zufallsgröße auf $(\Omega, \mathcal{A}, P)$ mit Werten in einem messbaren Raum $(\Omega_i, \mathcal{A}_i)$, so heißt die Familie $\{X_i : i \in I\}$ *stochastisch unabhängig* (kurz: die Zufallsgrößen X_i, $i \in I$, sind unabhängig), wenn die Familie $\{\sigma(X_i) : i \in I\}$ der erzeugten σ-Algebren im Sinne von (a) unabhängig ist.

Der folgende Satz zeigt, dass man sich beim Nachweis der entscheidenden Eigenschaft $(*)$ aus der Definition auf $\cap$-stabile Erzeugendensysteme beschränken kann.

SATZ 4.29 *Es seien $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum, $I \neq \emptyset$, und für jedes $i \in I$ $\mathcal{A}_i$ eine Unter-σ-Algebra von $\mathcal{A}$ mit $\cap$-stabilem Erzeugendensystem $\mathcal{E}_i$. Gilt dann*

$$P\Big(\bigcap_{k=1}^{n} E_{j_k}\Big) \;=\; \prod_{k=1}^{n} P(E_{j_k})$$

für alle endlichen $J = \{j_1, \ldots, j_n\} \subset I$ und alle $E_{j_k} \in \mathcal{E}_{j_k}$, $k = 1, \ldots, n$, so sind $\mathcal{A}_i$, $i \in I$, stochastisch unabhängig.

BEWEIS: Sei $J = \{j_1, \ldots, j_n\} \subset I$. Sei $\mathcal{D}_{j_1}$ die Menge aller $A \in \mathcal{A}_{j_1}$ mit

$$P(A \cap E_{j_2} \cap \ldots \cap E_{j_n}) \;=\; P(A)\, P(E_{j_1}) \ldots P(E_{j_n})$$

für alle $E_{j_2} \in \mathcal{E}_{j_2}, \ldots, E_{j_n} \in \mathcal{E}_{j_n}$. Man sieht leicht, dass $\mathcal{D}_{j_1}$ ein Dynkin-System ist. Da D_{j_1} den $\cap$-stabilen Erzeuger $\mathcal{E}_{j_1}$ von $\mathcal{A}_{j_1}$ enthält, gilt also $\mathcal{D}_{j_1} = \mathcal{A}_{j_1}$ nach Satz 4.7 (b). Im zweiten Schritt sei $\mathcal{D}_{j_2}$ die Menge aller $A \in \mathcal{A}_{j_2}$ mit

$$P(A_{j_1} \cap A \cap E_{j_3} \cap \ldots \cap E_{j_n}) \;=\; P(A_{j_1})\, P(A)\, P(E_{j_3}) \ldots P(E_{j_n})$$

für alle $E_{j_3} \in \mathcal{E}_{j_3}, \ldots, E_{j_n} \in \mathcal{E}_{j_n}$. Man sieht wieder, dass $\mathcal{D}_{j_2}$ ein Dynkin-System ist, das nach dem bereits bewiesenen Teil $\mathcal{E}_{j_2}$ enthält, und es folgt wieder $\mathcal{D}_{j_2} = \mathcal{A}_{j_2}$. Nach insgesamt n Schritten dieser Art erhält man die gewünschte Beziehung

$$P(A_{j_1} \cap \ldots \cap A_{j_n}) \;=\; P(A_{j_1}) \ldots P(A_{j_n})$$

für alle $A_{j_1} \in \mathcal{A}_{j_1}, \ldots, A_{j_n} \in \mathcal{A}_{j_n}$. □

Bei einer diskreten Zufallsgröße X bilden die Mengen $X^{-1}(\{x\})$, $x \in \mathrm{Bild}(X)$, ein $\cap$-stabiles Erzeugendensystem von $\sigma(X)$. Satz 3.17 zeigt also, dass Teil (b) der Definition 4.28 zu Definition 3.16 'abwärtskompatibel' ist.

Der Zugang über σ-Algebren bietet Vorteile, beispielsweise beim Beweis des folgenden Satzes, der grob gesprochen besagt, dass Funktionen unabhängiger Zufallsgrößen wieder unabhängig sind.

SATZ 4.30 *Für jedes $i \in I$ seien X_i eine Zufallsgröße mit Werten in $(\Omega_i, \mathcal{A}_i)$, $(\Omega_i', \mathcal{A}_i')$ ein weiterer meßbarer Raum und $g_i : \Omega_i \to \Omega_i'$ eine $(\mathcal{A}_i, \mathcal{A}_i')$-messbare Abbildung. Ist dann $\{X_i : i \in I\}$ eine unabhängige Familie, so ist auch $\{Y_i : i \in I\}$ mit $Y_i := g_i(X_i)$ unabhängig.*

BEWEIS: $\sigma(Y_i) \subset \sigma(X_i)$. □

BEISPIEL 4.31 Es sei $(\Omega, \mathcal{A}, P) = \big([0,1), \mathcal{B}_{[0,1)}, \mathrm{unif}(0,1)\big)$. Für jedes $n \in \mathbb{N}$ werde $X_n = \Omega \to \{0,1\}$ definiert durch

$$X_n(\omega) := \lfloor 2^n \omega \rfloor - 2\lfloor 2^{n-1}\omega \rfloor.$$

Dann gilt $\omega = \sum_{n=1}^{\infty} 2^{-n} X_n(\omega)$ — die Folge $0.X_1(\omega)X_2(\omega)X_3(\omega)\ldots$ ist also eine (mehr oder weniger: die) Binärdarstellung von ω.
Für alle $k_1, \ldots, k_n \in \{0,1\}$ gilt

$$P(X_1 = k_1, \ldots, X_n = k_n) \;=\; P\Big(\sum_{l=1}^{n} 2^{-l} k_l \le \omega < \sum_{l=1}^{n} 2^{-l} k_l + 2^{-n}\Big) \;=\; 2^{-n},$$

denn das Intervall besteht aus allen $\omega \in [0,1)$, deren Binärdarstellung mit den Ziffern (bits) $k_1, \ldots, k_n$ beginnt. Für beliebige $i_1 < i_2 < \ldots < i_n$ erhält man somit

$$\begin{aligned} &P(X_{i_1} = 1, \ldots, X_{i_n} = 1) \\ &\quad= \sum_{\substack{(k_1,\ldots,k_{i_n}) \in \{0,1\}^{i_n} \\ k_{i_j} = 1 \text{ für } j = 1,\ldots,n}} P(X_1 = k_1, X_2 = k_2, \ldots, X_{i_n} = k_n) \\ &\quad= 2^{-i_n} \#\big\{(k_1, \ldots, k_{i_n}) \in \{0,1\}^{i_n} : k_{i_j} = 1 \text{ für } j = 1, \ldots, n\big\} \\ &\quad= 2^{-i_n}\, 2^{i_n - n} \quad \text{(denn genau } n \text{ Positionen sind festgelegt)} \\ &\quad= 2^{-n}. \end{aligned}$$

Insbesondere folgt $P(X_{i_j} = 1) = 1/2$ und damit insgesamt

$$P(X_{i_1} = 1, \ldots, X_{i_n} = 1) \;=\; P(X_{i_1} = 1) \ldots P(X_{i_n} = 1).$$

Da $\big\{X_i^{-1}(\{1\})\big\}$ ein $\cap$-stabiles Erzeugendensystem von $\sigma(X_i)$ ist, haben wir damit die Unabhängigkeit der Zufallsvariablen $X_1, X_2, X_3, \ldots$ gezeigt. Außerdem gilt $\mathcal{L}(X_i) = \mathrm{Bin}(1, 1/2)$, die gesamte Konstruktion kann also als Modell für den unendlich oft wiederholten Wurf einer fairen Münze dienen. Umgekehrt ließe sich aus einer unendlichen Folge von Münzwürfen $k_1, k_2, \ldots$ eine auf $[0,1)$ gleichverteilte Zahl x durch $x := \sum_{i=1}^{\infty} k_i 2^{-i}$ konstruieren! ◁

Wir betrachten nun den Fall reellwertiger Zufallsgrößen etwas näher. Sind X und Y unabhängige Zufallsvariablen mit Verteilungsfunktionen F_X und F_Y, so gilt

$$P(X \le x, Y \le y) \;=\; P(X \le x)P(Y \le y) \;=\; F_X(x)F_Y(y)$$

für alle $x, y \in \mathbb{R}$. Definiert man die *gemeinsame Verteilungsfunktion* von zwei (beliebigen) Zufallsvariablen X und Y durch

$$F_{X,Y} = \mathbb{R}^2 \to \mathbb{R}, \quad F_{X,Y}(x,y) := P(X \le x, Y \le y),$$

so erhält man, dass bei Unabhängigkeit die gemeinsame Verteilungsfunktion das Produkt der einzelnen Verteilungsfunktionen ist, d.h.

$$F_{X,Y}(x,y) = F_X(x)\, F_Y(y) \qquad \text{für alle } x, y \in \mathbb{R}.$$

Die Mengen $(-\infty, x]$, $x \in \mathbb{R}$, bilden nach Satz 4.3 ein $\cap$-stabiles Erzeugendensystem von $\mathcal{B}(\mathbb{R})$, also folgt mit Satz 4.29 auch umgekehrt die Unabhängigkeit von X und Y aus dieser Darstellung.

Sind X und Y stetige Zufallsvariablen mit Dichten f_X, f_Y, d.h. insbesondere

$$F_X(x) = \int_{-\infty}^{x} f_X(y)\, dy, \quad F_Y(y) = \int_{-\infty}^{y} f_Y(z)\, dz,$$

so erhält man bei Unabhängigkeit

$$F_{X,Y}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_X(u) f_Y(v)\, du\, dv.$$

In naheliegender Verallgemeinerung des eindimensionalen Falles nennt man $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}_+$ eine *gemeinsame Dichte* von X und Y, wenn

$$P\big((X,Y) \in A\big) \;=\; \iint_A f_{X,Y}(x,y)\, dx\, dy$$

für 'hinreichend viele' $A \subset \mathbb{R}^2$ gilt (in der Vorlesung Stochastik II wird dies präzisiert). Insbesondere hat man bei unabhängigen Zufallsvariablen X, Y mit Dichten f_X, f_Y

$$f_{X,Y}(x,y) = f_X(x) f_Y(y),$$

die Analogie zum diskreten Fall (Satz 3.17) ist offensichtlich.

Mit gemeinsamen Dichtefunktionen lassen sich auch beispielsweise Erwartungswerte von Funktionen von mehreren Zufallsvariablen ausrechnen; wir beschränken uns wie oben auf den Fall von zwei Zufallsvariablen X und Y. Zur Erinnerung: Sind X und Y diskrete Zufallsgrößen mit gemeinsamer Massenfunktion

$p_{X,Y}(x,y) = P(X = x, Y = y)$, so gilt unter der Voraussetzung, dass die Summe absolut konvergiert,

$$Eg(X,Y) = \sum_{x\in\mathrm{Bild}(X)} \sum_{y\in\mathrm{Bild}(Y)} g(x,y)\, p_{X,Y}(x,y).$$

Ganz analog hat man in der stetigen Situation

$$Eg(X,Y) = \iint g(x,y)\, f_{X,Y}(x,y)\, dx\, dy$$

(Genaueres, beispielsweise zur Messbarkeit von g, wird in der Vorlesung Stochastik II besprochen). Hiermit erhält man u.a. eine Variante der Multiplikationsregel für unabhängige stetige Zufallsvariablen X, Y:

$$\begin{aligned} EXY &= \int\int xy\, f_X(x) f_Y(y)\, dx\, dy \\ &= \Big(\int x f_X(x)\, dx\Big)\Big(\int y f_Y(y)\, dy\Big) = (EX)\,(EY), \end{aligned}$$

man vergleiche dies mit Satz 3.18. Auch Begriffe wie Kovarianz etc. lassen sich auf diese Weise auf den stetigen Fall übertragen.

In der Maßtheorie (siehe die Vorlesung mit diesem Namen, aber auch den Beginn der Stochastik II) wird gezeigt, dass sowohl der diskrete als auch der stetige Fall Spezialfälle einer allgemeinen Theorie sind. Es gibt übrigens durchaus auch Zufallsvariable, die weder diskret noch stetig sind — ein Beispiel wird in den Übungen behandelt.

Mit dem obenstehenden sind die möglichen Analogiebetrachtungen bei weitem nicht erschöpft; die Faltung beispielsweise wird in den Übungsaufgaben behandelt.

Beispiel 4.32 Die Lebensdauer X einer Glühbirne vom Typ A sei exponentialverteilt mit Parameter λ_A, Y sei die Lebensdauer einer Glühbirne vom Typ B, ebenfalls exponentialverteilt, nun mit Parameter λ_B. Wir setzen voraus, dass die Zufallsvariablen X und Y unabhängig sind. Mit welcher Wahrscheinlichkeit brennt die B-Birne länger als die A-Birne? Die obigen Überlegungen führen auf

$$\begin{aligned} P(X < Y) &= P\big((X,Y) \in \{(x,y) \in \mathbb{R}^2 : x < y\}\big) \\ &= \iint_{\{(x,y)\in\mathbb{R}^2 :\, x<y\}} f_{X,Y}(x,y)\, dy\, dx \\ &= \iint_{\{(x,y)\in\mathbb{R}^2 :\, x<y\}} \lambda_A\, e^{-\lambda_A y}\, \lambda_B\, e^{-\lambda_B x}\, dy\, dx \end{aligned}$$

$$
\begin{aligned}
&= \int_0^\infty \Big(\int_x^\infty \lambda_B \, e^{-\lambda_B y} \, dy \Big) \lambda_A \, e^{-\lambda_A x} \, dx \\
&= \lambda_A \int_0^\infty e^{-\lambda_B x} e^{-\lambda_A x} dx \;=\; \frac{\lambda_A}{\lambda_A + \lambda_B} \, .
\end{aligned}
$$

$\lhd$

5. Grundbegriffe der mathematischen Statistik

5.1 Allgemeines. In der Wahrscheinlichkeitstheorie geht man von einem Modell $(\Omega, \mathcal{A}, P)$ für ein Zufallsexperiment aus und berechnet beispielsweise die Wahrscheinlichkeit eines Ereignisses A. In der Statistik soll man, nun ausgehend von den bei der Ausführung des Experiments gewonnenen Daten, eine Aussage über das zugehörige P machen (P ist also unbekannt). Beim zehnfachen Münzwurf ist beispielsweise eine typische wahrscheinlichkeitstheoretische Frage:

> Mit welcher Wahrscheinlichkeit kommt achtmal Kopf, wenn die Münze fair ist?

Typische statistische Fragestellungen wären in dieser Situation:

> Es kam achtmal Kopf. Welchen Wert hat p, die Wahrscheinlichkeit für Kopf? Ist die Münze fair, d.h. gilt $p = 1/2$?

Klar: Die Beobachtung $x = 8$ lässt die exakte Bestimmung von p nicht zu — auf der Basis von zufälligen Beobachtungen lassen sich i.a. keine absolut sicheren (nicht-trivialen) Schlüsse ziehen ('you can't make a silk purse out of a sow's ear').

Der formale Rahmen für die hier zu betrachtenden statistischen Fragestellungen besteht aus einem messbaren Raum $(\mathcal{X}, \mathcal{A})$, dem *Stichprobenraum*, der die möglichen Datenwerte x enthält; auf $(\mathcal{X}, \mathcal{A})$ hat man eine Familie $\mathcal{P}$ von Wahrscheinlichkeitsmaßen, die in Frage kommenden Verteilungen für die Daten (aus dem Zusammenhang sollte immer klar hervorgehen, ob sich das Symbol $\mathcal{P}$ auf eine Familie von Wahrscheinlichkeitsmaßen oder auf die Potenzmengenbildung bezieht). Diese Familie kann die Klasse *aller* Wahrscheinlichkeitsmaße auf dem Stichprobenraum sein, hat aber meistens eine bestimmte Struktur. Häufig ist $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, mit $\Theta \subset \mathbb{R}^d$, ein d-dimensionale *parametrische Familie*, Θ heißt dann die *Parametermenge*. Die Daten $x \in \mathcal{X}$ können als Realisierungen einer Zufallsgröße $X : \Omega \to \mathcal{X}$ mit unbekannter Verteilung $\mathcal{L}(X) \in \mathcal{P}$ betrachtet werden. Wird beispielsweise beim zehnfachen Münzwurf nur die Anzahl der 'Kopf'-Würfe beobachtet, so könnte man

$$\mathcal{X} = \{0, 1, \ldots, 10\},\ \mathcal{A} = \mathcal{P}(\mathcal{X}),\ \Theta = [0, 1],\ P_\theta = \mathrm{Bin}(10, \theta)$$

wählen. Einen besonders wichtigen Spezialfall der allgemeinen Situation erhält man, wenn die Daten durch unabhängige Wiederholungen eines Zufallsexperiments gewonnen werden, also $x = (x_1, \ldots, x_n)$ gilt, wobei x_i das Ergebnis der

i-ten Wiederholung ist. Man spricht dann von (den Werten) einer *Stichprobe vom Umfang n* aus einer Verteilung.

Wir betrachten die drei hauptsächlichen statistischen Verfahren: Schätzer, Tests und Konfidenzbereiche.

5.2 Schätztheorie. Ein *Schätzer* (auch: *Schätzfunktion*) ist eine Abbildung $\hat{\theta} : \mathcal{X} \to \Theta$, die jeder Beobachtung x einen Schätzwert $\hat{\theta} = \hat{\theta}(x)$ für den unbekannten Parameter θ zuordnet. Im Münzwurfbeispiel ist $\hat{\theta} := x/10$ ein naheliegender Schätzer.

Wie erhält man (gute) Schätzfunktionen? Ein plausibles und sehr wichtiges Prinzip besteht darin, dass man den Wert $\hat{\theta}$ wählt, unter dem die Beobachtung x die größte (infinitesimale) Wahrscheinlichkeit hat. Dies ist die *Likelihood-Methode*. Konkret nennen wir im diskreten Fall die Funktion

$$l(\,\cdot\,|x) : \Theta \to \mathbb{R}, \quad \theta \mapsto P_\theta(\{x\}),$$

die *Likelihood-Funktion* zur Beobachtung x. Hat $\hat{\theta} : \mathcal{X} \to \Theta$ die Eigenschaft

$$l\big(\hat{\theta}(x)\big|x\big) \;=\; \sup\big\{l(\theta|x) : \theta \in \Theta\big\} \quad \text{für alle } x \in \mathcal{X},$$

so nennen wir $\hat{\theta}$ einen *Maximum-Likelihood-Schätzer* für θ. Geht es in dieser Situation nicht um θ selbst, sondern um einen hiervon abhängenden Wert $\eta = g(\theta)$, so nennen wir $\hat{\eta} := g(\hat{\theta})$ den Maximum-Likelihood-Schätzer für η.

Es können natürlich allerlei Schwierigkeiten auftreten; beispielsweise wird das Supremum möglicherweise nicht angenommen, oder es ist nicht eindeutig. Bei der praktischen Anwendung ist es häufig bequemer, den Logarithmus der Wahrscheinlichkeit, also die *Log-Likelihood-Funktion*, zu maximieren.

Beispiel 5.1 (Das Capture-Recapture-Problem)
Ein See enthalte eine unbekannte Anzahl N von Fischen. Es werden M Fische gefangen, markiert, und wieder freigelassen. Nach einer gewissen Zeit werden n Fische gefangen, unter diesen befinden sich x markierte. Wie sollte man N schätzen?

Unter gewissen Voraussetzungen (Fische 'vermischen sich' etc.) erscheint das folgende Modell vernünftig: M und n sind bekannt, N ist der unbekannte Parameter (aus $\{M, M+1, M+2, \ldots\}$), und $\mathcal{X} = \{0, \ldots, n\}$ ist der Stichprobenraum. Die Beobachtung ist hypergeometrisch verteilt mit Parametern N, M und n, also

$$P_N(\{x\}) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{N}$$

Dann gilt

$$\frac{P_N(\{x\})}{P_{N-1}(\{x\})} \;=\; \frac{\binom{M}{x}\binom{N-M}{n-x}\binom{N-1}{n}}{\binom{N}{n}\binom{M}{x}\binom{N-1-M}{n-x}} \;=\; \frac{(N-M)(N-n)}{N(n-M-n+x)}$$

Hieraus folgt

$$\begin{aligned} P_N(\{x\}) > P_{N-1}(\{x\}) &\iff (N-M)(N-n) > N(N-M-n+x) \\ &\iff nM > Nx, \end{aligned}$$

also wird $N \to P_N(\{x\})$ maximal für $\hat{N} := \lfloor \frac{nM}{x} \rfloor$. Im Falle $nM/x \in \mathbb{N}$ wird das Maximum in $\hat{N}$ und $\hat{N}-1$ angenommen.

Man kann auch direkter argumentieren, dass der Anteil x/n der markierten Fische im Fang ungefähr übereinstimmen sollte mit dem Anteil M/N der markierten Fische im See. Konsequente Anwendung des Prinzips führt bei Beobachtung $x = 0$ auf den Schätzwert $N = \infty$ (nicht besonders realistisch, da dann kein Platz mehr für das Wasser bleibt). ◁

Bei einer Stichprobe vom Umfang n aus einer Verteilung mit Massenfunktion $p(\,\cdot\,|\theta)$ erhält man (siehe die Bemerkungen nach Satz 3.17) als Likelihood-Funktion

$$l(\theta|x) \;=\; l(\theta|x_1,\ldots,x_n) \;=\; \prod_{i=1}^{n} p(x_i|\theta).$$

Besonders dann, wenn das Maximum nach der Methode ‘Ableiten und Nullsetzen’ gefunden werden soll, erweist sich der Übergang zur Log-Likelihood-Funktion als sinnvoll.

Bei der *Momentenmethode* werden die *Momente der Stichprobe*,

$$\frac{1}{n}\sum_{i=1}^{n} x_i, \quad \frac{1}{n}\sum_{i=1}^{n} x_i^2, \quad \frac{1}{n}\sum_{i=1}^{n} x_i^3, \;\ldots$$

den ‘theoretischen’ Momenten $E_\theta X$, $E_\theta X^2$, $E_\theta X^3, \ldots$ (die ja von θ abhängen) gleichgesetzt, und die entstehenden Gleichungen werden nach θ aufgelöst. Man nimmt so viele Gleichungen, wie man braucht, um nach θ auflösen zu können. Hat man nur eine einzige Beobachtung x, so würde diese Methode auf die Gleichung $x = E_\theta X$ führen, beim Capture-Recapture-Problem in Verbindung mit der aus Beispiel 3.24(b) bekannten Formel für den Erwartungswert zur hypergeometrischen Verteilung wieder auf den Schätzer $\hat{N} \approx Mn/x$.

BEISPIEL 5.2 Ein Zufallsexperiment, in dem ein bestimmtes Ereignis A die Wahrscheinlichkeit θ hat, wird n-mal unabhängig wiederholt; θ ist zu schätzen. Schreiben wir 1 für das Eintreten von A und sonst 0, so sind die gewonnenen Daten Elemente von $\mathcal{X} = \{0,1\}^n$ und als Klasse der möglichen Verteilungen ergibt sich $\mathcal{P} = \{P_\theta : 0 \le \theta \le 1\}$, wobei zu P_θ die Massenfunktion

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

mit $k := \#\{1 \le i \le n : x_i = 1\}$ gehört. Zu gegebener Zahl k von Erfolgen erhält man also die Likelihood-Funktion $l(\theta) = \theta^k(1-\theta)^{n-k}$. Wir betrachten die Randfälle separat: Bei $k = 0$ erhält man das (eindeutige, globale) Maximum in $\hat\theta = 0$, bei $k = n$ in $\hat\theta = 1$. In den Fällen $k \in \{1,\ldots,n-1\}$ ist $l(0) = l(1) = 0$, $l(\theta|x) > 0$ auf $0 < \theta < 1$, und das Maximum kann über die Ableitung der Log-Likelihood-Funktion gefunden werden: Mit

$$\frac{\partial}{\partial\theta}\log l(\theta) \;=\; -\frac{n-k}{1-\theta} + \frac{k}{\theta}$$

führt dies auf den Maximum-Likelihood-Schätzer $\hat\theta = k/n$. Wegen

$$E_\theta X_i \;=\; 0\cdot(1-\theta) \,+\, 1\cdot\theta \;=\; \theta$$

führt die Momentenmethode auf den Ansatz $\frac{1}{n}\sum_{i=1}^n x_i = \theta$, also ebenfalls auf den Schätzer $\hat\theta = k/n$. Es ist natürlich auch intuitiv naheliegend, die unbekannte Wahrscheinlichkeit von A durch die relative Häufigkeit des Eintretens von A zu schätzen. ◁

Wie verfährt man im nicht-diskreten Fall? Hat man eine Stichprobe vom Umfang n aus einer Verteilung mit Dichtefunktion $f(\,\cdot\,|\theta)$, so bietet es sich an, anstelle der 'richtigen' Wahrscheinlichkeiten die 'infinitesimalen' Wahrscheinlichkeiten zu verwenden, also die gemeinsame Massenfunktion durch die gemeinsame Dichtefunktion zu ersetzen. Mit den Resultaten von Abschnitt 4.7 erhält man dann als Likelihood-Funktion

$$l(\theta|x) \;=\; l(\theta|x_1,\ldots,x_n) \;=\; \prod_{i=1}^{n} f(x_i|\theta).$$

BEISPIEL 5.3 Als Beispiel für eine stetige Situation mit mehrdimensionalem Parameterraum betrachten wir eine Stichprobe $X_1,\ldots,X_n$ aus der Normalverteilung $N(\mu,\sigma^2)$ mit unbekanntem $\mu \in \mathbb{R}$ und unbekanntem $\sigma^2 > 0$. Wir haben

$$f_{X_i}(x_i|\mu,\sigma^2) \;=\; \frac{1}{\sqrt{2\pi\sigma^2}}\,\exp\Big(-\frac{1}{2\sigma^2}(x_i-\mu)^2\Big),$$

erhalten also als gemeinsame Dichte in $x = (x_1, \ldots, x_n)$

$$f(x|\mu,\sigma^2) \;=\; \prod_{i=1}^{n} f_{X_i}(x_i|\mu,\sigma^2) \;=\; \Big(\frac{1}{\sqrt{2\pi\sigma^2}}\Big)^n \exp\Big(-\sum_{i=1}^{n}\frac{1}{2\sigma^2}(x_i-\mu)^2\Big)$$

und damit

$$\log\, l(\mu,\sigma^2) \;=\; -\frac{n}{2}\,\log(2\pi\sigma^2) \;-\; \frac{1}{2\sigma^2}\,\sum_{i=1}^{n}(x_i-\mu)^2.$$

Für jedes feste $\sigma^2 > 0$ wird dies als Funktion von μ durch den Stichprobenmittelwert $\bar{x}_n := \frac{1}{n}\sum_{i=1}^{n} x_i$ maximiert. Die Funktion

$$\sigma^2 \;\to\; -\frac{n}{2}\,\log(2\pi\sigma^2) \;-\; \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2$$

wiederum wird maximal in $\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2$. Damit erhält man die Maximum-Likelihood-Schätzer

$$\hat{\mu} = \bar{x}_n, \quad \widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2.$$

◁

Beispiel 5.4 In den bisherigen Beispielen war die Verteilung durch den zu schätzenden Parameter festgelegt — dies muss nicht unbedingt so sein. Will man beispielsweise in der Stichprobensituation den Erwartungswert der Zufallsvariablen schätzen, so führt die Momentenmethode auf den Schätzer $\bar{x}_n$. Bei der Maximum-Likelihood-Methode sind genauere Annahmen an die Verteilung nötig. Die Varianz wird häufig durch die *Stichprobenvarianz*

$$S_n^2 \;=\; \frac{1}{n-1}\,\sum_{i=1}^{n}(x_i-\bar{x}_n)^2$$

geschätzt. Mit $\mathrm{var}(X_i) = EX_i^2 - (EX_i)^2$ würde die Momentenmethode auf den Schätzer

$$\frac{1}{n}\sum_{i=1}^{n}x_i^2 - \Big(\frac{1}{n}\sum_{i=1}^{n}x_i\Big)^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2$$

führen (dieses Beispiel wird in einer Übungsaufgabe näher betrachtet). ◁

Bei den bisherigen Beispielen war der Ausgangspunkt stets eine Stichprobe aus einer festen Verteilung. In der statistischen Praxis stößt man schnell an die Grenzen dieses Modells; beispielsweise geht es häufig um die Abhängigkeit der Beobachtungen von anderen Größen. Wir behandeln exemplarisch eine qualitative und eine quantitative solche Situation.

BEISPIEL 5.5 (Zweistichprobenproblem) Angenommen, wir haben zwei Typen A und B von Glühbirnen mit jeweils exponentialverteilten Lebensdauern, Typ A mit Parameter λ_A und Typ B mit Parameter λ_B. Es werden m Exemplare des ersten und n Exemplare des zweiten Typs untersucht; man beobachtet die Lebensdauern $x_1, \ldots, x_m$ in der ersten und $y_1, \ldots, y_n$ in der zweiten Gruppe. Die Daten $x_1, \ldots, x_m, y_1, \ldots, y_n$ fassen wir als Realisierungen von unabhängigen Zufallsvariablen $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ auf, mit

$$X_i \sim \mathrm{Exp}(\lambda_A) \;\text{ für }\; i = 1, \ldots, m,\; Y_j \sim \mathrm{Exp}(\lambda_B) \;\text{ für }\; j = 1, \ldots, n.$$

Aus der gemeinsamen Dichte ergibt sich die Loglikelihood-Funktion

$$\begin{aligned}
&\log l(\lambda_A, \lambda_B | x_1, \ldots, x_m, y_1, \ldots, y_n) \\
&\qquad = \log\Big(\prod_{i=1}^{m} \lambda_A e^{-\lambda_A x_i} \prod_{j=1}^{n} \lambda_B e^{-\lambda_B y_j}\Big) \\
&\qquad = m \log(\lambda_A) - \lambda_A \sum_{i=1}^{m} x_i + n \log(\lambda_B) - \lambda_B \sum_{j=1}^{n} y_j.
\end{aligned}$$

Dies wird in

$$\begin{pmatrix} \hat\lambda_A \\ \hat\lambda_B \end{pmatrix} = \begin{pmatrix} 1/\bar x_m \\ 1/\bar y_n \end{pmatrix} \quad \text{mit } \bar x_m := \frac{1}{m}\sum_{i=1}^{m} x_i,\; \bar y_n := \frac{1}{n}\sum_{j=1}^{n} y_j$$

maximal. Für das Verhältnis $\theta = EX_i/EY_j = \lambda_B/\lambda_A$ der mittleren Lebensdauern erhält man so den Maximum-Likelihood-Schätzer $\hat\theta = \bar x_m/\bar y_n$. Auch eine entsprechende Variante der Momentenmethode würde auf diesen Schätzer führen. ◁

BEISPIEL 5.6 (Einfache lineare Regression) Unsere Beobachtungen $y_1, \ldots, y_n$ (die abhängigen Variablen, 'response') betrachten wir als Realisierungen der unabhängigen Zufallsvariabeln $Y_1, \ldots, Y_n$; zu jedem Y_i gehört eine Hilfsgröße (unabhängige Variable, Einstellvariable, 'covariate') x_i. Wir setzen voraus, dass der 'systematische Teil' EY_i affin-linear von dieser Größe abhängt,

$$EY_i = \alpha + \beta x_i \quad \text{für } i = 1, \ldots, n,$$

und interessieren uns für die unbekannten Parameter α und β (Achsenabschnitt und Steigung der Regressionsgeraden). Typische Beispiele sind die Abhängigkeit des Ernteertrags von der eingebrachten Düngemittelmenge oder auch das Klausurergebnis in Abhängigkeit von der in den Hausübungen erreichten Punktzahl; dabei ist eine affin-lineare Abhängigkeit in der Regel (bei nicht zu großen Bereichen für die Hilfsvariable) eine brauchbare Näherung.

Bei der auf Gauß zurückgehenden *Methode der kleinsten Quadrate* werden α und β durch die Werte $\hat{\alpha}$ und $\hat{\beta}$ geschätzt, die die Summe der quadrierten Abweichungen der beobachteten Werte der abhängigen Variablen von ihrem Erwartungswert unter dem Modell mit diesen Parametern, also die Funktion

$$(\alpha, \beta) \;\mapsto\; \sum_{i=1}^{n} \big(y_i \,-\, (\alpha + \beta x_i)\big)^2,$$

minimieren. Diese Idee kann als Anpassung der Momentenmethode angesehen werden: EY_i wird durch y_i ersetzt, an die Stelle der Auflösung nach α und β tritt die Approximation bzgl. des euklidischen Abstands. Eine etwas mühsame Rechnung führt auf

$$\begin{aligned} \hat{\alpha} &= \frac{(\sum_{i=1}^n x_i^2)(\sum_{i=1}^n y_i) - (\sum_{i=1}^n x_i)(\sum_{i=1}^n x_i y_i)}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}, \\ \hat{\beta} &= \frac{n\sum_{i=1}^n x_i y_i - (\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}. \end{aligned}$$

Setzt man zusätzlich voraus, dass die Y_i's normalverteilt sind, alle mit derselben (unbekannten) Varianz σ^2, so kann man Likelihood-Methoden verwenden: Um den Maximum-Likelihood-Schätzer für $(\alpha, \beta, \sigma^2)$ zu erhalten, müssen wir die Funktion

$$\begin{aligned} (\alpha, \beta, \sigma^2) \;&\mapsto\; \log\Big(\prod_{i=1}^{n} \phi\big(y_i \big| \alpha x_i + \beta, \sigma^2\big)\Big) \\ &=\; -\frac{n}{2}\,\log(2\pi\sigma^2) \,-\, \frac{1}{2\sigma^2}\,\sum_{i=1}^{n}\big(y_i - (\alpha + \beta x_i)\big)^2 \end{aligned}$$

maximieren (siehe auch Beispiel 5.3). Für die Parameter α und β ist dies äquivalent zu dem obigen Minimierungsproblem bei der Methode der kleinsten Quadrate, man erhält also dieselben Schätzer. ◁

Weitere Beispiele werden in den Übungen besprochen.

Wie beurteilt man die Qualität von Schätzfunktionen? Unser formales Modell geht von einem 'Hintergrundwahrscheinlichkeitsraum' $(\Omega, \mathcal{A}', \mathbb{P})$ aus; die

beobachteten Daten x werden als Werte (Realisierungen) einer Zufallsgröße $X : \Omega \to \mathcal{X}$ betrachtet (also: Großbuchstaben stehen für die Abbildung selbst, kleine Buchstaben für ihre Werte — eine Konvention, die wir allerdings nicht stets einhalten werden ...). Die Verteilung $\mathcal{L}(X)$ von X ist ein unbekanntes Element P von $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. Schätzfunktionen sind Abbildungen vom Datenraum $\mathcal{X}$ in den Parameterraum Θ. Im Falle $\Theta \subset \mathbb{R}$ ist $\hat{\theta}(X)$ in der Regel messbar (wir setzen dies in Zukunft stillschweigend voraus), also eine Zufallsvariable, deren Erwartungswert die Lage der Verteilung des Schätzers beschreibt. Verteilung und damit auch Erwartungswert hängen natürlich von der unbekannten Verteilung von X ab: Wir schreiben $E_\theta \hat{\theta}(X)$ oder kurz $E_\theta \hat{\theta}$ für den Erwartungswert von $\hat{\theta}(X)$ unter der Voraussetzung, dass $\mathcal{L}(X) = P_\theta$ gilt, also θ der wahre Parameter ist.

Ist $\Theta \subset \mathbb{R}$ oder betrachtet man allgemeiner eine reellwertige Parameterfunktion $g(\theta)$, so kann man die Differenz $\hat{\theta} - \theta$ bzw. $g(\hat{\theta}) - g(\theta)$ bilden. Wünschenswerte Eigenschaften eines Schätzers beziehen sich darauf, dass diese Differenz — die ja eine Zufallsgröße ist — in irgendeinem Sinne klein ist.

DEFINITION 5.7 Es sei $\hat{\eta}$ ein (messbarer) Schätzer für eine reellwertige Parameterfunktion $\eta = g(\theta)$. Wir setzen voraus, dass die im folgenden verwendeten Erwartungswerte existieren.

(i) Der Schätzer $\hat{\eta}$ heißt *erwartungstreu* (Englisch: *unbiased*) für $\eta = g(\theta)$, wenn gilt:

$$E_\theta \hat{\eta} = g(\theta) \quad \text{für alle } \theta \in \Theta,$$

die Differenz $E_\theta \hat{\eta} - g(\theta)$ ist der *systematische Fehler* oder *Bias* von $\hat{\eta}$.

(ii) Die *mittlere quadratische Abweichung* $\mathrm{MSE}(\,\cdot\,; \hat{\eta})$ von $\hat{\eta}$ wird definiert durch

$$\mathrm{MSE}(\theta; \hat{\eta}) := E_\theta\big(\hat{\eta} - g(\theta)\big)^2.$$

(MSE ist die Abkürzung für 'mean squared error').

Bei einem erwartungstreuen Schätzer ist der mittlere quadratische Fehler offensichtlich gleich der Varianz. Allgemein gilt

$$\mathrm{MSE}(\theta; \hat{\theta}) = \big(E_\theta \hat{\theta} - \theta\big)^2 + \mathrm{var}_\theta(\hat{\theta}).$$

BEISPIEL 5.8 Es seien $\mathcal{X} = \{0, \ldots, n\}$, $\Theta = (0, 1)$ und $P_\theta = \mathrm{Bin}(n, \theta)$. (Dies ist die aus Beispiel 5.2 bekannte Situation, wenn man dort nur die Anzahl k der Erfolge festhält.) Der Schätzer $\hat{\theta} = X/n$ ist offensichtlich erwartungstreu, denn X hat unter P_θ den Erwartungswert $n\theta$. Als mittleren quadratischen Fehler erhält man

$$\mathrm{MSE}(\theta; \hat{\theta}) = \mathrm{var}_\theta(\hat{\theta}) = \frac{1}{n^2}\,\mathrm{var}_\theta(X) = \frac{1}{n^2}\,n\theta(1-\theta) = \frac{\theta(1-\theta)}{n}.$$

Man kann zeigen, dass dieser Schätzer unter allen erwartungstreuen Schätzern für θ gleichmässig in $\theta \in (0,1)$ die kleinste mittlere quadratische Abweichung hat. (Dies gilt sogar im Rand: im Falle $\theta = 0$, $\theta = 1$ hat $\hat{\theta}$ den MSE 0, was nicht zu unterbieten ist.)

Was passiert, wenn man auch nicht-erwartungstreue Schätzer in die Konkurrenz aufnimmt? Klar: der 'entartete' Schätzer $\theta \equiv \theta_0$ für ein festes $\theta_0 \in \Theta$ hat MSE 0 in θ_0 (eine stehengebliebene Uhr zeigt zweimal am Tag die genaue Zeit an). Interessanter ist der Schätzer $\hat{\theta}_A := (X+1)/(n+2)$, der vermeidet, dass die Wahrscheinlichkeit durch 0 bzw. 1 geschätzt wird, wenn das interessierende Ereignis gar nicht bzw. immer eintritt. Man erhält

$$E_\theta \hat{\theta}_A \;=\; \frac{1}{n+2}(E_\theta X \,+\, 1) \;=\; \frac{n\theta + 1}{n+2}\,,$$

insbesondere ist $\hat{\theta}_A$ nicht erwartungstreu. Eine etwas längere Rechnung (oder Maple) liefert

$$E_\theta(\hat{\theta}_A - \theta)^2 \;=\; \frac{1 + (n-4)\theta(1-\theta)}{(n+2)^2},$$

und ein Vergleich der Funktionen zeigt, dass keiner der beiden Schätzer einen gleichmässig kleineren mittleren quadratischen Fehler hat als der andere. ◁

Beispiel 5.9 Es sei $X_1, \ldots, X_n$ eine Stichprobe aus unif$(0,\theta)$, der Gleichverteilung auf dem Intervall $(0,\theta)$ (siehe Abschnitt 4.5.1). Dann gilt $E_\theta X_i = \theta/2$, die Momentenmethode führt also auf $\hat{\theta}_{\mathrm{MM}} = 2\bar{X}_n$. Für die zugehörigen Dichten gilt $f(x|\theta) = 1/\theta$ für $0 \le x \le \theta$, $f(x|\theta) = 0$ sonst, also erhält man die Likelihood-Funktion

$$l(\theta) \;=\; \begin{cases} \theta^{-n}, & \text{falls } \theta \ge \max\{x_1, \ldots, x_n\}, \\ 0, & \text{sonst.} \end{cases}$$

Hier wird das globale Maximum auf dem Rand angenommen und man erhält $\hat{\theta}_{\mathrm{ML}} = \max\{X_1, \ldots, X_n\}$.

Welcher Schätzer ist besser? Es gilt $E_\theta X_i = \theta/2$, also

$$E_\theta \hat{\theta}_{\mathrm{MM}} \;=\; 2 \cdot \frac{1}{n} \sum_{i=1}^{n} E_\theta X_i \;=\; \theta,$$

d.h. $\hat{\theta}_{\mathrm{MM}}$ ist erwartungstreu. Als Verteilungsfunktion G_θ des Maximum-Likelihood-Schätzers ergibt sich

$$\begin{aligned} G_\theta(x) \;&=\; P_\theta(\hat{\theta}_{\mathrm{ML}} \le x) \\ &=\; P_\theta(X_1 \le x, \ldots, X_n \le x) \\ &=\; P_\theta(X_1 \le x) \cdot \ldots \cdot P_\theta(X_n \le x) \\ &=\; \Big(\frac{x}{\theta}\Big)^n \end{aligned}$$

für $0 \le x \le \theta$; für $x < 0$ gilt $G_\theta(x) = 0$ und für $x > \theta$ erhält man $G_\theta(x) = 1$. Eine zugehörige Dichte ist

$$g_\theta(x) \;=\; \begin{cases} \frac{1}{\theta} n (\frac{x}{\theta})^{n-1} & , 0 \le x \le \theta, \\ 0 & , \text{sonst}, \end{cases}$$

also folgt

$$E_\theta \hat\theta_{\mathrm{ML}} \;=\; \int x\, g_\theta(x)\, dx \;=\; \int_0^\theta x\, \frac{1}{\theta}\, n \left(\frac{x}{\theta}\right)^{n-1} dx \;=\; \frac{n}{n+1}\, \theta,$$

dieser Schätzer ist also nicht erwartungstreu — allerdings ist der systematische Fehler bei großem n klein. Für die mittleren quadratischen Abweichungen erhält man

$$\mathrm{MSE}(\hat\theta_{\mathrm{MM}}; \theta) \;=\; \mathrm{var}_\theta(\hat\theta_{\mathrm{MM}}) \;=\; \frac{4}{n^2} \sum_{i=1}^n \mathrm{var}_\theta(X_i) \;=\; \frac{\theta^2}{3n},$$

denn es gilt $\frac{1}{\theta} X_i \sim \mathrm{unif}(0,1)$ und damit $\mathrm{var}_\theta(X_i/\theta) = 1/12$ (siehe hierzu Beispiel 5.12 (i)). Beim Maximum-Likelihood-Schätzer erhält man

$$E_\theta\, \hat\theta^2_{\mathrm{ML}} \;=\; \int_0^\theta x^2 \frac{1}{\theta}\, n \left(\frac{x}{\theta}\right)^{n-1} dx \;=\; \frac{n}{n+2}\, \theta^2,$$

also

$$\begin{aligned} \mathrm{MSE}(\hat\theta_{\mathrm{ML}}; \theta) \;&=\; E_\theta\, \hat\theta^2_{\mathrm{ML}} \;-\; 2\theta E_\theta\, \hat\theta_{\mathrm{ML}} \;+\; \theta^2 \\ &=\; \frac{n}{n+2}\, \theta^2 - 2\theta\, \frac{n}{n+1}\, \theta + \theta^2 \;=\; \frac{2\theta^2}{(n+2)(n+1)}. \end{aligned}$$

Dies ist stets kleiner oder gleich dem für $\hat\theta_{\mathrm{MM}}$ erhaltenen Wert, echt kleiner ab $n = 3$ und bei großem n sehr viel kleiner! Ist man also bereit, einen (kleinen) systematischen Fehler zu akzeptieren, so wird man $\hat\theta_{\mathrm{ML}}$ bevorzugen. In einer Übungsaufgabe wird ein dritter Schätzer behandelt, der aus $\hat\theta_{\mathrm{ML}}$ hervorgeht und Erwartungstreue mit kleiner mittlerer quadratischer Abweichung verbindet. ◁

5.3 Tests. Es sei wieder $\mathcal{P}$ eine Familie von Wahrscheinlichkeitsmaßen auf $(\mathcal{X}, \mathcal{A})$. Oft soll anhand der Daten entschieden werden, ob die tatsächliche Verteilung P in einer vorgegebenen Teilfamilie $\mathcal{P}_0$ von $\mathcal{P}$ liegt, d.h. man will die Hypothese $H : P \in \mathcal{P}_0$ testen. Bei einer parametrisierten Familie $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ lässt sich die Teilfamilie über eine Teilmenge Θ_0 des Parameterraums Θ charakterisieren; die Hypothese lautet dann $H : \theta \in \Theta_0$, wobei θ für den 'wahren' Parameter steht. $K : \theta \in \Theta - \Theta_0$ (bzw. $K : \mathcal{P} - \mathcal{P}_0$) bezeichnet man als *Alternative*; man kann H und K auch als Zerlegung von Θ auffassen. H heißt *einfach* im Falle $\#\mathcal{P}_0 = 1$ bzw. $\#\Theta_0 = 1$ und *zusammengesetzt* sonst; analoge Bezeichnungen werden auch bei K verwendet.

DEFINITION 5.10 Eine (messbare) Funktion $\phi : \mathcal{X} \to [0,1]$ heißt *(randomisierte) Testfunktion zum Signifikanzniveau* α, kurz: Test zum Niveau α, wenn gilt:

$$E_P\phi(X) \leq \alpha \quad \text{für alle } P \in \mathcal{P}_0.$$

Die Abbildung $P \to E_P\phi(X)$ ist die *Gütefunktion* oder auch *Operationscharakteristik* des Tests; im parametrischen Fall ist dies

$$\beta : \Theta \to [0,1], \quad \beta(\theta) := E_\theta\phi(X).$$

Interpretation: Bei Vorliegen der Beobachtung x wird H mit Wahrscheinlichkeit $\phi(x)$ verworfen, also wird bei einem Test zum Niveau α die Wahrscheinlichkeit für eine irrtümliche Verwerfung der Hypothese nicht größer als α. Für α sind die Werte 0.1, 0.05, 0.01 und 0.001 gebräuchlich. Bei Tests geht es also darum, eine vorgegebene Hypothese anhand der Daten entweder zu verwerfen oder nicht zu verwerfen (beachte: 'nicht verwerfen' ist nicht dasselbe wie 'als richtig bewiesen'!). In der Regel wird man nicht-randomisierte Tests verwenden, bei denen also ϕ nur die Werte 0 und 1 annimmt. Die Menge $\{x \in \mathcal{X} : \phi(x) = 1\}$ ist dann der *Ablehnungsbereich* eines solchen Tests. Dieser wird häufig über eine *Testgröße* (auch: *Teststatistik*) T beschrieben, die die Eigenschaft hat, dass große Werte von T gegen H sprechen. In der Tat liefert eine solche Testgröße gleich eine ganze Familie von nicht-randomisierten Tests ϕ_c über

$$\phi_c(X) = \begin{cases} 1, & T(x) \geq c, \\ 0, & T(x) < c. \end{cases}$$

Man nennt in dieser Situation c den *kritischen Wert*.

Um diese Begriffe zu illustrieren, betrachten wir die folgende einfache Situation: Eine Münze wird zehnmal geworfen, θ bezeichne die unbekannte Wahrscheinlichkeit für Kopf, und es soll $H : \theta = 1/2$ getestet werden. Man ist also an der Hypothese interessiert, dass die Münze fair ist. Schreibt man wieder 1 für Kopf, 0 für Zahl und X_n für das Ergebnis des n-ten Wurfes, so liegt als Testgröße

$$T(X_1, \ldots, X_{10}) = \Big| \sum_{i=1}^{10} X_i - 5 \Big|$$

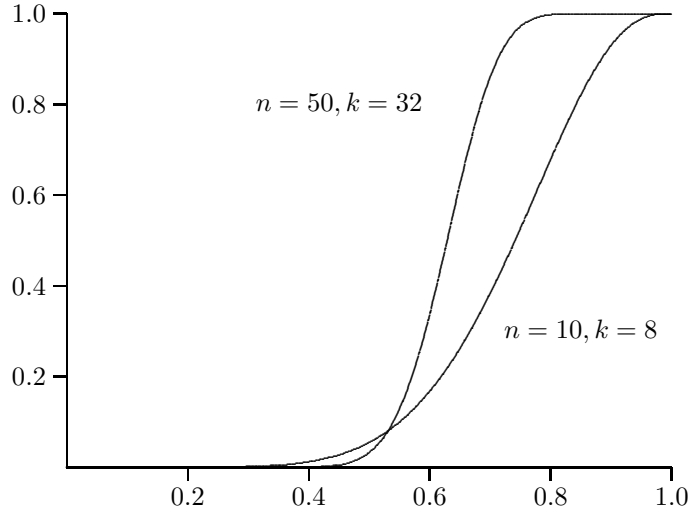
nahe: Große Werte von T sind unwahrscheinlich, wenn die Hypothese richtig ist. Angenommen, wir lehnen ab, wenn $T \geq 4$ gilt, d.h. wir wählen den kritischen Wert $c = 4$. Dies bedeutet, dass wir die Hypothese genau dann ablehnen, wenn 'Kopf' 0, 1, 9 oder 10mal vorkommt. Ist H richtig, so hat dieses Ereignis die Wahrscheinlichkeit

$$P_{0\cdot5}(T \geq 4) = \left(\binom{10}{0} + \binom{10}{1} + \binom{10}{9} + \binom{10}{10} \right) \cdot 2^{-10} = \frac{22}{1024} \approx 0.0215.$$

Dieses Verfahren würde also einen Test zum Niveau $\alpha = 0.05$, aber nicht zum Niveau $\alpha = 0.01$ liefern. Ganz allgemein gilt in dieser Situation

$$P_\theta(T \geq 4) = \binom{10}{0}\theta^0(1-\theta)^{10-0} + \binom{10}{1}\theta^1(1-\theta)^{10-1} + \binom{10}{9}\theta^9(1-\theta)^{10-9} + \binom{10}{10}\theta^{10}(1-\theta)^{10-10}.$$

Bei $\theta = 0.9$ beispielsweise erhält man den Wert 0.7361 und bei $\theta = 0.6$ den Wert 0.0480. Dies bedeutet, dass der Test bei $\theta = 0.9$ mit Wahrscheinlichkeit $1 - 0.7361 = 0.2639$ zu einer falschen Entscheidung führt, bei $\theta = 0.6$ immerhin mit Wahrscheinlichkeit 0.952!



Gütefunktionen zu zwei Tests:
$H_0: \ \theta \leq 0.5$ wird bei n Versuchswiederholungen verworfen,
wenn die Anzahl der Erfolge größer oder gleich k ist.

Analog kann man bei der einseitigen Hypothese $H : \theta \leq 1/2$ verfahren. Geht man ganz allgemein von n (statt wie oben speziell von $n = 10$) Versuchswiederholungen aus, so bietet sich die Variable $T(X_1, \ldots, X_n) = \sum_{i=1}^n X_i$ als Testgröße an, d.h. wir verwerfen die Hypothese, dass 'Kopf' mit einer Wahrscheinlichkeit kleiner oder gleich $1/2$ erscheint, wenn in n Würfen die Anzahl der 'Kopf'-Resultate eine bestimmte Schranke überschreitet. Im obigen Diagramm sind für zwei solche Tests, einmal bei $n = 10$ und kritischem Wert 8, und einmal bei $n = 50$ und kritischem Wert 32, die Gütefunktionen eingezeichnet.

Bei Tests geht es um nur zwei Entscheidungen: H wird verworfen oder H wird nicht verworfen. Als Folge hiervon gibt es zwei Fehlerarten:

- *Fehler 1. Art*: Die Hypothese wird verworfen, obwohl sie richtig ist.
- *Fehler 2. Art*: Die Hypothese wird nicht verworfen, obwohl sie falsch ist.

Für das Verständnis und den korrekten Gebrauch klassischer statistischer Tests ist die Unsymmetrie (nur für einen Typ Fehlentscheidung wird die Wahrscheinlichkeit begrenzt) ein sehr wichtiger Aspekt: Man hat in der Regel keine (brauchbare) Fehlerschranke für den Fehler zweiter Art. Es bietet sich ein Vergleich mit dem juristischen Prinzip 'im Zweifel für den Angeklagten' an: Eine Verurteilung soll nur bei hinreichend sicherer Beweislage erfolgen, ein Freispruch ist somit kein Unschuldsbeweis. Bei Tests: 'absence of evidence is not evidence of absence', eine Nicht-Ablehnung ist kein Beleg dafür, dass die Hypothese stimmt.

Die Wahrscheinlichkeit für eine falsche Entscheidung hängt natürlich von dem unbekannten wahren Parameter θ ab. Bei einem Test zum Niveau α darf die Wahrscheinlichkeit für einen Fehler 1. Art den Wert α nicht übersteigen. Alle Fehlerwahrscheinlichkeiten lassen sich aus der Gütefunktion ablesen. Man wird nun versuchen, bei einer vorgegebenen Schranke für den Fehler 1. Art einen Test zu finden, bei dem die Wahrscheinlichkeiten für einen Fehler 2. Art möglichst gleichmäßig minimiert werden. Bei einfacher Hypothese und einfacher Alternative (also bei $\#\mathcal{P} = 2$) kann man dieses Optimierungsproblem leicht lösen.

Satz 5.11 (Das Neyman-Pearson Lemma)
Es sei $\mathcal{P} = \{P_0, P_1\}$ und $\alpha \in (0,1)$. Wir setzen voraus, dass P_0 und P_1 entweder beide diskret sind oder beide ein Dichte haben, und schreiben p_0, p_1 für die Massenfunktionen im ersten und f_0, f_1 für die Dichten im zweiten Fall. Dann existieren ein $c \geq 0$ und ein $\gamma \in [0,1]$ mit

$$P_0(p_1 > cp_0) + \gamma P_0(p_1 = cp_0) = \alpha \text{ bzw. } P_0(f_1 > cf_0) + \gamma P_0(f_1 = cf_0) = \alpha$$

im diskreten bzw. stetigen Fall, und der Neyman-Pearson-Test $\phi : \mathcal{X} \to [0,1]$,

$$\phi(x) = \begin{cases} 1, & > \\ \gamma, & p_1(x) = cp_0(x) \\ 0, & < \end{cases} \quad \text{bzw.} \quad \phi(x) = \begin{cases} 1, & > \\ \gamma, & f_1(x) = cf_0(x) \\ 0, & < \end{cases}$$

im diskreten bzw. stetigen Fall ist ein Test zum Niveau α für $H : P = P_0$, der unter allen solchen Tests die kleinste Wahrscheinlichkeit für einen Fehler 2. Art hat.

BEWEIS: Wir betrachten nur den diskreten Fall. Der Beweis für den stetigen Fall verläuft sehr ähnlich, im wesentlichen müssen einige Summen durch Integrale ersetzt werden.

Wir können p_0 und p_1 als Zufallsvariablen auf dem Wahrscheinlichkeitsraum $(\mathcal{X}, \mathcal{A}, P_0)$ auffassen und erhalten beispielsweise

$$P_0(p_0 > 0) \;=\; \sum_{x\in\mathcal{X},\, p_0(x)>0} p_0(x) \;=\; \sum_{x\in\mathcal{X}} p_0(x) \;=\; P_0(\mathcal{X}) \;=\; 1.$$

Es sei c das $(1-\alpha)$-Quantil zur Verteilung von q,

$$q(x) \;:=\; \begin{cases} p_1(x)/p_0(x), & \text{falls } p_0(x) > 0, \\ 0, & \text{sonst.} \end{cases}$$

Aus unseren allgemeinen Betrachtungen zu Quantilfunktionen (Lemma 4.19, Übungsaufgaben) folgt dann, dass

$$P_0(q > c) \;\le\; \alpha \;\le\; P_0(q \ge c)$$

gilt. Wir setzen $\gamma := 0$ im Falle $P_0(q = c) = 0$ und

$$\gamma \;:=\; \frac{\alpha - P_0(q > c)}{P_0(q = c)}$$

sonst. Mit diesen Werten erhält man

$$\begin{aligned} P_0(p_1 > cp_0) + \gamma P(p_1 = cp_0) \;&=\; P_0(p_1 > cp_0, p_0 > 0) + \gamma P(p_1 = cp_0, p_0 > 0) \\ &=\; P_0(q > c) + \gamma P(q = c) \\ &=\; \alpha, \end{aligned}$$

womit der erste Teil der Behauptung bewiesen wäre.

Für den Beweis des zweiten (und interessanteren) Teils sei $\tilde\phi$ irgendein Test zum Niveau α für $H : P = P_0$. Wir setzen

$$A := \big\{x \in \mathcal{X} :\, \phi(x) > \tilde\phi(x)\big\}, \quad B := \big\{x \in \mathcal{X} : \phi(x) < \tilde\phi(x)\big\}.$$

Auf A ist $\phi > 0$, also $p_1 \ge cp_0$, auf B ist $\phi(x) < 1$, also $p_1 \le cp_0$. Damit folgt

$$\begin{aligned} E_1\phi(X) - E_1\tilde\phi(X) \;&=\; \sum_{x\in\mathcal{X}} \big(\phi(x) - \tilde\phi(x)\big) p_1(x) \\ &=\; \sum_{x\in A} \big(\phi(x) - \tilde\phi(x)\big) p_1(x) + \sum_{x\in B} \big(\phi(x) - \tilde\phi(x)\big) p_1(x) \\ &\ge\; \sum_{x\in A} \big(\phi(x) - \tilde\phi(x)\big) cp_0(x) + \sum_{x\in B} \big(\phi(x) - \tilde\phi(x)\big) cp_0(x) \\ &=\; c\sum_{x\in\mathcal{X}} \big(\phi(x) - \tilde\phi(x)\big) p_0(x) \\ &=\; c\big(E_0\phi(X) - E_0\tilde\phi(X)\big) \\ &\ge\; 0, \end{aligned}$$

denn $E_0\phi(X) = \alpha$, $E_0\tilde\phi(X) \le \alpha$. □

Der optimale Test hängt also nur über das Verhältnis p_1/p_0 bzw. f_1/f_0, den sogenannten *Likelihood-Quotienten*, von x ab. Der Ablehnungsbereich entsteht dadurch, dass man die x-Werte mit den größten Likelihood-Quotienten zusammenfasst, soweit dies die Fehlerschranke erlaubt. Dies ist eine auch intuitiv naheliegende Vorgehensweise.

BEISPIEL 5.12 Wie in Beispiel 5.2 sei $\mathcal{X} = \{0,1\}^n$,

$$p(x|\theta) = \theta^{T(x)}(1-\theta)^{n-T(x)} \quad \text{mit } T(x) = \sum_{i=1}^{n} x_i.$$

Wir betrachten zunächst die Familie $\mathcal{P} = \{P_{\theta_0}, P_{\theta_1}\}$ mit $0 < \theta_0 < \theta_1 < 1$ fest. Als Verhältnis der Massenfunktionen ergibt sich

$$\frac{p_1(x)}{p_0(x)} \;=\; \Big(\frac{1-\theta_1}{1-\theta_0}\Big)^{n-T(x)} \Big(\frac{\theta_1}{\theta_0}\Big)^{T(x)}.$$

Wegen $\theta_1 > \theta_0$ ist dies eine streng monoton wachsende Funktion von $T(x)$, d.h. zu jedem c existiert ein $\tilde{c}$ mit der Eigenschaft, dass

$$\begin{array}{c} > \\ p_1(x) = cp_0(x) \\ < \end{array} \quad \Longleftrightarrow \quad \begin{array}{c} > \\ T(x) = \tilde{c} \\ < \end{array}$$

für alle $x \in \mathcal{X}$ gilt. Nach dem Neymann-Pearson-Lemma ist also der beste Test für θ_0 gegen θ_1 von der Form

$$\phi(x) = \begin{cases} 1, & > \\ \gamma, & \sum_{i=1}^n x_i = \tilde{c} \\ 0, & < \end{cases},$$

wobei $\tilde{c}$ und $\gamma \in [0,1]$ bestimmt werden aus

$$P_{\theta_0}\Big(\sum_{i=1}^{n} X_i > \tilde{c}\Big) \;+\; \gamma\, P_{\theta_0}\Big(\sum_{i=1}^{n} X_i = \tilde{c}\Big) \;=\; \alpha.$$

(Die Überlegung, dass streng monoton wachsende Transformationen der Testgröße bei entsprechender Transformation des kritischen Werts den Test unverändert lassen, kann bei Rechnungen sehr hilfreich sein.) Man beachte nun, dass in der Beschreibung des Tests θ_1 nicht mehr auftritt; nur $\theta_1 > \theta_0$ wurde in der Herleitung verwendet. Die Hypothese $H : \theta = \theta_0$ gegen $K : \theta = \tilde{\theta}_1$ würde auf denselben Test führen, wenn nur $\tilde{\theta}_1 > \theta_0$ gilt. Dies zeigt, dass ϕ unter

allen Tests zum Niveau α für $H : \theta = \theta_0$ gegen $K : \theta > \theta_0$ gleichmäßig die Fehlerwahrscheinlichkeiten 2. Art minimiert, ϕ ist also ein *gleichmäßig bester Test* zum Niveau α für $\theta = \theta_0$ gegen $\theta > \theta_0$. Es kommt sogar noch besser: Jeder Test zum Niveau α für $H : \theta \leq \theta_0$ gegen $K : \theta > \theta_0$ ist auch ein Test zum Niveau α für $H : \theta = \theta_0$ gegen $K : \theta > \theta_0$. Da $E_\theta \phi$ eine monoton wachsende Funktion von θ ist, hält ϕ auch in dieser größeren Hypothese das Niveau α ein, minimiert also auch in dieser Klasse gleichmäßig die Fehlerwahrscheinlichkeiten zweiter Art. Gelegentlich lassen sich also mit Hilfe des Neyman-Pearson-Lemmas optimale Tests sogar bei zusammengesetzten Hypothesen und Alternativen bestimmen. ◁

BEISPIEL 5.13 Die Zufallsvariablen $X_1, \ldots, X_n$ seien unabhängig und exponentialverteilt mit unbekanntem Parameter $\theta > 0$. Anhand der Realisierungen soll

$$H : \theta = \theta_0 \quad \text{gegen} \quad K : \theta = \theta_1$$

getestet werden. Wir betrachten den Fall $\theta_1 > \theta_0$. Die Dichtefunktion zu $X = (X_1, \ldots, X_n)$ ist

$$f(x|\theta) \;=\; \prod_{i=1}^{n} \theta e^{-\theta x_i} \;=\; \theta^n e^{-\theta s_n} \quad \text{mit } s_n := \sum_{i=1}^{n} x_i \,.$$

Wie in Beispiel 5.12 ist für den optimalen Test nur die Realisierung s_n der Summe $S_n = \sum_{i=1}^n X_i$ der Zufallsvariablen relevant. Satz 5.11 führt mit $f_i(x) = f(x|\theta_i)$, $i = 0, 1$, auf die Testgröße

$$\frac{f_1(x)}{f_0(x)} \;=\; \Big(\frac{\theta_1}{\theta_0}\Big)^n e^{-(\theta_1 - \theta_0) s_n}.$$

Wegen $\theta_1 > \theta_0$ ist dies eine streng monoton fallende Funktion von s_n, der Neyman-Pearson-Test also von der Form

$$\phi(x) \;=\; \begin{cases} 1, & < \\ \gamma, & \sum_{i=1}^n x_i = \tilde{c} \\ 0, & > \end{cases},$$

wobei wieder $\tilde{c}$ und $\gamma \in [0, 1]$ bestimmt werden aus

$$P_0(S_n < \tilde{c}) \;+\; \gamma\, P_0(S_n = \tilde{c}) \;=\; \alpha.$$

Unter P_0 ist S_n $\Gamma(n, \theta_0)$-verteilt, insbesondere gilt also $P_0(S_n = c) = 0$ für alle $c \in \mathbb{R}$ und eine Randomisierung wird nicht benötigt. Der zweite Parameter der Gammaverteilung repräsentiert nur eine Umskalierung, insbesondere ist

$\theta_0 S_n$ unter der Hypothese $\Gamma(n,1)$-verteilt. Einer Tafel für die unvollständige Gammafunktion entnimmt man den Wert c mit

$$\int_0^{\theta_0 c} x^{n-1} e^{-x}\, dx \;=\; \alpha\, \Gamma(n)$$

(alternativ kann beispielsweise im Computeralgebra-Programm `Maple` die linke Seite mit `GAMMA(n,c)` berechnet werden), dieses c ist der kritische Wert bei Signifikanzniveau α. Wie im letzten Beispiel ergibt sich auch hier für alle Alternativwerte $\theta_1 > \theta_0$ derselbe Test, und die Wahrscheinlichkeit für eine Ablehnung wird mit fallendem θ kleiner, d.h. der Neyman-Pearson-Test ist sogar der gleichmäßig beste Test zum Niveau α für $H : \theta \le \theta_0$ gegen $K : \theta > \theta_0$. ◁

Hat man ganz allgemein eine parametrische Familie $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ von für die Beobachtungen in Frage kommenden Verteilungen (durchaus mit mehrdimensionalem Parameterraum Θ), so lassen sich Hypothese und Alternative durch Teilmengen von Θ beschreiben, d.h. man möchte

$$H : \theta \in \Theta_0 \quad \text{gegen} \quad K : \theta \in \Theta_1 := \Theta - \Theta_0$$

testen. Sind die Verteilungen P_θ, $\theta \in \Theta$, alle diskret oder alle stetig, so machen die bisher behandelten Ideen das folgende Vorgehen plausibel: Schätze θ durch die Werte, die die Likelihood-Funktion $\theta \mapsto l(\theta|x)$ (wobei wieder $l(\theta|x) = p(x|\theta)$ im diskreten und $l(\theta|x) = f(x|\theta)$ im stetigen Fall) auf Θ_0 bzw. Θ_1 maximieren und verwende den dann erhaltenen Dichtequotienten als Testgröße. Dies führt auf den *Likelihood-Quotienten-Test* (oder kurz LQ-Test), der ablehnt, wenn die Testgröße

$$T_{\mathrm{LQ}}(x) \;=\; \frac{\sup_{\theta\in\Theta_1} l(\theta|x)}{\sup_{\theta\in\Theta_0} l(\theta|x)}$$

einen durch die Forderung

$$\sup_{\theta\in\Theta_0} P_\theta(T \ge c) \;=\; \alpha$$

festgelegten kritischen Wert c übersteigt (man kann auch hier wieder randomisieren, wenn beispielsweise im diskreten Fall ein solches c nicht existiert).

Beispiel 5.14 Wir gehen aus von einer Stichprobe $X_1, \ldots, X_n$ aus einer Normalverteilung $N(\mu, \sigma^2)$ mit unbekannten $\mu \in \mathbb{R}$, $\sigma^2 > 0$ und wollen

$$H : \mu = \mu_0 \quad \text{gegen} \quad K : \mu \ne \mu_0$$

zum Niveau α testen (μ_0 und α sind vorgegeben). Dies passt in den oben beschriebenen Rahmen, mit $\theta = (\mu, \sigma^2)$,

$$\Theta = \mathbb{R} \times (0,\infty), \quad \Theta_0 = \{\mu_0\} \times (0,\infty), \quad \Theta_1 = \big(\mathbb{R} \setminus \{\mu_0\}\big) \times (0,\infty).$$

Zur Bestimmmung des LQ-Tests müssen wir die Funktion

$$l(\theta|x) \;=\; (2\pi\sigma^2)^{-n/2} \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2\Big)$$

auf Θ_1 bzw. Θ_0 maximieren. Da diese Funktion stetig ist und Θ_1 dicht liegt in Θ, gilt

$$\sup_{\theta\in\Theta_1} l(x|\theta) \;=\; \sup_{\theta\in\Theta} l(x|\theta)$$

und mit den Rechnungen aus Beispiel 5.3 (die ML-Schätzer sind $\hat\mu = \bar x_n$ und $\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^n (x_i - \bar x_n)^2$) folgt

$$\sup_{\theta\in\Theta_1} l(x|\theta) \;=\; (2\pi\widehat{\sigma^2})^{-n/2}\, e^{-n/2}.$$

Zur Bestimmung des Nenners der Testgröße muss l auf Θ_0 maximiert werden, wodurch $\mu = \mu_0$ festgelegt ist. Das Maximum der Funktion

$$\sigma^2 \;\mapsto\; (2\pi\sigma^2)^{-n/2} \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu_0)^2\Big)$$

wird in $\widetilde{\sigma^2} := \frac{1}{n}\sum_{i=1}^n (x_i-\mu_0)^2$ angenommen, also gilt

$$\sup_{\theta\in\Theta_0} l(x|\theta) \;=\; (2\pi\widetilde{\sigma^2})^{-n/2}\, e^{-n/2}$$

und man erhält insgesamt die Testgröße

$$\begin{aligned} T_{\mathrm{LQ}}(x) \;=\; \left(\frac{\widetilde{\sigma^2}}{\widehat{\sigma^2}}\right)^{n/2} &= \left(\frac{\sum_{i=1}^n (x_i - \bar x_n + \bar x_n - \mu_0)^2}{\sum_{i=1}^n (x_i-\bar x_n)^2}\right)^{n/2} \\ &= \left(1 + \frac{(\bar x_n - \mu_0)^2}{\widehat{\sigma^2}}\right)^{n/2}. \end{aligned}$$

Dies ist offensichtlich eine streng monoton wachsende Funktion von

$$T(x) \;=\; \frac{|\bar x_n - \mu_0|}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n (x_i - \bar x_n)^2}},$$

man erhält also denselben Test, wenn man als Testgröße T verwendet. Dies ergibt den *zweiseitigen t-Test* zur Hypothese $\mu = \mu_0$ bei Stichproben aus der Normalverteilung mit unbekannter Varianz.

Zur praktischen Ausführbarkeit muss allerdings noch die Verteilung der Testgröße unter der Hypothese bestimmt werden. Da die Hypothese nun aus mehr als einem Wert besteht, ist zunächst nicht einmal klar, ob nicht sogar mehrere Verteilungen, abhängig von dem unbekanntem σ^2, erscheinen. Zumindest diese Frage können wir bereits jetzt beantworten: Sind $X_1, \dots, X_n$ unabhängig und $N(\mu_0, \sigma^2)$-verteilt, so sind die Zufallsvariablen $Y_1, \dots, Y_n$ mit $Y_i := (X_i - \mu_0)/\sigma$ unabhängig (Satz 4.30) und $N(0,1)$-verteilt (Lemma 4.23 (c)). Man überprüft leicht, dass mit $\bar{Y}_n := \frac{1}{n}\sum_{i=1}^n Y_i$

$$T(X_1, \dots, X_n) \;=\; \frac{|\bar{Y}_n|}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y}_n)^2}}$$

gilt. Auf der rechten Seite sind μ_0 und σ^2 verschwunden, $T(X)$ hat also unter allen Verteilungen, für die die Hypothese richtig ist, eine feste Verteilung; diese hängt nicht von μ_0 ab. Es stellt sich heraus, dass diese Größe, nach Beseitigung der Betragsstriche, die *t-Verteilung mit $n-1$ Freiheitsgraden* hat; dies ist die Verteilung mit der Dichte

$$x \;\mapsto\; \frac{1}{\sqrt{\pi(n-1)}}\,\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})}\left(1 + \frac{x^2}{n-1}\right)^{-n/2}, \quad -\infty < x < \infty$$

(genaueres in der Vorlesung Stochastik II). ◁

Bemerkung 5.15 (a) Klassische Tests laufen in den folgenden Schritten ab: Zunächst wird die Hypothese festgelegt, dann eine geeignete Testgröße T gewählt. (Grob gilt, dass große Werte von T gegen die Hypothese sprechen sollen. Die Testgröße bestimmt letztlich, welche Abweichungen von der Hypothese der Test bevorzugt entdeckt; die Wahl sollte daher von der Alternative abhängen.) Bei nicht-randomisierten Tests mit einem Ablehnungsbereich von der Form $\{x \in \mathcal{X} : T(x) \geq c\}$ geht das Signifikanzniveau α nur über den kritischen Wert $c = c(\alpha)$ ein. Dieses Signifikanzniveau wird nun vor Ausführung des Experiments festgelegt, und nach Erhebung der Daten x und Berechnung von $T(x)$ die Entscheidung (Ablehnung/keine Ablehnung) festgehalten; bei Ablehnung der Hypothese $H : \mu \leq 0$ beispielsweise in der Form 'die Aussage $\mu > 0$ ist statistisch auf dem Niveau α abgesichert'. Hieraus geht nicht hervor, ob nicht vielleicht sogar für ein kleineres α auch eine Ablehnung erzielt worden wäre oder ob nicht ein weniger stringentes α doch eine Ablehnung geliefert hätte. Man gibt daher häufig anstelle eines Signifikanzniveaus den *p-Wert* der

Beobachtung x an: Dies ist der kleinste α-Wert, der noch zu einer Ablehnung der Hypothese geführt hätte. Der p-Wert ist somit die maximale Wahrscheinlichkeit, unter der Hypothese, dass die Testgröße mindestens so groß ist wie der tatsächlich beobachtete Wert. Der Übergang von einem festgelegten Signifikanzniveau zu p-Werten vermeidet einen Informationsverlust und überlässt letztlich dem Anwender die Wahl des Signifikanzniveaus.

(b) Wie aus dem Beweis zu Satz 5.11 hervorgeht, dient Randomisierung der Ausschöpfung der zugelassenen Fehlerwahrscheinlichkeit 1. Art. Als konkretes Beispiel betrachten wir die Hypothese, dass 'Kopf' bei einer gegebenen Münze höchstens mit Wahrscheinlichkeit $1/2$ erscheint. Soll dies durch zehnmaligen Wurf überprüft werden, so führt Beispiel 5.12 auf die Anzahl T der 'Kopf'-Würfe als Testgröße. Es gilt $P_{0.5}(T \geq 9) = 0.0108\ldots$, $P_{0.5}(T \geq 8) = 0.0546\ldots$, also ist der beste Test zum Niveau $\alpha = 0.05$ wegen

$$\gamma \;=\; \frac{\alpha - P_{0.5}(T \geq 9)}{P_{0.5}(T = 8)} \;=\; 0.89\ldots$$

von der Form

$$\phi(x) = \begin{cases} 1, & > \\ 0.89\ldots, & \sum_{i=1}^n x_i = 8 \\ 0, & < \end{cases} .$$

Wird nun die Münze zehnmal geworfen, so ist man nur im Falle $T < 8$ oder $T > 8$ fertig: Bei $T = 8$ wird ein weiteres, vom bisherigen Geschehen unabhängiges Zufallsexperiment ausgeführt, in dem mit Wahrscheinlichkeit $0.89\ldots$ ein bestimmtes Ereignis A eintritt. Erscheint tatsächlich A, so wird die Hypothese abgelehnt, sonst nicht.

Randomisierung wird von vielen Praktikern als mathematische Spielerei angesehen. Im Sinne von Teil (a) würde man beim Erhalt von achtmal 'Kopf' stattdessen angeben, dass man mit diesem Resultat bei $\alpha \geq 0.0108\ldots$ eine Ablehnung erhalten hätte. ◁

5.4 Konfidenzbereiche. Die Daten x seien wieder Realisierungen einer Zufallsgröße X, deren Verteilung ein unbekanntes Element einer vorgegebenen Familie $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ ist. Neben dem direkten Schätzen des Parameters θ und dem Testen von Aussagen über θ ist die Konstruktion von Konfidenzbereichen das dritte Standardverfahren der Statistik, man spricht hier auch von *Bereichsschätzern*. Jedem $x \in \mathcal{X}$ wird hierbei eine Teilmenge $C(x)$ des Parameterraums Θ zugeordnet. Gilt

$$P_\theta\big(C(X) \ni \theta\big) \;\geq\; 1 - \alpha \quad \text{für alle } \theta \in \Theta,$$

so nennt man $C(X)$ ein $100(1-\alpha)$-*prozentiges Konfidenzgebiet für* θ. Natürlich muss $\{x \in \mathcal{X} : C(x) \ni \theta\}$ für alle $\theta \in \Theta$ eine messbare Teilmenge des

Stichprobenraums sein. Ist $C(X)$ ein Intervall, so spricht man naheliegenderweise von einem *Konfidenzintervall*, bei $C(X) = (-\infty, \overline{\theta}(X)]$ nennt man $\overline{\theta}(X)$ eine *obere Konfidenzschranke zum Niveau* $1-\alpha$ etc.. Für α sind wieder die Werte $0.1, 0.05, 0.01, 0.001$ gebräuchlich. Wie bei Schätzern ist man auch hier u.U. nicht an dem gesamten Parameter θ, sondern nur an einem Teil $\eta = g(\theta)$ interessiert; die Ausdehnung dieser Konzepte auf solche Parameterfunktionen dürfte klar sein.

Beispiel 5.16 Ist $X_1, \ldots, X_n$ eine Stichprobe aus der Exponentialverteilung mit unbekanntem Parameter $\theta > 0$, so sind die Zufallsvariablen $\theta X_1, \ldots, \theta X_n$ unabhängig und exponentialverteilt mit Parameter 1, und nach einer Übungsaufgabe ist $Y := \min\{\theta X_1, \ldots, \theta X_n\}$ dann exponentialverteilt mit Parameter n. Es gilt also

$$P_\theta\Big(\theta \geq \frac{z}{\min\{X_1, \ldots, X_n\}}\Big) \;=\; P_\theta\big(\theta \min\{X_1, \ldots, X_n\} \geq z\big) \;=\; e^{-nz}$$

für alle $\theta \in \Theta = (0, \infty)$ und alle $z > 0$. Wählt man nun z in Abhängigkeit vom Stichprobenumfang n und dem gewählten Konfidenzniveau α so, dass $e^{-nz} = 1 - \alpha$ gilt, so erhält man mit

$$\underline{\theta}(X) \;=\; \frac{-\frac{1}{n}\log(1-\alpha)}{\min\{X_1, \ldots, X_n\}}$$

eine $100(1-\alpha)\%$-Konfidenzunterschranke für θ. ◁

Ein Konfidenzbereich $C(X)$ ist eine zufällige Menge, die den unbekannten Parameter θ mit einer bestimmten Wahrscheinlichkeit, dem Konfidenzniveau, überdeckt (enthält). Setzt man für X die Daten x ein, so erhält man eine Realisierung des Konfidenzbereichs, die den unbekannten Parameter entweder enthält oder nicht enthält. Ergibt sich beispielsweise das Intervall $[2.5, 3.1]$, so wird häufig, *aber falsch*, formuliert: 'das Intervall $[2.5, 3.1]$ enthält den unbekannten Parameter θ mit Wahrscheinlichkeit 0.95'. Ein ähnliches Missverständnis ist auch bei Anwendern statistischer Tests weit verbreitet: Wird eine Hypothese auf dem Niveau α abgelehnt, so heißt dies nicht, dass sie mit Wahrscheinlichkeit $1-\alpha$ falsch ist. Zur Verdeutlichung betrachten wir einen analogen Sachverhalt beim Würfelwurf: Die Augenzahl X nimmt mit Wahrscheinlichkeit $1/6$ den Wert 2 an — wurde geworfen und beispielsweise der Wert $x = 5$ erhalten, so heißt dies nicht, dass 5 mit Wahrscheinlichkeit $1/6$ gleich 2 ist! Es bleibt dem Experimentator natürlich unbenommen, Konfidenzintervalle mit subjektiven Wahrscheinlichkeiten im Sinne von Abschnitt 1 dieser Vorlesung zu verbinden und somit zu einer Aussage der Form 'die Stärke

meines Glaubens daran, dass das Intervall $[2.5, 3.1]$ den unbekannten Parameter θ enthält, hat den Wert 0.9' zu kommen.

Zwischen den Ablehnungsbereichen von Tests einfacher Hypothesen und Konfidenzbereichen besteht ein gelegentlich nützlicher Zusammenhang.

SATZ 5.17 *Für jedes $\theta_0 \in \Theta$ sei $A(\theta_0) \subset \mathcal{X}$ Ablehnungsbereich eines nichtrandomisierten Tests zum Niveau α für $H : \theta = \theta_0$ gegen $K : \theta \neq \theta_0$. Dann ist C,*

$$C(X) := \{\theta \in \Theta : X \notin A(\theta)\}$$

ein Konfidenzbereich zum Niveau $1 - \alpha$ für θ.

BEWEIS: Die Aussage ergibt sich sofort aus

$$P_\theta\big(C(X) \ni \theta\big) \;=\; P_\theta\big(X \notin A(\theta)\big) \;=\; 1 - P_\theta\big(X \in A(\theta)\big) \;\geq\; 1 - \alpha.$$

□

Eine weitere im Zusammenhang mit der Konstruktion von Konfidenzbereichen sehr nüzliche Idee ist die des *Pivots* (Englisch für 'Drehpunkt'): Hat man eine Funktion $h : \mathcal{X} \times \Theta \to \mathcal{Y}$ mit den Eigenschaften, dass erstens die Verteilung Q von $h(X, \theta)$ bei $\mathcal{L}(X) = P_\theta$ nicht von θ abhängt und dass zweitens Mengen der Form $\{x \in \mathcal{X} : h(x, \theta) \in A\}$ nach θ aufgelöst werden können (hier hat man oft eine Art 'Drehung'), so erhält man durch $C(X)$ mit $C(x) := \{\theta \in \Theta : h(x, \theta) \in A\}$ einen $100(1-\alpha)\%$-Konfidenzbereich, wenn man für A eine Menge mit $Q(A) \geq 1 - \alpha$ wählt. In Beispiel 5.16 ist $h(x, \theta) := \theta \min\{x_1, \ldots, x_n\}$ ein solcher Pivot, ein anderer (und besserer) ist $h(x, \theta) := \theta \sum_{i=1}^n x_i$.

Der Zusammenhang von Tests und Konfidenzintervallen, die Idee des Pivots und schließlich der Umgang mit Parameterfunktionen werden im folgenden Beispiel illustriert, bei dem es um Konfidenzbereiche für den Mittelwert bei normalverteilten Größen geht.

BEISPIEL 5.18 Es sei $X_1, \ldots, X_n$ eine Stichprobe aus $N(\mu, \sigma^2)$, wobei sowohl μ als auch $\sigma^2 (> 0)$ als unbekannt betrachtet werden. Es seien wieder

$$\bar{X}_n \;=\; \frac{1}{n} \sum_{i=1}^n X_i, \quad S_n^2 \;=\; \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2,$$

der Stichprobenmittelwert und die Stichprobenvarianz. Bereits beim t-Test in Beispiel 5.14 wurde verwendet, dass $\sqrt{n}(\bar{X}_n - \mu)/S_n$ eine t-Verteilung mit $n-1$ Freiheitsgraden hat. Bezeichnet wieder $t_{n-1;1-\alpha}$ das $(1-\alpha)$-Quantil zu dieser Verteilung, so gilt daher

$$P_{\mu,\sigma^2}\Big(\sqrt{n}\,\frac{\bar{X}_n - \mu}{S_n} \leq t_{n-1;1-\alpha}\Big) \;=\; 1 - \alpha \quad \text{für alle } \mu \in \mathbb{R},\, \sigma^2 > 0.$$

Unter Verwendung der einfachen Umformung

$$\sqrt{n}\,\frac{\bar{X}_n - \mu}{S_n} \;\le\; t_{n-1;1-\alpha} \quad\Longleftrightarrow\quad \mu \;\ge\; \bar{X}_n - \frac{1}{\sqrt{n}}\,S_n\,t_{n-1;1-\alpha}$$

(dies entspricht der oben erwähnten Auflösung oder 'Drehung') folgt hieraus, dass

$$\underline{\mu} \;=\; \bar{X}_n \;-\; \frac{1}{\sqrt{n}}\,S_n\,t_{n-1;1-\alpha}$$

eine $100(1-\alpha)\%$-Konfidenzunterschranke für μ ist. Ganz analog sieht man, dass

$$\Big[\bar{X}_n \;-\; \frac{1}{\sqrt{n}}\,S_n\,t_{n-1;1-\alpha/2}\,,\;\bar{X}_n \;+\; \frac{1}{\sqrt{n}}\,S_n\,t_{n-1;1-\alpha/2}\Big]$$

ein $100(1-\alpha)\%$-Konfidenzintervall für μ ist. ◁

Die obigen Beispiele beziehen sich alle auf stetige Verteilungen. In der Tat sind Konfidenzintervalle bei diskreten Verteilungen oft ein recht mühsames Geschäft. Wir bringen ein Beispiel, Konfidenzintervalle für Wahrscheinlichkeiten, bei dem asymptotische Überlegungen zu einer Vereinfachung führen.

Beispiel 5.19 Es seien wieder einmal $X_1, X_2, \ldots$ unabhängig und $\mathrm{Bin}(1,\theta)$-verteilt mit unbekanntem $\theta \in (0,1)$. Wir verwenden $\bar{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$ als Schätzer für θ (siehe auch Beispiel 5.2). Nach dem Satz von de Moivre-Laplace (Satz 4.24) gilt mit $S_n = \sum_{i=1}^n X_i = n\bar{X}_n$

$$\lim_{n\to\infty} P_\theta\Big(a \le \frac{S_n - n\theta}{\sqrt{n\theta(1-\theta)}} \le b\Big) \;=\; \Phi(b) - \Phi(a),$$

wobei wieder Φ die Verteilungsfunktion zur Standardnormalverteilung bezeichnet. Ist u_α das zugehörige α-Quantil, also $\Phi(u_\alpha) = \alpha$, so folgt mit $b := u_{1-\alpha/2}$, $a := -b$ bei großem n

$$P_\theta\Big(-u_{1-\alpha/2} \le \frac{S_n - n\theta}{\sqrt{n\theta(1-\theta)}} \le u_{1-\alpha/2}\Big) \;\approx\; 1-\alpha,$$

denn $\Phi(-u_{1-\alpha/2}) = 1-\Phi(u_{1-\alpha/2}) = 1-(1-\alpha/2) = \alpha/2$. Wegen $\theta(1-\theta) \le 1/4$ gilt

$$-u_{1-\alpha/2} \;\le\; \frac{S_n - n\theta}{\sqrt{n\theta(1-\theta)}} \;\le u_{1-\alpha/2}$$

$$\Longrightarrow \quad \bar{X}_n - \frac{u_{1-\alpha/2}}{2\sqrt{n}} \;\le\; \theta \;\le\; \bar{X}_n + \frac{n_{1-\alpha/2}}{2\sqrt{n}},$$

also ergibt sich

$$\Big[\bar{X}_n - \frac{1}{2\sqrt{n}}\, u_{1-\alpha/2}\,,\, \bar{X}_n + \frac{1}{2\sqrt{n}} u_{1-\alpha/2}\Big]$$

als (asymptotisches, konservatives) $100(1-\alpha)\%$-Konfidenzintervall für θ.
Bemerkenswert ist hier, dass die Länge des Intervalls mit $1/\sqrt{n}$ fällt; für eine weitere Dezimalstelle müsste man also den Stichprobenumfang verhundertfachen. Numerisches Beispiel: Soll bei einer Wahl ein Konfidenzintervall für die Anzahl der Stimmen einer Partei von der Form 'Prozentsatz in Stichprobe $\pm$ 1%' auf dem Niveau 0.95 erhalten werden, so muss

$$\frac{1}{2\sqrt{n}}\, u_{0.975} \;\leq\; 0.01$$

gelten. Mit $u_{0.975} = 1.96\ldots$ ergibt sich $n \geq 9604$; bei $\pm 0.1\%$ würde man schon $n \geq 960400$ benötigen. (Bei Umfragen werden in der Regel kompliziertere Verfahren verwendet, die von zusätzlicher Information, beispielsweise über das Wahlverhalten bestimmter Personenkreise, Gebrauch machen.) $\lhd$