Vorlesungsnotizen

Einführung in die Stochastik

Hanspeter Schmidli

Mathematisches Institut
der Universität zu Köln

Inhaltsverzeichnis

1. Diskrete Wahrscheinlichkeitsräume

1.1. Grundbegriffe

1.1.1. Der Raum der Elementarereignisse

Sei Ω eine abzählbare Menge, das heisst, Ω ist endlich, oder man kann die Elemente wie $\mathbb{N}$ numerieren.

Beispiele

- Würfeln mit einem Würfel

$$\Omega = \{1, 2, 3, 4, 5, 6\} .$$

- Wurf von zwei nicht unterscheidbaren Münzen

$$\Omega = \{(K, K), (K, Z), (Z, Z)\} \qquad \text{oder} \qquad \Omega = \{(0, 0), (0, 1), (1, 1)\} .$$

- Wurf von zwei unterscheidbaren Münzen

$$\Omega = \{(K, K), (K, Z), (Z, K), (Z, Z)\}$$

oder

$$\Omega = \{(0, 0), (0, 1), (1, 0), (1, 1)\} .$$

- Zahlenlotto

$$\Omega = \{(1, 2, 3, 4, 5, 6), (1, 2, 3, 4, 5, 7), \ldots, (44, 45, 46, 47, 48, 49)\} .$$

- Anzahl Würfe eines Würfels, bis das erste Mal '6' erscheint

$$\Omega = \{1, 2, 3, 4, \ldots\} = \mathbb{N} \setminus \{0\} .$$

- Anzahl Schadenfälle, die einer KFZ-Versicherung in einem bestimmten Jahr gemeldet werden

$$\Omega = \{0, 1, 2, 3, \ldots\} = \mathbb{N} .$$

- Ziehen von 3 Kugeln aus einer Urne mit 5 schwarzen und 6 roten Kugeln

$$\Omega = \{(S,S,S),(S,S,R),(S,R,R),(R,R,R)\} .$$

Falls es auf die Reihenfolge darauf ankommt

$$\Omega = \{(S,S,S),(S,S,R),(S,R,S),(S,R,R),\\ (R,S,S),(R,S,R),(R,R,S),(R,R,R)\} .$$

Manchmal kann es sinnvoll sein, einen Raum Ω zu definieren, der mehr als die möglichen Ereignisse enthält. Zum Beispiel, falls man die Anzahl Sterbefälle durch Hufschlag in den Ställen der preussischen Armee in einem bestimmten Jahr beschreiben will, nimmt man $\Omega = \mathbb{N}$, obwohl es nur eine endliche Anzahl von Einwohnern Preussens gab. Das Problem ist, dass man (früher) nur sehr mühsam die Wahrscheinlichkeiten exakt berechnen konnte.

Definition 1.1. *Ein* **Ereignis** *ist eine Teilmenge $A \subset \Omega$. Also, $\mathcal{F} = \{A : A \subset \Omega\}$ ist die Klasse der Ereignisse. Wir sagen A* **tritt ein**, *falls ein Elementarereignis aus A eintritt.*

1.1.2. Die Wahrscheinlichkeitsverteilung

Wir gewichten nun die Elemente $\{\omega_1, \omega_2, \omega_3, \ldots\} = \Omega$.

Definition 1.2. *Eine Funktion $p : \Omega \to [0,1]$ heisst* **Wahrscheinlichkeit**, *falls*

$$\sum_{\omega \in \Omega} p(\omega) = 1 .$$

Die Abbildung

$$\mathbb{P} : \mathcal{F} \to [0,1] , \quad A \mapsto \sum_{\omega \in A} p(\omega)$$

heisst **Wahrscheinlichkeitsverteilung**. *Den Raum $(\Omega, \mathcal{F}, \mathbb{P})$ nennen wir* **Wahrscheinlichkeitsraum**.

Die Intuition hinter einer Wahrscheinlichkeit ist die Folgende. Könnte man ein Zufallsexperiment n mal unabhängig voneinander durchführen, wobei n sehr sehr gross ist, dann würde das Elementarereignis ω ungefähr $p(\omega)n$ mal vorkommen. $p(\omega)$ ist also der Anteil an den n Experimenten, bei denen ω eintritt. Das Ereignis Ω kommt dann genau n mal vor, das heisst, $\mathbb{P}[\Omega]$ muss gleich 1 sein.

Beispiele

- Würfeln mit einem Würfel
$$p(\omega) = \tfrac{1}{6} .$$
- Wurf von zwei nicht unterscheidbaren Münzen
$$p((K,K)) = p((Z,Z)) = \tfrac{1}{4} , \quad p((K,Z)) = \tfrac{1}{2} .$$
- Wurf von zwei unterscheidbaren Münzen
$$p(\omega) = \tfrac{1}{4}; .$$
- Zahlenlotto
$$p(\omega) = \frac{1}{13\,983\,816} .$$
- Anzahl Würfe eines Würfels, bis das erste Mal '6' erscheint
$$p(n) = \frac{5^{n-1}}{6^n} .$$
- Ziehen von 3 Kugeln aus einer Urne mit 5 schwarzen und 6 roten Kugeln
$$p((S,S,S)) = \frac{2}{33} , \quad p((S,S,R)) = \frac{12}{33} ,$$
$$p((S,R,R)) = \frac{15}{33} , \quad p((R,R,R)) = \frac{4}{33} .$$

Hilfssatz 1.3. *Es gelten die folgenden Regeln:*

i) *Seien* $\{A_i\}$ *Mengen, so dass* $A_i \cap A_j = \emptyset$ *für* $i \neq j$. *Dann gilt*
$$\mathbb{P}[\cup A_i] = \sum_i \mathbb{P}[A_i] .$$

ii)
$$\mathbb{P}[A^c] = 1 - \mathbb{P}[A] .$$

iii) *Falls* $A \subset B$, *dann gilt*
$$\mathbb{P}[B \setminus A] = \mathbb{P}[B] - \mathbb{P}[A] .$$
Insbesondere gilt, $\mathbb{P}[A] \leq \mathbb{P}[B]$.

iv)
$$\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B] .$$
Insbesondere ist $\mathbb{P}[A \cup B] \leq \mathbb{P}[A] + \mathbb{P}[B]$.

v) $$\mathbb{P}[\cup_{i=1}^n A_i] = \sum_{k=1}^n (-1)^{k-1} \sum_{1\le i_1<i_2<\cdots<i_k\le n} \mathbb{P}[\cap_{j=1}^k A_{i_j}] \ .$$

vi) *Seien $A_1 \subset A_2 \subset \cdots$ eine unendliche Anzahl Mengen. Dann gilt*

$$\mathbb{P}[\cup_{i=1}^\infty A_i] = \lim_{i\to\infty} \mathbb{P}[A_i] \ .$$

vii) *Seien $A_1 \supset A_2 \supset \cdots$ eine unendliche Anzahl Mengen. Dann gilt*

$$\mathbb{P}[\cap_{i=1}^\infty A_i] = \lim_{i\to\infty} \mathbb{P}[A_i] \ .$$

Beweis.

i) Wir haben

$$\mathbb{P}[\cup A_i] = \sum_{\omega\in\cup A_i} p(\omega) = \sum_i \sum_{\omega\in A_i} p(\omega) = \sum_i \mathbb{P}[A_i] \ .$$

ii) Wir haben $A \cap A^c = \emptyset$ und $A \cup A^c = \Omega$. Also gilt

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

iii) Falls $A \subset B$ gilt $B = A \cup (B \setminus A)$ und $A \cap (B \setminus A) = \emptyset$. Also haben wir

$$\mathbb{P}[B] = \mathbb{P}[A] + \mathbb{P}[B \setminus A] \ .$$

iv) Die Aussage folgt aus

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

v) Die Aussage folgt aus

$$\cup_{i=1}^{n+1} A_i = (\cup_{i=1}^n A_i) \cup A_{n+1}$$

durch vollständige Induktion.

vi) Seien $A_0 = \emptyset$ und $B_n = A_n \setminus A_{n-1}$. Wir haben $\cup_{i=1}^\infty A_i = \cup_{i=1}^\infty B_i$, und $B_i \cap B_j = \emptyset$ für $i \neq j$. Also gilt

$$\mathbb{P}[\cup_{i=1}^\infty B_i] = \sum_{i=1}^\infty \mathbb{P}[B_i] = \lim_{n\to\infty} \sum_{i=1}^n \mathbb{P}[B_i] = \lim_{n\to\infty} \mathbb{P}[A_n] \ .$$

vii) Die Aussage folgt aus

$$\cap_{i=1}^\infty A_i = \Omega \setminus (\cup_{i=1}^\infty A_i^c) \ .$$

□

Beispiel

Lotterie, de Moivre (1718) Ein klassisches Problem ist das Folgende: In einer Lotterie ist die Chance auf einen Preis 1:39, dh. $p = 1/40$. Wie viele Lose muss man kaufen, um mindestens die gleiche Chance auf einen Preis wie auf lauter Nieten zu haben? Kauft man n Lose, so ist $\Omega = \{0,1\}^n$ und wir wählen die Verteilung

$$p(\omega) = \prod_{i=1}^{n} p^{\omega_i}(1-p)^{1-\omega_i} \ .$$

Diese Verteilung ist eine gute Approximation zur Binomialverteilung, wenn die Anzahl der Lose gross ist. Die Wahrscheinlichkeit, keinen Preis zu gewinnen ist $(1-p)^n$, und wir suchen nun n, so dass $(1-p)^n \leq \frac{1}{2}$. Die Lösung ist somit $n \geq \log\frac{1}{2}/\log(1-p) = -(\log 2/\log(1-p))$. In unserem Problem erhalten wir $n \geq -(\log 2/\log(39/40)) = 27.3779$, also $n = 28$.

de Moivre (1718) Oft kann man auch mit einer alternativen Betrachtung zur Lösung kommen. In einer Lotterie gibt es 40 000 Lose. Jemand kauft sich 8 000 Lose, wobei sich die Person für drei der Preise interessiert. Was ist die Wahrscheinlichkeit, dass mindestens einer der drei gewünschten Preise gewonnen wird? Betrachten wir die Wahrscheinlichkeit, keinen der Preise zu gewinnen. Wir haben nun die 8 000 gekauften Lose, und die 32 000 übrigen Lose. Wir ziehen nun die drei Lose, die einen der drei Preise gewinnen. Dann müssen die Preise in die 32 000 übrigen Lose fallen. Somit erhalten wir die Wahrscheinlichkeit

$$\frac{32\,000}{40\,000} \cdot \frac{31\,999}{39\,999} \cdot \frac{31\,998}{39\,998} = 0.511990 \ .$$

Die gesuchte Wahrscheinlichkeit wird damit

$$1 - 0.511990 = 0.488010 \ .$$

1.2. Laplace-Modelle

Sei nun Ω endlich und $p(\omega)$ konstant. Man nennt diese Verteilung auch **Gleichverteilung** auf Ω. Bezeichnen wir mit $|A|$ die Mächtigkeit der Menge A. Also erhalten wir

$$p(\omega) = \frac{1}{|\Omega|} \ , \qquad \mathbb{P}[A] = \frac{|A|}{|\Omega|} = \frac{\#\text{ günstige Fälle}}{\#\text{ mögliche Fälle}} \ .$$

Warnung: Der Laplace-Ansatz ist nicht immer sinnvoll. Bekannt ist der Fehler von d'Alembert. Der Wurf von zwei nicht unterscheidbaren Münzen gibt 3 mögliche

Fälle. Aber $p(\omega) = \frac{1}{3}$ gibt falsche Wahrscheinlichkeiten. Sind die Münzen unterscheidbar, erhältt man das korrekte Resultat.

Sei nun S eine Menge, und $\Omega = S^n$ für ein $n \in \mathbb{N}$. Haben wir $|S| = N$, dann ist $|\Omega| = N^n$. Wir nehmen $n \leq N$ an. Sei $\Omega_0 = \{\omega \in \Omega : s_i \neq s_j, i \neq j\}$ und $\Omega_1 = \{\{s_1, s_2, \ldots, s_n\} : (s_1, \ldots, s_n) \in \Omega_0\}$. Wir können diese Elementarereignisse als Urnenmodell interpretieren. S ist eine Menge von Kugeln in einer Urne. Wir ziehen n Kugeln. In Ω legen wir die Kugeln nach dem Ziehen wieder zurück, und können daher eine Kugel mehrmals ziehen. In Ω_0 legen wir die Kugeln nicht zurück, merken uns aber die Reihenfolge, in der wir die Kugeln gezogen haben. In Ω_1 legen wir die Kugeln nicht zurück, und interessieren uns auch nicht für die Reihenfolge. Wir erhalten somit

$$|\Omega_0| = \frac{N!}{(N-n)!}\,, \qquad |\Omega_1| = \binom{N}{n} = \frac{N!}{n!(N-n)!}\,.$$

Mehr zur Kombinatorik steht im Anhang B.

Beispiele

Geburtstagsproblem Es seien n Personen in einem Raum. Wie gross ist die Wahrscheinlichkeit, dass mindestens zwei Personen am selben Tag Geburtstag haben? Wir wählen $S = \{1, 2, \ldots, 365\}$ und die Gleichverteilung auf Ω. Dann suchen wir $A = \Omega_0^c$. Wir haben

$$|\Omega_0| = \frac{365!}{(365-n)!}\,, \qquad |\Omega| = 365^n\,.$$

wobei wir natürlich $n \leq 365$ annehmen. Also erhalten wir

$$\mathbb{P}[A] = 1 - \mathbb{P}[\Omega_0] = 1 - \frac{365!}{365^n(365-n)!}\,.$$

Ein paar Werte stehen in der untenstehenden Tabelle.

n	10	20	23	40	100	150
$\mathbb{P}[A]$	0.11695	0.41144	0.5073	0.89123	1	1
$\mathbb{P}[A]/\mathbb{P}[A^c]$	0.13244	0.69906	1.0297	8.1939	$3.2547\,10^6$	$4.08\,10^{15}$

Garderobeproblem (Montmart, 1708) n Mäntel werden zufällig an die n Personen verteilt. Wie gross ist die Wahrscheinlichkeit, dass niemand seinen Mantel erhält. Wir wählen Ω die Menge aller möglichen Permutationen von $\{1, 2, \ldots, n\}$

und $\mathbb{P}$ die Gleichverteilung. Wir haben $|\Omega| = n!$. Bezeichnen wir mit $A_i = \{\omega \in \Omega : \omega(i) = i\}$ das Ereignis, dass die i-te Person ihren Mantel erhält. Dann ist

$$\begin{aligned}\mathbb{P}[\text{mind. 1}] = \mathbb{P}[\cup_i A_i] &= \sum_{k=1}^{n}(-1)^{k+1} \sum_{1\le i_1<i_2<\cdots<i_k\le n} \mathbb{P}[A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}] \\ &= \sum_{k=1}^{n}(-1)^{k+1} \sum_{1\le i_1<i_2<\cdots<i_k\le n} \frac{(n-k)!}{n!} = \sum_{k=1}^{n}(-1)^{k+1}\binom{n}{k}\frac{(n-k)!}{n!} \\ &= \sum_{k=1}^{n}(-1)^{k+1}\frac{1}{k!}\,.\end{aligned}$$

Somit gilt

$$\mathbb{P}[\text{keine}] = 1 - \mathbb{P}[\cup_i A_i] = \sum_{k=0}^{n}(-1)^k\frac{1}{k!} \stackrel{n\to\infty}{\longrightarrow} \mathrm{e}^{-1} = 0.367894\,.$$

Meinungsumfrage Seien N Kugeln in einer Urne, $K \le N$ rote und $N - K$ schwarze Kugeln. Wir nehmen eine Stichprobe der Grösse $n \le N$. Sei Ω die Menge der Stichproben, $\mathbb{P}$ die Gleichverteilung und A_k das Ereignis, dass genau $k \le \min\{K, n\}$ rote Kugeln in der Stichprobe sind. Wir haben

$$|\Omega| = \binom{N}{n}\,, \qquad\qquad |A_k| = \binom{K}{k}\binom{N-K}{n-k}\,.$$

Das ergibt

$$\mathbb{P}[A_k] = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}\,.$$

Diese Verteilung heisst **hypergeometrische Verteilung**. Schreiben wir

$$\begin{aligned}\mathbb{P}[A_k] &= \frac{K!(N-K)!n!(N-n)!}{k!(K-k)!(n-k)!(N-K-(n-k))!N!} \\ &= \binom{n}{k}\frac{K(K-1)\cdots(K-k+1)}{N(N-1)\cdots(N-k+1)} \\ &\qquad \times \frac{(N-K)(N-K-1)\cdots(N-K-(n-k)+1)}{(N-k)(N-k-1)\cdots(N-n+1)}\,.\end{aligned}$$

Falls wir nun N gegen unendlich gehen lassen, so dass $K/N \to p$ konstant ist, dann konvergiert $(K-a)/(N-b)$ nach p und $(N-K-a)(N-b)$ nach $1-p$ für jedes feste a, b. Somit erhalten wir im Grenzwert

$$\mathbb{P}[A_k] \longrightarrow \binom{n}{k}p^k(1-p)^{n-k}\,.$$

Diese Verteilung heisst **Binomialverteilung** mit Parametern n und $p \in [0, 1]$, und dient als Approximation für die korrekte Verteilung, falls N gross ist.

1.3. Zufallsvariablen

1.3.1. Definition der Zufallsvariable

Sei E ein Raum, z.B. $E = \mathbb{R}$, $E = [-\infty, \infty]$, $E = \mathbb{R}^d$.

Definition 1.4. *Eine Funktion $X : \Omega \to E$, $\omega \mapsto X(\omega)$ heisst (E-wertige)* **Zufallsvariable**. *Wir schreiben kurz X, wenn wir $X(\omega)$ meinen.*

X ist somit ein zufälliger Wert.

Beispiele

- Sei $\Omega = \{1, 2, \ldots, 6\}^3$ die Werte, die man beim Würfeln mit drei Würfeln erhält. Interessiert man sich für die Augensumme, ist $X(\omega) = \omega_1 + \omega_2 + \omega_3$. Interessiert man sich für die Anzahl Sechsen, ist $Y = \mathbb{1}_{\omega_1=6} + \mathbb{1}_{\omega_2=6} + \mathbb{1}_{\omega_3=6}$ eine weitere Zufallsvariable. Mit $\mathbb{1}_A$ bezeichnen wir die Zufallsvariable, die den Wert 1 gibt, falls $\omega \in A$ und 0 sonst. Durch $Z = (X, Y)$ ist eine weitere Zufallsvariable definiert.

- Betrachtet man eine Lebensversicherung mit n versicherten Personen. Dann ist τ_i, die Anzahl Jahre, die Person i noch zu leben hat, eine Zufallsvariable. Hat jede Person eine reine Risiko-Lebensversicherung abgeschlossen (d.h. sie bekommt nur Geld, wenn sie bis zum Auslaufszeitpunkt T_i stirbt), dann ist $X = \sum_{i=1}^{n} \mathbb{1}_{\tau_i \leq T_i}$, die Anzahl Versicherungsfälle, auch eine Zufallsvariable.

- Beim Zahlenlotto ist der Gewinn X einer Person eine Zufallsvariable, die von den gezogenen (zufälligen) Zahlen abhängt. Eine weitere Zufallsvariable ist die Anzahl der Personen, die genau drei der gezogenen Zahlen auf ihrem Lottoschein haben.

Definition 1.5. *Wir schreiben kurz $\{X \in A\}$ für die Menge $\{\omega \in \Omega : X(\omega) \in A\}$. Die Gewichtung $\mu(x) = \mathbb{P}[X = x]$ nennen wir* **Verteilung** *der Zufallsvariable X.*

1.3.2. Der Erwartungswert

Nehmen wir nun an, $E \subset [-\infty, \infty]$. Eine wichtige Kennzahl der Zufallsvariable ist in der folgenden Definition gegeben.

Definition 1.6. *Der Wert*

$$\mathbb{E}[X] = \sum_{\Omega} X(\omega)p(\omega)$$

heisst **Erwartungswert** *der Zufallsvariable X, falls die Summe sinnvoll ist. Das heisst*

$$\sum_{\substack{\omega\in\Omega\\ X(\omega)>0}} X(\omega)p(\omega) < \infty \qquad \textit{oder} \qquad \sum_{\substack{\omega\in\Omega\\ X(\omega)<0}} X(\omega)p(\omega) > -\infty\ .$$

Wir verwenden die Konvention $\infty \cdot 0 = 0$.

Sei $X(\Omega) = \{x : \mathbb{P}[X = x] > 0\}$. Dann können wir

$$\mathbb{E}[X] = \sum_{x\in X(\Omega)} x\mu(x)$$

schreiben.

Der Erwartungswert kann folgendermassen motiviert werden. Nehmen wir an, dass X einen Gewinn in einem Spiel darstellt. Nehmen wir weiter an, wir spielen das Spiel n-mal, wobei n sehr gross ist. Dann haben wir ungefähr $np(\omega)$ mal den Gewinn $X(\omega)$. Zählen wir die Gewinne zusammen, erhalten wir ungefähr den Gewinn $n\mathbb{E}[X]$. Im Durchschnitt ist der Gewinn also $\mathbb{E}[X]$. Wir werden diese Interpretation später beweisen (Satz 3.4).

Hilfssatz 1.7. *Der Erwartungswert ist linear und positiv, das heisst für zwei Zufallsvariablen X und Y und Zahlen $\alpha, \beta \in \mathbb{R}$ gilt $\mathbb{E}[\alpha X + \beta Y] = \alpha\mathbb{E}[X] + \beta\mathbb{E}[Y]$. Falls $X \geq 0$, so gilt $\mathbb{E}[X] \geq 0$.* □

Beispiele

- Betrachten wir das Problem, wenn wir mit n Würfeln würfeln. Sei X die Anzahl geworfener Augen. Sei zuerst $n = 1$. Dann erhalten wir

  $$\mathbb{E}[X] = \tfrac{1}{6}(1 + 2 + \cdots + 6) = 3.5\ .$$

  Betrachten wir nun n Würfel, und sei X_i die Augenzahl des i-ten Würfels. Aus der Linearität erhalten wir

  $$\mathbb{E}[X] = \mathbb{E}\Big[\sum_{i=1}^{n} X_i\Big] = \sum_{i=1}^{n} \mathbb{E}[X_i] = 3.5n\ .$$

- Sei beim Garderobeproblem X die Anzahl der Personen, die Ihren eigenen Mantel erhalten. Wenn genau k Personen ihren eigenen Mantel erhalten, können wir diese Personen auf $\binom{n}{k}$ Arten wählen. Die Ausgewählten bekommen ihren Mantel, und die restlichen $n-k$ Personen müssen fremde Mäntel erhalten. Also haben wir

$$\mathbb{P}[X=k] = \frac{\binom{n}{k}(n-k)!\sum_{j=0}^{n-k}(-1)^j/j!}{n!} = \frac{1}{k!}\sum_{j=0}^{n-k}\frac{(-1)^j}{j!} .$$

  Also erhalten wir für den Erwartungswert

$$\mathbb{E}[X] = \sum_{k=0}^{n}\frac{k}{k!}\sum_{j=0}^{n-k}\frac{(-1)^j}{j!} = \sum_{k=1}^{n}\frac{1}{(k-1)!}\sum_{j=0}^{n-k}\frac{(-1)^j}{j!} .$$

  Alternativ könnte man $X_i = 1$ setzen, falls die i-te Person ihren Mantel erhält, und 0 sonst. Dann ist

$$\mathbb{E}[X_i] = \mathbb{P}[X_1 = 1] = \frac{(n-1)!}{n!} = \frac{1}{n} ,$$

  da es $n-1$ Möglichkeiten gibt, die anderen $n-1$ Mäntel zu verteilen. Also erhalten wir

$$\mathbb{E}[X] = \mathbb{E}\Big[\sum_{i=1}^{n} X_i\Big] = \sum_{i=1}^{n}\frac{1}{n} = 1 .$$

- **Petersburger Paradox** (D. Bernoulli) Wir besuchen das Kasino und bezahlen c Euro Eintritt. Wir werfen eine faire Münze so lange, bis das erste Mal Zahl erscheint. Sei T die Anzahl Würfe. Als Gewinn erhalten wir 2^T Euro ausbezahlt. Um k Mal zu werfen, muss $k-1$ Mal Kopf erscheinen, und das letzte Mal Zahl. Betrachten wir die Gleichverteilung auf k Würfen, haben wir 2^k Möglichkeiten, aber nur 1 günstige. Das heisst, $\mathbb{P}[T=k] = 2^{-k}$. Wir erhalten also für den Erwartungwert des Gewinnes $X = 2^T$,

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} 2^k \mathbb{P}[T=k] = \sum_{k=1}^{\infty} 2^k 2^{-k} = \sum_{k=1}^{\infty} 1 = \infty .$$

  Fragt man aber die Leute, wieviel sie maximal Eintritt für dieses Spiel bezahlen würden, nennen alle relativ kleine Zahlen. Daniel Bernoulli schliesst daraus, dass die Leute nicht nach dem erwarteten Gewinn schauen, sondern nach dem erwarteten Nutzen. Das heisst, man hat eine konkave wachsende Nutzenfunktion $u(x)$, und spielt das Spiel, falls

$$u(c) \leq \mathbb{E}[u(X)] = \sum_{k=1}^{\infty} u(2^k)\mathbb{P}[T=k] .$$

Dieses Prinzip wird heute noch in der Ökonomie verwendet, um das Verhalten von Händlern im Markt zu studieren. Das Prinzip gilt aber sicher nicht, falls man Glücksspiele betrachtet, da sonst niemand mitspielen würde. Ist zum Beispiel $u(x) = \sqrt{x}$, so erhalten wir

$$\begin{aligned}\mathbb{E}[u(X)] &= \sum_{k=1}^{\infty} \sqrt{2^k}\mathbb{P}[T=k] = \sum_{k=1}^{\infty} 2^{k/2}2^{-k} = \sum_{k=1}^{\infty} 2^{-k/2} \\ &= \frac{1}{\sqrt{2}-1} = \sqrt{2}+1 = 2.41421\ .\end{aligned}$$

Der maximale Betrag, den die Leute also zahlen würden, ist $(\sqrt{2}+1)^2 = 5.8284$.

- Betrachten wir die Meinungsumfrage. Sei X die Anzahl der roten Kugeln, die wir in n Versuchen ziehen. Nehmen wir an, dass $n \leq \min\{K, N-K\}$. Somit ist

$$\begin{aligned}\mathbb{E}[X] &= \sum_{k=0}^{n} k\mathbb{P}[A_k] = \sum_{k=1}^{n} k\frac{K!(N-K)!n!(N-n)!}{k!(K-k)!(n-k)!(N-K-(n-k))!N!} \\ &= \sum_{k=1}^{n} \frac{K!(N-K)!n!(N-n)!}{(k-1)!(K-k)!(n-k)!(N-K-(n-k))!N!} \\ &= \sum_{k=0}^{n-1} \frac{K!(N-K)!n!(N-n)!}{k!(K-1-k)!(n-1-k)!(N-K-(n-1-k))!N!} \\ &= \sum_{k=0}^{n-1} \frac{K!(N-1-(K-1))!n!(N-1-(n-1))!}{k!(K-1-k)!(n-1-k)!(N-1-(K-1)-(n-1-k))!N!} \\ &= \frac{nK}{N}\sum_{k=0}^{n^*} \frac{K^*!(N^*-K^*)!n^*!(N^*-n^*)!}{k!(K^*-k)!(n^*-k)!(N^*-K^*-(n^*-k))!N^*!}\ ,\end{aligned}$$

  wobei $N^* = N-1$, $K^* = K-1$ und $n^* = n-1$. Die Terme unter der Summe sind die Wahrscheinlichkeiten von A_k^*, wobei wir N, K, n durch N^*, K^*, n^* ersetzen. Somit ist die Summe 1, und $\mathbb{E}[X] = nK/N$.

  Einfacher können wir das Problem lösen, wenn wir $X_i = 1$ setzen, falls die i-te Kugel rot ist, und 0 sonst. Da wir N Kugeln ziehen können, und K davon rot sind, haben wir $\mathbb{E}[X_i] = K/N$. Also erhalten wir

$$\mathbb{E}[X] = \mathbb{E}\Big[\sum_{i=1}^{n} X_i\Big] = \sum_{i=1}^{n} \mathbb{E}[X_i] = n\frac{K}{N}\ .$$

- **Pascals Wette** Es gibt zwei Möglichkeiten, Gott existiert oder Gott existiert nicht. Falls Gott existiert, ist der Gewinn eines gläubigen Menschen unendlich,

der eines ungläubigen Menschen ist $-a$ für ein $a > 0$, das heisst, ein Ungläubiger muss Busse zahlen. Falls Gott nicht existiert, ist der Gewinn eines Gläubigen $-b$, da er auf die Freuden des Lebens verzichtet, der Gewinn eines Ungläubigen ist 0. Also ist der Erwartungswert eines Gläubigen unendlich, der eines Ungläubigen endlich.

- **de Moivre [4, Problem XII]** n Spieler spielen folgendes Spiel. Auf dem Tisch befindet sich der Betrag S. Der erste Spieler wirft zwei unterscheidbare Münzen. Wirft er zweimal Kopf, so erhält er den ganzen Betrag auf dem Tisch, und das Spiel ist beendet. Wirft er zweimal Zahl, so muss er den Betrag S zusätzlich auf den Tisch legen, das heisst, den Betrag verdoppeln. Zeigt die erste Münze Kopf und die zweite Zahl, so erhält der erste Spieler $\frac{1}{2}S$, das heisst, den halben Betrag auf dem Tisch. Zeigt die erste Münze Zahl und die zweite Kopf, so erhält der Spieler nichts, und muss auch nichts bezahlen. Ist das Spiel nicht beendet, so kommt der nächste Spieler an die Reihe, und das Spiel wird nach den selben Regeln fortgesetzt. Nachdem alle Spieler einmal gespielt haben, ist der erste Spieler wieder an der Reihe, sofern das Spiel noch nicht beendet wurde.

  De Moivre macht folgende Überlegungen, um den Erwartungswert des Gewinnes des ersten Spielers zu finden. Nehmen wir zuerst an, dass unendlich viele Spieler mitspielen. Dann kommt der erste Spieler nur einmal an die Reihe. Der erwartete Gewinn des ersten Spielers wird dann

  $$\tfrac{1}{4}S - \tfrac{1}{4}S + \tfrac{1}{4}S/2 + \tfrac{1}{4}0 = S/8 \; .$$

  Der erste Spieler könnte nun, anstatt zu spielen, den Betrag $S/8$ vom Tisch nehmen, und das Spiel den weiteren Spielern überlassen. Dann würde sich auf dem Tisch noch der Betrag $\frac{7}{8}S$ befinden. Der zweite Spieler könnte sich auch dafür entscheiden, seinen Anteil, also $\frac{1}{8} \cdot \frac{7}{8}S$ vom Tisch zu nehmen, und es wäre noch $(\frac{7}{8})^2 S$ auf dem Tisch. Der k-te Spieler würde dann $\frac{1}{8}(\frac{7}{8})^{k-1}S$ erhalten, wenn er statt dem Spiel das Geld wählt. Spielen jetzt n Spieler mit, so kann sich der erste Spieler jedes n-te Mal den Betrag nehmen, also

  $$\sum_{k=0}^{\infty} \tfrac{1}{8}(\tfrac{7}{8})^{kn} S = \tfrac{1}{8} \frac{S}{1 - (\frac{7}{8})^n} = \frac{8^{n-1} S}{8^n - 7^n} \; .$$

  Spielen also 2 Spieler, so ist dies $\frac{8}{15}S$, und der zweite Spieler kann $\frac{7}{15}S$ erwarten. Spielen drei Spieler, so kann der erste Spieler $\frac{64}{169}S$, der zweite $\frac{56}{169}S$, und der dritte $\frac{49}{169}S$ erwarten.

Hilfssatz 1.8. *Sei $\mathbb{P}[X \in \mathbb{N}] = 1$, das heisst X nimmt nur Werte aus $\mathbb{N}$ an. Dann gilt*

$$\mathbb{E}[X] = \sum_{k=0}^{\infty} \mathbb{P}[X > k] \, .$$

Beweis. Wir haben

$$\begin{aligned}\mathbb{E}[X] &= \sum_{k=1}^{\infty} k\mathbb{P}[X = k] = \sum_{k=1}^{\infty}\sum_{\ell=0}^{k-1} \mathbb{P}[X = k] = \sum_{\ell=0}^{\infty}\sum_{k=\ell+1}^{\infty} \mathbb{P}[X = k] \\ &= \sum_{\ell=0}^{\infty} \mathbb{P}[X \geq \ell + 1] = \sum_{\ell=0}^{\infty} \mathbb{P}[X > \ell] \, .\end{aligned}$$

□

Sei T eine Wartezeit mit der **geometrischen Verteilung** $\mathbb{P}[T > k] = q^k$. Dann ist der Erwartungswert

$$\mathbb{E}[T] = \sum_{k=0}^{\infty} q^k = \frac{1}{1-q} \, .$$

1.4. Irrfahrten

1.4.1. Definition

Betrachten wir N Perioden, und setzen $\Omega = \{-1, 1\}^N$. Sei X_i das Resultat der i-ten Periode. Eine **Irrfahrt** ist der Prozess $S_0 = 0$ und $S_n = \sum_{i=1}^{n} X_i$, der Ort der Irrfahrt nach der n-ten Periode. Zwei mögliche Pfade einer Irrfahrt sind in Abbildung 1.1 gegeben.

Wir verwenden die Gleichverteilung auf Ω. Wir wollen nun die Verteilung von S_n bestimmen. Wir bemerken zuerst, dass zu geraden Zeitpunkten nur gerade Werte angenommen werden können, zu ungeraden Zeitpunkten nur ungerade Werte. Um zum Zeitpunkt n im Punkt k sein zu können, muss $\frac{1}{2}(n+k)$ mal die Eins und $\frac{1}{2}(n-k)$ mal die -1 aufgetreten sein. Die letzten $N - n$ Werte sind ohne Bedeutung. Wir haben $|\Omega| = 2^N$. Es gibt $\binom{n}{(n+k)/2}$ Möglichkeiten, die Stellen zu wählen, an denen 1 auftritt. Wir haben 2^{N-n} Möglichkeiten, die letzten $N-n$ Stellen zu besetzen. Also erhalten wir

$$\mathbb{P}[S_n = k] = \frac{\binom{n}{(n+k)/2} 2^{N-n}}{2^N} = \binom{n}{(n+k)/2} 2^{-n} \, ,$$

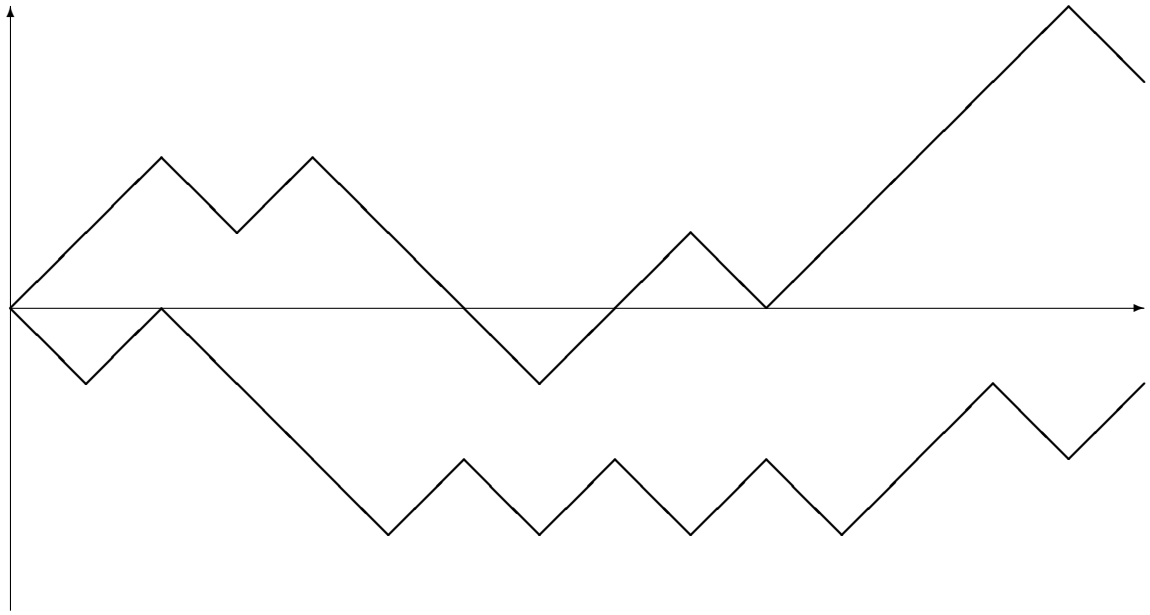

Abbildung 1.1: *Zwei Pfade einer Irrfahrt*

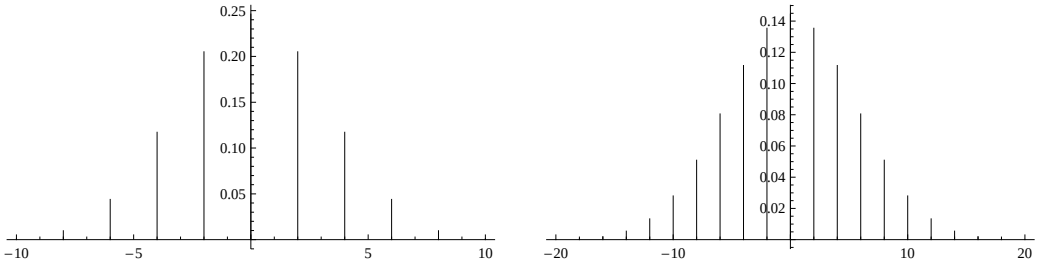

Abbildung 1.2: *Verteilung von S_{10} und S_{30}*

falls $n + k \in \{0, 2, \ldots, 2n\}$, und 0 sonst. Wir bemerken, dass $\mathbb{P}[S_n = k] = \mathbb{P}[S_n = -k]$. Insbesondere erhalten wir

$$\mathbb{E}[S_n] = \sum_{k=-n}^{n} k\mathbb{P}[S_n = k] = 0 \; .$$

Dies kann man auch aus

$$\mathbb{P}[X_i = 1] = \frac{2^{N-1}}{2^N} = \tfrac{1}{2} = \mathbb{P}[X_i = -1] \; ,$$

und daher $\mathbb{E}[X_i] = \frac{1}{2} \cdot 1 + \frac{1}{2}(-1) = 0$ schliessen.

Sei nun n fest. Dann schliessen wir aus der *Stirling-Formel*

$$n! = n^{n+1/2}\mathrm{e}^{-n+\varepsilon_n}\sqrt{2\pi} \; , \qquad \frac{1}{12n+1} < \varepsilon_n < \frac{1}{12n} \; ,$$

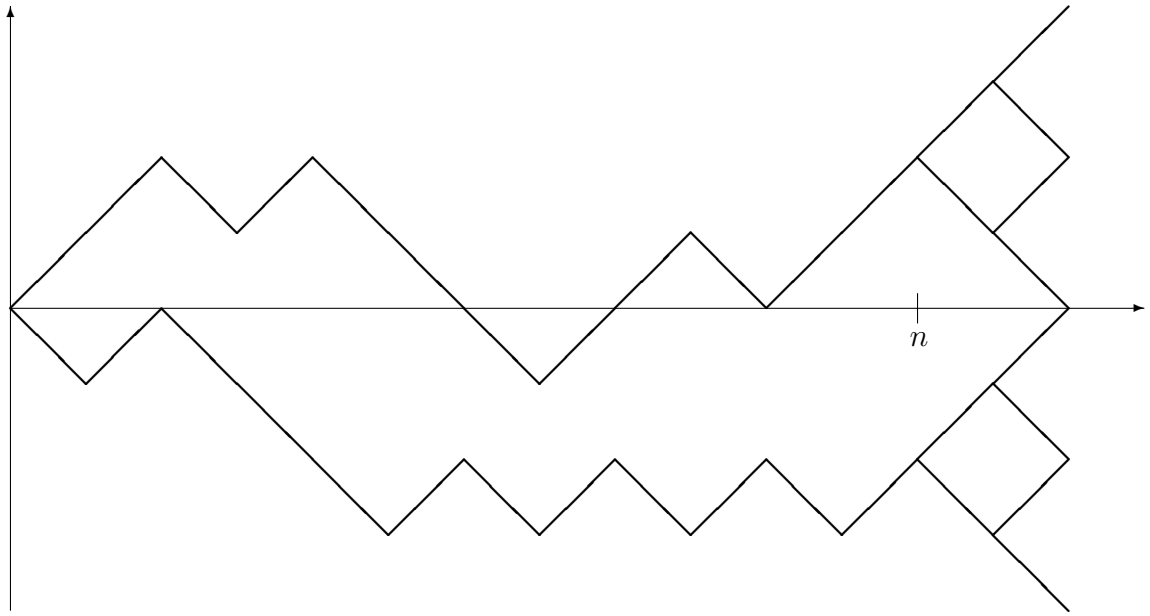

Abbildung 1.3: *Beobachtbares Ereignis bis zum Zeitpunkt n*

dass

$$\begin{aligned}\mathbb{P}[S_n = k] &\approx \frac{n^{n+1/2}\mathrm{e}^{-n}\sqrt{2\pi}}{((n+k)/2)^{(n+k)/2+1/2}((n-k)/2)^{(n-k)/2+1/2}\mathrm{e}^{-n}2\pi}2^{-n} \\ &= \Big(\frac{n^2}{(n+k)(n-k)}\Big)^{n/2+1/2}\Big(\frac{n-k}{n+k}\Big)^{k/2}\sqrt{\frac{2}{\pi n}}\,.\end{aligned}$$

Wir sehen, dass letzterer Ausdruck für jedes feste k gegen Null konvergiert. Insbesondere konvergiert $\mathbb{P}[S_n \in [a,b]]$ für jedes feste $a < b$ gegen Null. Somit breitet sich für wachsendes n die Verteilung immer mehr aus.

1.4.2. Spielsysteme

Wir können die Irrfahrt $\{S_n\}$ als Bilanzentwicklung in einem Glücksspiel betrachten, bei dem man immer 1 auf $+1$ setzt. Wir wollen nun erlauben, dass der Spieler zu jedem Zeitpunkt n einen Betrag auf 1 oder -1 setzen kann, den er erst zum Zeitpunkt n bestimmt. Falls der Betrag vom bisherigen Verlauf des Spiels abhängt, wird der gesetzte Betrag zufällig. Der Spieler kann aber nicht zukünftige Ereignisse voraussehen. Wir definieren daher

Definition 1.9. *Ein Ereignis $A \subset \Omega$ heisst* **beobachtbar** *zum Zeitpunkt n, wenn $A = \{\omega : (X_1(\omega), X_2(\omega), \ldots, X_n(\omega)) \in A_n\}$, für ein $A_n \subset \{-1,1\}^n$. Wir definieren $\mathcal{F}_n$ als die Menge aller bis zum Zeitpunkt n beobachtbaren Ereignisse.*

Mit einem Pfad bis zur Zeit n müssen alle möglichen Pfade ab dem Zeitpunkt n in A sein, siehe auch Abbildung 1.3. Es gilt $\{\emptyset, \Omega\} = \mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_N = \mathcal{F}$. Es ist einfach zu sehen, dass wenn $A_1 \in \mathcal{F}_n$ und $A_2 \in \mathcal{F}_n$, dann sind auch $A_1 \cup A_2 \in \mathcal{F}_n$ (mindestens eines der Ereignisse tritt ein), $A_1 \cap A_2 \in \mathcal{F}_n$ (beide Ereignisse treten ein) und $A_1^c \in \mathcal{F}_n$ (das Ereignis tritt nicht ein).

Hilfssatz 1.10. *Sei $A \in \mathcal{F}_n$ und $n < N$. Dann ist*

$$\mathbb{P}[A \cap \{X_{n+1} = 1\}] = \mathbb{P}[A \cap \{X_{n+1} = -1\}] = \tfrac{1}{2}\mathbb{P}[A] \, .$$

Beweis. Wir haben $|A|/2^{N-n}$ Möglichkeiten (für jeden Pfad von 0 bis n gibt es 2^{N-n} Fortsetzungen von n nach N), die ersten n Koordinaten von ω zu wählen, und 2^{N-n-1} Möglichkeiten die letzten $N-n-1$ Koordinaten zu wählen. Dies gibt $|A \cap \{X_{n+1} = 1\}| = 2^{N-n-1}|A|/2^{N-n} = |A|/2$. Somit erhalten wir

$$\mathbb{P}[A \cap \{X_{n+1} = 1\}] = \frac{|A|/2}{2^N} = \tfrac{1}{2}\frac{|A|}{2^N} = \tfrac{1}{2}\mathbb{P}[A] \, .$$

□

Wir lassen einen Spieler nun einen Betrag V_n im Zeitpunkt $n-1$ auf 1 setzen. Falls der Spieler auf -1 setzen will, wählt er einen negativen Betrag.

Definition 1.11. *Ein Spielsystem ist eine Familie $\{V_n : n = 1, 2, \ldots N\}$ von reellen Zufallsvariablen, so dass $\{V_n = c\} \in \mathcal{F}_{n-1}$.*

Der Ertrag in Periode n wird dann $V_n X_n$, die Gesamtbilanz wird somit $S_n^V = \sum_{i=1}^n V_i X_i$.

Satz 1.12. (Unmöglichkeit gewinnträchtiger Spielsysteme) *Für jedes Spielsystem $\{V_n\}$ gilt $\mathbb{E}[S_n^V] = 0$.*

Beweis. Es genügt zu zeigen, dass $\mathbb{E}[V_n X_n] = 0$. Seien c_i die möglichen Werte von V_n und $A_i = \{V_n = c_i\} \in \mathcal{F}_{n-1}$. Dann haben wir $\mathbb{P}[A_i \cap \{X_n = 1\}] = \mathbb{P}[A_i \cap \{X_n = -1\}] = \frac{1}{2}\mathbb{P}[A_i]$. Also

$$\mathbb{E}[V_n X_n] = \sum_i (c_i \mathbb{P}[A_i \cap \{X_n = 1\}] - c_i \mathbb{P}[A_i \cap \{X_n = -1\}]) = 0 \, .$$

□

Definition 1.13. *Eine* **Stoppzeit** *ist eine Abbildung* $T : \Omega \to \{0, 1, \ldots, N\}$ *mit der Eigenschaft, dass* $\{T = n\} \in \mathcal{F}_n$.

Wir stoppen zur Zeit n also nur auf Grund der Information, die wir bis zum Zeitpunkt n haben. Die Bilanz bis zur Zeit T ist $S_T^V = \sum_{i=1}^T V_i X_i$.

Korollar 1.14. (Stoppsatz) *Für jede Stoppzeit* T *und jedes Spielsystem* $\{V_n\}$ *gilt* $\mathbb{E}[S_T^V] = 0$.

Beweis. Wir haben, dass $\tilde{V}_n = V_n \mathbb{1}_{T \geq n}$ ein Spielsystem ist, da für $c \neq 0$

$$\{\tilde{V}_n = c\} = \{V_n = c\} \cap (\cap_{t=0}^{n-1} \{T = t\}^c) \in \mathcal{F}_{n-1} \ ,$$

und

$$\{\tilde{V}_n = 0\} = \{V_n = 0\} \cup (\cup_{t=0}^{n-1} \{T = t\}) \in \mathcal{F}_{n-1} \ .$$

Also gilt $\mathbb{E}[S_T^V] = \mathbb{E}[S_N^{\tilde{V}}] = 0$. □

Insbesondere gibt folgendes Spielsystem keinen Gewinn. Man setzt zuerst 1 auf 1. Falls man gewinnt, stoppt man. Ansonsten verdoppelt man den Einsatz. Auf diese Weise macht man schliesslich einen Gewinn von 1. Diese Verdoppelungsstrategie funktioniert aber nicht. Um sicher zu gewinnen, muss man unendlich lange spielen können, und unendlich Geld zur Verfügung haben. Spielt man nur eine begrenzte Zeit, gewinnt man mit "hoher" Wahrscheinlichkeit 1, mit einer kleinen Wahrscheinlichkeit verliert man alles. Um sicher n-Mal spielen zu können, braucht man das Kapital $2^n - 1$.

Korollar 1.15. *Für jede Stoppzeit* T *gilt* $\mathbb{E}[S_T^2] = \mathbb{E}[T]$.

Beweis. Wählen wir $V_n = 2S_{n-1}$. Dann ist

$$V_n X_n = 2S_{n-1}X_n = (S_{n-1} + X_n)^2 - S_{n-1}^2 - X_n^2 = S_n^2 - S_{n-1}^2 - 1 \ .$$

Insbesondere ist

$$S_T^V = \sum_{n=1}^T V_n X_n = S_T^2 - T \ .$$

Die Behauptung folgt nun aus dem Stoppsatz. □

Sei $c \in \mathbb{Z}$ und $T_c = \inf\{n \geq 1 : S_n = c\}$. Wir setzen $T = T_c \wedge N = \min\{T_c, N\}$. Wir haben $T \geq 1$, und für $k < N$,

$$\{T = k\} = \{S_1 \neq c, S_2 \neq c, \ldots, S_{k-1} \neq c, S_k = c\} \in \mathcal{F}_k \; .$$

Weiter gilt

$$\{T = N\} = \{S_1 \neq c, S_2 \neq c, \ldots, S_{N-1} \neq c\} \in \mathcal{F}_{N-1} \subset \mathcal{F}_N \; .$$

Also ist T eine Stoppzeit. Wir finden also

$$0 = \mathbb{E}[S_T] = \mathbb{E}[S_T \mathbb{1}_{T_c<N}] + \mathbb{E}[S_N \mathbb{1}_{T=N}] = c\mathbb{P}[T_c < N] + \mathbb{E}[S_N \mathbb{1}_{T=N}] \; .$$

1.4.3. Das Ruinproblem

Sei $a < 0 < b$ und $N > \max\{|a|, b\}$. Ein Spieler A mit Anfangskapital $|a|$ spielt gegen Spieler B mit Anfangskapital b. Sei $T = \min\{T_a, T_b, N\}$. Wie oben folgt, dass T eine Stoppzeit ist. Falls $\{T_a < T_b\}$, dann ist Spieler A ruiniert. Aus dem Stoppsatz schliessen wir

$$0 = \mathbb{E}[S_T] = a\mathbb{P}[T = T_a] + b\mathbb{P}[T = T_b] + \mathbb{E}[S_N \mathbb{1}_{T=N\neq\min\{T_a,T_b\}}] \; .$$

Wir wollen nun $N \to \infty$ gehen lassen. Zuerst sehen wir, dass

$$|\mathbb{E}[S_N \mathbb{1}_{T=N\neq\min\{T_a,T_b\}}]| \leq \max\{|a|, b\}\mathbb{P}[S_N \in [a, b]] \to 0 \; ,$$

was wir in Abschnitt 1.4.1 bewiesen haben. Es gilt, dass $\mathbb{P}[T = T_a] = \mathbb{P}[T_a < T_b, T_a \leq N]$ eine monotone Folge in N ist. Also konvergiert $\mathbb{P}[T = T_a]$ zu einem Wert r_a. Dann muss $\mathbb{P}[T_b = T]$ nach $1 - r_a$ konvergieren. Wir schliessen aus $0 = ar_a + b(1 - r_a)$, dass

$$r_a = \frac{b}{b-a} \; , \qquad\qquad 1 - r_a = \frac{|a|}{b-a} \; .$$

Weiter folgern wir aus

$$\mathbb{E}[T] = \mathbb{E}[S_T^2] = a^2\mathbb{P}[T = T_a] + b^2\mathbb{P}[T = T_b] + \mathbb{E}[S_N^2 \mathbb{1}_{\min\{T_a,T_b\}>N}] \; ,$$

und daraus, dass der letzte Term gegen Null konvergiert, dass

$$\lim_{N\to\infty} \mathbb{E}[T] = a^2\frac{b}{b-a} + b^2\frac{|a|}{b-a} = |a|b \; .$$

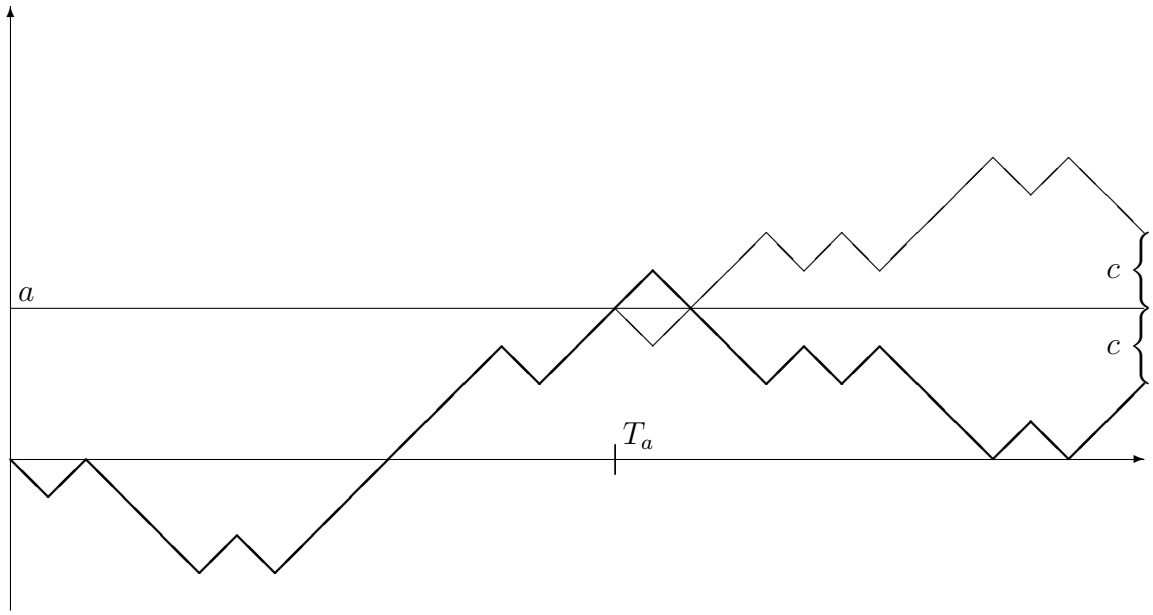

Abbildung 1.4: *Reflektionsprinzip*

1.4.4. Das Reflektionsprinzip

Hilfssatz 1.16. *Wir haben für $a > 0$ und $c \geq 0$,*

$$\mathbb{P}[T_a \leq N, S_N = a - c] = \mathbb{P}[S_N = a + c] \ .$$

Beweis. Die Anzahl Pfade von (T_a, a) nach $(N, a - c)$ ist gleich der Anzahl Pfade von (T_a, a) nach $(N, a + c)$, siehe Abbildung 1.4. Da auf $\{S_N = a + c\}$ das Ereignis $\{T_a \leq N\}$ sicher eintritt, ist das Resultat bewiesen. □

Daraus können wir jetzt die Verteilung von T_a bestimmen.

Satz 1.17. *Wir haben für $a > 0$ und $n < N$*

$$\mathbb{P}[T_a \leq n] = \mathbb{P}[S_n \notin [-a, a - 1]] \ ,$$

und daher

$$\mathbb{P}[T_a = n] = \tfrac{1}{2}(\mathbb{P}[S_{n-1} = a - 1] - \mathbb{P}[S_{n-1} = a + 1]) = \frac{a}{n}\mathbb{P}[S_n = a] \ .$$

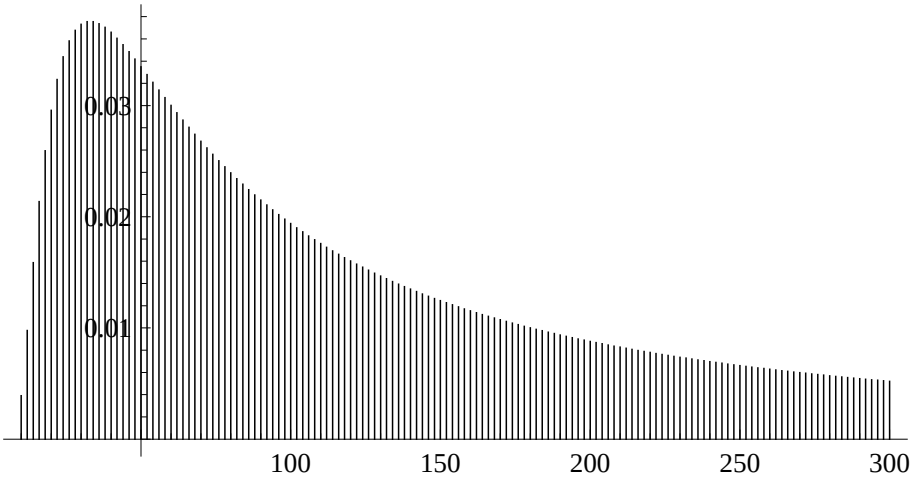

Abbildung 1.5: *Verteilung von* T_{10}.

Beweis. Für $b \geq a$ gilt $\mathbb{P}[S_n = b, T_a \leq n] = \mathbb{P}[S_n = b]$. Also folgt

$$\begin{aligned}\mathbb{P}[T_a \leq n] &= \sum_b \mathbb{P}[S_n = b, T_a \leq n] = \sum_{b \geq a} \mathbb{P}[S_n = b] + \sum_{c>0} \mathbb{P}[S_n = a - c, T_a \leq n] \\ &= \sum_{b \geq a} \mathbb{P}[S_n = b] + \sum_{c>0} \mathbb{P}[S_n = a + c] = \mathbb{P}[S_n \geq a] + \mathbb{P}[S_n > a] \\ &= \mathbb{P}[S_n \geq a] + \mathbb{P}[S_n < -a] \; .\end{aligned}$$

Dies beweist die erste Aussage.

Aus der ersten Aussage erhalten wir

$$\begin{aligned}\mathbb{P}[T_a = n] &= \mathbb{P}[T_a \leq n] - \mathbb{P}[T_a \leq n-1] \\ &= \mathbb{P}[S_n \notin [-a, a-1]] - \mathbb{P}[S_{n-1} \notin [-a, a-1]] \\ &= \mathbb{P}[S_{n-1} \notin [-a-1, a]] + \tfrac{1}{2}\mathbb{P}[S_{n-1} \in [-a-1, -a]] + \tfrac{1}{2}\mathbb{P}[S_{n-1} \in [a-1, a]] \\ &\qquad - (\mathbb{P}[S_{n-1} \notin [-a-1, a]] + \mathbb{P}[S_{n-1} = -a-1] + \mathbb{P}[S_{n-1} = a]) \\ &= \tfrac{1}{2}(\mathbb{P}[S_{n-1} = -a] + \mathbb{P}[S_{n-1} = a-1] - \mathbb{P}[S_{n-1} = -a-1] - \mathbb{P}[S_{n-1} = a]) \\ &= \tfrac{1}{2}(\mathbb{P}[S_{n-1} = a-1] - \mathbb{P}[S_{n-1} = a+1]) \; .\end{aligned}$$

Wir haben die Formel (falls $n + a$ gerade ist)

$$\begin{aligned}\mathbb{P}[S_{n-1} = a-1] &= \binom{n-1}{(n+a)/2-1} 2^{-n+1} = \frac{n+a}{n}\binom{n}{(n+a)/2} 2^{-n} \\ &= \frac{n+a}{n}\mathbb{P}[S_n = a] \; ,\end{aligned}$$

und

$$\mathbb{P}[S_{n-1} = a+1] = \binom{n-1}{(n+a)/2} 2^{-n+1} = \frac{n-a}{n}\binom{n}{(n+a)/2} 2^{-n} = \frac{n-a}{n}\mathbb{P}[S_n = a] \; .$$

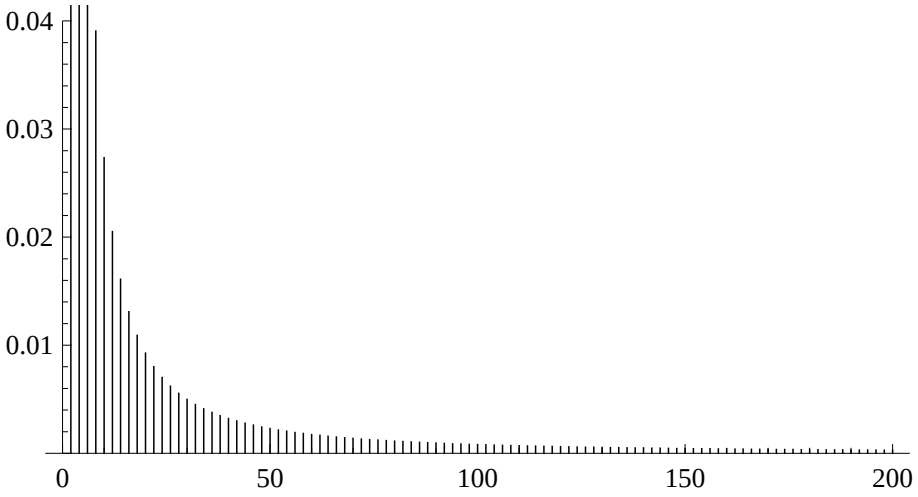

Abbildung 1.6: *Verteilung von* T_0.

Die Differenz ergibt die letzte Aussage. □

Insbesondere erhalten wir, da die Irrfahrt das Intervall $[-a, a-1]$ verlässt,

$$\lim_{n\to\infty} \mathbb{P}[T_a \leq n] = \lim_{n\to\infty} \mathbb{P}[S_n \notin [-a, a-1]] = 1 \ .$$

Somit erreicht die Irrfahrt den Wert a in endlicher Zeit.

Korollar 1.18. *Für die Rückkehrzeit nach* 0 *erhalten wir*

$$\mathbb{P}[T_0 > 2n] = \mathbb{P}[S_{2n} = 0] \ .$$

Beweis. Wir erhalten

$$\begin{aligned}\mathbb{P}[T_0 > 2n] &= \mathbb{P}[S_1 \neq 0, \ldots, S_{2n} \neq 0] = 2\mathbb{P}[S_1 > 0, \ldots, S_{2n} > 0] \\ &= \mathbb{P}[S_1 > -1, \ldots, S_{2n-1} > -1] = \mathbb{P}[T_{-1} > 2n-1] = \mathbb{P}[T_1 > 2n-1] \\ &= \mathbb{P}[T_1 > 2n] = \mathbb{P}[S_{2n} \in [-1, 0]] = \mathbb{P}[S_{2n} = 0] \ .\end{aligned}$$

□

Wir sehen, dass $\lim_{n\to\infty} \mathbb{P}[T_0 > 2n] = 0$. Also kehrt die Irrfahrt in endlicher Zeit nach 0 zurück. Aber

$$\mathbb{E}[T_0] = \sum_{n=0}^{\infty} \mathbb{P}[T_0 > n] = 2\sum_{n=0}^{\infty} \mathbb{P}[T_0 > 2n] = 2\sum_{n=0}^{\infty} \binom{2n}{n} 2^{-2n} \geq 2\sum_{n=1}^{\infty} \frac{1}{\sqrt{\pi n}} \mathrm{e}^{-1/6} \ .$$

Es folgt $\mathbb{E}[T_0] = \infty$. Das heisst, dass man im Schnitt sehr lange warten muss, bis die Irrfahrt nach 0 zurückkehrt. Man sieht auch in Abbildung 1.6, dass die Verteilung in der Nähe von 0 konzentriert ist, aber für grosse n sehr langsam abfällt.

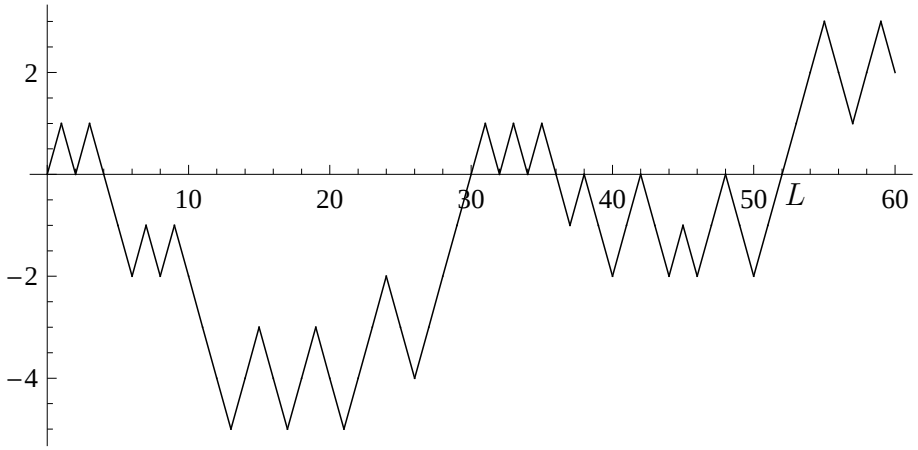

Abbildung 1.7: *Definition von L.*

Korollar 1.19. *Wir haben für $a > 0$*

$$\mathbb{P}[S_n = a, T_0 > n] = \frac{a}{n}\mathbb{P}[S_n = a] \; .$$

Beweis. Durchschreiten wir den Pfad rückwärts, starten wir in a und kehren erst im Zeitpunkt n nach Null zurück. Das ist das selbe wie in 0 zu starten und den Punkt $-a$ im Zeitpunkt n das erste Mal zu erreichen. Wir erhalten also

$$\mathbb{P}[S_n = a, T_0 > n] = \mathbb{P}[S_n = -a, T_{-a} = n] = \mathbb{P}[T_a = n] = \frac{a}{n}\mathbb{P}[S_n = a] \; .$$

□

1.4.5. Das arcsin Gesetz

Sei $L(\omega) = \max\{0 \leq n \leq 2N : S_n(\omega) = 0\}$ der Zeitpunkt des letzten Besuches in 0 vor $2N$. Im Glückspiel ist $L(\omega)$ der Zeitpunkt, seit dem der Leader in Führung ist. Da die Irrfahrt immer wieder nach 0 zurückkehrt, würde man erwarten, dass $\mathbb{P}[L \leq N] \to 0$, wenn N gegen Unendlich geht. Es zeigt sich aber, dass die Verteilung von L symmetrisch ist.

Satz 1.20. (arcsin-Gesetz) *Die Verteilung von L ist gegeben durch*

$$\mathbb{P}[L = 2n] = \mathbb{P}[S_{2n} = 0]\mathbb{P}[S_{2(N-n)} = 0] = \binom{2n}{n}\binom{2(N-n)}{N-n}2^{-2N} \; .$$

Diese Verteilung heisst **diskrete arcsin Verteilung**.

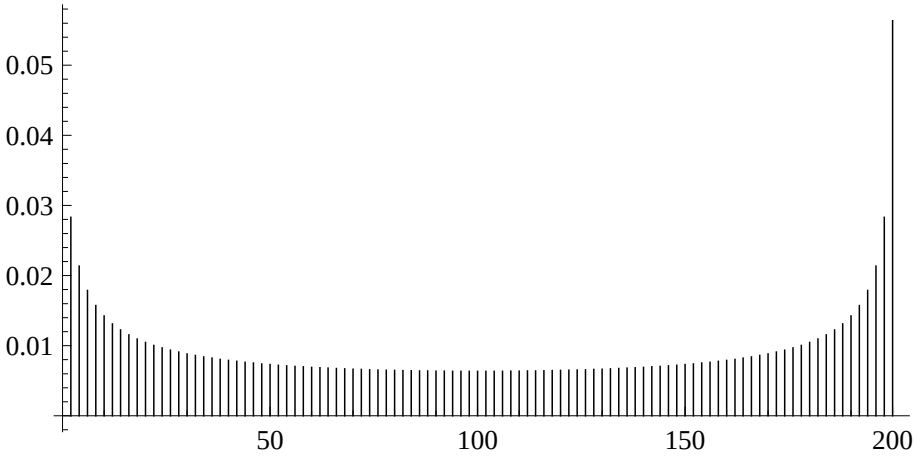

Abbildung 1.8: *Verteilung von L.*

Beweis. Die Anzahl Pfade mit $L = 2n$ sind die Anzahl Pfade der Länge $2n$ mit $S_{2n} = 0$ mal die Anzahl Pfade der Länge $2N - 2n$, die nicht nach 0 zurückkehren. Somit erhalten wir

$$\mathbb{P}[L = 2n] = \mathbb{P}[S_{2n} = 0]\mathbb{P}[T_0 > 2N - 2n] = \mathbb{P}[S_{2n} = 0]\mathbb{P}[S_{2N-2n} = 0] \; .$$

Einsetzen der Formel ergibt den letzten Ausdruck. □

Bemerkung. Wir haben gesehen, dass $\mathbb{P}[S_{2n} = 0] \approx 1/\sqrt{\pi n}$, also ist $\mathbb{P}[L = 2n] \approx 1/(\pi\sqrt{n(N-n)})$. Definieren wir $f(x) = 1/(\pi\sqrt{x(1-x)})$, erhalten wir

$$\mathbb{P}\Big[\frac{L}{2N} \leq x\Big] \approx \sum_{n \leq xN} \frac{1}{N} f(n/N) \approx \int_0^x f(y)\, \mathrm{d}y = \frac{2}{\pi} \arcsin\sqrt{x} \; .$$

Daher kommt der Name arcsin-Verteilung. ■

1.4.6. Das Gesetz vom iterierten Logarithmus

Wir könnten uns nun noch dafür interessieren, in welchem Gebiet sich die Pfade $\{S_n\}$ bewegen. Betrachten wir nun den Fall $N = \infty$. Das heisst, wir arbeiten mit einem stetigen Wahrscheinlichkeitsraum, den wir in Kapitel 2 einführen werden. Dann gilt das folgende Resultat, siehe auch Satz 3.5:

Satz 1.21. (Gesetz vom iterierten Logarithmus) *Es gilt mit Wahrscheinlichkeit* 1

$$\overline{\lim_{n\to\infty}} \frac{S_n}{\sqrt{2n\log\log n}} = -\underline{\lim_{n\to\infty}} \frac{S_n}{\sqrt{2n\log\log n}} = 1\ .$$

□

1.5. Bedingte Wahrscheinlichkeiten

1.5.1. Definition

Wir arbeiten weiter mit einem diskreten Wahrscheinlichkeitsraum.

Definition 1.22. *Sei B ein Ereignis mit $\mathbb{P}[B] > 0$. Der Ausdruck*

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

heisst **bedingte Wahrscheinlichkeit** *des Ereignisses A bezüglich B.*

$\mathbb{P}[A \mid B]$ ist das relative Gewicht von $A \cap B$ am Ereignis B. Wenn wir schon wissen (oder annehmen), dass B eintritt, dann wird, falls A eintritt, auch $A \cap B$ eintreten. Wenn wir nun ein Experiment sehr oft durchführen, und dann nur die Experimente betrachten, bei denen B eingetreten ist, dann ist der relative Anteil der Experimente, bei denen auch A eintritt, $\mathbb{P}[A \mid B]$. Zum Beispiel, wir wissen, dass ein medizinischer Test positiv ist (B). Wir interesieren uns für das Ereignis 'krank' (A).

In einem Laplace-Modell, sind die möglichen Fälle $|B|$, falls B eintritt. Die Anzahl der günstigen Fälle, bei denen auch A eintritt, sind $|A \cap B|$. Also haben wir

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} = \frac{|A \cap B|/|\Omega|}{|B|/|\Omega|} = \frac{|A \cap B|}{|B|}\ ,$$

was mit der Intuition übereinstimmt.

Es ist einfach zu sehen, dass $p'(\omega) = \mathbb{P}[\omega \mid B]$ eine Wahrscheinlichkeitsverteilung ist. Somit gelten für $\mathbb{P}[\cdot \mid B]$ die selben Regeln wie für $\mathbb{P}[\cdot]$. Wir haben weiter, falls $B \subset A$, dass $\mathbb{P}[A \mid B] = 1$. Falls $B \subset A^c$, dann gilt $\mathbb{P}[A \mid B] = 0$.

Beispiele

- Wir werfen zwei Würfel. Der erste Würfel zeigt 4 an. Die Wahrscheinlichkeit, dass der zweite Würfel 6 zeigt, ist dann

$$\mathbb{P}[X_2 = 6 \mid X_1 = 4] = \frac{\mathbb{P}[X_1 = 4, X_2 = 6]}{\mathbb{P}[X_1 = 4]} = \frac{1/36}{1/6} = \frac{1}{6} \, .$$

  Die Verteilung der Augenanzahl des zweiten Würfels wird durch die Information über den ersten Würfel nicht geändert. Anders ist es, falls wir wissen, dass die Summe der Augen 9 beträgt. Da es vier Möglichkeiten gibt, wie 9 entstehen kann, ist $\mathbb{P}[X_1 + X_2 = 9] = 4/36 = 1/9$. Somit haben wir

$$\mathbb{P}[X_2 = 6 \mid X_1 + X_2 = 9] = \frac{\mathbb{P}[X_1 = 3, X_2 = 6]}{\mathbb{P}[X_1 + X_2 = 9]} = \frac{1/36}{1/9} = \frac{1}{4} \, .$$

  Es kann aber auch sein, dass die Augensummen keine Information über den zweiten Würfel enthält. So ist

$$\mathbb{P}[X_2 = 6 \mid X_1 + X_2 = 7] = \frac{\mathbb{P}[X_1 = 1, X_2 = 6]}{\mathbb{P}[X_1 + X_2 = 7]} = \frac{1/36}{1/6} = \frac{1}{6} \, .$$

- Jemand wirft mit verbundenen Augen zwei Münzen. Er erhält die Information, dass mindestens einmal Kopf geworfen wurde. Dann ist die Wahrscheinlichkeit, dass zweimal Kopf geworfen wurde

$$\mathbb{P}[(K,K) \mid \text{mind. 1x Kopf}] = \mathbb{P}[(K,K) \mid \{(K,K),(K,Z),(Z,K)\}] = \frac{1/4}{3/4} = \frac{1}{3} \, .$$

  Viele Menschen würden sagen, dass es zwei Möglichkeiten für die andere Münze gibt, und dass damit die Wahrscheinlichkeit 1/2 sein sollte. Dieses Argument ist falsch, da man nicht die Information erhält, welche Münze Kopf zeigt.

- Bei der Irrfahrt erhalten wir für $a \neq 0$, $n + a$ gerade und $n \geq |a|$

$$\mathbb{P}[T_a = n \mid S_n = a] = \frac{\mathbb{P}[T_a = n, S_n = a]}{\mathbb{P}[S_n = a]} = \frac{\mathbb{P}[T_a = n]}{\mathbb{P}[S_n = a]} = \frac{|a|}{n} \, .$$

  Also gibt es eine einfache Formel, dass, falls $S_n = a$, dies auch der erste Besuch in a ist. Weiter haben wir

$$\mathbb{P}[T_0 > n \mid S_n = a] = \frac{\mathbb{P}[T_0 > n, S_n = a]}{\mathbb{P}[S_n = a]} = \frac{|a|}{n} \, .$$

- Bei einer Wahl zwischen zwei Kandidaten erhält der erste Kandidat a Stimmen, der zweite Kandidat b Stimmen. Es sei $a > b$. Wie gross ist die Wahrscheinlichkeit, dass der erste Kandidat während der ganzen Auszählung in Führung liegt. Sei $N = a + b$. Setzen wir $X_i = 1$, falls die i-te ausgezählte Stimme dem ersten Kandidaten gehört, und $X_i = -1$ sonst. Definieren wir $S_n = X_1 + \cdots + X_n$, dann müssen wir alle Pfade bestimmen, bei denen $S_n > 0$ für alle n, wobei wir wissen, dass $S_N = a - b$. Dies ist das gleiche Problem wie bei der Irrfahrt. Wir haben daher

$$\begin{aligned}\mathbb{P}[\text{immer in Führung}] &= \mathbb{P}[S_1 > 0, \ldots, S_N > 0 \mid S_N = a - b] \\ &= \mathbb{P}[T_0 > N \mid S_N = a - b] = \frac{a-b}{N} = \frac{a-b}{a+b}\,.\end{aligned}$$

Definition 1.23. *Sei X eine Zufallsvariable und B ein Ereignis mit $\mathbb{P}[B] > 0$. Der Ausdruck*

$$\mathbb{E}[X \mid B] = \sum_{\omega \in \Omega} X(\omega)\mathbb{P}[\omega \mid B]$$

heisst **bedingter Erwartungswert** *bezüglich B.*

Betrachten wir eine Irrfahrt. Wir haben schon im Hilfssatz 1.10 bewiesen, dass

$$\begin{aligned}&\mathbb{P}[X_{n+1} = 1 \mid S_1 = s_1, S_2 = s_2, \ldots, S_n = s_n] \\ &\quad = \mathbb{P}[X_{n+1} = -1 \mid S_1 = s_1, S_2 = s_2, \ldots, S_n = s_n] \\ &\quad = \frac{\mathbb{P}[S_1 = s_1, S_2 = s_2, \ldots, S_n = s_n, X_{n+1} = -1]}{\mathbb{P}[S_1 = s_1, S_2 = s_2, \ldots, S_n = s_n]} = \tfrac{1}{2}\,.\end{aligned}$$

Somit erhalten wir aus dem bisherigen Verlauf keine Informationen. Insbesondere haben wir $\mathbb{E}[X_{n+1} \mid S_1 = s_1, S_2 = s_2, \ldots, S_n = s_n] = 0$. Anders sieht es aus, falls wir den Endstand $S_N = a$ kennen. Wir erhalten

$$\begin{aligned}\mathbb{P}[X_i = 1 \mid S_N = a] = \frac{\mathbb{P}[S_N = a, X_i = 1]}{\mathbb{P}[S_N = a]} &= \frac{\binom{N-1}{(N+a)/2-1}2^{-N}}{\binom{N}{(N+a)/2}2^{-N}} = \frac{(N+a)/2}{N} \\ &= \frac{1}{2} + \frac{a}{2N}\,.\end{aligned}$$

Kennen wir auch die Vorgeschichte, müssen wir alle Pfade von s_n nach a statt der Pfade von 0 nach a betrachten, also haben wir

$$\mathbb{P}[X_{n+1} = 1 \mid S_N = a, S_1 = s_1, \ldots, S_n = s_n] = \frac{1}{2} + \frac{a - s_n}{2(N-n)}\,.$$

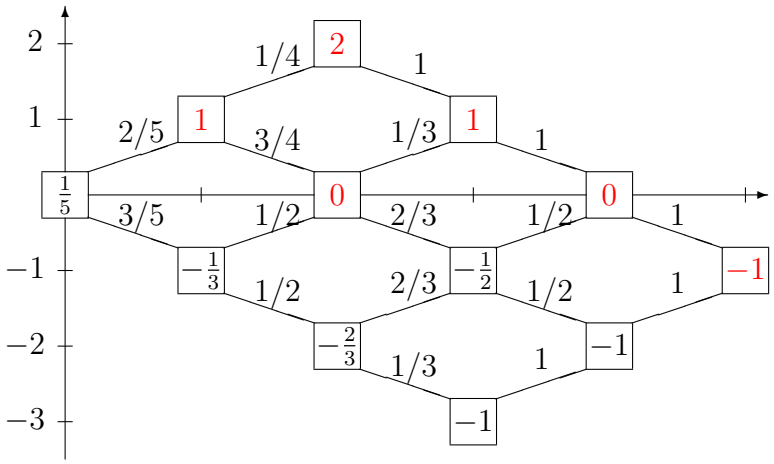

Abbildung 1.9: *Dynamische Programmierung*

Der bedingte Erwartungswert wird dann

$$\begin{aligned}\mathbb{E}[X_{n+1} \mid S_N = a, S_1 = s_1, \ldots, S_n = s_n] \\ = 1\Big(\frac{1}{2} + \frac{a - s_n}{2(N-n)}\Big) + (-1)\Big(\frac{1}{2} - \frac{a - s_n}{2(N-n)}\Big) = \frac{a - s_n}{N - n}\,.\end{aligned}$$

Die Irrfahrt hat also eine Drift, falls $s_n \neq a$.

Durch die Insider-Information können wir Gewinnspiele ausnutzen, um positive Gewinnerwartungen zu erhalten. Es ist klar, dass wir im Falle $a \geq 0$ einen Gewinn erzielen können. Ist z.B. $a > 0$, werden wir das Spiel sicher nicht stoppen, wenn $S_t \leq a$. Auch im Falle $a = 0$, gibt die Stoppzeit $T = \inf\{n \leq N : S_n = 1\} \wedge N$ eine positive Gewinnerwartung. Aber sogar im Falle $a < 0$, kann es sein, dass wir eine Stoppzeit T konstruieren können, für die $\mathbb{E}[S_T] > 0$ ("bad news is better than no news").

Nehmen wir an, $N = 5$ und $S_N = a = -1$. Zum Zeitpunkt 4 ist entweder $S_4 = 0$ oder $S_4 = -2$. Im ersten Fall werden wir stoppen, im zweiten Fall weiterspielen. Der Wert der Strategie im ersten Fall ist 0, im zweiten Fall -1. Zur Zeit $n = 3$ ist $S_3 \in \{1, -1, -3\}$. Wir werden somit stoppen, falls $S_3 = 1$, und sonst weiterspielen. Schlechter als -1 kann es ja nicht werden, und falls $S_3 = -1$ haben wir ja immer noch die Möglichkeit, dass $S_4 = 0$, also ist es besser weiterzuspielen. Der Wert der optimalen Stoppstrategie ist somit 1, falls $S_3 = 1$, $-1/2 = \frac{1}{2}0 + \frac{1}{2}(-1)$, falls $S_3 = -1$. So können wir die optimale Strategie rekursiv bestimmen. Dieses Verfahren heisst **dynamische Programmierung**. Wir haben $V(a, N) = a$ und

$$V(s, n) = \max\{s, \mathbb{E}[V(S_{n+1}, n+1) \mid S_n = s, S_N = a]\}\,.$$

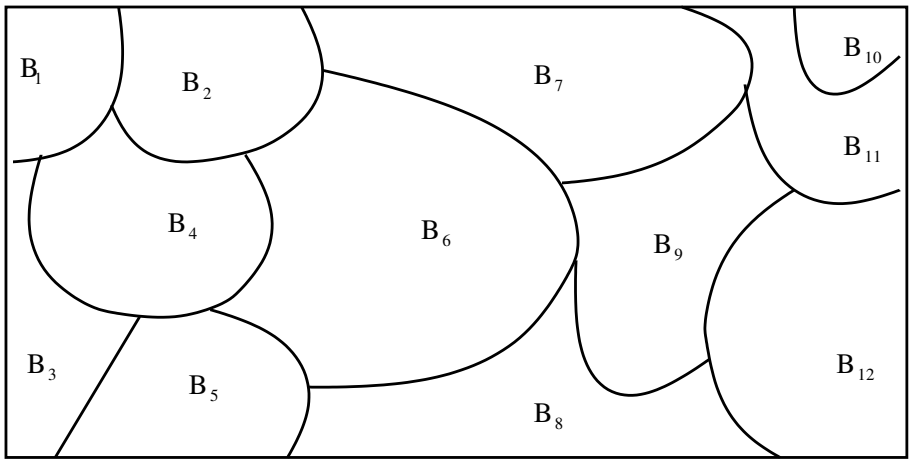

Abbildung 1.10: *Zerlegung von* Ω

Die bedingten Übergangswahrscheinlichkeiten sind in Abbildung 1.9 dargestellt. Damit lässt sich die Funktionen $V(s,n)$ berechnen. Eine optimale Strategie erhalten wir, wenn $V(s,n) = s$. Daher ist in unserem Beispiel eine mögliche optimal Strategie

$$T = \begin{cases} 1, & \text{falls } S_1 = 1, \\ 3, & \text{falls } S_1 = -1 \text{ und } S_3 = 1, \\ 4, & \text{falls } S_1 = -1,\ S_3 = -1 \text{ und } S_4 = 0, \\ 5, & \text{sonst.} \end{cases}$$

Der erwartete Gewinn ist dann $\mathbb{E}[S_T \mid S_5 = -1] = V(0,0) = 1/5$.

1.5.2. Berechnung von absoluten Wahrscheinlichkeiten aus bedingten

Wir haben eine Urne mit n schwarzen und k roten Kugeln. Wir ziehen eine Kugel und legen dann zwei der gleichen Farbe zurück. Dann ziehen wir eine Kugel. Wir wollen dann wissen, wie gross die Wahrscheinlichkeit ist, beim zweiten Zug eine schwarze Kugel zu ziehen.

Wenn wir nun wissen, dass beim ersten Mal eine schwarze Kugel gezogen wurde, kennen wir die Wahrscheinlichkeit $(n+1)/(n+k+1)$. Wurde beim ersten Mal eine rote Kugel gezogen, ist die Wahrscheinlichkeit $n/(n+k+1)$. Wie wir nun unser Problem lösen, sagt der folgende

Satz 1.24. *Sei* $\{B_i\}$ *eine Zerlegung von* Ω*, das heisst,* $B_i \cap B_j = \emptyset$*, falls* $i \neq j$*, und* $\cup_i B_i = \Omega$*. Dann gilt für jedes Ereignis* A*,*

$$\mathbb{P}[A] = \sum_{i:\mathbb{P}[B_i]\neq 0} \mathbb{P}[A \mid B_i]\mathbb{P}[B_i] \,.$$

Beweis. Wir können $A = A \cap \Omega = A \cap (\cup_i B_i) = \cup_i (A \cap B_i)$ schreiben und haben $(A \cap B_i) \cap (A \cap B_j) = \emptyset$ für $i \neq j$. Daher gilt

$$\mathbb{P}[A] = \sum_i \mathbb{P}[A \cap B_i] = \sum_{i:\mathbb{P}[B_i]\neq 0} \mathbb{P}[A \cap B_i] = \sum_{i:\mathbb{P}[B_i]\neq 0} \mathbb{P}[A \mid B_i]\mathbb{P}[B_i] \; .$$

□

Beispiele

- Im oben betrachteten Beispiel sei X_1 die erste gezogene Kugel und X_2 die zweite gezogene Kugel. Dann haben wir

$$\begin{aligned}\mathbb{P}[X_2 = S] &= \mathbb{P}[X_2 = S \mid X_1 = S]\mathbb{P}[X_1 = S] + \mathbb{P}[X_2 = S \mid X_1 = R]\mathbb{P}[X_1 = R] \\ &= \frac{n+1}{n+k+1}\,\frac{n}{n+k} + \frac{n}{n+k+1}\,\frac{k}{n+k} = \frac{n}{n+k} \; .\end{aligned}$$

- Seien n schwarze und n rote Kugeln auf zwei Urnen verteilt. In der ersten Urne seien k schwarze und ℓ rote Kugeln. Wir wählen die erste Urne mit Wahrscheinlichkeit p, die zweite Urne mit Wahrscheinlichkeit $1-p$ und ziehen aus dieser Urne eine Kugel. Die Wahrscheinlichkeit einer schwarzen Kugel ist dann

$$\mathbb{P}[S] = \mathbb{P}[S \mid 1]\mathbb{P}[1] + \mathbb{P}[S \mid 2]\mathbb{P}[2] = \frac{k}{k+\ell}p + \frac{n-k}{2n-k-\ell}(1-p) \; .$$

  Sogar im Falle $p = \frac{1}{2}$ ist diese Wahrscheinlichkeit verschieden von $\frac{1}{2}$ (n von $2n$ Kugeln sind schwarz), falls $k \neq \ell$ und $n \neq k + \ell$.

- Bei einer Irrfahrt wollen wir $\mathbb{P}[T_0 > 2n]$ bestimmen. Wir erhalten dann

$$\mathbb{P}[T_0 > 2n] = \sum_{j=-n}^{n} \mathbb{P}[T_0 > 2n \mid S_{2n} = 2j]\mathbb{P}[S_{2n} = 2j] = \sum_{j=-n}^{n} \frac{|j|}{n}\binom{2n}{n+j}2^{-2n} \; .$$

Satz 1.25. *Seien A_1, A_2, …,A_n Ereignisse. Dann gilt*

$$\mathbb{P}[A_1 \cap A_2 \cap \cdots \cap A_n] = \mathbb{P}[A_1]\mathbb{P}[A_2 \mid A_1]\mathbb{P}[A_3 \mid A_1 \cap A_2] \cdots \mathbb{P}[A_n \mid A_1 \cap \cdots \cap A_{n-1}] \; .$$

Beweis. Wir beweisen den Satz mittels vollständiger Induktion. Die Formel gilt für $n = 1$. Falls sie für n gilt, haben wir

$$\begin{aligned}\mathbb{P}[A_1 \cap \cdots \cap A_n \cap A_{n+1}] &= \mathbb{P}[(A_1 \cap \cdots \cap A_n) \cap A_{n+1}] \\ &= \mathbb{P}[A_{n+1} \mid A_1 \cap \cdots \cap A_n]\mathbb{P}[A_1 \cap \cdots \cap A_n] \; .\end{aligned}$$

Einsetzen der Induktionsvoraussetzung ergibt die Aussage. □

Sei nun $\Omega = \{(x_0, x_1, \ldots, x_N) : x_i \in E\}$. Wir nennen dann die Familie von Zufallsvariablen $\{X_i : 0 \leq i \leq N\}$ einen **stochastischen Prozess**. Die Wahrscheinlichkeitsfunktion lässt sich dann schreiben als

$$\begin{aligned}&\mathbb{P}[X_0 = x_0, \ldots, X_N = x_N] \\ &= \mathbb{P}[X_0 = x_0]\mathbb{P}[X_1 = x_1 \mid X_0 = x_0] \cdots \mathbb{P}[X_N = x_N \mid X_0 = x_0, \ldots, X_{N-1} = x_{N-1}] \; .\end{aligned}$$

Der stochastische Prozess ist also festgelegt durch die **Startverteilung** $\mathbb{P}[X_0 = x_0]$ und die bedingten Wahrscheinlichkeiten $\mathbb{P}[X_n = x_n \mid X_0 = x_0, \ldots, X_{n-1} = x_{n-1}]$. Bei der Irrfahrt haben wir $\mathbb{P}[S_0 = 0] = 1$ und

$$\mathbb{P}[S_n = s_n \mid S_0 = 0, \ldots, S_{n-1} = s_{n-1}] = \tfrac{1}{2}\mathbb{1}_{s_n \in \{s_{n-1}-1, s_{n-1}+1\}} \; .$$

Dies ist besonders einfach, da der nächste Wert nicht vom ganzen Pfad abhängt, sondern nur vom letzten Wert. Einen stochastischen Prozess mit dieser Eigenschaft, das heisst

$$\mathbb{P}[X_n = x_n \mid X_0 = x_0, \ldots, X_{n-1} = x_{n-1}] = \mathbb{P}[X_n = x_n \mid X_{n-1} = x_{n-1}] \; ,$$

nennt man **Markov-Prozess**.

In der Lebensversicherungsmathematik tabelliert man die Wahrscheinlichkeiten $p_x = \mathbb{P}[T > x + 1 \mid T > x]$, wobei T die Lebenszeit einer Frau oder eines Mannes ist. Ist nun eine Person bei Versicherungsbeginn x Jahre alt, lässt sich die Verteilung des Alters beim Tod aus diesen bedingten Wahrscheinlichkeiten berechnen

$$\mathbb{P}[T \in (t, t+1] \mid T > x] = (1 - p_t) \prod_{s=x}^{t-1} p_s \; .$$

Insbesondere gilt

$$\mathbb{P}[T > t \mid T > x] = \prod_{s=x}^{t-1} p_s \; .$$

Einen Spezialfall erhalten wir, falls wir $p_s = p$ konstant wählen. Dann ist

$$\mathbb{P}[T > t \mid T > x] = p^{t-x} \; .$$

Dies ist die **geometrische Verteilung** mit Parameter $p \in (0, 1)$, $\mathbb{P}[T > t] = p^t$. Wir haben nämlich

$$\mathbb{P}[T > t \mid T > x] = \frac{\mathbb{P}[T > t, T > x]}{\mathbb{P}[T > x]} = \frac{\mathbb{P}[T > t]}{\mathbb{P}[T > x]} = \frac{p^t}{p^x} = p^{t-x} \; .$$

Also, $\mathbb{P}[T > t+s \mid T > t] = p^s = \mathbb{P}[T > s]$. Die geometrische Verteilung hat kein Gedächtnis. Die Verteilung der Restwartezeit ändert sich nicht, wenn wir schon t Zeiteinheiten gewartet haben. Für die geometrische Verteilung gilt

$$\mathbb{P}[T = t] = \mathbb{P}[T > t-1] - \mathbb{P}[T > t] = (1-p)p^{t-1} ,$$

wobei $t \geq 1$.

1.5.3. Die Bayes'sche Regel

Folgende Regel ist oft nützlich.

Satz 1.26. *Sei $\{B_i\}$ eine Zerlegung von Ω und A ein Ereignis mit $\mathbb{P}[A] > 0$. Dann gilt für alle i,*

$$\mathbb{P}[B_i \mid A] = \frac{\mathbb{P}[A \mid B_i]\mathbb{P}[B_i]}{\sum_{j:\mathbb{P}[B_j]>0} \mathbb{P}[A \mid B_j]\mathbb{P}[B_j]} .$$

Beweis. Nach der Definition gilt

$$\mathbb{P}[B_i \mid A] = \frac{\mathbb{P}[A \cap B_i]}{\mathbb{P}[A]} .$$

Der Zähler lässt sich mit Hilfe von Satz 1.25 umschreiben, der Nenner mit Hilfe von Satz 1.24. □

Beispiele

- Von 1000 Personen haben 3 eine bestimmte Krankheit. Nennen wir B das Ereignis, 'Sie haben die Krankheit'. Ein Test auf die Krankheit ist in 95% positiv, wenn eine Person die Krankheit hat. Der Test ist in 99% der Fälle negativ, wenn eine Person die Krankheit nicht hat. Nennen wir das Ereignis, 'Sie haben einen positiven Test' A. Nehmen wir an, Ihr Test ist positiv. Dann ist die Wahrscheinlichkeit, dass Sie die Krankheit haben

$$\mathbb{P}[B \mid A] = \frac{\mathbb{P}[A \mid B]\mathbb{P}[B]}{\mathbb{P}[A \mid B]\mathbb{P}[B] + \mathbb{P}[A \mid B^c]\mathbb{P}[B^c]} = \frac{0.95 \cdot 0.003}{0.95 \cdot 0.003 + 0.01 \cdot 0.997}$$
$$= \frac{285}{1282} = 0.2223 .$$

  Die Wahrscheinlichkeit, dass Sie die Krankheit haben, wenn Ihr Test positiv ist, ist somit etwa 22%. Dies liegt daran, dass die Krankheit selten ist. Ein positiver Test ist also kein Grund zur Panik. Die Industrie würde nun den Test mit dem Argument verkaufen, dass die Wahrscheinlichkeit, dass Sie die Krankheit haben, durch ein positives Testergebnis auf das 74-fache steigt.

- Auf einem Übermittlungskanal werden Nachrichten aus einem Alphabet I gesendet, und Nachrichten aus einem Alphabet J empfangen. Zum Beispiel, $I = \{0, 1, 2, \ldots, 15\}$, $J = \{0, 1, 2, \ldots, 127\}$. Das Ereignis A_j besagt, dass j empfangen wurde, das Ereignis B_i besagt, dass i gesendet wurde. Man kennt die relative Buchstabenhäufigkeit $\mathbb{P}[B_i]$, und die bedingten Wahrscheinlichkeiten $\mathbb{P}[A_j \mid B_i]$. Da *Rauschen* die Übertragung stört, kann man das gesendete Signal nicht eindeutig bestimmen. Wir suchen nun eine *Dekodierung* $\phi : J \to I$, so dass das Ereignis $C_\phi = \{\cup_j (A_j \cap B_{\phi(j)})\}$ maximale Wahrscheinlichkeit hat. C_ϕ ist das Ereignis, dass richtig dekodiert wurde. Als Erstes müssen wir aus der Bayes'schen Formel die Wahrscheinlichkeiten

$$\mathbb{P}[B_i \mid A_j] = \frac{\mathbb{P}[A_j \mid B_i]\mathbb{P}[B_i]}{\sum_k \mathbb{P}[A_j \mid B_k]\mathbb{P}[B_k]}$$

berechnen. Wir bemerken, dass

$$\mathbb{P}[C_\phi] = \sum_j \mathbb{P}[B_{\phi(j)} \mid A_j]\mathbb{P}[A_j] \,,$$

wobei wir

$$\mathbb{P}[A_j] = \sum_i \mathbb{P}[A_j \mid B_i]\mathbb{P}[B_i]$$

haben. Die Lösung des Problems ist also $\phi(j)$ so zu wählen, dass

$$\mathbb{P}[B_{\phi(j)} \mid A_j] = \max_i \mathbb{P}[B_i \mid A_j] \,.$$

Wir wählen also das B_i, das am wahrscheinlichsten gesendet wurde, falls man A_j empfangen hat.

Nehmen wir an, $I = J = \{0, 1\}$. Wir benötigen $\alpha = \mathbb{P}[B_1]$, $p_0 = \mathbb{P}[A_0 \mid B_0]$ und $p_1 = \mathbb{P}[A_1 \mid B_1]$. Wir finden somit

$$\begin{aligned}
\mathbb{P}[B_0 \mid A_0] &= \frac{p_0(1-\alpha)}{p_0(1-\alpha) + (1-p_1)\alpha} \,, \\
\mathbb{P}[B_1 \mid A_0] &= \frac{(1-p_1)\alpha}{p_0(1-\alpha) + (1-p_1)\alpha} \,, \\
\mathbb{P}[B_0 \mid A_1] &= \frac{(1-p_0)(1-\alpha)}{(1-p_0)(1-\alpha) + p_1\alpha} \,, \\
\mathbb{P}[B_1 \mid A_1] &= \frac{p_1\alpha}{(1-p_0)(1-\alpha) + p_1\alpha} \,.
\end{aligned}$$

Falls 0 empfangen wird, wählen wir 0, wenn $p_0(1-\alpha) > (1-p_1)\alpha$, also $\alpha < p_0/(1+p_0-p_1)$. Falls 1 empfangen wird, wählen wir 1, falls $p_1\alpha > (1-p_0)(1-\alpha)$,

das heisst falls $\alpha > (1-p_0)/(1+p_1-p_0)$. Nehmen wir an, dass $\frac{1}{2} < p_i < 1$, dann gilt immer

$$\frac{1-p_0}{1+p_1-p_0} < \frac{p_0}{1+p_0-p_1} \; .$$

Wir wählen folgende Dekodierung

$$(\phi(0),\phi(1)) = \begin{cases} (0,0) \; , & \text{falls } \alpha \leq \frac{1-p_0}{1+p_1-p_0}, \\ (0,1) \; , & \text{falls } \frac{1-p_0}{1+p_1-p_0} < \alpha < \frac{p_0}{1+p_0-p_1}, \\ (1,1) \; , & \text{sonst.} \end{cases}$$

Wir sehen also, falls 1 zu oft oder zu selten gesendet wird, ist eine Dekodierung nicht möglich.

1.6. Unabhängigkeit

1.6.1. Definition von unabhängigen Ereignissen

Ein Ereignis B wird nicht durch ein Ereignis A beeinflusst, falls $\mathbb{P}[B \mid A] = \mathbb{P}[B]$. Dies können wir auch schreiben als $\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B]$. Daher machen wir folgende

Definition 1.27. *Sei $\{A_i : i \in I\}$ eine Kollektion von Ereignissen. Wir sagen $\{A_i : i \in I\}$ ist* **(stochastisch) unabhängig**, *falls*

$$\forall J \subset I \text{ (endlich)} \quad \Longrightarrow \quad \mathbb{P}[\cap_{j\in J} A_j] = \prod_{j\in J} \mathbb{P}[A_j] \; .$$

Die Unabhängigkeit ist also keine Eigenschaft der Ereignisse, sondern der Wahrscheinlichkeitsverteilung. Wir bemerken auch, dass die paarweise Unabhängigkeit nicht die Unabhängigkeit impliziert. Sei zum Beispiel

$$\Omega = \{(1,1,1),(1,2,3),(1,3,2),(2,1,3),(2,2,2),(2,3,1),(3,1,2),(3,2,1),(3,3,3)\} \; .$$

Wir beschreiben ω durch die drei Zufallsvariablen (X_1, X_2, X_3). Wählen wir die Gleichverteilung auf Ω, so sind die abhängigen Ereignisse $\{X_1 = a\}$, $\{X_2 = b\}$, $\{X_3 = c\}$ paarweise unabhängig. Es genügt auch nicht die Definition für $J = I$ zu fordern. Hat zum Beispiel ein Ereignis die Wahrscheinlichkeit 0, so folgt

$$\mathbb{P}[\cap_{j\in I} A_j] = 0 = \prod_{j\in I} \mathbb{P}[A_j] \; .$$

Hilfssatz 1.28. *Seien $\{A_i : i \in I\}$ unabhängig. Für die Ereignisse $\{B_i : i \in I\}$ gelte, $B_i = A_i$, oder $B_i = A_i^c$. Dann sind auch B_i unabhängig.*

Beweis. Es genügt die Aussage für endliches I zu zeigen. Falls $B_i = A_i$ für alle i, ist die Aussage trivial. Durch Induktion nach der Anzahl i, für die $B_i = A_i^c$, genügt es die Aussage zu zeigen, falls $B_i = A_i^c$ und $B_j = A_j$ für alle $j \neq i$. Wir haben

$$\begin{aligned}\mathbb{P}[A_i^c \cap (\cap_{j\neq i} A_j)] &= \mathbb{P}[\cap_{j\neq i} A_j] - \mathbb{P}[A_i \cap (\cap_{j\neq i} A_j)] = \prod_{j\neq i} \mathbb{P}[A_j] - \mathbb{P}[A_i] \prod_{j\neq i} \mathbb{P}[A_j] \\ &= (1 - \mathbb{P}[A_i]) \prod_{j\neq i} \mathbb{P}[A_j] = \mathbb{P}[A_i^c] \prod_{j\neq i} \mathbb{P}[A_j] \ .\end{aligned}$$

□

Ein Beispiel ist die Irrfahrt. Hier sind die Ereignisse $\{\{X_i = 1\} : 1 \leq i \leq N\}\}$ unabhängig, da

$$\mathbb{P}[X_{n_1} = 1, \ldots, X_{n_k} = 1] = \frac{2^{N-k}}{2^N} = 2^{-k} = \mathbb{P}[X_{n_1} = 1] \cdots \mathbb{P}[X_{n_k} = 1]$$

gilt. Wir können einige oder alle der 1'en durch -1 ersetzen, und somit sind die Ereignisse $\{\{X_i = x_i\} : 1 \leq i \leq N\}$ unabhängig.

1.6.2. Unabhängige und identisch verteilte $\{0, 1\}$ Experimente

Sei $N \in \mathbb{N}$, und $\Omega = \{(x_1, x_2, \ldots, x_N\} : x_i \in \{0, 1\}\}$. Wir bezeichnen mit X_i das Ergebnis des i-ten Experiments. Wir nehmen an, dass die Ereignisse $\{X_i = 1 : 1 \leq i \leq N\}$ unabhängig sind, und dass $\mathbb{P}[X_i = 1] = p$. Im Laplace-Modell ist $p = \frac{1}{2}$. Sei $k = \sum_{i=1}^N x_i$, so erhalten wir

$$\mathbb{P}[X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N] = \prod_{i=1}^N \mathbb{P}[X_i = x_i] = p^k (1-p)^{N-k} \ .$$

Wir wollen nun die Anzahl Erfolge $(X_i = 1)$ zählen, $S_N = \sum_{i=1}^N X_i$. Für den Erwartungswert erhalten wir sofort $\mathbb{E}[S_N] = \sum_{i=1}^N \mathbb{E}[X_i] = Np$. Wir wollen nun die Verteilung von S_N bestimmen. Damit $S_N = k$, müssen wir alle ω finden, für die $S_N(\omega) = k$. Jedes dieser ω hat die gleiche Wahrscheinlichkeit $p(\omega) = p^k(1-p)^{N-k}$. Also müssen wir noch die Anzahl solcher ω's finden. Aus einer Menge von N Elementen müssen wir k wählen, die wir gleich 1 setzen. Daher haben wir $\binom{N}{k}$ Möglichkeiten, diese Elemente zu wählen. Also haben wir

$$\mathbb{P}[S_n = k] = \binom{N}{k} p^k (1-p)^{N-k} \ .$$

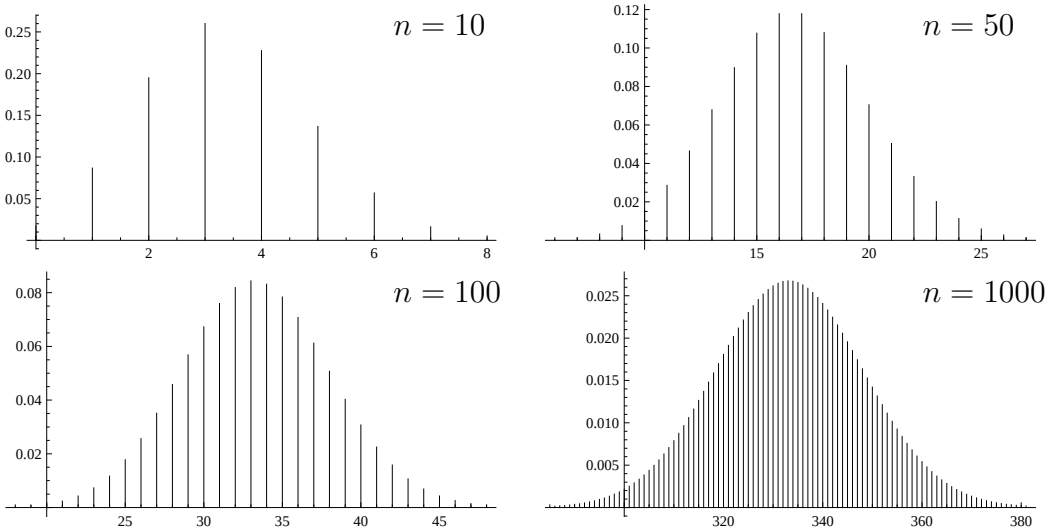

Abbildung 1.11: *Binomialverteilung für* $p = 1/3$ *und* $n = 10, 50, 100, 1000$.

Dies ist die **Binomialverteilung** mit Parametern N und p. Für grosse N ist die Berechnung der Wahrscheinlichkeiten schwierig. Betrachten wir die "standardisierte" Verteilung

$$Z_N = \frac{S_N - Np}{\sqrt{Np(1-p)}} \,,$$

so nähert die sich immer mehr einer bestimmten Kurve an (Abbildung 1.11). Wir werden später im Abschnitt 3.4 beweisen, wie diese Kurve aussieht. Wir haben daher die Näherungsformel

$$\mathbb{P}[S_n \leq c] = \mathbb{P}\Big[\frac{S_N - Np}{\sqrt{Np(1-p)}} \leq \frac{c - Np}{\sqrt{Np(1-p)}}\Big] \approx \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}\, \mathrm{d}x \,,$$

wobei $z = (c - Np)/(\sqrt{Np(1-p)})$. Diese Integrale kann man in Tabellen finden.

Nehmen wir an, wir wollen ein Signal s über einen Datenkanal übertragen. Senden wir s ist die Wahrscheinlichkeit, dass auch s empfangen wird gleich $p \in (\frac{1}{2}, 1)$. Um die Wahrscheinlichkeit einer korrekten Übertragung zu erhöhen, senden wir das Signal N mal, und nehmen an, dass die einzelnen Übertragungen unabhängig sind. Wir wählen dann das Signal, das öfters auftritt. Wir haben dann also die Wahrscheinlichkeit einer korrekten Übertragung

$$\mathbb{P}[S_N > N/2] = \sum_{k=\lfloor N/2 \rfloor + 1}^{N} \binom{N}{k} p^k (1-p)^{N-k} \,.$$

Die folgende Tabelle zeigt ein paar Wahrscheinlichkeiten $\mathbb{P}[S_N > N/2]$ für verschiedene N und p.

N	3	5	7	9	19	49	99	199
$p = 0.95$	0.9928	0.9988	0.9998	1	1	1	1	1
$p = 0.9$	0.972	0.9914	0.9973	0.9991	1	1	1	1
$p = 0.7$	0.784	0.8369	0.874	0.9012	0.9674	0.9983	1	1
$p = 0.6$	0.648	0.6826	0.7102	0.7334	0.8139	0.9224	0.9781	0.9978
$p = 0.55$	0.5748	0.5931	0.6083	0.6214	0.6710	0.7597	0.8413	0.9216
$p = 0.51$	0.515	0.5187	0.5219	0.5246	0.5352	0.556	0.5791	0.6112

Ist p nahe bei 1, hat man für kleines N schon eine hohe Wahrscheinlichkeit, das richtige Ergebnis zu finden. Ist p nahe bei $\frac{1}{2}$, so muss man N gross wählen. Für $p = \frac{1}{2}$ ist sogar die Folge unabhängig vom Eingangssignal, und eine korrekte Dekodierung ist unmöglich.

Sei nun $T = \inf\{k : X_k = 1\}$ die Wartezeit auf den ersten Erfolg. Für die Verteilung erhalten wir

$$\mathbb{P}[T > k] = \mathbb{P}[X_1 = 0, \dots, X_k = 0] = (1-p)^k \ .$$

Die Wartezeit auf den ersten Erfolg hat also eine **geometrische Verteilung**.

In der Praxis wird man den Erfolgsparameter p nicht kennen. Aber die Folge $\{X_1, \dots, X_n\}$ enthält dann Information über den Parameter. Treten zum Beispiel viele Erfolge auf, wird man vermuten, dass p gross ist. Wir werden später im Kapitel 4 sehen, wie man p schätzen kann. Eine Möglichkeit zur Modellierung der Experimente ist der Ansatz von Laplace,

$$\mathbb{P}[A] = \int_0^1 \mathbb{P}_p[A] \, \mathrm{d}p \ ,$$

wobei $\mathbb{P}_p[A]$ die Verteilung bezeichnet, falls p der richtige Parameter ist. Man kann einfach zeigen, dass $\mathbb{P}$ eine Wahrscheinlichkeitsverteilung ist. Wir haben

$$\mathbb{P}[X_i = 1] = \int_0^1 p \, \mathrm{d}p = \tfrac{1}{2} \ .$$

Sei $k = x_1 + \cdots + x_n$, so erhalten wir

$$\mathbb{P}[X_1 = x_1, \dots, X_n = x_n] = \int_0^1 p^k (1-p)^{n-k} \, \mathrm{d}p = \frac{k!(n-k)!}{(n+1)!} \ .$$

Wir sehen, dass nun die Ereignisse $\{X_i = x_i\}$ abhängig sind. Berechnen wir die bedingte Verteilung des nächsten Experiments

$$\mathbb{P}[X_{n+1} = 1 \mid X_1 = x_1, \dots, X_n = x_n] = \frac{\dfrac{(k+1)!(n-k)!}{(n+2)!}}{\dfrac{k!(n-k)!}{(n+1)!}} = \frac{k+1}{n+2} \ .$$

Wir können das Resultat als

$$\frac{k+1}{n+2} = \frac{n}{n+2}\,\frac{k}{n} + \Big(1 - \frac{n}{n+2}\Big)\frac{1}{2}$$

schreiben. Das ist eine konvexe Kombination der bisherigen Erfolgsquote und der a priori Wahrscheinlichkeit $\frac{1}{2}$. Je mehr Erfahrung wir haben, desto stärker wird die bisherige Erfolgsquote gewichtet. In der Statistik heisst diese Methode **Bayes'sche Statistik**, in der Versicherungsmathematik heisst sie **Kredibilität**.

Aus den Wahrscheinlichkeiten können wir sehen, dass man aus dem bisherigen Verlauf der Experimente lernt. Laplace hat die Methode angewandt, um die Wahrscheinlichkeit zu berechnen, dass morgen die Sonne wieder aufgeht. Aus der Bibel hat er die Anzahl Tage seit der Erschaffung der Erde berechnet. Er hat dann als Wahrscheinlichkeit, dass morgen die Sonne wieder aufgeht ($k = n$), $(n+1)/(n+2)$ bekommen. Heute (5.11.2019) ist somit die Wahrscheinlichkeit, dass die Sonne morgen nicht aufgeht $4.54648 \cdot 10^{-7}$.

Die Wahrscheinlichkeit $(k+1)/(n+2)$ können wir auch in einem Urnenmodell bekommen, wenn in der Urne $k+1$ rote Kugeln und $n-k+1$ schwarze Kugeln liegen. Wir starten also mit einer Urne mit je einer roten und einer schwarzen Kugel. Jedes Mal, wenn wir eine Kugel ziehen, legen wir zwei der gezogenen Farbe zurück. Dies ergibt das gleiche Modell,

$$\begin{aligned} \mathbb{P}[X_1 = R] &= \frac{1}{2}\,, \\ \mathbb{P}[X_{n+1} = R \mid X_1 = x_1, \dots, X_n = x_n] &= \frac{k+1}{n+2}\,, \end{aligned}$$

wobei k die Anzahl Experimente bezeichnet, bei denen eine rote Kugel gezogen wurde. Aus Satz 1.25 folgt, dass dies wirklich das gleiche Modell ist.

1.6.3. Von der Binomial- zur Poisson-Verteilung

Oft hat man Situationen, in denen N sehr gross, aber p sehr klein ist. Zum Beispiel:

- Bei einer Lebensversicherung sind viele Leute versichert. Dass der Einzelne stirbt ist eher unwahrscheinlich.
- In einem radioaktiven Isotop hat es sehr viele Atome. Dass ein einzelnes Atom zerfällt, ist aber unwahrscheinlich.

- In den Ställen der preussischen Armee arbeiten viele Leute. Dass ein Einzelner durch Hufschlag stirbt, ist unwahrscheinlich. Bortkiewics ermittelte aus Daten von 10 Kavallerieregimentern aus 20 Jahren, dass pro Regiment und Jahr im Durchschnitt 0.61 Personen durch Hufschlag sterben. Ein Vergleich der unten eingeführten Approximation mit den Daten ergibt

Todesfälle	Anzahl Jahre	Häufigkeit	Approximation
0	109	0.545	0.543
1	65	0.325	0.331
2	22	0.110	0.101
3	3	0.015	0.021
4	1	0.005	0.003

Wir interessieren uns nun für die Verteilung der Anzahl Erfolge. Die exakte Verteilung ist die Binomialverteilung. Falls N gross ist, ist die Verteilung aber aufwändig zu berechnen.

Wir wählen nun N und setzen $p = \lambda/N$. Dann ist der Erwartungswert der Binomialverteilung konstant λ. Wir lassen nun N gegen Unendlich gehen, also geht p gegen Null. Wir erhalten dann

$$\begin{aligned}\mathbb{P}[S_N = k] &= \binom{N}{k}\Big(\frac{\lambda}{N}\Big)^k\Big(1-\frac{\lambda}{N}\Big)^{N-k}\\ &= \frac{\lambda^k}{k!}\frac{N(N-1)\cdots(N-k+1)}{N^k}\Big(1-\frac{\lambda}{N}\Big)^{N-k} \to \frac{\lambda^k}{k!}1\mathrm{e}^{-\lambda}\,.\end{aligned}$$

Der Grenzwert ist wirklich eine Wahrscheinlichkeitsverteilung, und heisst **Poisson-Verteilung** mit Parameter λ.

Wir hatten im Binomialmodell, dass $\lambda = N\lambda/N$ die erwartete Anzahl Erfolge ist. Dies gilt auch im Grenzmodell

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

Beispiel 1.29. (Bell Systems, 1926) In $N = 267$ Serien von je 515 Anrufen wurde die Anzahl der Fehlverbindungen beobachtet. Die folgende Tabelle zeigt die Anzahl Serien mit k Fehlverbindungen

k	≤ 2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	≥ 17
N_k	1	5	11	14	22	43	31	40	35	20	18	12	7	6	2	0
Poi	2	5	10	18	26	33	36	35	31	24	18	12	7	4	2	2

Insgesamt gab es 2334 Fehlverbindungen, pro Serie also $\lambda = 8.74$. Benutzen wir ein Poisson-Modell, dann ist die relative Häufigkeit von k Fehlverbindungen $\mathrm{e}^{-\lambda}\lambda^k/k!$. Der theoretische Wert für die Häufigkeit ist dann also $N\mathrm{e}^{-\lambda}\lambda^k/k!$. Wie wir aus der Tabelle sehen können, weicht dieser Wert nicht stark von den beobachteten Häufigkeiten ab. Wenn auch dieser Wert nicht genau mit der Wirklichkeit übereinstimmt, bietet er doch der Telefongesellschaft eine Möglichkeit, relativ gute Vorhersagen zu machen. ■

Definition 1.30. *Eine Serie von Zufallsvariablen $\{X_i : i \in I\}$ heisst* **unabhängig**, *falls die Ereignisse $\{\{X_i = c_i\} : i \in I\}$ für alle c_i unabhängig sind.*

Die Poisson-Verteilung hat folgende Eigenschaft.

Hilfssatz 1.31. *Die Zufallsvariablen X_1 und X_2 seien unabhängig und Poisson-verteilt mit Parametern λ_1 und λ_2, respektive. Sei $X = X_1 + X_2$ und $\lambda = \lambda_1 + \lambda_2$.*

i) *Die Variable X ist Poisson-verteilt mit Parameter λ.*

ii) *Die bedingte Verteilung von X_1 bezüglich des Ereignisses $\{X = n\}$ ist die Binomialverteilung mit Parametern $p = \lambda_1/\lambda$ und n.*

Beweis. i) Wir haben

$$\begin{aligned}\mathbb{P}[X = n] &= \sum_{k=0}^{n} \mathbb{P}[X_1 = k, X_2 = n-k] = \sum_{k=0}^{n} \frac{\lambda_1^k}{k!}\mathrm{e}^{-\lambda_1}\frac{\lambda_2^{n-k}}{(n-k)!}\mathrm{e}^{-\lambda_2} \\ &= \frac{\lambda^n}{n!}\mathrm{e}^{-\lambda}\sum_{k=0}^{n}\binom{n}{k}\Big(\frac{\lambda_1}{\lambda}\Big)^k\Big(\frac{\lambda_2}{\lambda}\Big)^{n-k} = \frac{\lambda^n}{n!}\mathrm{e}^{-\lambda}\Big(\frac{\lambda_1}{\lambda}+\frac{\lambda_2}{\lambda}\Big)^n = \frac{\lambda^n}{n!}\mathrm{e}^{-\lambda}\,.\end{aligned}$$

ii) Die Definition der bedingten Verteilung ist für $k \le n$

$$\begin{aligned}\mathbb{P}[X_1 = k \mid X = n] &= \frac{\mathbb{P}[X_1 = k, X_2 = n-k]}{\mathbb{P}[X = n]} = \frac{n!}{\lambda^n}\mathrm{e}^{\lambda}\frac{\lambda_1^k}{k!}\mathrm{e}^{-\lambda_1}\frac{\lambda_2^{n-k}}{(n-k)!}\mathrm{e}^{-\lambda_2} \\ &= \binom{n}{k}p^k(1-p)^{n-k}\,.\end{aligned}$$

□

In einem gewissen Sinne kann man die Aussagen auch umkehren. Sei X Poisson-veteilt mit Parameter λ und seien $\{I_k : k \in \mathbb{N}\}$ unabhängige Variablen (auch unabhängig von X) mit $\mathbb{P}[I_k = 1] = 1 - \mathbb{P}[I_k = 0] = p \in (0, 1)$. Wir bilden nun

$$X_1 = \sum_{k=1}^{X} I_k \ , \qquad\qquad X_2 = \sum_{k=1}^{X} (1 - I_k) \ .$$

Dann sind X_1 und X_2 unabhängige Poisson-verteilte Variablen mit Parameter $\lambda_1 = \lambda p$ und $\lambda_2 = \lambda(1 - p)$. In der Tat,

$$\mathbb{P}[X_1 = x_1, X_2 = x_2] = \frac{\lambda^{x_1+x_2}}{(x_1 + x_2)!} \mathrm{e}^{-\lambda} \binom{x_1 + x_2}{x_1} p^{x_1} (1 - p)^{x_2} = \frac{\lambda_1^{x_1}}{x_1!} \mathrm{e}^{-\lambda_1} \frac{\lambda_2^{x_2}}{x_2!} \mathrm{e}^{-\lambda_2} \ .$$

Man bemerke, dass dieses Beispiel einen Raum mit überabzählbar vielen Elementarereignissen benötigt. Die gesuchte Wahrscheinlichkeit hingegen passt in unser Modell, wenn wir nur endlich viele (z.B. $X_1 + X_2$) $\{I_k\}$ in unserem Raum zulassen.

2. Stetige Wahrscheinlichkeitsräume

Der Rahmen eines diskreten Wahrscheinlichkeitsraumes ist oft zu klein. Wir haben zum Beispiel bei der Irrfahrt jeweils den Grenzwert $N \to \infty$ betrachtet. Es wäre besser, einen unendlichen Zeithorizont $N = \infty$ zu wählen, aber der Raum $\{x_1, x_2, \ldots\}$ ist überabzählbar. Wir möchten aber die Wahrscheinlichkeiten auf dem grösseren Raum kennen, wie zum Beispiel $\mathbb{P}[\lim_{N\to\infty} N^{-1}S_N \leq a]$, da wir nicht immer sicher sein können, dass $N \to \infty$ auch die Wahrscheinlichkeit liefert, die wir intuitiv erwarten. Weiter zeigt es sich, dass es nicht möglich ist, die Wahrscheinlichkeitsverteilung "sinnvoll" auf alle Teilmengen von $\{-1, 1\}^{\mathbb{N}}$ zu erweitern, so dass die Verteilung auf Ereignisse in endlicher Zeit mit den in Abschnitt 1.4 verwendeten übereinstimmt.

Die Statistik hat oft Fragen zu klären, ob gewisse Annahmen sinnvoll sind oder nicht, oder welche Parameter am "wahrscheinlichsten" sind. Ist die Anzahl der Daten gross, wird der Rechenaufwand zu gross. Man betrachtet daher die Verteilung einer Test-Statistik, die so normiert ist, dass die endliche Statistik gegen eine bestimmte Wahrscheinlichkeitsverteilung konvergiert, falls die Datenanzahl gegen unendlich konvergiert. Die Grenzverteilung ist dann normalerweise eine Verteilung auf $\mathbb{R}$, das auch überabzählbar ist. Wir müssen daher unseren Begriff des Wahrscheinlichkeitsraumes verallgemeinern.

2.1. Allgemeine Wahrscheinlichkeitsräume

2.1.1. Die Axiome von Kolmogorov

Wir erlauben nun, dass $\Omega \neq \emptyset$ eine beliebige nichtleere Menge ist. Zuerst müssen wir bestimmen, welche Ereignisse wir zulassen. Es gibt nämlich Situationen, siehe Beispiel 2.3, in denen man keine geeignete Wahrscheinlichkeitsverteilung auf allen Teilmengen von Ω definieren kann. Weiter ist es manchmal nicht wünschenswert, den Raum der Ereignisse zu gross zu wählen. Wir definieren zuerst die Eigenschaften, die die Klasse der zulässigen Ereignisse haben soll.

Definition 2.1. *Sei $\mathfrak{A}$ eine Klasse von Teilmengen von Ω. $\mathfrak{A}$ heisst* **σ-Algebra**, *falls*

i) $\emptyset \in \mathfrak{A}$ *(das heisst, $\mathfrak{A}$ kann nicht leer sein).*

ii) *Ist $A \in \mathfrak{A}$, dann ist auch $A^c \in \mathfrak{A}$.*

iii) *Sind* $A_1, A_2, \ldots \in \mathfrak{A}$*, dann ist auch* $\cup_n A_n \in \mathfrak{A}$.

Wir verlangen also, dass die Kollektion $\mathfrak{A}$ unter abzählbaren Mengenoperationen abgeschlossen ist. Die kleinste mögliche σ-Algebra ist $\mathfrak{A} = \{\emptyset, \Omega\}$. Wir bemerken, dass für $A_1, A_2, \ldots \in \mathfrak{A}$ auch $A_i^c \in \mathfrak{A}$. Damit ist $\cup_i A_i^c \in \mathfrak{A}$, und somit

$$\bigcap_i A_i = \Big(\bigcup_i A_i^c\Big)^c \in \mathfrak{A}\,.$$

Wir wählen nun eine σ-Algebra $\mathcal{F}$ von zulässigen Ereignissen. Wir sagen $(\Omega, \mathcal{F})$ ist ein **messbarer Raum**.

Definition 2.2. *Sei* $(\Omega, \mathcal{F})$ *ein messbarer Raum und* $\mathbb{P} : \mathcal{F} \to [0,1]$ *eine Funktion auf* $\mathcal{F}$*. Die Funktion* $\mathbb{P}$ *heisst* **Wahrscheinlichkeitsmass** *auf* $(\Omega, \mathcal{F})$*, falls*

i) $\mathbb{P}[\Omega] = 1$ *(das Mass ist normiert).*

ii) *Seien* $A_1, A_2, \ldots \in \mathcal{F}$*, so dass* $A_i \cap A_j = \emptyset$ *für alle* $i \neq j$*, dann gilt*

$$\mathbb{P}[\cup_i A_i] = \sum_i \mathbb{P}[A_i]$$

(das Mass ist σ*-additiv).*

Ist $(\Omega, \mathcal{F})$ ein messbarer Raum und $\mathbb{P}$ ein Wahrscheinlichkeitsmass auf $(\Omega, \mathcal{F})$, dann nennen wir $(\Omega, \mathcal{F}, \mathbb{P})$ einen **Wahrscheinlichkeitsraum**. Wir bemerken, dass unsere Definition in diskreten Räumen einen Wahrscheinlichkeitsraum ergibt.

Wählen wir $A_1 = \Omega$ und $A_k = \emptyset$ für $k \geq 2$, so erhalten wir

$$1 = \mathbb{P}[\Omega] = \mathbb{P}[\Omega] + \sum_{k=2}^{\infty} \mathbb{P}[\emptyset] = 1 + \sum_{k=2}^{\infty} \mathbb{P}[\emptyset]\,.$$

Es folgt, dass $\mathbb{P}[\emptyset] = 0$. Insbesonder gilt ii) auch für eine endliche Anzahl von Mengen.

Ist $\mathfrak{A}_0$ eine Kollektion von Teilmengen von Ω, dann ist

$$\mathfrak{A} = \sigma(\mathfrak{A}_0) := \bigcap_{\substack{\mathfrak{B}\,\sigma\text{-Algebra} \\ \mathfrak{A}_0 \subset \mathfrak{B}}} \mathfrak{B}$$

eine σ-Algebra. Es ist somit die kleinste σ-Algebra, die $\mathfrak{A}_0$ enthält.

Beispiel 2.3. Betrachten wir $\Omega = [0,1]$. Wir konstruieren $\mathcal{F}$ so, dass alle abgeschlossenen Intervalle $[a,b]$ für $0 \leq a < b \leq 1$ in $\mathcal{F}$ sind. Es folgt dann, dass alle Intervalle (links/rechts offen/abgeschlossen) in $\mathcal{F}$ sind. Wir definieren dann $\mathcal{F}$ als die kleinste σ-Algebra, die alle $[a,b]$ enthält. Diese σ-Algebra heisst **Borel-σ-Algebra**. Die Borel-σ-Algebra existiert. Insbesondere sind die Mengen $\{\omega\} \in \mathcal{F}$, für alle $\omega \in \Omega = [0,1]$. Es gibt aber Teilmengen von $[0,1]$, die nicht in $\mathcal{F}$ sind.

Wir definieren nun die Wahrscheinlichkeitsfunktion mit der Eigenschaft, dass $\mathbb{P}[[a,b]] = b - a$. Dieses Mass heisst **Lebesgue-Mass** auf $[0,1]$, und existiert. Man kann zeigen, dass es kein Wahrscheinlichkeitsmass auf der Menge aller Teilmengen von $[0,1]$ gibt, das mit dem Lebesgue-Mass verträglich ist; das heisst, für das $\mathbb{P}[[a,b]] = b-a$ für alle $a < b$ gilt. Es folgt, dass $\mathbb{P}[\{\omega\}] = 0$ für alle ω. Insbesondere hat jede abzählbare Menge A die Wahrscheinlichkeit

$$\mathbb{P}[A] = \mathbb{P}[\cup_{\omega \in A} \{\omega\}] = \sum_{\omega \in A} \mathbb{P}[\{\omega\}] = 0 \, .$$

Um ein Beispiel zu konstruieren, das zeigt, dass das Lebeguesmass sich nicht auf allen Teilmengen von $\Omega = [0,1)$ konstruieren kann, betrachten wir folgendes Beispiel. Wir sagen $x \sim y$, falls $x - y \in \mathbb{Q}$. Dies ist eine Äquivalenzrelation. Aus dem Auswahlsaxiom folgt, dass wir aus jeder Äquivalenzklasse genau ein Element wählen können. Nennen wir diese Menge A. Für $q \in \mathbb{Q} \cap [0,1)$ können wir $A_q = \{a + q - \lfloor a + q \rfloor : a \in A\}$ bilden, wobei $\lfloor x \rfloor$ den Ganzzahlteil von x bezeichnet. Wir haben dann $[0,1) = \cup_q A_q$. Es ist klar, dass $A_q \cap A_r = \emptyset$ für $q \neq r$. Aus der Symmetrie folgt $1 = \mathbb{P}[\Omega] = \sum_q \mathbb{P}[A_q] = \infty \mathbb{P}[A]$. Somit müsste $\mathbb{P}[A] = 0$ gelten. Dann wäre aber auch $\mathbb{P}[\Omega] = \sum_q 0 = 0$. Somit kann A keine messbare Menge sein. ■

2.1.2. Einfache Folgerungen

Hilfssatz 2.4. *Die Aussagen von Hilfssatz 1.3 gelten auch für allgemeine Wahrscheinlichkeitsräume.* □

Korollar 2.5. *Für $A_1, A_2, \ldots \in \mathcal{F}$ gilt*

$$\mathbb{P}[\cup_i A_i] \leq \sum_i \mathbb{P}[A_i] \, .$$

Beweis. Dies folgt aus

$$\mathbb{P}[\cup_i A_i] = \lim_{n\to\infty} \mathbb{P}[\cup_{i=1}^n A_i] \leq \lim_{n\to\infty} \sum_{i=1}^n \mathbb{P}[A_i] = \sum_i \mathbb{P}[A_i] \; .$$

□

Ein Ereignis von besonderer Bedeutung ist

$A_\infty = \cap_{n\in\mathbb{N}} \cup_{k\geq n} A_k$ *unendlich viele der Ereignisse* A_k *treten ein.*

Also, für jedes n gibt es ein $k \geq n$, so dass A_k eintritt.

Wir definieren **Unabhängigkeit** von Ereignissen wie im diskreten Fall,

$$\forall J \subset I \text{ (endlich)} \quad \Longrightarrow \quad \mathbb{P}[\cap_{j\in J} A_j] = \prod_{j\in J} \mathbb{P}[A_j] \; .$$

Satz 2.6. (Lemma von Borel–Cantelli) *Es gelten folgende Aussagen:*

i) *Falls* $\sum_{i=1}^\infty \mathbb{P}[A_k] < \infty$, *dann gilt* $\mathbb{P}[A_\infty] = 0$, *das heisst, nur endlich viele der* A_k *treten ein.*

ii) *Sind* $\{A_i : i \in \mathbb{N}\}$ *unabhängig und* $\sum_{i=1}^\infty \mathbb{P}[A_i] = \infty$, *dann gilt* $\mathbb{P}[A_\infty] = 1$.

Bemerkung. Die Unabhängigkeit in ii) ist wichtig. Seien $\{X_i : i \in \mathbb{N}\}$ unabhängige Zufallsvariablen mit $\mathbb{P}[X_i = 0] = \mathbb{P}[X_i = 1] = \frac{1}{2}$, dann gilt für $A_i = \{X_0 = 1, X_i = 1\}$, dass $\sum_{i=1}^\infty \mathbb{P}[A_i] = \sum_{i=1}^\infty \frac{1}{4} = \infty$, aber $\mathbb{P}[A_\infty] \leq \mathbb{P}[X_0 = 1] = \frac{1}{2}$. ■

Beweis. i) Wir haben $\cup_{k\geq n+1} A_k \subset \cup_{k\geq n} A_k$. Daher gilt nach Hilfssatz 1.3 vii)

$$\mathbb{P}[A_\infty] = \lim_{n\to\infty} \mathbb{P}[\cup_{k\geq n} A_k] \leq \lim_{n\to\infty} \sum_{k=n}^\infty \mathbb{P}[A_k] = 0 \; .$$

ii) Es gilt

$$A_\infty^c = \cup_{n\in\mathbb{N}} \cap_{k\geq n} A_k^c \; ,$$

und somit nach Hilfssatz 1.3 vi)

$$\mathbb{P}[A_\infty^c] = \lim_{n\to\infty} \mathbb{P}[\cap_{k\geq n} A_k^c] \; .$$

Wir haben die Abschätzung

$$\mathbb{P}[\cap_{k\geq n} A_k^c] = \lim_{m\to\infty} \mathbb{P}[\cap_{k=n}^m A_k^c] = \lim_{m\to\infty} \prod_{k=n}^m (1 - \mathbb{P}[A_k])$$
$$\leq \lim_{m\to\infty} \exp\Big\{ - \sum_{k=n}^m \mathbb{P}[A_k] \Big\} = 0 \; .$$

Dies beweist die Behauptung. □

In der Bioinformatik gibt es 4 Aminobasen, die in DNA vorkommen. Die Basen haben die Namen A, C, G und T. Nehmen wir an, dass ein DNA eine zufällige Anordnung der vier Buchstaben $\{x_1, x_2, \ldots, x_N\}$ ist. Da N sehr gross ist, wählen wir $N = \infty$. In unserem Modell seien die Zufallsvariablen $\{X_i : i \geq 1\}$ unabhängig. Wir bezeichnen mit $p_A = \mathbb{P}[X_i = A]$, etc. die entsprechenden Wahrscheinlichkeiten. Wir nehmen an, dass alle Wahrscheinlichkeiten strikt positiv sind. Sei $n \in \mathbb{N}$ und $\{x_1, x_2, \ldots, x_n\}$ ein bestimmtes Wort, das aus den 4 Buchstaben gebildet werden kann. Dann gilt

Proposition 2.7. *Das Wort $\{x_1, x_2, \ldots, x_n\}$ taucht mit Wahrscheinlichkeit 1 unendlich oft im Text auf.*

Beweis. Betrachten wir die Ereignisse

$$A_k = \{X_{(k-1)n+1} = x_1, X_{(k-1)n+2} = x_2, \ldots, X_{kn} = x_n\} \; .$$

Diese Ereignisse sind unabhängig. Wir haben

$$\mathbb{P}[A_k] = \prod_{i=1}^n p_{x_i} > 0 \; .$$

Also ist $\sum_k \mathbb{P}[A_k] = \infty$. Das Borel–Cantelli-Lemma beweist die Behauptung. □

2.1.3. Transformation von Wahrscheinlichkeitsräumen

Manchmal kann man Resultat von einem Wahrscheinlichkeitsraum auf einen anderen übertragen. Sei $(\Omega', \mathcal{F}', \mathbb{P}')$ ein Wahrscheinlichkeitsraum, und $(\Omega, \mathcal{F})$ ein messbarer Raum.

Definition 2.8. *Wir sagen eine Abbildung $\varphi : \Omega' \to \Omega$ ist* **messbar**, *falls für alle $A \in \mathcal{F}$ gilt, dass*

$$\varphi^{-1}(A) := \{\omega' \in \Omega' : \varphi(\omega') \in A\} \in \mathcal{F}' .$$

Ist nun $\mathcal{F} = \sigma(\mathfrak{A}_0)$ für eine Kollektion $\mathfrak{A}_0$ von Teilmengen von Ω, dann genügt es die Eigenschaft für die Kollektion $\{\varphi^{-1}(A) : A \in \mathfrak{A}_0\}$ zu überprüfen.

Satz 2.9. *Ist $\varphi : \Omega' \to \Omega$ eine messbare Abbildung, dann ist durch*

$$\mathbb{P}[A] = \mathbb{P}' \circ \varphi^{-1}[A] = \mathbb{P}'[\varphi^{-1}(A)]$$

ein Wahrscheinlichkeitsmass auf $(\Omega, \mathcal{F})$ definiert.

Beweis. Wir haben

$$\mathbb{P}[\Omega] = \mathbb{P}'[\varphi^{-1}(\Omega)] = \mathbb{P}'[\Omega'] = 1 .$$

Ist nun $\{A_i\}$ eine Kollektion von Mengen mit $A_i \cap A_j = \emptyset$ für $i \neq j$, dann ist $\varphi^{-1}(A_i) \cap \varphi^{-1}(A_j) = \emptyset$. Weiter gilt

$$\mathbb{P}[\cup_i A_i] = \mathbb{P}'[\varphi^{-1}(\cup_i A_i)] = \mathbb{P}'[\cup_i \varphi^{-1}(A_i)] = \sum_i \mathbb{P}'[\varphi^{-1}(A_i)] = \sum_i \mathbb{P}[A_i] .$$

□

Sei $\Omega' = [0, 1]$, $\mathcal{F}'$ die Borel-σ-Algebra auf $[0, 1]$ und $\mathbb{P}'$ das Lebesguemass. Sei Ω die Menge aller binären $\{0, 1\}$ Folgen. Wir wählen die σ-Algebra $\mathcal{F}$, die durch die Ereignisse $\{X_i = 1\}$ erzeugt wird. Wir ordnen nun jeder Zahl $x \in [0, 1]$ die Folge $(x_1, x_2, \ldots)$ zu, für die $x = \sum_{k=1}^{\infty} x_k 2^{-k}$. Wir haben also

$$x_n = 0 \Longleftrightarrow x \in [2k2^{-n}, (2k + 1)2^{-n}) \text{ für ein } k \in \{0, 1, \ldots, 2^{n-1} - 1\}.$$

Das heisst,

$$\varphi^{-1}(\{X_n = 0\}) = \bigcup_{k=0}^{2^{n-1}-1} [2k2^{-n}, (2k + 1)2^{-n}) \in \mathcal{F}' .$$

Somit ist $\varphi(x)$ messbar. Setzen wir nun $\mathbb{P}$ als das Bild der Gleichverteilung, erhalten wir

$$\mathbb{P}[X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n] = \mathbb{P}'\Big[\Big[\sum_{k=1}^{n} x_k 2^{-k}, \sum_{k=1}^{n} x_k 2^{-k} + 2^{-n}\Big)\Big] = 2^{-n} .$$

Wir erhalten also die "Gleichverteilung" auf der Menge von unendlich vielen Würfen einer fairen Münze. Umgekehrt können wir aus der Existenz eines Wahrscheinlichkeitsmasses für unendlich viele Würfe einer fairen Münze die Existenz des Lebesguemasses beweisen.

2.2. Zufallsvariable und ihre Verteilungen

Bezeichnen wir mit $\mathfrak{B}^1 = \sigma(\{(-\infty, a] : a \in \mathbb{R}\})$ die Borel-σ-Algebra auf $\mathbb{R}$. Diese σ-Algebra enthält alle Intervalle, alle offenen und alle abgeschlossenen Mengen. Sei nun $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum.

Definition 2.10. *Eine (reelle)* **Zufallsvariable** *ist eine messbare Abbildung $X : \Omega \to \mathbb{R}$. Die Funktion $F_X(x) := \mathbb{P}[X \leq x]$ heisst* **Verteilungsfunktion** *der Zufallsvariable X.*

Durch die Verteilungsfunktion lassen sich alle Wahrscheinlichkeiten $\mathbb{P}[X \in A]$ mit $A \in \mathfrak{B}^1$ bestimmen, da die Ereignisse $\{(-\infty, b]\}$ die Borel-σ-Algebra erzeugen. Insbesondere haben wir $\mathbb{P}[X \in (a, b]] = F_X(b) - F_X(a)$.

Hilfssatz 2.11.

i) *Eine Verteilungsfunktion $F(x)$ hat die folgenden Eigenschaften*

   a) *$F(x)$ ist wachsend.*

   b) *$F(x)$ ist rechtsstetig.*

   c) *Es gilt $\lim_{x\to-\infty} F(x) = 0$ und $\lim_{x\to\infty} F(x) = 1$.*

ii) *Sei $F(x)$ eine Funktion, die die Eigenschaften* a) – c) *hat. Dann gibt es einen Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \mathbb{P})$ und eine Zufallsvariable X, die die Verteilungsfunktion $F(x)$ hat.*

Beweis. i) a) Sei $x \leq y$. Da $\{X \leq x\} \subset \{X \leq y\}$, gilt $F(x) = \mathbb{P}[X \leq x] \leq \mathbb{P}[X \leq y] = F(y)$.
b) Sei $\{h_n\}$ eine Folge von strikt positiven Zahlen, die gegen Null konvergiert. Wir setzen $\tilde{h}_n = \sup_{k\geq n} h_k \geq h_n$. Dann konvergiert auch $\tilde{h}_n$ nach Null. Weiter ist $\{X \leq x\} = \cap_n \{X \leq x + \tilde{h}_n\}$. Also haben wir

$$F(x) = \mathbb{P}[X \leq x] = \lim_{n\to\infty} \mathbb{P}[X \leq x + \tilde{h}_n] = \lim_{n\to\infty} F(x + \tilde{h}_n) \, .$$

Da $F(x) \leq F(x + h_n) \leq F(x + \tilde{h}_n)$ folgt auch $\lim_{n\to\infty} F(x + h_n) = F(x)$.
c) Sei $\{x_n\}$ eine Folge, die gegen $-\infty$ konvergiert. Wir setzen $y_n = \sup_{k\geq n} x_n$. Dann konvergiert $\{y_n\}$ monoton gegen $-\infty$. Weiter ist $\cap_n \{X \leq y_n\} = \emptyset$. Also erhalten wir

$$\overline{\lim_{n\to\infty}} F(x_n) \leq \lim_{n\to\infty} F(y_n) = \lim_{n\to\infty} \mathbb{P}[X \leq y_n] = \mathbb{P}[\emptyset] = 0 \, ,$$

also $\lim_{n\to\infty} F(x_n) = 0$. Analog folgt $\lim_{x\to\infty} F(x) = 1$.
ii) Sei $\Omega' = [0,1]$, $\mathcal{F}'$ die Borel-σ-Algebra auf Ω' und $\mathbb{P}$ die Gleichverteilung (Lebesguemass auf $[0,1]$). Da $F(x)$ wachsend ist, können wir die Umkehrabbildung

$$F^{-1}(\omega') = \inf\{x \in \mathbb{R} : F(x) > \omega'\}$$

definieren. Aus der Definition und der Rechtsstetigkeit schliessen wir

$$\{\omega' \in [0, F(x))\} \subset \{F^{-1}(\omega') \leq x\} \subset \{\omega' \in [0, F(x)]\} \ .$$

Somit ist $\{F^{-1}(\omega') \leq x\} = [0, F(x))$ oder $\{F^{-1}(\omega') \leq x\} = [0, F(x)]$. Das heisst, F^{-1} ist eine messbare Abbildung von $[0,1]$ nach $\mathbb{R}$, also eine Zufallsvariable. Die Verteilungsfunktion ist

$$F(x) = \mathbb{P}[\omega' \in [0, F(x))] \leq \mathbb{P}[F^{-1}(\omega') \leq x] \leq \mathbb{P}[\omega' \in [0, F(x)]] = F(x) \ .$$

□

2.11 Die obige Beweismethode hat auch eine praktische Anwendung. Auf einem Computer lassen sich *Pseudo-Zufallszahlen* $\{U_n\}$ erzeugen. Diese Zufallszahlen nähern die Gleichverteilung auf $[0,1]$ an. Wollen wir nun Zufallsvariablen $\{X_n\}$ mit der Verteilungsfunktion $F(x)$ erzeugen, so können wir $X_n = F^{-1}(U_n)$ setzen.

Beispiele

- Sei X ein $\{0,1\}$ Experiment mit Erfolgsparameter p. Dann ist

$$F(x) = \begin{cases} 0, & \text{falls } x < 0, \\ 1-p, & \text{falls } 0 \leq x < 1, \\ 1, & \text{falls } x \geq 1. \end{cases}$$

- Sei X eine Binomialverteilte Zufallsvariable mit Parameter n und p. Dann haben wir

$$F(x) = \sum_{k=0}^{\lfloor x \rfloor} \binom{n}{k} p^k (1-p)^{n-k} \ .$$

  Hier verwenden wir die Konvention, dass $\binom{n}{k} = 0$, falls $n < k$.

- Sei X Poissonverteilt mit Parameter λ. Dann haben wir

$$F(x) = \sum_{k=0}^{\lfloor x \rfloor} \frac{\lambda^k}{k!} \mathrm{e}^{-\lambda} \ .$$

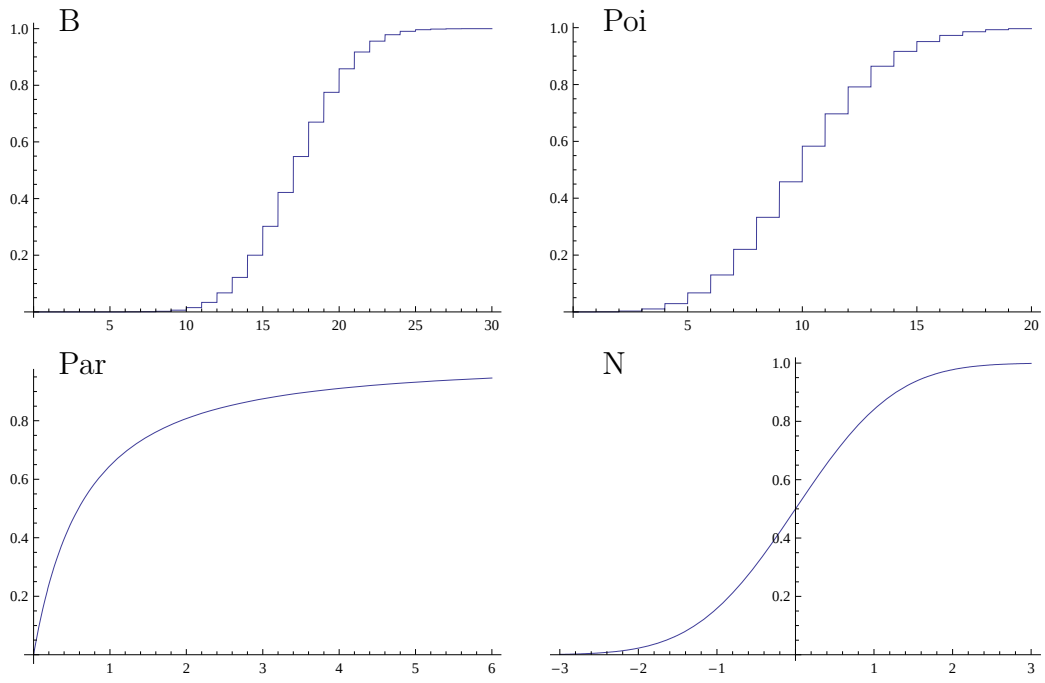

Abbildung 2.1: *Die Verteilungsfunktionen der Binomialverteilung (B), Poissonverteilung (Poi), Paretoverteilung (Pa) und Normalverteilung (N)*

- Die Funktion $(1-(1+x/\beta)^{-\alpha})\mathbb{1}_{x>0}$ mit $\alpha, \beta > 0$ ist eine Verteilungsfunktion, und heisst **Pareto-Verteilung** mit Parameter α und β. Diese Verteilung ist populär in der Versicherungsmathematik, und wird zum Beispiel zur Modellierung von Katastrophenschäden verwendet. Sie hat die folgende Eigenschaft. Nehmen wir an, wir wissen, dass $\{X > x_0\}$. Dann hat $X - x_0$ die bedingte Verteilung

$$\begin{aligned}\mathbb{P}[X - x_0 \le y \mid X > x_0] &= \frac{\mathbb{P}[x_0 < X \le x_0 + y]}{\mathbb{P}[X > x_0]} = \frac{F(x_0+y) - F(x_0)}{1 - F(x_0)} \\ &= \frac{(1 + x_0/\beta)^{-\alpha} - (1 + (x_0+y)/\beta)^{-\alpha}}{(1 + x_0/\beta)^{-\alpha}} \\ &= 1 - \Big(\frac{\beta + x_0 + y}{\beta + x_0}\Big)^{-\alpha} = 1 - \Big(1 + \frac{y}{\beta + x_0}\Big)^{-\alpha},\end{aligned}$$

  wobei $y \ge 0$. Also erhalten wir wieder eine Pareto-Verteilung.

- Die Funktion $F(x) = (1 - \mathrm{e}^{-\alpha x})\mathbb{1}_{x>0}$ mit $\alpha > 0$ ist eine Verteilungsfunktion und heisst **Exponentialverteilung** mit Parameter α. Wissen wir, dass $\{X > x_0\}$, dann hat $X - x_0$ die Verteilung (für $y \ge 0$)

$$\mathbb{P}[X - x_0 \le y \mid X > x_0] = \frac{F(x_0+y) - F(x_0)}{1 - F(x_0)} = \frac{\mathrm{e}^{-\alpha x_0} - \mathrm{e}^{-\alpha(x_0+y)}}{\mathrm{e}^{-\alpha x_0}} = 1 - \mathrm{e}^{-\alpha y}.$$

Die Exponentialverteilung hat somit, analog zur geometrischen Verteilung im diskreten Fall, kein Gedächtnis.

- Die Funktion
$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-y^2/2}\, \mathrm{d}y$$
ist eine Verteilungsfunktion. Sie heisst **standard Normalverteilung**. Betrachten wir nun die Variable $\mu + \sigma X$ mit $\sigma > 0$. Die hat die Verteilung
$$\begin{aligned}\mathbb{P}[\mu + \sigma X \le x] &= \mathbb{P}\Big[X \le \frac{x-\mu}{\sigma}\Big] = \int_{-\infty}^{(x-\mu)/\sigma} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-y^2/2}\, \mathrm{d}y \\ &= \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-(z-\mu)^2/(2\sigma^2)} \frac{\mathrm{d}z}{\sigma} = \int_{-\infty}^{x} \frac{1}{\sqrt{2\sigma^2\pi}} \mathrm{e}^{-(z-\mu)^2/(2\sigma^2)}\, \mathrm{d}z\ .\end{aligned}$$
Diese Verteilung heisst Normalverteilung mit Mittelwert μ und Varianz σ^2.

Wir sehen in den Beispielen zwei grundsätzlich verschiedene Typen von Verteilungen. Die diskreten Verteilungen haben Verteilungsfunktionen, die stückweise konstant sind. Zwischen diesen konstanten Teilen gibt es Sprünge. Wir haben dann eine abzählbare Menge von Punkten $\{x_1, x_2, \ldots\}$, an denen die Verteilungsfunktion einen Sprung der Höhe p_k hat. Wir können die Verteilung durch die Paare $\{(x_k, p_k)\}$ charakterisieren. Bei diesem Typ haben wir, dass die ganze Masse auf abzählbar viele Punkte verteilt ist.

Der zweite Typ hat eine stetige Verteilungsfunktion. Die betrachteten Verteilungsfunktionen liessen sich alle, wie in der folgenden Definition schreiben.

Definition 2.12. *Eine Verteilung heisst* **absolutstetig**, *falls sich die Verteilungsfunktion als* $F(x) = \int_{-\infty}^{x} f(z)\, \mathrm{d}z$ *schreiben lässt, wobei* $f(z)$ *eine messbare Funktion ist. Die Funktion* $f(x)$ *heisst* **Dichtefunktion** *der Verteilung.*

Es ist einfach zu zeigen, dass jede (stückweise) stetige Funktion messbar ist, und somit kann jede positive stückweise stetige Funktion, deren Integral über die reellen Zahlen 1 ergibt, als Dichte benutzt werden.

Für die Pareto-Verteilung erhalten wir durch Differenzierung die Dichtefunktion $f(x) = \alpha\beta^\alpha(\beta + x)^{-\alpha-1}\mathbb{1}_{x>0}$. Die Exponentialverteilung hat die Dichtefunktion $f(x) = \alpha \mathrm{e}^{-\alpha x}\mathbb{1}_{x>0}$, und die Normalverteilung hat die Dichtefunktion $f(x) = \mathrm{e}^{-(x-\mu)^2/(2\sigma^2)}/\sqrt{2\sigma^2\pi}$. Die Gleichverteilung auf $[0,1]$ hat die Dichte $f(x) = \mathbb{1}_{0<x<1}$. Generell kann man die Gleichverteilung auf $[a,b]$ mit $a < b$ definieren. Die entsprechende Dichte ist $f(x) = (b-a)^{-1}\mathbb{1}_{a<x<b}$.

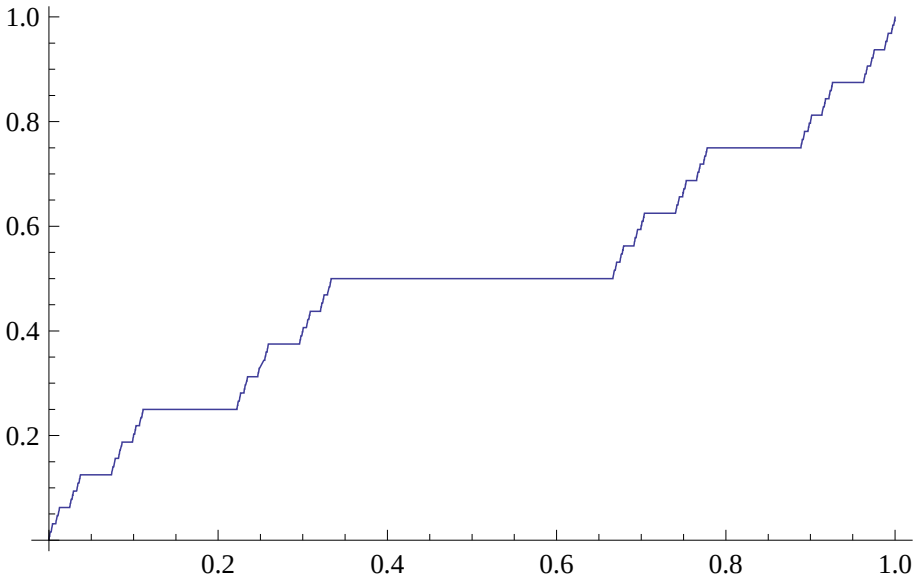

Abbildung 2.2: *Singuläre Verteilungsfunktion*

Neben den diskreten und den absolutstetigen Verteilungen gibt es noch einen dritten Typ, die **singulären** Verteilungen. Diese Verteilungen sind stetig, aber haben an allen Stellen, an denen $F(x)$ differenzierbar ist, die Ableitung 0. Sie lassen sich somit nicht mit einer Dichtefunktion schreiben. Wir werden diese Verteilungen nicht weiter betrachten, da sie für praktische Anwendungen nicht benützt werden. Wir können nämlich jede Verteilung durch eine absolutstetige Verteilung approximieren. Da man aus Daten nicht ersehen kann, ob eine Verteilung absolutstetig oder singulär ist, genügt es diskrete, absolutstetige und Mischungen dieser beiden Typen zu betrachten. Wir geben aber ein Beispiel für eine singuläre Verteilung.

Sei $x \in [0,1]$. Dann können wir jede Zahl im Dreiersystem darstellen, das heisst, wir schreiben $x = \sum_{k=1}^{\infty} x_k(x) 3^{-k}$ mit $x_k(x) \in \{0,1,2\}$. Sei $T(x) = \inf\{k : x_k(x) = 1\}$. Wir definieren nun die Abbildung

$$F(x) = \sum_{k=1}^{T(x)} \mathbb{1}_{x_k(x)\geq 1} 2^{-k} \ .$$

Das heisst, wir setzen $F(x) = \frac{1}{2}$ auf $[\frac{1}{3}, \frac{2}{3}]$, $F(x) = \frac{1}{4}$ auf $[\frac{1}{9}, \frac{2}{9}]$ und $F(x) = \frac{3}{4}$ auf $[\frac{7}{9}, \frac{8}{9}]$. Auf diese Art unterteilen wir die verbleibenden Intervalle in drei Teile, und setzen die Funktion im mittleren Teil auf den Mittelwert zwischen dem linken und dem rechten Rand des Intervalls. Diese Abbildung ist steigend und stetig. Man kann zeigen, dass die Funktion nicht absolutstetig sein kann. Der Graph der Verteilungsfunktion ist in Abbildung 2.2 gegeben.

Generell ist eine Verteilungsfunktion eine Mischung aus den oben beschriebenen drei Typen

$$F(x) = \alpha F_1(x) + \beta F_2(x) + (1-\alpha-\beta)F_3(x) \ ,$$

wobei $\alpha, \beta \geq 0$ und $\alpha + \beta \leq 1$, $F_1(x)$ ist eine diskrete Verteilungsfunktion, $F_2(x)$ ist eine absolutstetige Verteilungsfunktion und $F_3(x)$ ist eine singuläre Verteilungsfunktion.

2.3. Erwartungswerte

Sei X eine Zufallsvariable auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \mathbb{P})$. Für viele Probleme ist es wichtig zu wissen, welchen Wert man von X im Mittel erwarten kann. Im diskreten Fall hat der Erwartungswert diese Funktion. Wir wollen nun den Erwartungswert auf die stetigen Modelle verallgemeinern.

Für eine diskrete Zufallsvariable ist es natürlich, den Erwartungswert wie im diskreten Modell zu berechnen

$$\mathbb{E}[X] = \sum_k p_k x_k \ ,$$

sofern die rechte Seite wohldefiniert ist. Für den stetigen Fall diskretisieren wir die Verteilungsfunktion, und nehmen zuerst an, dass die Zufallsvariable $|X| \leq c$ beschränkt ist,

$$\sum_{k=-n+1}^{n} \frac{kc}{n} \mathbb{P}[(k-1)c/n < X \leq kc/n] = \sum_{k=-n}^{n} \frac{kc}{n}(F(kc/n) - F((k-1)c/n)) \ .$$

Bilden wir den Grenzwert $n \to \infty$, so konvergiert der Ausdruck. Den Grenzwert bezeichnen wir als Erwartungswert.

Ist X unbeschränkt und $X \geq 0$, so können wir den Erwartungswert $\mathbb{E}[\min\{X, n\}]$ bilden. Dieser Erwartungswert ist wachsend in n, und somit existiert ein Grenzwert in $[0, \infty]$. Wir nennen diesen Wert dann Erwartungswert. Für beliebiges X teilen wir X in Positivteil $X^+ = \max\{X, 0\}$ und Negativteil $X^- = \max\{-X, 0\}$ auf. Wir haben dann $X = X^+ - X^-$ und $|X| = X^+ + X^-$. Gilt $\mathbb{E}[X^+] < \infty$ oder $\mathbb{E}[X^-] < \infty$, so definieren wir $\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-]$. Ist $\mathbb{E}[X^+] = \mathbb{E}[X^-] = \infty$, so lässt sich kein sinnvoller Erwartungswert definieren.

Ist die Zufallsvariable absolutstetig, dann erhalten wir die Formel

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

vorausgesetzt, dass die rechte Seite wohldefiniert ist. Ist $F(x) = \alpha F_1(x)+(1-\alpha)F_2(x)$ mit $\alpha \in (0,1)$ mit $F_1(x)$ einer diskreten Verteilungsfunktion gegeben durch $\{(x_k, p_k)\}$ und $F_2(x)$ einer absolutstetigen Verteilungsfunktion mit Dichtefunktion $f(x)$, dann ist

$$\mathbb{E}[X] = \alpha \sum_k p_k x_k + (1-\alpha) \int_{-\infty}^{\infty} x f(x)\, \mathrm{d}x \; .$$

Damit wir nicht zwischen diskreten und stetigen Variablen unterscheiden müssen, schreiben wir für den Erwartungswert

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \, \mathrm{d}F(x) \; .$$

Definition 2.13. *Eine Zufallsvariable heisst* **integrierbar**, *falls* $\mathbb{E}[|X|] < \infty$.

Der Erwartungswert hat folgende Eigenschaften:

- *Linearität*, das heisst
$$\mathbb{E}\Big[\sum_{k=1}^{n} c_i X_i\Big] = \sum_{k=1}^{n} c_i \mathbb{E}[X_i] \; .$$
- *Monotonie*, das heisst $\mathbb{E}[X] \leq \mathbb{E}[Y]$, falls $X \leq Y$.
- *Monotone Stetigkeit*, das heisst, falls $X_1 \leq X_2 \leq \cdots$ mit $\mathbb{E}[|X_1|] < \infty$, so gilt für $X = \lim_{n\to\infty} X_n$
$$\mathbb{E}[X] = \lim_{n\to\infty} \mathbb{E}[X_n] \; .$$
- *Beschränkte Konvergenz*, das heisst, sind $\{X_n\}$ Zufallsvariablen, so dass der Erwartungswert $\mathbb{E}[\sup_n |X_n|] < \infty$ endlich ist und $X = \lim_{n\to\infty} X_n$ existiert, dann gilt
$$\mathbb{E}[X] = \lim_{n\to\infty} \mathbb{E}[X_n] \; .$$

Ist X eine Zufallsvariable mit Verteilungsfunktion $F(x)$, so ist für jede messbare Funktion $h : \mathbb{R} \to \mathbb{R}$ die Grösse $h(X)$ auch eine Zufallsvariable. Der Erwartungswert lässt sich dann berechnen als

$$\mathbb{E}[h(X)] = \int_{-\infty}^{\infty} h(x)\, \mathrm{d}F(x) = \begin{cases} \sum_k h(x_k) p_k \; , & \text{falls } F(x) \text{ diskret ist,} \\ \int_{-\infty}^{\infty} h(x) f(x)\, \mathrm{d}x \; , & \text{falls } F(x) \text{ absolutstetig ist.} \end{cases}$$

Spezialfälle In den folgenden Spezialfällen wird angenommen, dass die entsprechenden Erwartungswerte existieren.

- Für den positiven Teil $X^+ = \max\{X, 0\}$ erhalten wir

$$\mathbb{E}[X^+] = \int_0^\infty x \, \mathrm{d}F(x) = \begin{cases} \sum_{k:x_k>0} x_k p_k \ , & \text{falls } F(x) \text{ diskret ist,} \\ \int_0^\infty x f(x) \, \mathrm{d}x \ , & \text{falls } F(x) \text{ absolutstetig ist.} \end{cases}$$

  Eine analoge Formel gilt für $X^- = \max\{-X, 0\}$. Aus diesen beiden Formeln erhalten wir $\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-]$ und $\mathbb{E}[|X|] = \mathbb{E}[X^+] + \mathbb{E}[X^-]$.

- Das p-te Moment von X ist definiert als

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

  wobei $p \in \mathbb{N}$. Ist $X \geq 0$, so kann das p-te Moment auch für $p \in [0, \infty)$ definiert werden. Ist $p \in \mathbb{N}$, dann heisst $\mathbb{E}[(X - \mathbb{E}[X])^p]$ das p-te zentrale Moment von X. Eine besondere Kennzahl ist die **Varianz** von X,

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

  Die Varianz ist eine Kennzahl dafür, wie stark X um den Erwartungswert fluktuiert. Da die Grösse quadriert ist, betrachtet man oft auch die **Standardabweichung** $\sqrt{\mathrm{Var}[X]}$.

- Die **momenterzeugende Funktion** ist definiert als $M_X(r) = \mathbb{E}[\mathrm{e}^{rX}]$. Der Name kommt daher, dass $M_X^{(p)}(r) = \mathbb{E}[X^p \mathrm{e}^{rX}]$ (die p-te Ableitung), und daher das p-te Moment $M_X^{(p)}(0) = \mathbb{E}[X^p]$ aus der momenterzeugenden Funktion erhalten werden kann.

Beispiele

- *Binomialverteilung* Für den Erwartungswert erhalten wir

$$\mathbb{E}[X] = \sum_{k=0}^n k \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=1}^n n \binom{n-1}{k-1} p^k (1-p)^{n-k} = np \ .$$

  Das zweite Moment wird

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

  Für die Varianz erhalten wir $\mathrm{Var}[X] = np - np^2 = np(1-p)$.

- *Gleichverteilung auf* $[a,b]$ Das p-te Moment wird

$$\mathbb{E}[X^p] = \frac{1}{b-a}\int_a^b x^p \,\mathrm{d}x = \frac{1}{p+1}\frac{b^{p+1}-a^{p+1}}{b-a} .$$

Insbesondere ist $\mathbb{E}[X] = \frac{1}{2}(a+b)$ und $\mathbb{E}[X^2] = \frac{1}{3}(b^2+ab+a^2)$. Damit wird $\mathrm{Var}[X] = \frac{1}{12}(b-a)^2$.

- *Exponentialverteilung* Für $p \in \mathbb{N}\setminus\{0\}$ erhalten wir

$$\mathbb{E}[X^p] = \int_0^\infty x^p \alpha \mathrm{e}^{-\alpha x}\,\mathrm{d}x = \int_0^\infty p x^{p-1}\mathrm{e}^{-\alpha x}\,\mathrm{d}x = \frac{p}{\alpha}\mathbb{E}[X^{p-1}] .$$

Durch Induktion ergibt sich $\mathbb{E}[X^p] = \alpha^{-p}p!$. Also ist $\mathbb{E}[X] = \alpha^{-1}$, $\mathbb{E}[X^2] = 2\alpha^{-2}$, $\mathrm{Var}[X] = \alpha^{-2}$. Für beliebiges $p > 0$ können wir die Momente ausdrücken durch $\mathbb{E}[X^p] = \alpha^{-p}\Gamma(p+1)$, wobei

$$\Gamma(x) := \int_0^\infty y^{x-1}\mathrm{e}^{-y}\,\mathrm{d}y$$

die Gamma-Funktion bezeichnet.

- *Normalverteilung* Für die Standardnormalverteilung ergibt sich für $p \in \mathbb{N}$

$$\mathbb{E}[X^p] = \int_{-\infty}^\infty x^p \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2}\,\mathrm{d}x .$$

Ist p ungerade, ergibt sich aus der Symmetrie, dass $\mathbb{E}[X^p] = 0$. Für allgemeines $p > 0$ erhalten wir

$$\begin{aligned}\mathbb{E}[|X|^p] &= \int_{-\infty}^\infty |x|^p \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2}\,\mathrm{d}x = 2\int_0^\infty x^p \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2}\,\mathrm{d}x \\ &= \frac{1}{\sqrt{\pi}}2^{p/2}\int_0^\infty z^{\frac{1}{2}(p-1)}\mathrm{e}^{-z}\,\mathrm{d}z = \frac{1}{\sqrt{\pi}}2^{p/2}\Gamma(\tfrac{1}{2}(p+1)) .\end{aligned}$$

Aus $1 = \mathbb{E}[|X|^0] = \Gamma(\frac{1}{2})/\sqrt{\pi}$ können wir schliessen, dass $\Gamma(\frac{1}{2}) = \sqrt{\pi}$. Mit Hilfe der Formeln $\Gamma(x+1) = x\Gamma(x)$ und $\Gamma(n+1) = n!$ für $n \in \mathbb{N}$ erhalten wir $\mathbb{E}[|X|] = \sqrt{2/\pi}$, $\mathbb{E}[X^2] = (2/\sqrt{\pi})\frac{1}{2}\Gamma(\frac{1}{2}) = 1$, $\mathbb{E}[|X|^3] = 2\sqrt{2/\pi}$, $\mathbb{E}[X^4] = 3$.

Für die Normalverteilung mit Mittelwert μ und Varianz σ^2, $Y = \mu + \sigma^2 X$, erhalten wir die Momente am einfachsten über die binomischen Formeln. So ist $\mathbb{E}[Y] = \mathbb{E}[\mu+\sigma X] = \mu$ und $\mathbb{E}[Y^2] = \mathbb{E}[(\mu+\sigma X)^2] = \mu^2+2\mu\sigma 0+\sigma^2 1 = \sigma^2+\mu^2$. Also haben wir die Varianz $\mathrm{Var}[Y] = \sigma^2$.

2.4. Ungleichungen

Manchmal benötigt man nicht den exakten Erwartungswert oder die exakte Wahrscheinlichkeit, sondern eine Abschätzung ist ausreichend. Die Abschätzungen lassen sich oft leichter berechnen, als der Erwartungswert oder die Wahrscheinlichkeit.

Hilfssatz 2.14. (Jensens Ungleichung) *Ist X eine Zufallsvariable mit endlichem Erwartungswert und $u : \mathbb{R} \to \mathbb{R}$ eine konvexe Funktion, so dass $\mathbb{E}[u(X)]$ existiert. Dann gilt $\mathbb{E}[u(X)] \geq u(\mathbb{E}[X])$. Ist $\mathbb{P}[X = \mathbb{E}[X]] < 1$ und $u(x)$ strikt konvex, dann gilt die strikte Ungleichung.*

Bemerkung. Ist $u(x)$ konkav, so ist $-u(x)$ konvex. Also gilt in diesem Fall $\mathbb{E}[u(x)] \leq u(\mathbb{E}[X])$. ■

Beweis. Für eine konvexe Funktion gibt es für jeden Punkt x_0 eine Gerade $\ell(x) = u(x_0) + k(x_0)(x - x_0)$, die $u(x)$ in x_0 berührt, so dass $\ell(x) \leq u(x)$. Setzen wir $x_0 = \mathbb{E}[X]$. Dann erhalten wir

$$u(\mathbb{E}[X]) = u(x_0) = u(x_0) + k(x_0)(\mathbb{E}[X] - x_0) = \mathbb{E}[\ell(X)] \leq \mathbb{E}[u(X)] \; .$$

Ist $u(x)$ strikt konvex, so ist $u(x) > \ell(x)$ für $x \neq x_0$. Analog folgt dann die strikte Ungleichung. □

Wir erhalten die Ungleichungen $\mathbb{E}[|X|] \geq |\mathbb{E}[X]|$ und $\mathbb{E}[X^2] \geq \mathbb{E}[X]^2$.

Korollar 2.15. *Ist $0 < p \leq q$, so gilt*

$$\mathbb{E}[|X|^p]^{1/p} \leq \mathbb{E}[|X|^q]^{1/q} \; .$$

Beweis. Die Funktion $u(x) = x^{q/p}$ ist konvex auf $[0, \infty)$. Somit erhalten wir

$$\mathbb{E}[|X|^q] = \mathbb{E}[(|X|^p)^{q/p}] \geq \mathbb{E}[|X|^p]^{q/p} \; .$$

Dies ist äquivalent zur Behauptung. □

Hat man Informationen über Erwartungswert oder Varianz, lassen sich auch Wahrscheinlichkeiten abschätzen.

Hilfssatz 2.16. *Sei $h(x)$ eine positive wachsende Funktion. Dann gilt*

$$h(c)\mathbb{P}[X \geq c] \leq \mathbb{E}[h(X)] \; .$$

Beweis. Da $h(x)$ wachsend und positiv ist, gilt $h(c)\mathbb{1}_{X\geq c} \leq h(X)$. Nimmt man den Erwartungswert, folgt die Aussage. □

Korollar 2.17. (Markov-Ungleichung) *Sei $c > 0$ und $\mathbb{E}[|X|] < \infty$. Dann gilt*

$$\mathbb{P}[|X| \geq c] \leq c^{-1}\mathbb{E}[|X|] \; .$$

Beweis. Wählen wir in Hilfssatz 2.16 $h(x) = x$, und wenden dies auf die Zufallsvariable $|X|$ an, so folgt die Aussage. □

Ist X eine positive Zufallsvariable mit $\mathbb{E}[X] = 0$, erhalten wir $\mathbb{P}[X \geq c] = 0$, also ist $\mathbb{P}[X = 0] = 1$.

Korollar 2.18. (Chebychev-Ungleichung) *Sei $\mathbb{E}[X^2] < \infty$ und $c > 0$. Dann gilt*

$$\mathbb{P}[|X - \mathbb{E}[X]| \geq c] \leq c^{-2}\,\mathrm{Var}[X] \; . \tag{2.1}$$

Beweis. Wir wenden Hilfssatz 2.16 mit $h(x) = x^2$ auf die Zufallsvariable $|X - \mathbb{E}[X]|$ an. □

Korollar 2.19. *Sei $r > 0$, so dass die momenterzeugende Funktion $M_X(r) = \mathbb{E}[\mathrm{e}^{rX}]$ existiert. Dann gilt*

$$\mathbb{P}[X \geq c] \leq \mathrm{e}^{-rc}M_X(r) = \exp\{-(rc - \log M_X(r))\} \; .$$

Beweis. Dies folgt aus Hilfssatz 2.16 mit $h(x) = \mathrm{e}^{rx}$. □

Die obige Ungleichung ist wichtig in der Theorie der *grossen Abweichungen*. Man wählt r so, dass $rc - \log M_X(r)$ maximal wird.

Hilfssatz 2.20. (Cauchy–Schwarz-Ungleichung) *Seien X und Y Zufallsvariablen mit $\mathbb{E}[X^2+Y^2]<\infty$. Dann gilt*

$$(\mathbb{E}[XY])^2 \le \mathbb{E}[X^2]\mathbb{E}[Y^2] \; .$$

Beweis. Ist $\mathbb{E}[Y^2]=0$, so haben wir oben gesehen, dass $Y=0$ gelten muss. In diesem Fall gilt die Ungleichung trivialerweise. Nehmen wir also $\mathbb{E}[Y^2]>0$ an. Für jedes $\alpha\in\mathbb{R}$ haben wir

$$0 \le \mathbb{E}[(X-\alpha Y)^2] = \mathbb{E}[X^2] - 2\alpha\mathbb{E}[XY] + \alpha^2\mathbb{E}[Y^2] \; .$$

Die rechte Seite wird minimal für $\alpha=\mathbb{E}[XY]/\mathbb{E}[Y^2]$. Setzen wir diesen Wert ein, erhalten wir

$$0 \le \mathbb{E}[X^2] - \frac{(\mathbb{E}[XY])^2}{\mathbb{E}[Y^2]} \; .$$

Dies ist äquivalent zur Behauptung. □

Korollar 2.21. (Ungleichung von Cantelli) *Sei X eine Zufallsvariable und $\mathbb{E}[X^2]<\infty$. Dann gilt für jedes $c\ge 0$*

$$\mathbb{P}[X \ge \mathbb{E}[X]+c] \le \frac{\mathrm{Var}[X]}{c^2+\mathrm{Var}[X]} \; .$$

Beweis. Wir dürfen $\mathbb{E}[X]=0$ annehmen. Aus der Cauchy–Schwarz-Ungleichung (Lemma 2.20) erhalten wir

$$\begin{aligned} c^2 = (\mathbb{E}[c-X])^2 &\le (\mathbb{E}[(c-X)\mathbb{1}_{X<c}])^2 \le \mathbb{E}[(c-X)^2]\mathbb{E}[\mathbb{1}^2_{X<c}] \\ &= (c^2+\mathrm{Var}[X])\mathbb{P}[X<c] = (c^2+\mathrm{Var}[X])(1-\mathbb{P}[X\ge c]) \; . \end{aligned}$$

Auflösen nach $\mathbb{P}[X\ge c]$ gibt die Behauptung. □

2.5. Varianz, Kovarianz, lineare Prognose

Wir wollen nun Rechenregeln für die Varianz finden. Seien $a,b\in\mathbb{R}$ und $\mathbb{E}[X^2]<\infty$. Dann gilt

$$\mathrm{Var}[aX+b] = \mathbb{E}[(aX+b-\mathbb{E}[aX+b])^2] = \mathbb{E}[\{a(X-\mathbb{E}[X])\}^2] = a^2\,\mathrm{Var}[X] \; .$$

Seien X und Y zwei Zufallsvariablen. Dann gilt

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

Wir machen daher folgende

Definition 2.22. *Die Grösse*

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

heisst **Kovarianz** *von X und Y. Wir sagen, X und Y sind* **unkorreliert**, *falls* $\mathrm{Cov}[X,Y]=0$.

Es folgt sofort, dass $\mathrm{Cov}[X,X]=\mathrm{Var}[X]$.

Wir wollen nun die Kovarianz berechnen, falls X und Y unabhängig sind. Wir werden die Unabhängigkeit von Zufallsvariablen erst in Definition 2.24 definieren. Für die Berechnungen unten benutzen wir aber nur die diskrete Version (Definition 1.30).

Nehmen wir zuerst an, X und Y seien diskrete Variablen. Wir charakterisieren sie durch $\{x_i,p_i)\}$ und $\{(y_i,q_i)\}$. Dann gilt

$$\begin{aligned}\mathbb{E}[XY] &= \sum_k\sum_j x_k y_j \mathbb{P}[X=x_k, Y=y_j] = \sum_k\sum_j x_k y_j p_k q_j \\ &= \Big(\sum_k x_k p_k\Big)\Big(\sum_j y_j q_j\Big) = \mathbb{E}[X]\mathbb{E}[Y]\ .\end{aligned}$$

Sei nun X eine stetige und Y eine diskrete Variable. Wir nehmen zuerst an, dass $|X|$ und $|Y|$ durch c beschränkt sind. Dann haben wir

$$\sum_{k,j}\frac{kc}{n}y_j\mathbb{P}[(k-1)c/n<X\le kc/n, Y=y_j] = \sum_k\frac{kc}{n}\mathbb{P}[(k-1)c/n<X\le kc/n]\mathbb{E}[Y]$$

wie im diskreten Fall. Lassen wir $n\to\infty$ erhalten wir

$$\mathbb{E}[XY] = \int_{-c}^{c} x\,\mathrm{d}F(x)\mathbb{E}[Y] = \mathbb{E}[X]\mathbb{E}[Y]\ .$$

Sind nun $X,Y\ge 0$ positive Zufallsvariablen, erhalten wir aus monotoner Konvergenz

$$\mathbb{E}[XY] = \lim_{n\to\infty}\mathbb{E}[X\mathbb{1}_{X\le n}Y\mathbb{1}_{Y\le n}] = \lim_{n\to\infty}\mathbb{E}[X\mathbb{1}_{X\le n}]\mathbb{E}[Y\mathbb{1}_{Y\le n}] = \mathbb{E}[X]\mathbb{E}[Y]\ .$$

Für beliebige X,Y haben wir

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

Analog folgt die Formel $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$, falls beide X und Y stetig sind. Wir haben also, $\mathrm{Cov}[X, Y] = 0$, falls X und Y unabhängig sind.

Für *unabhängige* Zufallsvariablen gilt also

$$\mathrm{Var}\Big[\sum_{k=1}^{n} X_k\Big] = \sum_{k=1}^{n} \mathrm{Var}[X_k] \; .$$

Sind $\{X_k\}$ unabhängige $\{0, 1\}$ Experimente mit Parameter p. Dann ist $\mathbb{E}[X_i] = p$ und $\mathbb{E}[X_i^2] = \mathbb{E}[X_i] = p$. Also ist die Varianz $\mathrm{Var}[X_i] = p - p^2 = p(1-p)$. Aus der Summenformel erhalten wir $\mathrm{Var}[S_n] = \sum_{k=1}^{n} \mathrm{Var}[X_i] = np(1-p)$, was mit der Varianz der Binomialverteilung übereinstimmt.

Wir wollen nun Rechenregeln für die Kovarianz bestimmen. Wir erhalten aus der Symmetrie der Definition

$$\mathrm{Cov}[X, Y] = \mathrm{Cov}[Y, X] \; .$$

Weiter gilt für $a, b \in \mathbb{R}$

$$\mathrm{Cov}[X, aY + b] = \mathbb{E}[(X - \mathbb{E}[X])a(Y - \mathbb{E}[Y])] = a\,\mathrm{Cov}[X, Y] \; .$$

Ist Z eine weitere Zufallsvariable, erhalten wir

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

Definition 2.23. *Sei* $\mathrm{Var}[X]\,\mathrm{Var}[Y] > 0$. *Die Grösse*

$$\mathrm{Cor}[X, Y] := \frac{\mathrm{Cov}[X, Y]}{\sqrt{\mathrm{Var}[X]\,\mathrm{Var}[Y]}}$$

heisst **Korrelation** *von* X *und* Y.

Aus der Cauchy–Schwarz Ungleichung (Lemma 2.20) schliessen wir $\mathrm{Cor}[X, Y] \in [-1, 1]$.

Betrachten wir den Extremfall $\mathrm{Cor}[X, Y] = 1$. Wir dürfen $\mathbb{E}[X] = \mathbb{E}[Y] = 0$ annehmen. Dann haben wir

$$\mathbb{E}\left[\left(X - \sqrt{\frac{\mathrm{Var}[X]}{\mathrm{Var}[Y]}}\, Y\right)^2\right] = \mathrm{Var}[X] - \frac{\mathrm{Var}[X]}{\mathrm{Var}[Y]}\,\mathrm{Var}[Y] = 0 \; .$$

Somit gilt

$$\mathbb{P}\left[X = \sqrt{\frac{\mathrm{Var}[X]}{\mathrm{Var}[Y]}}\, Y\right] = 1 \; .$$

Analog folgt, falls $\mathrm{Cor}[X,Y] = -1$,

$$\mathbb{P}\left[X = -\sqrt{\frac{\mathrm{Var}[X]}{\mathrm{Var}[Y]}}\, Y\right] = 1 \; .$$

Betrachten wir das folgende Problem. Seien X, Y zwei Zufallsvariablen. Wir beobachten X und wollen nun Y vorhersagen. Wir verwenden eine *lineare Prognose* $\hat{Y} = aX + b$. Wir suchen nun Zahlen a, b, so dass der mittlere quadratische Fehler $\mathbb{E}[(\hat{Y} - Y)^2]$ minimal wird. Aus

$$\mathbb{E}[(\hat{Y} - Y)^2] = \mathrm{Var}[\hat{Y} - Y] + \mathbb{E}[\hat{Y} - Y]^2$$

und der Tatsache, dass $\mathrm{Var}[\hat{Y} - Y]$ nicht von b abhängt, schliessen wir, dass $\mathbb{E}[\hat{Y} - Y] = 0$, also $b = \mathbb{E}[Y] - a\mathbb{E}[X]$. Ist nun $\mathrm{Var}[X] = 0$, ist die beste Prognose $\hat{Y} = a\mathbb{E}[X] + b = \mathbb{E}[Y]$. Wir können also $\mathrm{Var}[X] > 0$ annehmen. Es bleibt

$$\begin{aligned}\mathbb{E}[(\hat{Y} - Y)^2] = \mathrm{Var}[\hat{Y} - Y] &= \mathrm{Var}[\hat{Y}] + \mathrm{Var}[Y] - 2\,\mathrm{Cov}[\hat{Y}, Y] \\ &= a^2\,\mathrm{Var}[X] + \mathrm{Var}[Y] - 2a\,\mathrm{Cov}[X,Y] \; .\end{aligned}$$

Dies ist minimal für $a = \mathrm{Cov}[X,Y]/\,\mathrm{Var}[X]$. Also haben wir die *optimale lineare Prognose*

$$\hat{Y} = \frac{\mathrm{Cov}[X,Y]}{\mathrm{Var}[X]} X + \mathbb{E}[Y] - \frac{\mathrm{Cov}[X,Y]}{\mathrm{Var}[X]}\mathbb{E}[X] = \mathbb{E}[Y] + \frac{\mathrm{Cov}[X,Y]}{\mathrm{Var}[X]}(X - \mathbb{E}[X]) \; .$$

Das Verfahren heisst **lineare Regression**. Ist $\mathrm{Var}[X] = \mathrm{Var}[Y]$, erhalten wir

$$\hat{Y} = \mathbb{E}[Y] + \mathrm{Cor}[X,Y](X - \mathbb{E}[X]) \; .$$

Francis Galton hat die Körpergrössen von Vätern und Söhnen untersucht. Er fand dann die Regressionsformel mit $a \in (0,1)$, das heisst positive Korrelation. Somit sind die Söhne von grossen Vätern auch gross, aber im Durchschnitt nicht so stark vom Mittelwert entfernt wie die Väter. Analog sind Söhne kleiner Väter auch klein, aber im Mittel näher beim Mittelwert als die Väter. Er nannte dies "regression to mediocrity" (Rückentwicklung zum Mittelmass). Daher kommt der Name "lineare Regression".

2.6. Die gemeinsame Verteilung von d Zufallsvariablen

Seien $X_1, X_2, \ldots, X_d$ eine Familie von Zufallsvariablen. Betrachten wir den Vektor $\boldsymbol{X} = (X_1, X_2, \ldots, X_d)$, dann ist $\boldsymbol{X} : \Omega \to \mathbb{R}^d$ eine Abbildung. Sei $\mathfrak{B}^d$ die von den Mengen $A_1 \times A_2 \times \cdots \times A_d$, $A_k \in \mathfrak{B}^1$, erzeugte σ-Algebra auf $\mathbb{R}^d$. Sie heisst **Borel-σ-Algebra** auf $\mathbb{R}^d$. Da

$$\{\boldsymbol{X} \in A_1 \times \cdots \times A_d\} = \cap_{k=1}^d \{X_k \in A_k\} \ ,$$

ist die Abbildung $\boldsymbol{X}$ eine messbare Abbildung von Ω nach $\mathbb{R}^d$. Die Borel-σ-Algebra $\mathfrak{B}^d$ wird erzeugt durch Mengen der Form $(-\infty, a_1] \times (-\infty, a_2] \times \cdots \times (-\infty, a_d]$. Es genügt daher die **gemeinsame Verteilungsfunktion**

$$F(x_1, x_2, \ldots, x_d) = \mathbb{P}[X_1 \leq x_1, \ldots, X_d \leq x_d]$$

zu kennen.

Die Verteilung von $\boldsymbol{X}$ heisst, analog zum eindimensionalen Fall, **absolutstetig**, wenn es eine messbare Funktion $f : \mathbb{R}^d \to \mathbb{R}$ gibt, so dass

$$F(x_1, \ldots, x_d) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_d} f(y_1, \ldots, y_d) \, \mathrm{d}y_d \cdots \, \mathrm{d}y_1 \ .$$

Die Dichtefunktion von $\boldsymbol{X}$ muss dann eine positive reelle Funktion auf $\mathbb{R}^d$ sein, so dass $\int_{\mathbb{R}^d} f(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} = 1$.

Aus der gemeinsamen Verteilungsfunktion $F(\boldsymbol{x})$ können wir auch die Verteilung von X_k bestimmen,

$$\begin{aligned} F_k(x_k) &= \mathbb{P}[X_k \leq x_k] \\ &= \mathbb{P}[X_1 < \infty, \ldots, X_{k-1} < \infty, X_k \leq x_k, X_{k+1} < \infty, \ldots, X_d < \infty] \\ &= F(\infty, \ldots, \infty, x_k, \infty, \ldots, \infty) \ . \end{aligned}$$

Analog lässt sich die gemeinsame Verteilung der k Zufallsvariablen $X_{i_1}, X_{i_2,}, \ldots, X_{i_k}$ für $1 \leq i_1 < i_2 < \cdots < i_k \leq d$ bestimmen.

Ist nun $\boldsymbol{X}$ absolutstetig, erhalten wir, dass auch X_k absolutstetig ist. Die Dichte von X_k ist dann

$$f_k(x_k) = \int_{x_1=-\infty}^{\infty} \cdots \int_{x_{k-1}=-\infty}^{\infty} \int_{x_{k+1}=-\infty}^{\infty} \cdots \int_{x_d=-\infty}^{\infty} f(\boldsymbol{x}) \, \mathrm{d}x_d \cdots \, \mathrm{d}x_{k+1} \, \mathrm{d}x_{k-1} \cdots \, \mathrm{d}x_1 \ .$$

Achtung: Es kann sein, dass alle Zufallsvariablen X_k eindimensional absolutstetig sind, aber mehrdimensional nicht absolutstetig sind.

Definition 2.24. *Die Zufallsvariablen $X_1, \dots, X_d$ heissen* **(stochastisch) unabhängig**, *falls*

$$F_{\boldsymbol{X}}(x_1, \dots, x_d) = F_1(x_1)F_2(x_2)\cdots F_d(x_d) \ .$$

Die Definition ist äquivalent zu

$$\mathbb{P}[\cap_{k=1}^d \{X_k \in A_k\}] = \prod_{k=1}^d \mathbb{P}[X_k \in A_k] \ ,$$

wobei $A_k \in \mathfrak{B}^1$ Borel Mengen sind. Weiter lässt sich zeigen, dass für messbare Funktionen $h_k : \mathbb{R} \to \mathbb{R}$, die Formel

$$\mathbb{E}\Big[\prod_{k=1}^d h_k(X_k)\Big] = \prod_{k=1}^d \mathbb{E}[h_k(X_k)]$$

gilt, falls $X_1, \dots, X_d$ unabhängig sind, siehe auch Abschnitt 2.5. Gilt umgekehrt die obige Formel für alle messbaren Funktion h_k, dann sind die Zufallsvariablen unabhängig. Dies folgt sofort, falls man $h_k(x_k) = \mathbb{1}_{x_k \le a_k}$ wählt. Insbesondere folgt für unabhängige Zufallsvariablen

$$\mathbb{E}\Big[\prod_{k=1}^d X_k\Big] = \prod_{k=1}^d \mathbb{E}[X_k] \ .$$

Wie wir schon vorher bewiesen haben, sind also unabhängige Zufallsvariablen unkorreliert.

Wie wir schon für Ereignisse bemerkt haben, impliziert paarweise Unabhängigkeit nicht stochastische Unabhängigkeit. Wir können auch aus der Unkorreliertheit nicht schliessen, dass zwei Zufallsvariablen unabhängig sind. Ist zum Beispiel X standardnormalverteilt, und $Y = X^2$, so sind X und Y nicht unabhängig. Zum Beispiel ist

$$\mathbb{P}[X > 1, Y > 1] = \mathbb{P}[X > 1] > \mathbb{P}[X > 1]\mathbb{P}[Y > 1] \ .$$

Aber $\mathbb{E}[XY] = \mathbb{E}[X^3] = 0 = \mathbb{E}[X]\mathbb{E}[Y]$, da $\mathbb{E}[X] = 0$.

Proposition 2.25. *Die absolutstetigen Zufallsvariablen $X_1, X_2, \dots, X_d$ sind genau dann unabhängig, wenn ihre gemeinsame Dichte sich als*

$$f(x_1, \dots, x_d) = f_1(x_1)\cdots f_k(x_k)$$

schreiben lässt.

Beweis. Lässt sich $f(\boldsymbol{x})$ als Produkt schreiben, dann gilt

$$F(\boldsymbol{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} \prod_{k=1}^{d} f_k(y_k)\, \mathrm{d}y_d \cdots\, \mathrm{d}y_1 = \prod_{k=1}^{d} \int_{-\infty}^{x_k} f_k(y_k)\, \mathrm{d}y_k = \prod_{k=1}^{d} F_k(x_k)\ .$$

Also sind die Zufallsvariablen unabhängig.

Seien die Zufallsvariablen nun unabhängig. Dann gilt

$$\begin{aligned}\int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} f(\boldsymbol{y})\, \mathrm{d}\boldsymbol{y} = F(\boldsymbol{x}) = \prod_{k=1}^{d} F_k(x_k) &= \prod_{k=1}^{d} \int_{-\infty}^{x_k} f_k(y_k)\, \mathrm{d}y_k \\ &= \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} \prod_{k=1}^{d} f_k(y_k)\, \mathrm{d}y_d \cdots\, \mathrm{d}y_1\ .\end{aligned}$$

Da die Formel für alle $\boldsymbol{x}$ (mit Ausnahme einer Menge mit Mass 0) gelten muss, folgt die Produktformel für die Dichtefunktion. □

Als nächstes betrachten wir Summen von unabhängigen Zufallsvariablen.

Hilfssatz 2.26. *Seien X_1 und X_2 unabhängige absolutstetige Zufallsvariablen mit Dichtefunktionen $f_1(x)$ und $f_2(x)$. Dann ist $X = X_1 + X_2$ absolutstetig mit Dichtefunktion*

$$f(x) = \int_{-\infty}^{\infty} f_1(z) f_2(x - z)\, \mathrm{d}z\ .$$

Bemerkung. Die Formel für $f(x)$ heisst **Faltung** von f_1 und f_2. Man schreibt oft kurz $f(x) = f_1 * f_2(x)$. Für die Verteilungsfunktionen schreiben wir $F(x) = F_1 * F_2(x)$. Sind $X_1, \ldots, X_d$ identisch und unabhängig verteilt, so schreiben wir kurz für die Verteilungsfunktion der Summe $X_1 + \cdots + X_d$, $F^{*d}(x)$, und für die Dichte $f^{*d}(x)$. ■

Beweis. Kennen wir X_1, dann muss $X_2 \leq x - X_1$ sein, damit $X \leq x$ gilt. Wir erhalten also

$$\begin{aligned}F(x) &= \int_{-\infty}^{\infty} \int_{-\infty}^{x - x_1} f(x_1, x_2)\, \mathrm{d}x_2\, \mathrm{d}x_1 = \int_{-\infty}^{\infty} \int_{-\infty}^{x} f(x_1, z - x_1)\, \mathrm{d}z\, \mathrm{d}x_1 \\ &= \int_{-\infty}^{x} \int_{-\infty}^{\infty} f(x_1, z - x_1)\, \mathrm{d}x_1\, \mathrm{d}z\ .\end{aligned}$$

Da $f(x_1, z - x_1) = f_1(x_1) f_2(z - x_1)$ folgt die Behauptung. □

Beispiele

- *Normalverteilung auf* $\mathbb{R}^d$ Sind $X_1, \ldots, X_d$ standardnormalverteilt und unabhängig, dann hat $\boldsymbol{X}$ die Dichte

$$f(\boldsymbol{x}) = (2\pi)^{-d/2} \exp\Big\{-\tfrac{1}{2}\sum_{k=1}^{d} x_k^2\Big\} .$$

Ist nun $\boldsymbol{A}$ eine $d \times d$ Matrix und $\boldsymbol{\mu}$ ein d-dimensionaler Vektor, dann können wir eine neue Zufallsvariable $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{\mu}$ definieren. Also haben wir

$$Y_k = \sum_{\ell=1}^{d} A_{k\ell} X_\ell + \mu_k .$$

Der Mittelwert ist dann $\mathbb{E}[Y_k] = \mu_k$. Die Kovarianzen erhalten wir aus

$$\Sigma_{ij} = \mathrm{Cov}[Y_i, Y_j] = \mathbb{E}\Big[\sum_{k=1}^{d} A_{ik}X_k \sum_{\ell=1}^{d} A_{j\ell}X_\ell\Big] = \sum_{k=1}^{d}\sum_{\ell=1}^{d} A_{ik}A_{j\ell}\mathbb{E}[X_kX_\ell]$$
$$= \sum_{k=1}^{d} A_{ik}A_{jk} = (\boldsymbol{A}\boldsymbol{A}^\top)_{ij} .$$

Betrachten wir nun den Fall, dass $\boldsymbol{\Sigma} = (\Sigma_{ij})$ nicht invertierbar ist. Dann gibt es einen Vektor $\boldsymbol{a} \in \mathbb{R}^d \setminus \{\boldsymbol{0}\}$, so dass $\boldsymbol{a}^\top\boldsymbol{A}\boldsymbol{A}^\top\boldsymbol{a} = 0$. Insbesondere ist

$$\mathrm{Var}[\boldsymbol{a}^\top\boldsymbol{Y}] = \boldsymbol{a}^\top\boldsymbol{A}\boldsymbol{A}^\top\boldsymbol{a} = 0 .$$

Also haben wir, dass $\boldsymbol{a}^\top\boldsymbol{Y} = \mathbb{E}[\boldsymbol{a}^\top\boldsymbol{Y}] = \boldsymbol{a}^\top\boldsymbol{\mu}$. Wir sehen also, dass sich eines der Y_k als Linearkombination der anderen schreiben lässt. Es genügt also, die gemeinsame Verteilung der anderen Y_i zu kennen. Nehmen wir daher an, dass $\boldsymbol{\Sigma}$ invertierbar ist, und setzen wir $\boldsymbol{C} = \boldsymbol{\Sigma}^{-1}$. Es muss dann gelten, dass $\boldsymbol{A}$ invertierbar ist. Also ist $\boldsymbol{C} = (\boldsymbol{A}^{-1})^\top\boldsymbol{A}^{-1}$. Wir erhalten dann

$$\sum_{k=1}^{d} x_k^2 = \boldsymbol{x}^\top\boldsymbol{x} = \boldsymbol{x}^\top\boldsymbol{A}^\top\boldsymbol{C}\boldsymbol{A}\boldsymbol{x} = (\boldsymbol{y}-\boldsymbol{\mu})^\top\boldsymbol{C}(\boldsymbol{y}-\boldsymbol{\mu}) .$$

Somit hat $\boldsymbol{Y}$ die Dichtefunktion

$$f_{\boldsymbol{Y}}(y_1, \ldots, y_d) = (2\pi)^{-d/2}(\det\boldsymbol{\Sigma})^{-1/2} \exp\{-\tfrac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\} ,$$

siehe auch Hilfssatz 2.27 unten. Diese Verteilung heisst **d-dimensionale Normalverteilung** mit Mittelwert $\boldsymbol{\mu}$ und Kovarianzmatrix $\boldsymbol{\Sigma}$. Wir sehen, dass die Normalverteilung durch Mittelwert und Kovarianzmatrix bestimmt ist.

Berechnen wir nun die Verteilung von $X = aX_1 + bX_2$. Dann sind aX_1 und bX_2 unabhängig und normalverteilt mit Mittelwert 0 und Varianz a^2, bzw. b^2. Die Dichte von X ist dann

$$f(x) = \frac{1}{2ab\pi} \int_{-\infty}^{\infty} \exp\Big\{-\tfrac{1}{2}\Big(\frac{z^2}{a^2} + \frac{(x-z)^2}{b^2}\Big)\Big\}\, \mathrm{d}z \ .$$

Wir schreiben

$$\begin{aligned} \frac{z^2}{a^2} + \frac{(x-z)^2}{b^2} &= \frac{(z - b^{-2}(a^{-2}+b^{-2})^{-1}x)^2}{(a^{-2}+b^{-2})^{-1}} + x^2(b^{-2} - b^{-4}(a^{-2}+b^{-2})^{-1}) \\ &= \frac{(z - b^{-2}(a^{-2}+b^{-2})^{-1}x)^2}{(a^{-2}+b^{-2})^{-1}} + \frac{x^2}{a^2+b^2} \ . \end{aligned}$$

Da

$$\frac{1}{\sqrt{2\pi(a^{-2}+b^{-2})^{-1}}} \int_{-\infty}^{\infty} \exp\Big\{-\tfrac{1}{2}\frac{(z - b^{-2}(a^{-2}+b^{-2})^{-1}x)^2}{(a^{-2}+b^{-2})^{-1}}\Big\}\, \mathrm{d}z = 1 \ ,$$

erhalten wir

$$f(x) = \frac{1}{\sqrt{2\pi(a^2+b^2)}} \exp\Big\{-\tfrac{1}{2}\frac{x^2}{a^2+b^2}\Big\} \ .$$

Also ist $aX_1 + bX_2$ normalverteilt mit Varianz $a^2 + b^2$. Wir sehen also, dass Y_k normalverteilt ist mit Mittelwert μ_k und Varianz Σ_{kk}. Insbesondere ist auch $Y_1 + Y_2$ normalverteilt mit Mittelwert $\mu_1 + \mu_2$ und Varianz $\Sigma_{11} + \Sigma_{22} + 2\Sigma_{12}$.

Wir können auch in umgekehrter Richtung vorgehen. Seien $\{Y_k : 1 \le k \le d\}$ d-dimensional normalverteilte Zufallsvariablen, so dass $\boldsymbol{\Sigma}$ invertierbar ist. Da $\boldsymbol{\Sigma}$ symmetrisch mit einer strikt positiven Diagonalen, gibt es eine symmetrische Matrix $\boldsymbol{A}$, so dass $\boldsymbol{A}^2 = \boldsymbol{\Sigma}$. Dann ist $\boldsymbol{X} = \boldsymbol{A}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu})$ d-dimensional standard normalverteilt. Wir sehen also, dass für multinomial normalverteilte Zufallsvariablen die gemeinsame Verteilungsfunktion aus den Mittelwerten und der Kovarianzmatrix folgt. Insbesondere folgt für multinomial normalverteilte Zufallsvariablen die Unabhängigkeit aus der Unkorreliertheit.

Aber aufgepasst. Sind X und Y zwei normalverteilte Zufallsvariablen, so lässt sich nicht daraus schliessen, dass der Vektor (X, Y) zweidimensional normalverteilt ist. Und damit folgt auch nicht die Unabhängigkeit nicht aus der Unkorreliertheit. Sei X standardnormalverteilt und Z unabhängig von X mit $\mathbb{P}[Z = 1] = \mathbb{P}[Z = -1] = \frac{1}{2}$. Setzen wir $Y = ZX$. Dann ist

$$\mathbb{P}[X > 1, Y > 1] = \mathbb{P}[X > 1, Z = 1] = \tfrac{1}{2}\mathbb{P}[X > 1] > (\mathbb{P}[X > 1])^2 \ ,$$

da $\mathbb{P}[X > 1] < \frac{1}{2}$. Somit sind X und Y abhängig. Wegen der Symmetrie der Normalverteilung, sind beide Randverteilungen normal. Für die Kovarianz erhalten wir

$$\mathbb{E}[XY] = \mathbb{E}[ZX^2] = \mathbb{E}[Z]\mathbb{E}[X^2] = 0 \cdot 1 = 0 \, .$$

Man kann auch Beispiele konstruieren, bei denen (X, Y) absolutstetig mit normalverteilten Randverteilungen sind. Sei $\Phi(x) = (2\pi)^{-1/2} \int_{-\infty}^{x} \mathrm{e}^{-y^2/2} \, \mathrm{d}y$ die Standard-Normalverteilung. Die gemeinsame Verteilung

$$F(x, y) = \frac{\Phi(x)\Phi(y)}{\Phi(x) + \Phi(y) - \Phi(x)\Phi(y)}$$

hat dann die Randverteilungen $\Phi(x)$. Die Verteilung ist absolutstetig mit der Dichte

$$f(x, y) = \frac{\mathrm{e}^{-(x^2+y^2)/2}\Phi(x)\Phi(y)}{\pi(\Phi(x) + \Phi(y) - \Phi(x)\Phi(y))^3} \, .$$

Da $\Phi(x)$ nicht in geschlossener Form dargestellt werden kann, kann es sich nicht um eine bivariate Normalverteilung handeln. Sei $(\tilde{X}, \tilde{Y})$ ein Vektor mit der Verteilung $F(x, y)$. Sei N gleichverteilt auf $\{1, 2, 3, 4\}$ und unabhängig von $(\tilde{X}, \tilde{Y})$. Definieren wir

$$(X, Y) = \begin{cases} (\tilde{X}, \tilde{Y}) \, , & \text{falls } N = 1, \\ (-\tilde{X}, \tilde{Y}) \, , & \text{falls } N = 2, \\ (\tilde{X}, -\tilde{Y}) \, , & \text{falls } N = 3, \\ (-\tilde{X}, -\tilde{Y}) \, , & \text{falls } N = 4. \end{cases} \, .$$

Dann hat auch (X, Y) normalverteilte Randverteilungen. Wegen der Symmetrie sind sogar X und Y unkorreliert. Aber X und Y sind nicht unabhängig.

- Seien $\{X_k\}$ unabhängig und gleichverteilt auf $[0, 1]$. Dann hat $X = X_1 + X_2$ die Dichte

$$f^{*2}(x) = \int_0^1 \mathbb{1}_{x-z \in [0,1]} \, \mathrm{d}z = \begin{cases} x \, , & \text{falls } 0 \le x \le 1, \\ 2 - x \, , & \text{falls } 1 < x \le 2, \\ 0 \, , & \text{sonst.} \end{cases}$$

Die Summe $Z = X_1 + X_2 + X_3$ hat die Dichte

$$f^{*3}(x) = \int_0^1 f^{*2}(x - z) \, \mathrm{d}z = \begin{cases} \frac{1}{2}x^2 \, , & \text{falls } 0 \le x \le 1, \\ \frac{3}{4} - (x - \frac{3}{2})^2 \, , & \text{falls } 1 < x \le 2, \\ \frac{1}{2}(3 - x)^2 \, , & \text{falls } 2 < x \le 3, \\ 0 \, , & \text{sonst.} \end{cases}$$

Die Dichten sind in Abbildung 2.3 dargestellt.

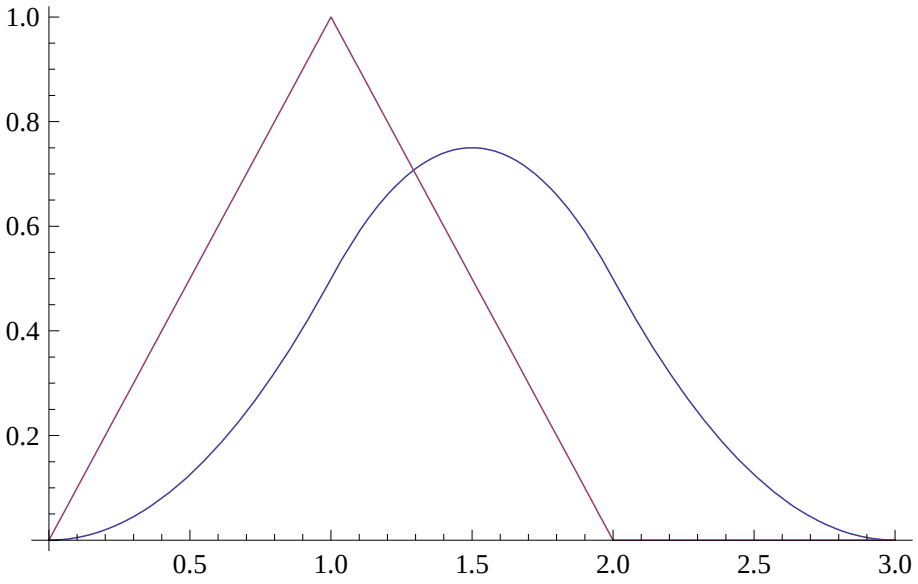

Abbildung 2.3: *Dichte der Faltung von gleichverteilten Zufallsvariablen*

- Seien $\{X_k\}$ Gamma-verteilt mit Parameter γ_k und α, das heisst, mit Dichtefunktion

$$f_k(x) = \frac{\alpha^{\gamma_k}}{\Gamma(\gamma_k)} x^{\gamma_k - 1} \mathrm{e}^{-\alpha x} \mathbb{1}_{x \geq 0} \, .$$

Ist $\gamma_k = 1$, so erhalten wir die Exponentialverteilung mit Parameter α. Wir erhalten für die Faltung für $x \geq 0$

$$\begin{aligned} f_1 * f_2(x) &= \int_{-\infty}^{\infty} f_1(z) f_2(x-z) \, \mathrm{d}z = \frac{\alpha^{\gamma_1+\gamma_2}}{\Gamma(\gamma_1)\Gamma(\gamma_2)} \int_0^x z^{\gamma_1 - 1} (x-z)^{\gamma_2 - 1} \mathrm{e}^{-\alpha x} \, \mathrm{d}z \\ &= \frac{\alpha^{\gamma_1+\gamma_2}}{\Gamma(\gamma_1)\Gamma(\gamma_2)} x^{\gamma_1+\gamma_2-1} \mathrm{e}^{-\alpha x} \int_0^1 y^{\gamma_1 - 1} (1-y)^{\gamma_2 - 1} \, \mathrm{d}y \\ &= \frac{\alpha^{\gamma_1+\gamma_2}}{\Gamma(\gamma_1+\gamma_2)} x^{\gamma_1+\gamma_2-1} \mathrm{e}^{-\alpha x} \, . \end{aligned}$$

Also ist $X_1 + X_2$ Gamma-verteilt mit Parameter $\gamma_1 + \gamma_2$ und α.

Nehmen wir nun $\gamma_k = \gamma$ an, erhalten wir für die Summe von d unabhängigen Gamma-verteilten Zufallsvariablen die Dichte

$$f^{*d}(x) = \frac{\alpha^{d\gamma}}{\Gamma(d\gamma)} x^{d\gamma - 1} \mathrm{e}^{-\alpha x} \, .$$

Insbesondere gilt für exponentialverteilte Zufallsvariablen ($\gamma = 1$)

$$f^{*d}(x) = \frac{\alpha^d}{\Gamma(d)} x^{d-1} \mathrm{e}^{-\alpha x} = \frac{\alpha^d}{(d-1)!} x^{d-1} \mathrm{e}^{-\alpha x} \, .$$

Eine wichtige Anwendung ist die Folgende. Eine Maschine hat eine Komponente, die eine Lebensdauer T_k mit einer Exponentialverteilung mit Parameter α hat. Verschiedene Komponenten haben eine unabhängige Lebensdauer. Ist die Komponente defekt, wird sie durch eine neue ersetzt. Sei N_t die Anzahl der Komponenten, die bis zum Zeitpunkt t ersetzt werden mussten. Wir wollen nun die Verteilung von N_t bestimmen. Sei $S_n = T_1 + \cdots + T_n$. Wir haben

$$\mathbb{P}[N_t = 0] = \mathbb{P}[T_1 > t] = \mathrm{e}^{-\alpha t} ,$$

und für $n \geq 1$,

$$\begin{aligned}\mathbb{P}[N_t = n] &= \mathbb{P}[S_n \leq t < S_{n+1}] = \mathbb{P}[S_n \leq t] - \mathbb{P}[S_{n+1} \leq t] \\ &= \int_0^t \Big(\frac{\alpha^n}{(n-1)!} z^{n-1} \mathrm{e}^{-\alpha z} - \frac{\alpha^{n+1}}{n!} z^n \mathrm{e}^{-\alpha z} \Big) \, \mathrm{d}z = \frac{\alpha^n}{n!} \int_0^t \frac{\mathrm{d}}{\mathrm{d}z} (z^n \mathrm{e}^{-\alpha z}) \, \mathrm{d}z \\ &= \frac{(\alpha t)^n}{n!} \mathrm{e}^{-\alpha t} . \end{aligned} \tag{2.2}$$

Also ist N_t Poissonverteilt mit Parameter αt. N_t hat also Mittelwert αt und Varianz αt.

Manchmal betrachtet man nicht die Zufallsvariablen selber, sondern eine Funktion davon. In folgendem Fall kann man die Dichte "einfach" erhalten.

Hilfssatz 2.27. *Sei $\boldsymbol{X}$ eine absolutstetige Zufallsvariable auf $\mathbb{R}^d$ mit der Dichte $f_X(\boldsymbol{x})$. Ferner sei $h : \mathbb{R}^d \to \mathbb{R}^d$ eine injektive Funktion mit Umkehrfunktion $u : h(\mathbb{R}^d) \to \mathbb{R}^d$, so dass u stetig differenzierbar ist. Sei $J(\boldsymbol{y})$ die Ableitung von u,*

$$J(\boldsymbol{y}) = \begin{pmatrix} \frac{\partial u_1}{\partial y_1} & \cdots & \frac{\partial u_1}{\partial y_d} \\ \vdots & & \vdots \\ \frac{\partial u_d}{\partial y_1} & \cdots & \frac{\partial u_d}{\partial y_d} \end{pmatrix} .$$

Dann ist $\boldsymbol{Y} = h(\boldsymbol{X})$ absolutstetig mit Dichte

$$f_Y(\boldsymbol{y}) = |\det J(\boldsymbol{y})| \, f(u(\boldsymbol{y})) \mathbb{1}_{y \in h(\mathbb{R}^d)} .$$

Beweis. Wir bemerken zuerst, dass $h(x)$ eine messbare Funktion ist, da $h(x)$ als Umkehrfunktion von $u(x)$ stetig ist. Sei $B \subset \mathbb{R}^d$ eine Borelmenge. Dann gilt für $\boldsymbol{x} = u(\boldsymbol{y})$

$$\mathbb{P}[\boldsymbol{Y} \in B] = \mathbb{P}[\boldsymbol{X} \in u(B)] = \int_{u(B)} \cdots \int f_X(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = \int_B \cdots \int |\det J(\boldsymbol{y})| f_X(u(\boldsymbol{y})) \, \mathrm{d}\boldsymbol{y} .$$

Somit ist $\boldsymbol{Y}$ absolutstetig mit Dichte $f_Y(\boldsymbol{y})$. □

Beispiel: Seien $(X_1, X_2) \in \mathbb{R}^2$ eine Zufallsvariable und $Y_1 = X_1 + X_2$, $Y_2 = X_1 - X_2$. Wir haben dann $h_1(x_1, x_2) = x_1 + x_2$ und $h_2(x_1, x_2) = x_1 - x_2$. Für die Funktion $u(y_1, y_2)$ erhalten wir $u_1(y_1, y_2) = \frac{1}{2}(y_1 + y_2)$ und $u_2(y_1, y_2) = \frac{1}{2}(y_1 - y_2)$ Die Ableitung ist dann

$$J(y_1, y_2) = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} .$$

Wir haben $\det J(\boldsymbol{y}) = -\frac{1}{2}$. Also erhalten wir für die Dichte von (Y_1, Y_2)

$$f_Y(y_1, y_2) = \tfrac{1}{2} f(\tfrac{1}{2}(y_1 + y_2), \tfrac{1}{2}(y_1 - y_2)) .$$

Integrieren wir bezüglich y_2 erhalten wir, für $y_2 = 2z - y_1$, dass $Y_1 = X_1 + X_2$ die Dichte

$$\int f_Y(y_1, y_2) \, \mathrm{d}y_2 = \int f_X(z, y_1 - z) \, \mathrm{d}z$$

hat. Dies stimmt mit der Formel aus Hilfssatz 2.26 überein.

2.7. Bedingte Verteilungen

Seien X, Y Zufallsvariablen mit der gemeinsamen Verteilung $F(x, y)$. Wir wollen nun das Problem betrachten, wie die Verteilung von X aussieht, falls wir Y beobachtet haben, das heisst, wir suchen $\mathbb{P}[X \leq x \mid Y = y]$. Ist $\mathbb{P}[Y = y] > 0$, dann können wir die früher eingeführte Formel

$$\mathbb{P}[X \leq x \mid Y = y] = \frac{\mathbb{P}[X \leq x, Y = y]}{\mathbb{P}[Y = y]}$$

verwenden. Wir wollen nun absolutsteige Verteilungen betrachten. Das Problem ist, dass $\mathbb{P}[Y = y] = 0$.

Nehmen wir an, dass die Dichte von Y stetig und an der Stelle y verschieden von Null ist, $f_Y(y) > 0$. Dann ist $\mathbb{P}[y - \varepsilon < Y < y + \varepsilon] > 0$. Wir können also die bedingte Verteilung

$$\mathbb{P}[X \leq x \mid y - \varepsilon < Y < y + \varepsilon] = \frac{\mathbb{P}[X \leq x, y - \varepsilon < Y < y + \varepsilon]}{\mathbb{P}[y - \varepsilon < Y < y + \varepsilon]}$$

berechnen. Wir schreiben

$$\frac{\int_{y-\varepsilon}^{y+\varepsilon} \int_{-\infty}^{x} f(v, w) \, \mathrm{d}v \, \mathrm{d}w}{\int_{y-\varepsilon}^{y+\varepsilon} f_Y(w) \, \mathrm{d}w} = \frac{\frac{1}{2\varepsilon} \int_{y-\varepsilon}^{y+\varepsilon} \int_{-\infty}^{x} f(v, w) \, \mathrm{d}v \, \mathrm{d}w}{\frac{1}{2\varepsilon} \int_{y-\varepsilon}^{y+\varepsilon} f_Y(w) \, \mathrm{d}w} .$$

Lassen wir nun ε nach 0 gehen, erhalten wir

$$\mathbb{P}[X \leq x \mid Y = y] = \frac{\int_{-\infty}^{x} f(v,y)\,\mathrm{d}v}{f_Y(y)} \ .$$

Wir sehen, die bedingte Verteilung von X gegeben $\{Y = y\}$ ist absolutstetig mit der Dichte

$$f(x \mid y) = \frac{f(x,y)}{f_Y(y)} \ .$$

Letztere Formel kann man für allgemeine Dichten $f(x,y)$ beweisen. Insbesondere ist für unabhängige X, Y, $f(x \mid y) = f_X(x)$. Weiter gilt die Bayes'sche Regel

$$f(y \mid x) = \frac{f(x,y)}{f_X(x)} = \frac{f(x \mid y) f_Y(y)}{f_X(x)} \ .$$

Die bedingte Verteilung (stetig und diskret) hat die folgenden beiden Eigenschaften:

i) Für jedes A ist die Abbildung $y \mapsto \mathbb{P}[X \in A \mid Y = y]$ messbar.

ii) Für jedes A, B gilt

$$\mathbb{P}[X \in A, Y \in B] = \int_B \mathbb{P}[X \in A \mid Y = y]\,\mathrm{d}F_Y(y) \ .$$

Generell kann man die bedingte Verteilung über die obigen zwei Eigenschaften definieren.

Beispiele

- Betrachten wir die zweidimensionale Normalverteilung

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\Big\{-\frac{1}{2(1-\rho^2)}\Big(\frac{\xi_X^2}{\sigma_X^2} - 2\rho\frac{\xi_X\xi_Y}{\sigma_X\sigma_Y} + \frac{\xi_Y^2}{\sigma_Y^2}\Big)\Big\} \ ,$$

  wobei $\xi_X = x - \mu_X$, $\xi_Y = y - \mu_Y$, $\mu_i \in \mathbb{R}$, $\sigma_i > 0$ $(i \in \{X, Y\})$ und $|\rho| < 1$. Für die bedingte Verteilung erhalten wir

$$\begin{aligned} f(x \mid y) &= \frac{\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\{-\frac{1}{2(1-\rho^2)}(\frac{\xi_X^2}{\sigma_X^2} - 2\rho\frac{\xi_X\xi_Y}{\sigma_X\sigma_Y} + \frac{\xi_Y^2}{\sigma_Y^2})\}}{\frac{1}{\sigma_Y\sqrt{2\pi}} \exp\{-\frac{1}{2}\frac{\xi_Y^2}{\sigma_Y^2}\}} \\ &= \frac{1}{\sigma_X\sqrt{2\pi(1-\rho^2)}} \exp\Big\{-\frac{1}{2\sigma_X^2(1-\rho)^2}\Big(\xi_X - \rho\frac{\sigma_X}{\sigma_Y}\xi_Y\Big)^2\Big\} \ . \end{aligned}$$

  Somit ist die bedingte Verteilung von X gegeben Y normalverteilt mit Varianz $(1-\rho^2)\sigma_X^2$ und Mittelwert $\mu_X + \rho\sigma_X(Y-\mu_Y)/\sigma_Y$. Das Resultat hätte man leichter erhalten können, wenn man $X = \mu_X + \sigma_X(\rho V + \sqrt{1-\rho^2}W)$ und $Y = \mu_Y + \sigma_Y V$ für unabhängige standardnormalverteilte V, W gesetzt hätte.

- Seien X und $\tilde{X}$ unabhängige exponential verteilte Zufallsvariablen mit Parameter α und $Y = X + \tilde{X}$. Die gemeinsame Dichte von X und Y ist

$$f(x,y) = \alpha^2 \exp\{-\alpha x - \alpha(y-x)\} \mathbb{I}_{0<x<y} = \alpha^2 \exp\{-\alpha y\} \mathbb{I}_{0<x<y} \, .$$

  Somit haben wir für die bedingte Wahrscheinlichkeit gegeben $\{Y = y\}$

$$f(x \mid y) = \frac{\alpha^2 \exp\{-\alpha y\}}{\alpha^2 y \exp\{-\alpha y\}} \mathbb{I}_{0<x<y} = \frac{1}{y} \mathbb{I}_{0<x<y} \, .$$

  Also ist X bedingt auf Y gleichverteilt auf $[0, Y]$.

- Betrachten wir ein Beispiel, in dem eine diskrete Verteilung mit einer absolutstetigen Verteilung gemischt wird. Sei Λ eine Gamma-verteilte Zufallsvariable mit Parametern γ und α. Gegeben $\{\Lambda = \lambda\}$ sei die bedingte Verteilung von N eine Poisson-Verteilung mit Parameter λ. Die unbedingte Verteilung von N ist dann

$$\mathbb{P}[N = n] = \int_0^\infty \frac{\lambda^n}{n!} \mathrm{e}^{-\lambda} \frac{\alpha^\gamma}{\Gamma(\gamma)} \lambda^{\gamma-1} \mathrm{e}^{-\alpha\lambda} \, \mathrm{d}\lambda = \frac{\Gamma(\gamma+n)}{n! \, \Gamma(\gamma)} \Big(\frac{\alpha}{\alpha+1}\Big)^\gamma \Big(\frac{1}{\alpha+1}\Big)^n \, .$$

  Diese Verteilung heisst **negative Binomialverteilung** mit Parametern γ und $p = 1/(\alpha+1)$. Haben wir nun $\{N = n\}$ beobachtet, folgt für die Verteilung von Λ

$$\begin{aligned} \mathbb{P}[\Lambda \le \ell \mid N = n] &= \frac{\int_0^\ell (n! \, \Gamma(\gamma))^{-1} \alpha^\gamma \lambda^{\gamma+n-1} \mathrm{e}^{-(\alpha+1)\lambda} \, \mathrm{d}\lambda}{(n! \, \Gamma(\gamma))^{-1} \Gamma(\gamma+n) \alpha^\gamma (\alpha+1)^{-(\gamma+n)}} \\ &= \int_0^\ell \frac{(\alpha+1)^{\gamma+n}}{\Gamma(\gamma+n)} \lambda^{\gamma+n-1} \mathrm{e}^{-(\alpha+1)\lambda} \, \mathrm{d}\lambda \, . \end{aligned}$$

  Somit ist Λ bedingt auf $\{N = n\}$ Gamma-verteilt mit Parametern $\gamma + n$ und $\alpha + 1$.

Definition 2.28. *Seien X und Y Zufallsvariablen und X sei absolutstetig. Die bedingte Erwartung von X gegeben Y ist die Zufallsvariable*

$$\mathbb{E}[X \mid Y] = \int x f(x \mid Y) \, \mathrm{d}x \, .$$

Man beachte, dass die rechte Seite von Y abhängt, also zufällig ist.

Wir haben die folgende Eigenschaft der bedingten Erwartung.

Hilfssatz 2.29. *Seien X und Y Zufallsvariablen und $h(y)$ eine messbare reelle Funktion, so dass die folgenden Erwartungswerte wohldefiniert sind. Dann gilt*

$$\mathbb{E}[h(Y)X] = \mathbb{E}[h(Y)\mathbb{E}[X \mid Y]] \, .$$

Beweis. Wir beweisen den Hilfssatz nur im absolutstetigen Fall. Dann haben wir

$$\begin{aligned}\mathbb{E}[h(Y)\mathbb{E}[X \mid Y]] &= \int h(y) \int x \frac{f(x,y)}{f_Y(y)} \, \mathrm{d}x f_Y(y) \, \mathrm{d}y \\ &= \iint h(y) x f(x,y) \, \mathrm{d}x \, \mathrm{d}y = \mathbb{E}[h(Y)X] \, .\end{aligned}$$

□

Wir haben nun die folgende Interpretation der bedingten Erwartung.

Proposition 2.30. *Seien X, Y Zufallsvariablen, so dass $\mathbb{E}[X^2] < \infty$. Für jede messbare Funktion $h(y)$ gilt*

$$\mathbb{E}[(X - \mathbb{E}[X \mid Y])^2] \leq \mathbb{E}[(X - h(Y))^2] \, .$$

Beweis. Wir können annehmen, dass $\mathbb{E}[(h(Y))^2] < \infty$. Wir erhalten

$$\begin{aligned}\mathbb{E}[(X - h(Y))^2] = \mathbb{E}[(X - \mathbb{E}[X \mid Y])^2] &+ \mathbb{E}[(\mathbb{E}[X \mid Y] - h(Y))^2] \\ &+ 2\mathbb{E}[(\mathbb{E}[X \mid Y] - h(Y))(X - \mathbb{E}[X \mid Y])] \, .\end{aligned}$$

Setzen wir $g(Y) = \mathbb{E}[X \mid Y] - h(Y)$, erhalten wir

$$\mathbb{E}[g(Y)(X - \mathbb{E}[X \mid Y])] = \mathbb{E}[g(Y)X] - \mathbb{E}[g(Y)\mathbb{E}[X \mid Y]] = 0 \, .$$

Somit ist $\mathbb{E}[(X - h(Y))^2] \geq \mathbb{E}[(X - \mathbb{E}[X \mid Y])^2]$ und Gleichheit gilt genau dann, wenn $\mathbb{P}[h(Y) = \mathbb{E}[X \mid Y]] = 1$. □

Wir können somit sagen, dass $\mathbb{E}[X \mid Y]$ die beste Prognose von X ist, wenn man Y beobachtet.

3. Grenzwertsätze

Sei nun $\{X_1, X_2, \ldots\}$ eine Folge von Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \mathbb{P})$. Wir interessieren uns nun für die Summe $S_n = X_1 + \cdots + X_n$, und vor allem für die Asymptotik $n \to \infty$. Zum einen wollen wir $n^{-1}S_n$ betrachten (Gesetz der grossen Zahl) und die Form der Verteilung von S_n bestimmen (zentraler Grenzwertsatz).

3.1. Schwaches Gesetz der grossen Zahl

Satz 3.1. *Seien $\mathbb{E}[X_i] = \mu$ unabhängig von i und die Varianzen im Schnitt beschränkt, $\sup_n n^{-1}\sum_{i=1}^n \mathrm{Var}[X_i] < \infty$. Sind die Zufallsvariablen $\{X_i\}$ unkorreliert, so gilt*

$$\lim_{n\to\infty} \mathbb{P}\Big[\Big|\frac{S_n}{n} - \mu\Big| \geq \varepsilon\Big] = 0$$

für jedes $\varepsilon > 0$.

Beweis. Wir haben $\mathbb{E}[n^{-1}S_n] = n^{-1}\sum_{k=1}^n \mathbb{E}[X_k] = \mu$, und, wegen der Unkorreliertheit, $\mathrm{Var}[n^{-1}S_n] = n^{-2}\sum_{k=1}^n \mathrm{Var}[X_k] \to 0$. Somit folgt das Resultat aus der Chebychev Ungleichung (2.1). □

Sind die Zufallsvariablen $\{X_k\}$ unabhängig und identisch verteilt, dann sind die Bedingungen des Satzes erfüllt. Machen wir Zufallsexperimente unabhängig voneinander, haben wir nun die Intuition, mit der wir Wahrscheinlichkeiten eingeführt haben, auch formal bewiesen.

Beispiele

- Für unabhängige $\{0,1\}$ Experimente mit Erfolgsparameter p hat Jacob Bernoulli 1713 durch kombinatorische Argumente bewiesen, dass $\mathbb{P}[|n^{-1}S_n - p| \geq \varepsilon] \to 0$. Ist n gross, hat man also ungefähr np Erfolge und $n(1-p)$ Misserfolge.

  Sei $f(x) : [0,1] \to \mathbb{R}$ eine stetige Funktion. Wir definieren die *Bernstein-Polynome*

  $$B_n(x) := \sum_{k=0}^n f(k/n)\binom{n}{k}x^k(1-x)^{n-k} .$$

  Wir erhalten dann die Abschätzung

  $$|B_n(p) - f(p)| = |\mathbb{E}[f(S_n/n) - f(p)]| \overset{\text{Jensen}}{\leq} \mathbb{E}[|f(S_n/n) - f(p)|] .$$

Setzen wir $\|f\| = \sup_x |f(x)|$, erhalten wir

$$|B_n(p) - f(p)| \leq 2\|f\|\mathbb{P}[|S_n/n - p| \geq \varepsilon] + \sup_{|x-y|\leq\varepsilon} |f(x) - f(y)|\mathbb{P}[|S_n/n - p| < \varepsilon] \; .$$

Aus dem schwachen Gesetz der grossen Zahl folgt, dass der erste Term gegen Null konvergiert. Aus der Chebychev Ungleichung kann man schliessen, dass die Konvergenz gleichmässig in p ist. Der zweite Term kann durch die Wahl von ε gleichmässig beliebig klein gemacht werden, da jede stetige Funktion gleichmässig stetig ist. Somit konvergieren die Bernstein Polynome gleichmässig gegen die Funktion $f(x)$.

- Seien $\{X_i\}$ unabhängige Experimente mit verschiedenen Erfolgsparameter p_i. Setzen wir $\tilde{X}_i = X_i - p_i$, dann haben die $\{\tilde{X}_i\}$ den gemeinsamen Mittelwert 0 und die Varianz $\mathrm{Var}[\tilde{X}_i] = \mathrm{Var}[X_i] = p_i(1-p_i) \leq \frac{1}{4}$. Also gilt

$$\mathbb{P}\Big[\Big|\frac{S_n}{n} - \frac{\sum_{k=1}^n p_k}{n}\Big| \geq \varepsilon\Big] = \mathbb{P}\Big[\Big|\frac{\sum_{k=1}^n \tilde{X}_k}{n}\Big| \geq \varepsilon\Big] \to 0 \; .$$

  Auch bei verschiedenen Erfolgsparameter nähert sich der Durchschnitt immer mehr dem Mittelwert an.

- In den Anwendungen braucht man oft einen Ausdruck der Form $\mathbb{E}[f(X)]$, wobei $f(x)$ eine stetige Funktion ist, und X eine Zufallsvariable (z.B. Optionspreis). Oft ist es schwer $\mathbb{E}[f(X)]$ auszurechnen, aber relativ einfach, X auf einem Computer zu simulieren. Man erzeugt sich dann n unabhängige Zufallsvariablen $\{X_k\}$ mit der gleichen Verteilung wie X. Da $n^{-1}\sum_{k=1}^n f(X_k)$ sich immer mehr $\mathbb{E}[f(X)]$ annähert, gibt dieses Verfahren mit hoher Wahrscheinlichkeit eine gute Approximation von $\mathbb{E}[f(X)]$. Dieses Verfahren heisst **Monte–Carlo Simulation**.

  Will man ein Integral $\int_0^1 f(x)\,\mathrm{d}x$ numerisch berechnen, hat man manchmal Probleme, falls $f(x)$ nicht eine schöne Funktion ist. Man bemerkt, dass für unabhängige und auf $[0,1]$ gleichverteilte Zufallsvariablen $\{X_k\}$ der Mittelwert $\mathbb{E}[f(X_k)] = \int_0^1 f(x)\,\mathrm{d}x$ gleich dem gesuchten Integral ist. Daher lässt sich das Integral mit der Monte–Carlo Simulation $n^{-1}\sum_{k=1}^n f(X_k)$ approximieren. Der Vorteil dieser Methode ist, dass die Integrationsdiskretisierung nicht regelmässig ist, das heisst, nicht $n^{-1}\sum_{k=1}^n f(k/n)$.

3.2. Konvergenzbegriffe

Seien nun $\{X_i\}$ und X Zufallsvariablen auf $(\Omega, \mathcal{F}, \mathbb{P})$. Wir definieren nun verschiedene Arten von Konvergenz von X_n nach X.

- **Stochastische Konvergenz** Wir sagen X_n konvergiert stochastisch gegen X, $X_n \xrightarrow{\mathbb{P}} X$, falls
$$\lim_{n\to\infty} \mathbb{P}[|X_n - X| \geq \varepsilon] = 0$$
für alle $\varepsilon > 0$.

- **Fast sichere Konvergenz** Wir sagen X_n konvergiert fast sicher gegen X, $X_n \to X$, falls
$$\mathbb{P}[\{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}] = 1 \, .$$

- **$\mathcal{L}^p$-Konvergenz**, $p \geq 1$ Wir sagen, X_n konvergiert in $\mathcal{L}^p$ gegen X, falls
$$\lim_{n\to\infty} \mathbb{E}[|X_n - X|^p] = 0 \, .$$

- **Konvergenz in Verteilung** Wir sagen X_n konvergiert in Verteilung gegen X, $X_n \xrightarrow{\mathrm{d}} X$, falls
$$\lim_{n\to\infty} \mathbb{P}[X_n \leq x] = \mathbb{P}[X \leq x]$$
für alle $x \in \mathbb{R}$, an denen $F_X(x)$ stetig ist. Dieser Konvergenzbegriff betrachtet nur die Verteilungen. Zum Beispiel sind $\{X_k\}$ unabhängig und identisch verteilt, dann konvergiert X_n in Verteilung gegen X_1. Dieser Konvergenzbegriff kann daher nur verwendet werden, wenn wir uns nicht für $\lim_{n\to\infty} X_n$ interessieren, sondern für die Verteilungen.

Wir wollen die Konvergenzbegriffe nun vergleichen. Wir konzentrieren uns dabei auf die ersten drei Begriffe, da der letzte Konvergenzbegriff von einer anderen Art ist.

Proposition 3.2.

i) *"Fast sichere Konvergenz" impliziert "stochastische Konvergenz."*

ii) *"$\mathcal{L}^p$-Konvergenz" impliziert "stochastische Konvergenz."*

iii) *Für $q > p$ impliziert "$\mathcal{L}^q$-Konvergenz" die "$\mathcal{L}^p$-Konvergenz."*

iv) *Ist $\mathbb{E}[(\sup_n |X_n|)^p] < \infty$, so folgt die "$\mathcal{L}^p$-Konvergenz" aus der "fast sicheren Konvergenz."*

v) *Sei für jedes $\varepsilon > 0$*
$$\sum_{n=1}^{\infty} \mathbb{P}[|X_n - X| \geq \varepsilon] < \infty \, .$$

Dann konvergiert X_n sowohl stochastisch als auch fast sicher gegen X. Insbesondere hat jede stochastisch konvergierende Folge eine fast sicher konvergierende Teilfolge.

Beweis. i) Die fast sichere Konvergenz ist gleichbedeutend mit

$$\mathbb{P}[\cap_k \cup_m \cap_{n\geq m}\{|X_n - X| \leq k^{-1}\}] = 1$$

(für alle k gibt es ein m, so dass für alle $n \geq m$, $|X_n - X| \leq k^{-1}$ gilt). Also gilt

$$\mathbb{P}[\cup_m \cap_{n\geq m} \{|X_n - X| \leq \ell^{-1}\}] \geq \mathbb{P}[\cap_k \cup_m \cap_{n\geq m}\{|X_n - X| \leq k^{-1}\}] = 1 \; .$$

Wegen der Monotonie in m haben wir weiter

$$1 = \mathbb{P}[\cup_m \cap_{n\geq m} \{|X_n - X| \leq k^{-1}\}] = \lim_{m\to\infty} \mathbb{P}[\cap_{n\geq m}\{|X_n - X| \leq k^{-1}\}]$$

für alle k. Wir können k^{-1} durch ε ersetzen. Also haben wir

$$\lim_{m\to\infty} \mathbb{P}[|X_m - X| \leq \varepsilon] \geq \lim_{m\to\infty} \mathbb{P}[\cap_{n\geq m}\{|X_n - X| \leq \varepsilon\}] = 1 \; .$$

Dies ist die stochastische Konvergenz.

ii) Dies folgt sofort mittels Hilfssatz 2.16 aus

$$\mathbb{P}[|X_n - X| \geq \varepsilon] \leq \frac{\mathbb{E}[|X_n - X|^p]}{\varepsilon^p} \; .$$

iii) Dies folgt sofort mittels Korollar 2.15 aus

$$\mathbb{E}[|X_n - X|^p]^{1/p} \leq \mathbb{E}[|X_n - X|^q]^{1/q} \; .$$

iv) Dies folgt aus der Eigenschaft der beschränkten Konvergenz.

v) Aus dem Borel–Cantelli-Lemma folgt, dass

$$\mathbb{P}[\{|X_n - X| \geq \varepsilon \text{ unendlich oft}\}] = 0 \; .$$

Sei nun $\{\varepsilon_m\}$ eine Folge von echt positiven Zahlen, die monoton gegen Null konvergiert. Dann ist

$$\begin{aligned}\mathbb{P}[\cup_m\{|X_n - X| \geq \varepsilon_m \text{ unendlich oft}\}] &\leq \sum_m \mathbb{P}[\{|X_n - X| \geq \varepsilon_m \text{ unendlich oft}\}] \\ &= 0 \; .\end{aligned}$$

Also konvergiert X_n fast sicher gegen X.

Konvergiert X_n stochastisch gegen X, so wählen wir eine steigende Folge n_k, so dass $\mathbb{P}[|X_{n_k} - X| \geq k^{-1}] < k^{-2}$. Dann erfüllt $\{X_{n_k} : k \in \mathbb{N}\}$ die Bedingung, und konvergiert somit fast sicher gegen X. □

Beispiele

- Sei $\mathbb{P}$ die Gleichverteilung auf $[0,1]$. Für $n \geq 1$ und $k \in \{0,1,\ldots,2^n-1\}$ definieren wir $Z_{n,k} = \mathbb{1}_{(k2^{-n},(k+1)2^{-n}]}$. Wir lassen nun $X_1 = Z_{1,0}$, $X_2 = Z_{1,1}$, $X_3 = Z_{2,0}$, etc., das heisst, wir zählen lexikographisch ab. Da immer wieder eine 1 auftritt, haben wir $\underline{\lim}\, X_n = 0$ und $\overline{\lim}\, X_n = 1$. Also kann X_n nicht fast sicher konvergieren. Aber für $\varepsilon \in (0,1)$ haben wir

$$\mathbb{P}[|Z_{n,k}| \geq \varepsilon] = \mathbb{P}[Z_{n,k} = 1] = \mathbb{E}[Z_{n,k}^p] = 2^{-n} \; .$$

  Somit konvergiert X_n stochastisch und in $\mathcal{L}^p$ gegen 0.

- Sei $\mathbb{P}$ die Gleichverteilung auf $[0,1]$. Wir definieren $X_n = 2^n \mathbb{1}_{[0,2^{-n}]}$. Da $\mathbb{P}[\omega > 0] = 1$, erhalten wir, dass X_n fast sicher gegen 0 konvergiert. Aber

$$\mathbb{E}[X_n] = 2^n \mathbb{P}[[0,2^{-n}]] = 2^n 2^{-n} = 1 \; .$$

  Somit konvergiert X_n nicht in $\mathcal{L}^1$ gegen 0, und damit auch nicht in $\mathcal{L}^p$.

Hilfssatz 3.3. *Seien $\{X_n\}$ und X Zufallsvariablen. Folgende Aussagen sind äquivalent:*

i) *X_n konvergiert in Verteilung gegen X.*

ii) *Für jede stetige beschränkte Funktion $f(x)$ gilt*

$$\lim_{n\to\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)] \; .$$

iii) *Für jede dreimal stetig differenzierbare und beschränkte Funktion $f(x)$ mit beschränkten ersten drei Ableitungen gilt*

$$\lim_{n\to\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)] \; .$$

Beweis. "i) $\Rightarrow$ ii)" Es gibt nur abzählbar viele Punkte, an denen $F(x)$ nicht stetig ist. Somit gilt für alle Intervalle der Form $(y,z]$, wobei $y < z$ und $F(x)$ ist stetig in y und z, dass die Aussage für Funktionen der Form $f(x) = c\mathbb{1}_{(y,z]}(x)$ gilt. Wir können nun so Ober- und Untersummen wie beim Riemann-Integral bilden, und die Aussage mit Hilfe der Monotonieeigenschaft des Erwartungswertes beweisen.
"ii) $\Rightarrow$ iii)" trivial.
"iii) $\Rightarrow$ i)" Betrachten wir die Funktion

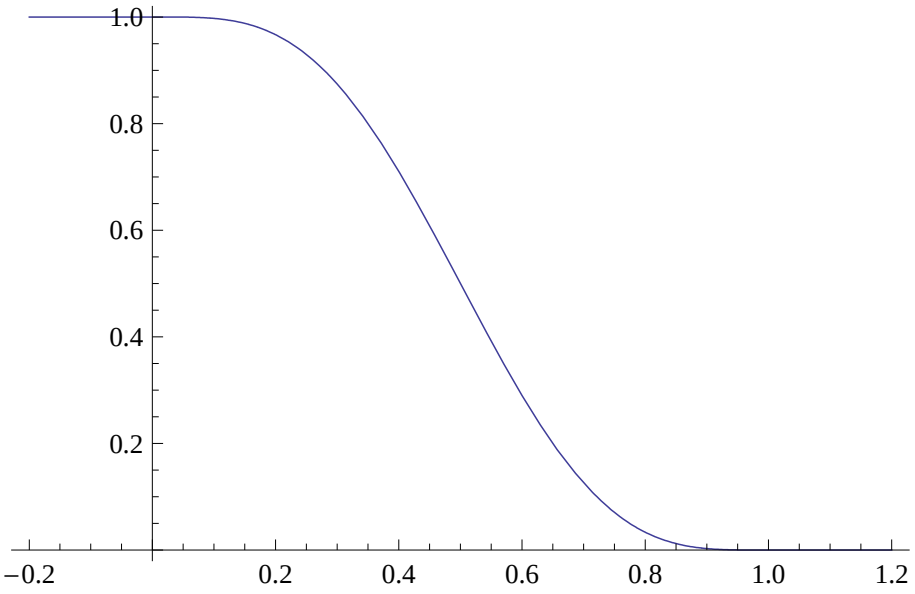

Abbildung 3.1: *Die Funktion* $P(x)$

$$P(x) = \begin{cases} 1\,, & \text{falls } x \le 0, \\ 0\,, & \text{falls } x \ge 1, \\ 20x^7 - 70x^6 + 84x^5 - 35x^4 + 1\,, & \text{sonst,} \end{cases}$$

dargestellt in Abbildung 3.1. Wir haben $P(0) = 1$ und $P(1) = 0$, die ersten drei Ableitungen in 0 sind Null, und die ersten drei Ableitungen in 1 sind 0. Somit ist $P(x)$ dreimal stetig differenzierbar. Aus $P'(x) = 140x^3(x-1)^3$ folgt, dass $P(x)$ im Intervall $[0,1]$ fallend ist. Sei y ein Punkt, an dem $F(x)$ stetig ist. Dann gilt

$$\mathbb{1}_{(-\infty,y]}(x) \le P((x-y)/\delta) \le \mathbb{1}_{(-\infty,y+\delta]}(x)\,.$$

Also erhalten wir

$$\overline{\lim_{n\to\infty}}\, F_n(y) \le \lim_{n\to\infty} \mathbb{E}[P((X_n - y)/\delta)] = \mathbb{E}[P((X-y)/\delta)] \le F(y+\delta)\,.$$

Da δ beliebig war, gilt $\overline{\lim}_{n\to\infty} F_n(y) \le F(y)$. Weiter gilt

$$\underline{\lim_{n\to\infty}}\, F_n(y) \ge \lim_{n\to\infty} \mathbb{E}[P((X_n - y + \delta)/\delta)] = \mathbb{E}[P((X-y+\delta)/\delta)] \ge F(y-\delta)\,.$$

Da δ beliebig war, gilt $\underline{\lim}_{n\to\infty} F_n(y) \ge F(y)$. □

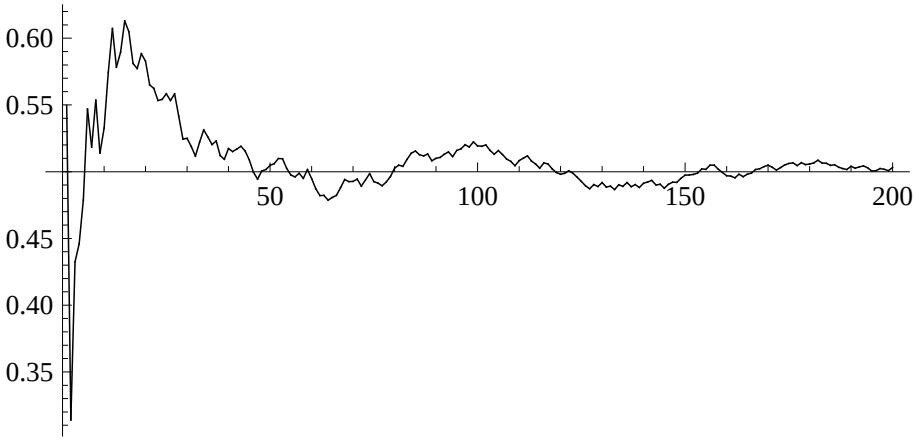

Abbildung 3.2: *Typisches S_n/n für uniform auf $[0,1]$ verteilte $\{X_k\}$*

3.3. Starkes Gesetz der grossen Zahl

Wir haben gesehen, dass unter dem schwachen Gesetz der grossen Zahl der Durchschnitt von n Zufallsvariablen mit hoher Wahrscheinlichkeit nahe bei ihrem Mittelwert liegt. Wir wollen dieses Gesetz nun verschärfen und betrachten daher S_n/n für ein ω. Ein möglicher Pfad ist in Abbildung 3.2 illustriert.

Satz 3.4. *Seien $\{X_k\}$ unabhängig mit festem Erwartungswert $\mathbb{E}[X_k] = \mu$. Weiter gelte eine der folgenden Bedingungen:*

i) *$\{X_k\}$ seien identisch verteilt.*

ii) *Es gelte $\sup_k \mathbb{E}[X_k^4] < \infty$.*

Dann konvergiert S_n/n fast sicher gegen μ.

Beweis. i) Teilen wir $X_k = X_k^+ - X_k^-$ in positiven und negativen Teil auf, dann genügt es, den Satz für positive Zufallsvariablen zu beweisen. Lassen wir $\tilde{X}_n = X_n \mathbb{1}_{X_n \leq n}$ und $\tilde{S}_n = \sum_{k=1}^n \tilde{X}_k$. Wir zeigen zuerst, dass

$$\lim_{n\to\infty} \frac{\tilde{S}_n - \mathbb{E}[\tilde{S}_n]}{n} = 0 \, .$$

Sei $\alpha > 1$ und $k_n = \lfloor \alpha^n \rfloor$. Wir wählen nun $\varepsilon > 0$. Es folgt aus der Chebychev-

Ungleichung

$$\sum_{n=1}^{\infty} \mathbb{P}\Big[\frac{|\tilde{S}_{k_n} - \mathbb{E}[\tilde{S}_{k_n}]|}{k_n} \geq \varepsilon\Big] \leq \sum_{n=1}^{\infty} \frac{\mathrm{Var}[\tilde{S}_{k_n}]}{\varepsilon^2 k_n^2} = \sum_{n=1}^{\infty} \frac{1}{\varepsilon^2 k_n^2} \sum_{m=1}^{k_n} \mathrm{Var}[\tilde{X}_m]$$
$$= \frac{1}{\varepsilon^2} \sum_{m=1}^{\infty} \mathrm{Var}[\tilde{X}_m] \sum_{n:k_n \geq m} \frac{1}{k_n^2} \, .$$

Da $\sum_{n=\ell}^{\infty} (\alpha^n)^{-2} = \alpha^{-2\ell}/(1-\alpha^{-2})$, gibt es eine Konstante c_α, so dass $\sum_{n:k_n \geq m} k_n^{-2} \leq c_\alpha m^{-2}$. Damit erhalten wir

$$\sum_{n=1}^{\infty} \mathbb{P}\Big[\frac{|\tilde{S}_{k_n} - \mathbb{E}[\tilde{S}_{k_n}]|}{k_n} \geq \varepsilon\Big] \leq \frac{c_\alpha}{\varepsilon^2} \sum_{m=1}^{\infty} \frac{\mathbb{E}[\tilde{X}_m^2]}{m^2} = \frac{c_\alpha}{\varepsilon^2} \sum_{m=1}^{\infty} m^{-2} \sum_{\ell=0}^{m-1} \int_\ell^{\ell+1} x^2 \, \mathrm{d}F(x)$$
$$= \frac{c_\alpha}{\varepsilon^2} \sum_{\ell=0}^{\infty} \sum_{m=\ell+1}^{\infty} m^{-2} \int_\ell^{\ell+1} x^2 \, \mathrm{d}F(x)$$
$$\leq A + \frac{c_\alpha}{\varepsilon^2} \sum_{\ell=1}^{\infty} \ell^{-1} \int_\ell^{\ell+1} x^2 \, \mathrm{d}F(x)$$
$$\leq A + \frac{2c_\alpha}{\varepsilon^2} \sum_{\ell=1}^{\infty} \int_\ell^{\ell+1} x \, \mathrm{d}F(x) < \infty \, ,$$

wobei A den Term für $\ell = 0$ bezeichnet. Somit schliessen wir aus Proposition 3.2 v), dass $k_n^{-1}(\tilde{S}_{k_n} - \mathbb{E}[\tilde{S}_{k_n}])$ fast sicher gegen 0 konvergiert. Da

$$\frac{\mathbb{E}[\tilde{S}_n]}{n} = \frac{1}{n} \sum_{k=1}^{n} \int_0^k x \, \mathrm{d}F(x) \to \int_0^\infty x \, \mathrm{d}F(x) = \mu \, ,$$

konvergiert also $k_n^{-1}\tilde{S}_{k_n}$ fast sicher gegen μ. Sei nun $n \in [k_m, k_{m+1})$. Dann gilt

$$\frac{k_m}{k_{m+1}} \frac{\tilde{S}_{k_m}}{k_m} = \frac{\tilde{S}_{k_m}}{k_{m+1}} \leq \frac{\tilde{S}_n}{n} \leq \frac{\tilde{S}_{k_{m+1}}}{k_m} = \frac{k_{m+1}}{k_m} \frac{\tilde{S}_{k_{m+1}}}{k_{m+1}} \, .$$

Lassen wir n gegen Unendlich streben, erhalten wir

$$\frac{1}{\alpha}\mu \leq \underline{\lim_{n\to\infty}} \frac{\tilde{S}_n}{n} \leq \overline{\lim_{n\to\infty}} \frac{\tilde{S}_n}{n} \leq \alpha\mu \, .$$

Da $\alpha > 1$ beliebig war, folgt dass $n^{-1}\tilde{S}_n \to \mu$.

Betrachten wir nun, wie oft das Ereignis $\{\tilde{X}_n \neq X_n\}$ eintritt. Wir haben

$$\sum_{n=1}^{\infty} \mathbb{P}[\tilde{X}_n \neq X_n] = \sum_{n=1}^{\infty} \mathbb{P}[X_n > n] = \sum_{n=1}^{\infty} \sum_{m=n}^{\infty} \mathbb{P}[X_n \in (m, m+1]]$$
$$= \sum_{m=1}^{\infty} \sum_{n=1}^{m} \mathbb{P}[X_n \in (m, m+1]] = \sum_{m=1}^{\infty} m\mathbb{P}[X_1 \in (m, m+1]]$$
$$\leq \mathbb{E}[X_1] < \infty \, .$$

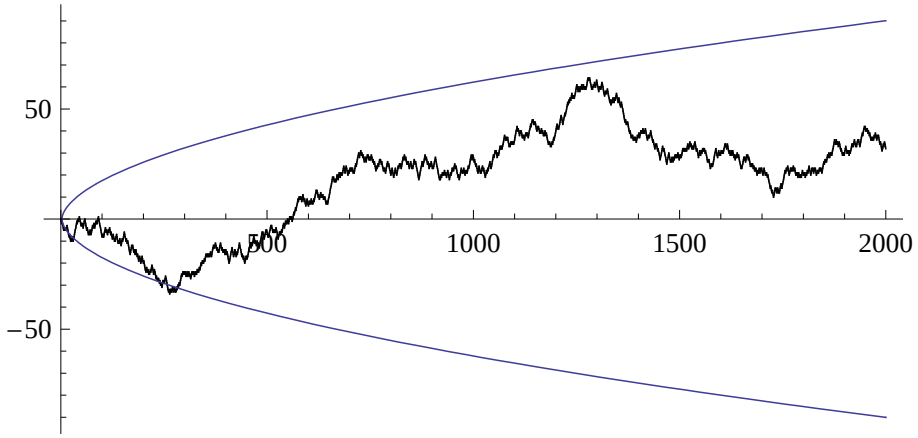

Abbildung 3.3: *Irrfahrt und die Grenzen des iterierten Logarithmus*

Somit folgt aus dem Borel–Cantelli Lemma, dass $\{\tilde{X}_n \neq X_n\}$ nur endlich oft eintritt. Insbesondere gilt, dass $n^{-1}S_n$ und $n^{-1}\tilde{S}_n$ den gleichen Grenzwert μ haben.
ii) Wir können ohne Beschränkung der Allgemeinheit annehmen, dass $\mu = 0$. Wir erhalten

$$\mathbb{E}[X_i^2]^2 \leq \mathbb{E}[X_i^4] \leq M = \sup_k \mathbb{E}[X_k^4] \; .$$

Für S_n ergibt sich die Abschätzung

$$\mathbb{E}[S_n^4] = \sum_{i,j,k,\ell=1}^{n} \mathbb{E}[X_i X_j X_k X_\ell] \leq nM + 6\frac{n(n-1)}{2}M + 0 \leq 3n^2 M \; .$$

Also haben wir mit Hilfe von Hilfssatz 2.16

$$\mathbb{P}\Big[\Big|\frac{S_n}{n}\Big| \geq \varepsilon\Big] \leq \frac{\mathbb{E}[(n^{-1}S_n)^4]}{\varepsilon^4} \leq \frac{3n^2 M}{\varepsilon^4 n^4} = \frac{3M}{\varepsilon^4 n^2} \; .$$

Letzterer Ausdruck ist summierbar, also können wir folgern aus Proposition 3.2 v), dass $n^{-1}S_n$ fast sicher gegen Null konvergiert. □

Für viele Situationen ist die Bedingung $\mathbb{E}[X_k^4] \leq M$ erfüllt. Zum Beispiel bei unabhängigen $\{0,1\}$ Experimenten mit Erfolgsparameter p_i. Somit erhält man, dass $n^{-1}S_n - n^{-1}\mathbb{E}[S_n]$ fast sicher gegen Null konvergiert.

Das starke Gesetz der grossen Zahl gibt uns damit eine Schranke, wie schnell die Summe S_n wachsen kann. Ist $\mu = 0$, finden wir dass $|S_n| \leq \varepsilon n$, falls n gross genug ist. Genauere Grenzen für S_n hat Alexander Jakowlewitsch Khintchine gefunden.

Satz 3.5. (Gesetz vom iterierten Logarithmus) *Seien $\{X_k\}$ unabhängig und identisch verteilt mit Mittelwert $\mathbb{E}[X_k] = 0$ und Varianz $\sigma^2 = \mathrm{Var}[X_k] < \infty$. Dann gilt*

$$\overline{\lim_{n\to\infty}} \frac{S_n}{\sqrt{2\sigma^2 n \log\log n}} = 1$$

und

$$\underline{\lim_{n\to\infty}} \frac{S_n}{\sqrt{2\sigma^2 n \log\log n}} = -1 \ .$$

□

3.4. Zentraler Grenzwertsatz

Aus dem Gesetz der grossen Zahl wissen wir, dass $n^{-1}S_n$ gegen den Mittelwert $n^{-1}\mathbb{E}[S_n]$ konvergiert. Zur Verwendung in der Statistik benötigen wir aber genauere Informationen über S_n. Wir wollen daher wissen, wie die Verteilung von S_n für grosse n aussieht. Nehmen wir an, dass $\{X_k\}$ unabhängig sind, Mittelwert μ_k und endliche Varianz σ_k^2 haben. Um die Verteilung studieren zu können, standardisieren wir nun die Zufallsvariable

$$S_n^* = \frac{S_n - \mathbb{E}[S_n]}{\sqrt{\mathrm{Var}[S_n]}} \ .$$

Dann ist $\mathbb{E}[S_n^*] = 0$ und $\mathrm{Var}[S_n^*] = 1$. Wir beweisen nun zwei Varianten des zentralen Grenzwertsatzes.

Satz 3.6. *Seien $\{X_n\}$ unabhängige Zufallsvariablen und es gelte eine der beiden folgenden Bedingungen:*

i) $\sup_n \mathbb{E}[|X_n^3] < \infty$ *und* $\underline{\lim}_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \mathrm{Var}[X_i] > 0$.

ii) $\{X_n\}$ *sind identisch verteilt mit Varianz* $\sigma^2 < \infty$.

Dann konvergiert S_n^* in Verteilung gegen die standard Normalverteilung

$$\lim_{n\to\infty} \mathbb{P}[S_n^* \le x] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{e}^{-y^2/2}\, \mathrm{d}y$$

für alle $x \in \mathbb{R}$.

Beweis. Wir dürfen $\mathbb{E}[X_n] = 0$ annehmen. Wir verwenden Hilfssatz 3.3. Sei $f(x)$ eine dreimal stetig differenzierbare beschränkte Funktion mit beschränkten ersten drei Ableitungen. Der Restterm in der Taylor-Formel

$$f(z+y) = f(z) + f'(z)y + \tfrac{1}{2}f''(z)y^2 + R(z,y)$$

lässt sich abschätzen durch

$$|R(z,y)| \le \tfrac{1}{6}|f'''(\tilde{z})|\,|y^3| \le C|y^3|\,,$$

oder durch

$$|R(z,y)| \le \tfrac{1}{2}|f''(\tilde{z}) - f''(z)|\,|y^2| \le \delta(y)|y^2|\,,$$

wobei $\tilde{z}$ zwischen z und $z+y$ liegt, $\delta(y)$ beschränkt ist und $\lim_{y\to 0}\delta(y) = 0$.

Definieren wir $Y_{i,n} = X_i/\sigma(S_n)$ und sei $\tilde{Y}_{i,n}$ eine normalverteilte Zufallsvariable mit Mittelwert 0 und Varianz $\sigma^2(Y_{i,n})$ unabhängig von den anderen Variablen. Dann ist $S_n^* = Y_{1,n} + \cdots + Y_{n,n}$, und $\tilde{S}_n = \tilde{Y}_{1,n} + \cdots + \tilde{Y}_{n,n}$ ist standard normalverteilt. Wir schreiben nun

$$\begin{aligned} f(S_n^*) - f(\tilde{S}_n) &= \sum_{k=1}^{n} f(Z_{k,n} + Y_{k,n}) - f(Z_{k,n} + \tilde{Y}_{k,n}) \\ &= \sum_{k=1}^{n} f'(Z_{k,n})(Y_{k,n} - \tilde{Y}_{k,n}) + \tfrac{1}{2}f''(Z_{k,n})(Y_{k,n}^2 - \tilde{Y}_{k,n}^2) \\ &\qquad + R(Z_{k,n}, Y_{k,n}) - R(Z_{k,n}, \tilde{Y}_{k,n})\,, \end{aligned}$$

wobei $Z_{k,n} = \tilde{Y}_{1,n} + \cdots + \tilde{Y}_{k-1,n} + Y_{k+1,n} + \cdots + Y_{n,n}$. Die Variablen $Z_{k,n}$, $Y_{k,n}$ und $\tilde{Y}_{k,n}$ sind unabhängig. Daher ist

$$\mathbb{E}[f'(Z_{k,n})(Y_{k,n} - \tilde{Y}_{k,n})] = \mathbb{E}[f'(Z_{k,n})](\mathbb{E}[Y_{k,n}] - \mathbb{E}[\tilde{Y}_{k,n}]) = 0\,.$$

Analog folgt, dass $\mathbb{E}[f''(Z_{k,n})(Y_{k,n}^2 - \tilde{Y}_{k,n}^2)] = 0$. Für den Mittelwert ergibt sich somit

$$\begin{aligned} |\mathbb{E}[f(S_n^*)] - \mathbb{E}[f(\tilde{S}_n)]| &= \Big|\sum_{k=1}^{n} \mathbb{E}[R(Z_{k,n}, Y_{k,n})] - \mathbb{E}[R(Z_{k,n}, \tilde{Y}_{k,n})]\Big| \\ &\le \sum_{k=1}^{n} \mathbb{E}[|R(Z_{k,n}, Y_{k,n})|] + \mathbb{E}[|R(Z_{k,n}, \tilde{Y}_{k,n})|]\,. \end{aligned}$$

Wir müssen nun die letzte Summe abschätzen.

i) Für das dritte Moment der Normalverteilung haben wir die Abschätzung

$$\mathbb{E}[|\tilde{Y}_{k,n}|^3] = \sqrt{\frac{8}{\pi}}\mathbb{E}[\tilde{Y}_{k,n}^2]^{3/2} \le 2\mathbb{E}[Y_{k,n}^2]^{3/2} \le 2\mathbb{E}[Y_{k,n}^3]\,.$$

Damit erhalten wir

$$|\mathbb{E}[f(S_n^*)] - \mathbb{E}[f(\tilde{S}_n)]| \leq \sum_{k=1}^{n} 3C\mathbb{E}[Y_{k,n}^3] \leq 3C \sum_{k=1}^{n} \frac{\mathbb{E}[|X_k|^3]}{\sigma^3(S_n)}$$
$$\leq \frac{3Cn \sup_k \mathbb{E}[|X_k|^3]}{\sigma^3(S_n)} = \frac{1}{\sqrt{n}} \frac{3C \sup_k \mathbb{E}[|X_k|^3]}{(n^{-1}\sigma^2(S_n))^{3/2}} .$$

Somit konvergiert der Ausdruck gegen Null.

ii) Wir erhalten wir die Abschätzung

$$|\mathbb{E}[f(S_n^*)] - \mathbb{E}[f(\tilde{S}_n)]| \leq \sum_{k=1}^{n} \mathbb{E}[\delta(Y_{k,n})Y_{k,n}^2] + \mathbb{E}[\delta(\tilde{Y}_{k,n})\tilde{Y}_{k,n}^2]$$
$$= n\Big(\mathbb{E}\Big[\delta\Big(\frac{X_1}{\sigma\sqrt{n}}\Big)\frac{X_1^2}{\sigma^2 n}\Big] + \mathbb{E}\Big[\delta\Big(\frac{\tilde{X}_1}{\sigma\sqrt{n}}\Big)\frac{\tilde{X}_1^2}{\sigma^2 n}\Big]\Big)$$
$$= \mathbb{E}\Big[\delta\Big(\frac{X_1}{\sigma\sqrt{n}}\Big)\frac{X_1^2}{\sigma^2}\Big] + \mathbb{E}\Big[\delta\Big(\frac{\tilde{X}_1}{\sigma\sqrt{n}}\Big)\frac{\tilde{X}_1^2}{\sigma^2}\Big] .$$

Das Resultat folgt nun mit beschränkter Konvergenz, da $\delta(y)$ beschränkt ist. □

Da die Normalverteilung als Grenzwert auftritt, nimmt diese Verteilung eine besondere Rolle ein. Man findet daher die Normalverteilung in Tabellenbüchern.

Der klassische Spezialfall sind unabhängige 0-1 Experimente mit Erfolgsparameter $0 < p < 1$

$$\lim_{n\to\infty} \mathbb{P}\Big[\frac{S_n - np}{\sqrt{np(1-p)}} \leq x\Big] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{e}^{-y^2/2}\, \mathrm{d}y .$$

Für $p = \frac{1}{2}$ wurde dies 1730 von Abraham de Moivre und für beliebiges p von Pierre-Simon Laplace 1812 gezeigt. Hier wurde direkt die exakte Wahrscheinlichkeit mit Hilfe der Sterlingschen Formel ausgewertet. De Moivre kannte aber die Integraldarstellung der Normalverteilung noch nicht.

Eine Anwendung könnte die Folgende sein. Jemand will im Kasino Roulette spielen. Hier ist $p = 18/37$. Er hat vor, an einem Abend 100 Mal zu spielen. Wieviel Geld muss der Spieler mitnehmen, um am Ende des Abends mit Wahrscheinlichkeit 99% keine Schulden zu haben? Wir formulieren das Problem mit 0-1 Experimenten, das heisst, der gewonnene Betrag ist $2S_n - n$. Wir suchen daher zuerst x, so dass

$$\mathbb{P}\Big[\frac{S_n - np}{\sqrt{np(1-p)}} \leq x\Big] = 0.01 .$$

Aus einer Tabelle finden wir $x = -2.3263$ für die Normalverteilung. Also ist das gesuchte Kapital

$$100 - 2\frac{1800}{37} + 2 \cdot 2.3263\sqrt{34200/1369} = 25.9572\,.$$

Der Spieler braucht also 26 Geldeinheiten. Dies ist auch das Resultat, das man bei exakter Berechnung erhält.

4. Schätztheorie

4.1. Die Problemstellung

Sein $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. In praktischen Situationen kann man $(\Omega, \mathcal{F})$ modellieren, aber man kennt $\mathbb{P}$ nicht. Was man hat, sind Daten, wie z.B. bereits aufgetretene Schäden bei einer Versicherung, oder die Tagesschlusskurse einer Aktie an der Börse. Wie soll man nun $\mathbb{P}$ wählen, damit das Mass mit den Daten konsistent ist?

Oft hat man eine Idee, welche Art der Verteilung in Frage kommt. Zum Beispiel verwendet man in der Versicherung oft Gamma- oder Pareto-Verteilungen. Ein beliebtes Modell in der Finanzmathematik (Black–Scholes-Modell) ergibt log-normalverteilte Aktienpreise, das heisst, der Logarithmus des Preises hat eine Normalverteilung. Schränkt man die möglichen Wahrscheinlichkeitsmasse auf diese Klassen ein, dann muss man nur noch die Parameter zu bestimmen. Sei also $\Theta \subset \mathbb{R}^d$ eine Menge von möglichen Parametern. Wir betrachten nun die Menge der Masse $\{\mathbb{P}_\theta : \theta \in \Theta\}$. Wir haben das Problem darauf reduziert, den richtigen Parameter θ zu schätzen, beziehungsweise eine Menge von mit den Daten verträglichen θ's zu finden.

Bezeichnen wir den "richtigen" Parameter mit θ_0. Seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen, die unter dem Mass $\mathbb{P}_{\theta_0}$ erzeugt wurden. Ein **Schätzer** ist nun eine Zufallsvariable $T = T(X_1, X_2, \ldots, X_n) : \mathbb{R}^n \to \Theta$.

Wir wollen natürlich nun nicht irgendeinen Schätzer, sondern einen möglichst "guten". Wir müssen daher Kriterien definieren, die uns gute Schätzer beschreiben.

Wir sagen T ist **konsistent**, falls T für $n \to \infty$ stochastisch gegen θ_0 konvergiert. Wir sagen T ist **stark konsistent**, falls T fast sicher gegen θ_0 konvergiert. Ein guter Schätzer sollte zumindest konsistent sein.

Wir sagen T ist **unverfälscht**, falls $\mathbb{E}_{\theta_0}[T] = \theta_0$. Diese Eigenschaft ist oft erwünscht, da man auf diese Weise hofft, schon für "kleines" n nahe beim richtigen Wert zu liegen.

Oft ist man nicht direkt daran interessiert, θ_0 zu schätzen. Man benötigt zum Beispiel eine Kennzahl $g(\theta_0)$. In diesem Fall sollte T gegen $g(\theta_0)$ konvergieren. Wir suchen dann einen Schätzer für $g(\theta_0)$. Insbesondere ist der Schätzer unverfälscht, falls $\mathbb{E}_{\theta_0}[T] = g(\theta_0)$.

Betrachten wir das Problem, den Parameter einer Exponentialverteilung zu

schätzen. Die mehrdimensionale Dichte der Daten $X_1, X_2, \ldots, X_n$ ist

$$f_{\boldsymbol{X}}(x_1, x_2, \ldots, x_n) = \prod_{k=1}^{n} \alpha \mathrm{e}^{-\alpha x_k} = \alpha^n \exp\Big\{-\alpha \sum_{k=1}^{n} x_k\Big\} .$$

Die Information, die also in den Daten steckt ist $\sum_{k=1}^n X_k$. Kennen wir also $\sum_{k=1}^n X_k$, so haben wir genau so viel Information über den unbekannten Parameter, wie wenn wir die einzelnen Daten kennen würden. Solche Statistiken nennen wir **suffizient**.

Allgemein sagen wir, eine Statistik $\mathcal{T}(X_1, \ldots, X_n) : \mathbb{R}^n \to \mathbb{R}^d$ ist **suffizient**, falls die bedingte Verteilung von $(X_1, \ldots, X_n)$ gegeben $\{\mathcal{T}(X_1, \ldots, X_n) = (t_1, \ldots, t_d)\}$ nicht vom Parameter θ abhängt.

Beispiel Seien $X_1, \ldots, X_n$ normalverteilt mit Mittelwert μ und Varianz σ^2. Die Dichtefunktion der Daten ist

$$\begin{aligned} f_{\boldsymbol{X}}(x_1, \ldots, x_n) &= \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(x_k-\mu)^2}{2\sigma^2}} \\ &= (2\pi\sigma^2)^{-n/2} \exp\Big\{\frac{2\mu \sum_{k=1}^n x_i - \sum_{k=1}^n x_k^2 - n\mu^2}{2\sigma^2}\Big\} . \end{aligned}$$

Somit ist $\mathcal{T}(X_1, \ldots, X_n) = (\sum_{k=1}^n X_i, \sum_{k=1}^n X_k^2)$ eine suffiziente Statistik

4.2. Schätzen von Kennzahlen

Betrachten wir zuerst ein einfacheres Problem. Wir haben Daten $X_1, X_2, \ldots, X_n$, die unabhängig und identisch verteilt sind. Wir benötigen nun, zum Beispiel um die Prämie eines Versicherungsbetrages zu berechnen, den Mittelwert μ und die Varianz σ^2 der zugrundeliegenden unbekannten Verteilung.

Wir sind also interessiert, die Grösse $g(\theta) = \mathbb{E}_\theta[f(X_1)]$ für eine Funktion $f : \mathbb{R} \to \mathbb{R}$ zu schätzen. Das starke Gesetz der grossen Zahl, legt uns den Schätzer

$$T = \frac{1}{n} \sum_{k=1}^{n} f(X_k)$$

nahe. Wir wissen, dass dieser Schätzer stark konsistent ist, und wegen der Linearität des Erwartungswertes, ist der Schätzer auch unverfälscht.

4.2.1. Der Erwartungswert

Falls wir den Erwartungswert schätzen wollen, ist der natürliche Schätzer der Mittelwert

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} X_k \ .$$

Dieser Schätzer heisst **empirischer Mittelwert**. Betrachten wir zum Beispiel 0-1 Experimente mit unbekanntem Erfolgsparameter p. Wir möchten nun sicherstellen, dass der Schätzwert nicht zu stark von p abweicht, das heisst, wir verlangen $\mathbb{P}_p[|\hat{\mu} - p| \leq \varepsilon] \geq 1 - \alpha$. Wir benutzen den zentralen Grenzwertsatz

$$\mathbb{P}_p[|\hat{\mu} - p| \leq \varepsilon] = \mathbb{P}_p\Big[\frac{|S_n - np|}{\sqrt{np(1-p)}} \leq \frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\Big] \approx \Phi(x) - \Phi(-x) = 2\Phi(x) - 1 \ , \tag{4.1}$$

wobei $x = \varepsilon\sqrt{n}/\sqrt{p(1-p)}$ und $\Phi(y)$ ist die Verteilung der Standardnormalverteilung.

Setzen wir $\varepsilon = 0.02$ und $\alpha = 0.05$, erhalten wir $x = 1.96$. Also benötigen wir

$$n \geq \frac{1.96^2 p(1-p)}{0.02^2} = 9604 p(1-p) \ .$$

Um nun ein explizites n zu erhalten, können wir p durch $\hat{\mu}$ ersetzen, oder den schlimmst möglichen Fall $p(1-p) = 0.25$ annehmen, was $n \geq 2401$ ergibt.

Man muss aber aufpassen, wenn man den Mittelwert mit dem empirischen Mittelwert schätzt. Fällt nämlich die Flanke $1 - F(x)$ zu langsam gegen Null, kann der empirische Mittelwert weit weg vom Erwartungswert liegen. Die folgende Tabelle zeigt 14 Simulationen von $\hat{\mu}$ für die Verteilungsfunktion $F(x) = 1 - (1+x)^{-1.1}$ und $n = 1000$.

3.99255	4.88243	4.52478	4.73841	4.24971	3.74427	6.89294
7.81461	4.96825	5.47598	5.14919	4.81354	19.4518	4.77443

Der Erwartungswert dieser Verteilung ist 10. Der empirische Mittelwert unterschätzt hier meistens den richtigen Wert massiv. Das liegt daran, dass grosse Werte für X_i nicht sehr wahrscheinlich sind, aber trotzdem stark zum Erwartungswert beitragen.

4.2.2. Die Varianz

Um die Varianz zu schätzen nehmen wir zuerst an, dass der Mittelwert μ bekannt ist. Dann ist $n^{-1}\sum_{k=1}^{n}(X_k - \mu)^2$ der natürliche Schätzer. Normalerweise kennt man

den Mittelwert aber nicht. Eine einfache Idee ist, den Mittelwert durch den Schätzer $\hat{\mu}$ zu ersetzen. Der Mittelwert dieses Schätzers ist dann aber

$$\mathbb{E}\Big[\frac{1}{n}\sum_{k=1}^{n}(X_k-\hat{\mu})^2\Big] = \frac{n-1}{n}\sigma^2 \; .$$

Im Mittel unterschätzt man also die korrekte Varianz. Daher verwendet man meistens den unverfälschten Schätzer

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k-\hat{\mu})^2 \; .$$

Dieser Schätzer heisst **empirische Varianz**, und ist nach dem Gesetz der grossen Zahl auch stark konsistent. Aber auch für die Varianz kann der Schätzer ganz falsche Resultate liefern, falls die Flanke langsam abfällt. Zum Beispiel ist $\hat{\sigma}^2$ endlich, wenn die Varianz nicht existiert.

4.3. Die Momentenmethode

Wir wollen nun einen d-dimensionalen Parameter $\theta \in \Theta$ einer Verteilungsfunktion $F_\theta(x)$ schätzen. Wir haben die unabhängigen und identisch verteilten Daten $X_1, \ldots, X_n$, und nehmen an, dass $\mathbb{E}[|X_1|^d] < \infty$. Die Momentenmethode schätzt nun die Momente

$$\hat{\mu}_i = \frac{1}{n}\sum_{k=1}^{n} X_k^i \; , \qquad i = 1, \ldots, d \; .$$

Wir wählen nun den Parameter θ, der die selben Momente hat; das heisst

$$\mathbb{E}_\theta[X_1^i] = \hat{\mu}_i \; , \qquad i = 1, \ldots, d \; .$$

Damit man die Methode anwenden kann, muss das obige Gleichungssystem eindeutig lösbar sein.

Da die Schätzer für $\hat{\mu}_i$ stark konsistent sind, erhält man sofort, dass der Momentenschätzer stark konsistent ist, sofern die Momente $\mathbb{E}_\theta[X_1^i]$ stetig in θ sind und die Momente den Parameter θ eindeutig bestimmen.

Beispiel: Nehmen wir an, dass X_k Gamma verteilt ist mit Parametern γ und α. Wir haben

$$\mathbb{E}[X_1] = \frac{\alpha^\gamma}{\Gamma(\gamma)}\int_0^\infty x x^{\gamma-1}\mathrm{e}^{-\alpha x}\,\mathrm{d}x = \frac{\alpha^\gamma}{\Gamma(\gamma)}\int_0^\infty x^{\gamma}\mathrm{e}^{-\alpha x}\,\mathrm{d}x = \frac{\alpha^\gamma\Gamma(\gamma+1)}{\Gamma(\gamma)\alpha^{\gamma+1}} = \frac{\gamma}{\alpha} \; ,$$

$$\mathbb{E}[X_1^2] = \frac{\alpha^\gamma}{\Gamma(\gamma)} \int_0^\infty x^{\gamma+1} \mathrm{e}^{-\alpha x} \, \mathrm{d}x = \frac{\alpha^\gamma \Gamma(\gamma+2)}{\Gamma(\gamma)\alpha^{\gamma+2}} = \frac{\gamma(\gamma+1)}{\alpha^2} \, .$$

Somit müssen wir das Gleichungssystem

$$\hat{\mu} = \frac{1}{n} \sum_{k=1}^n X_k = \frac{\gamma}{\alpha} \, , \qquad \hat{\mu}_2 = \frac{1}{n} \sum_{k=1}^n X_k^2 = \frac{\gamma(\gamma+1)}{\alpha^2}$$

lösen. Wir erhalten $1 + \gamma^{-1} = \hat{\mu}_2/\hat{\mu}^2$, also den Schätzer

$$\hat{\gamma} = \frac{\hat{\mu}^2}{\hat{\mu}_2 - \hat{\mu}^2} \, , \qquad \hat{\alpha} = \frac{\hat{\gamma}}{\hat{\mu}} = \frac{\hat{\mu}}{\hat{\mu}_2 - \hat{\mu}^2} \, .$$

Wir könnten $\hat{\mu}_2 - \hat{\mu}^2$ auch durch die empirische Varianz $\hat{\sigma}^2$ ersetzen.

Für die Gamma Verteilung funktioniert die Methode gut. Man muss aber aufpassen, falls die Flanke $1 - F_\theta(x)$ nur langsam gegen Null konvergiert, also wenn zum Beispiel nicht alle Momente von X_1 existieren.

4.4. Das Maximum-Likelihood-Prinzip

Nehmen wir wieder an, dass $X_1, \ldots, X_n$ unabhängig und identisch verteilt sind unter $\mathbb{P}_\theta$ mit Verteilungsfunktion $F_\theta(x)$ und, im absolutstetigen Fall, Dichte $f_\theta(x)$. Die Likelihoodfunktion ist definiert als

$$L_\theta(x_1, \ldots, x_n) = \begin{cases} \mathbb{P}_\theta[X_1 = x_1] \cdots \mathbb{P}_\theta[X_n = x_n] \, , & \text{falls } F_\theta(x) \text{ diskret,} \\ f_\theta(x_1) \cdots f_\theta(x_n) \, , & \text{falls } F_\theta(x) \text{ absolutstetig.} \end{cases}$$

Die Likelihoodfunktion ist also die "Wahrscheinlichkeit", dass die beobachteten Daten auftreten. Das Maximum-Likelihood-Prinzip wählt nun den Parameter, der die Likelihoodfunktion maximiert. Da der Likelihood ein Produkt ist, ist es oft einfacher $\log L_\theta(x_1, \ldots, x_n)$ zu maximieren.

Der Maximum-Likelihood-Schätzer ist sehr beliebt, da er sehr schöne Eigenschaften besitzt. Unter schwachen Voraussetzungen ist der Schätzer optimal in einem hier nicht näher beschriebenen Sinne. Weiter gilt unter schwachen Voraussetzungen an die Dichte, dass der Schätzer asymptotisch normalverteilt ist, das heisst, es gilt eine Art zentraler Grenzwertsatz. Dies ist vor allem wichtig, um Konfidenzintervalle zu finden, siehe Abschnitt 4.7.

Beispiele:

- Sei $\Theta = \mathbb{R} \times (0, \infty)$, und $\{X_i\}$ seien normalverteilt mit Parameter $\theta = (\mu, \sigma^2)$. Dann ist

$$\log L_\theta(x_1, \ldots, x_n) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{k=1}^{n} (x_k - \mu)^2 \ .$$

  Maximieren wir bezüglich μ folgt $\hat{\mu} = n^{-1} \sum_{k=1}^{n} x_k$, also der empirische Mittelwert. Maximiert man weiter bezüglich σ^2, erhält man $\hat{\sigma}^2 = n^{-1} \sum_{k=1}^{n} (x_k - \hat{\mu})^2$. Der Maximum-Likelihood-Schätzer stimmt also mit dem Momentenschätzer überein.

- Sei $\Theta = (0, \infty)$ und $F_\theta(x)$ die Gleichverteilung auf $(0, \theta)$. Dann wissen wir, dass $\theta \geq \max_k x_k$. Die Likelihoodfunktion wird dann $L_\theta(x_1, \ldots, x_n) = \theta^{-n} \mathbb{1}_{\theta \geq \max_k x_k}$. Somit erhalten wir den Schätzer

$$\hat{\theta} = \max_{1 \leq k \leq n} x_k \ .$$

  Da $\mathbb{P}_\theta[\hat{\theta} \leq t] = \prod_{k=1}^{n} \mathbb{P}_\theta[X_k \leq t] = (t/\theta)^n \wedge 1$, ergibt sich $\mathbb{E}_\theta[\hat{\theta}] = \theta n/(n+1)$. Aus dem Borel–Cantelli-Lemma ist einfach zu sehen, dass der Schätzer stark konsistent ist. Das zweite Moment des Schätzers ist $\mathbb{E}_\theta[\hat{\theta}^2] = \theta^2 n/(n+2)$. Also erhalten wir die Varianz $\mathrm{Var}_\theta[\hat{\theta}] = \theta^2 n/((n+1)^2(n+2))$. Vergleichen wir dies mit dem Momentenschätzer. Da $\mathbb{E}_\theta[X_1] = \theta/2$ ist der Momentenschätzer $2\hat{\mu}$. Man beachte, dass möglicherweise $2\hat{\mu} < \max_k x_k$. Die Varianz ist $\mathrm{Var}_\theta[2\hat{\mu}] = 4\,\mathrm{Var}_\theta[X_1]/n = \theta^2/(3n)$. Der Momentenschätzer fluktuiert also stärker als der Maximum-Likelihood-Schätzer.

- Die Dichtefunktion sei $f_\theta(x) = \frac{1}{2}\beta \mathrm{e}^{-\beta|x-m|}$. Die Verteilung hat Mittelwert m und Varianz $2/\beta^2$. Die Likelihoodfunktion ist gegeben durch

$$\log L_\theta(x_1, \ldots, x_n) = n \log(\beta/2) - \beta \sum_{k=1}^{n} |x_k - m| \ .$$

  Um den Schätzer für m zu erhalten, müssen wir $\sum_{k=1}^{n} |x_k - m|$ minimieren. Das heisst, $\hat{m}$ ist der **Median** von $x_1, \ldots x_n$. Das ist

$$\hat{m} = \begin{cases} x_{(\{n+1\}/2)} \ , & \text{falls } n \text{ ungerade,} \\ \in [x_{(n/2)}, x_{(n/2+1)})] \ , & \text{falls } n \text{ gerade.} \end{cases}$$

Hier ist $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ die Ordnungsstatistik (das heisst die geordneten Daten) der Beobachtungen. Für den zweiten Parameter erhalten wir

$$\hat{\beta} = \frac{n}{\sum_{k=1}^{n} |x_k - \hat{m}|} = \frac{1}{\frac{1}{n}\sum_{k=1}^{n} |x_k - \hat{m}|} \ .$$

- Seien $\{X_i\}$ Poissonverteilt mit Parameter λ. Dann ist der Likelihood gegeben durch

$$\log L_\lambda(x_1, \ldots, x_n) = \sum_{k=1}^{n} (x_k \log \lambda - \log x_k! - \lambda) = -n\lambda + \log \lambda \sum_{k=1}^{n} x_k - \sum_{k=1}^{n} \log x_k! \ .$$

Somit erhalten wir den Schätzer $\hat{\lambda} = n^{-1} \sum_{k=1}^{n} x_k = \hat{\mu}$. Dieser Schätzer stimmt also mit dem Momentenschätzer überein.

Bemerkung. Der Maximum-Likelihood-Schätzer funktioniert natürlich auch gut, falls die Daten abhängig sind. Dann ist die Likelihoodfunktion einfach die gemeinsame Wahrscheinlichkeit bzw. die gemeinsame Dichte der Daten. ■

4.5. Bayes'sche Statistik

Um die Güte eines Schätzers zu bestimmen, könnten wir eine Verlustfunktion definieren, $\rho(T, \theta)$. Dies ist der Verlust, wenn wir T schätzen, aber θ richtig ist. Eine populäre Wahl ist $\rho(T, \theta) = \|T - \theta\|^2$. Eine Möglichkeit ist das **Min-Max-Prinzip**, das zum Beispiel in der Spieltheorie angewendet wird. Man sucht T, so dass

$$\max_\theta \mathbb{E}_\theta[\rho(T, \theta)] = \min_{T'} \max_\theta \mathbb{E}_\theta[\rho(T', \theta)] \ ,$$

wobei das Minimum über alle möglichen Schätzer T' genommen wird.

In vielen Situationen hat man aber Erfahrung. Zum Beispiel bei einer Versicherung wurden schon vorher ähnliche Verträge abgeschlossen. Man hat daher Schätzungen, welche θ bei anderen Verträgen aufgetreten sind. Oder, wenn man schätzen will, wieviel Wasser in einer Stadt durch undichte Rohrleitungen verschwindet, bevor es beim Konsumenten ankommt. Da kennt man schon Werte aus anderen Städten. Das heisst, man "kennt" die Verteilung von θ. Nennen wir diese Verteilung $F_\theta(\vartheta)$. Man wird dann also die Grösse

$$\bar{\rho}(T) = \int \mathbb{E}_\vartheta[\rho(T, \vartheta)] \, \mathrm{d}F_\theta(\vartheta)$$

minimieren. Nehmen wir an, dass die Verteilungen absolutstetig seien. Sei $f_\theta(\vartheta)$ die Dichte von θ, und sei $f(\boldsymbol{x}|\vartheta)$ die bedingte Verteilung von $\boldsymbol{X} = (X_1, \ldots, X_n)$ gegeben $\theta = \vartheta$. Dann haben wir

$$\begin{aligned}\bar{\rho}(T) &= \iint \rho(T(\boldsymbol{x}), \vartheta) f(\boldsymbol{x} \mid \vartheta) \, \mathrm{d}\boldsymbol{x} f_\theta(\vartheta) \, \mathrm{d}\vartheta \\ &= \iint \rho(T(\boldsymbol{x}), \vartheta) \frac{f(\boldsymbol{x} \mid \vartheta) f_\theta(\vartheta)}{f_x(\boldsymbol{x})} \, \mathrm{d}\vartheta f_x(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \\ &= \iint \rho(T(\boldsymbol{x}), \vartheta) f(\vartheta \mid \boldsymbol{x}) \, \mathrm{d}\vartheta f_x(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \, ,\end{aligned}$$

wobei $f_x(\boldsymbol{x})$ die Dichte von $\boldsymbol{X}$ bezeichnet und $f(\vartheta \mid \boldsymbol{x})$ die bedingte Dichte von θ gegeben $\boldsymbol{X} = \boldsymbol{x}$. Der Ausdruck wird minimiert, indem man das innere Integral für jedes $\boldsymbol{x}$ minimiert. Das bedeutet: Man muss das Minimierungsproblem für jede **apriori** Verteilung $F_\theta(\vartheta)$ lösen, wobei T eine Konstante ist. Beobachtet man die Daten $\boldsymbol{X}$, nimmt man die **aposteriori** Verteilung $F_\theta(\vartheta \mid \boldsymbol{X})$ statt der apriori Verteilung. Da die Daten über die Bayes'sche Formel in den Schätzer eingehen, nennt man diese statistische Methode Bayes'sche Statistik.

Ist nun $\rho(T, \theta) = (T - \theta)^2$, dann folgt aus

$$\mathbb{E}[(T-\theta)^2] = \mathbb{E}[\{(T - \mathbb{E}[\theta]) - (\theta - \mathbb{E}[\theta])\}^2] = \mathbb{E}[(T - \mathbb{E}[\theta])^2] + \mathbb{E}[(\theta - \mathbb{E}[\theta])^2] \, ,$$

dass $T = \mathbb{E}[\theta]$ der optimale Schätzer ist, falls man keine Daten hat (also T eine Konstante ist). Beobachtet man Daten, folgt somit, dass $T(\boldsymbol{X}) = \mathbb{E}[\theta \mid \boldsymbol{X}]$ den erwarteten Verlust minimiert.

4.6. Die Informationsungleichung

Um nützlich zu sein, sollte ein Schätzer nahe beim wahren Wert liegen. Somit sollte der Mittelwert nicht zu stark vom wahren Wert abweichen und gleichzeitig die Varianz klein sein. Wir wollen daher die Varianz eines Schätzers T für den Parameter θ untersuchen.

Wir machen folgende Annahmen. $X_1, \ldots, X_n$ sind unabhängig und identisch verteilt mit nach θ differenzierbarer Dichte $f_\theta(x)$. Weiter sollen alle Dichten den selben Träger haben, das heisst, die Menge $\overline{\{x : f_\theta(x) > 0\}}$ soll unabhängig von θ sein.

Die Bedingungen sind zum Beispiel für die Normalverteilung erfüllt, aber nicht für die Gleichverteilung auf $[0, \theta]$, da der Träger das Intervall $[0, \theta]$ ist, und daher abhängig von θ. Wir definieren die **Fisher-Information** für θ

$$I(\theta) = \int \Big(\frac{\frac{\partial}{\partial\theta} f_\theta(x)}{f_\theta(x)} \Big)^2 f_\theta(x) \, \mathrm{d}x = \mathbb{E}_\theta \Big[\Big(\frac{\partial}{\partial\theta} \log f_\theta(X) \Big)^2 \Big] \, .$$

Wir bezeichnen der Einfachheit halber mit $f'_\theta(x) = \frac{\partial}{\partial\theta} f_\theta(x)$ die Ableitung nach θ. Wir haben

$$\mathbb{E}_\theta\Big[\frac{f'_\theta(X_1)}{f_\theta(X_1)}\Big] = \int \frac{f'_\theta(x)}{f_\theta(x)} f_\theta(x)\,\mathrm{d}x = \int f'_\theta(x)\,\mathrm{d}x = \Big(\int f_\theta(x)\,\mathrm{d}x\Big)' = 0\,.$$

Somit können wir die Fisher-Information als

$$I(\theta) = \mathbb{E}_\theta\Big[\Big(\frac{f'_\theta(X_1)}{f_\theta(X_1)}\Big)^2\Big] = \mathrm{Var}_\theta\Big[\frac{f'_\theta(X_1)}{f_\theta(X_1)}\Big]\,.$$

interpretieren.

Satz 4.1. (Rao–Cramér) *Seien die Annahmen oben erfüllt. Für jeden Schätzer T von θ gilt*

$$\mathrm{Var}_\theta[T] \geq \frac{(1+B'(\theta))^2}{nI(\theta)}\,,$$

wobei $B(\theta) = \mathbb{E}_\theta[T] - \theta$. Insbesondere, ist T unverfälscht, dann haben wir

$$\mathrm{Var}_\theta[T] \geq \frac{1}{nI(\theta)}\,.$$

Beweis. Den Spezialfall für unverfälschte Schätzer erhalten wir sofort aus $B(\theta) = 0$. Wir bemerken zuerst, dass

$$\begin{aligned}\theta + B(\theta) = \mathbb{E}_\theta[T] &= \int_{\mathbb{R}^n} T(x_1,\ldots,x_n) f_\theta(x_1)\cdots f_\theta(x_n)\,\mathrm{d}x_n\cdots\,\mathrm{d}x_1\\ &= \int_{\mathbb{R}^n} T(x_1,\ldots,x_n) L_\theta(x_1,\ldots,x_n)\,\mathrm{d}x_n\cdots\,\mathrm{d}x_1\,.\end{aligned}$$

Also ist

$$\begin{aligned}1 + B'(\theta) &= \int_{\mathbb{R}^n} T(x_1,\ldots,x_n) L'_\theta(x_1,\ldots,x_n)\,\mathrm{d}x_n\cdots\,\mathrm{d}x_1\\ &= \int_{\mathbb{R}^n} T(x_1,\ldots,x_n)\frac{L'_\theta(x_1,\ldots,x_n)}{L_\theta(x_1,\ldots,x_n)} L_\theta(x_1,\ldots,x_n)\,\mathrm{d}x_n\cdots\,\mathrm{d}x_1\\ &= \mathbb{E}_\theta\Big[T\frac{L'_\theta(X_1,\ldots,X_n)}{L_\theta(X_1,\ldots,X_n)}\Big] = \mathrm{Cov}_\theta\Big[T, \frac{L'_\theta(X_1,\ldots,X_n)}{L_\theta(X_1,\ldots,X_n)}\Big]\,,\end{aligned}$$

da

$$\mathbb{E}_\theta\Big[\frac{L'_\theta(X_1,\ldots,X_n)}{L_\theta(X_1,\ldots,X_n)}\Big] = \Big(\int_{\mathbb{R}^n} L_\theta(x_1,\ldots,x_n)\,\mathrm{d}x_n\cdots\,\mathrm{d}x_1\Big)' = 0\,.$$

Aus der Cauchy–Schwarz-Ungleichung (Hilfssatz 2.20) schliessen wir nun

$$(1+B'(\theta))^2 = \Big(\mathrm{Cov}_\theta\Big[T, \frac{L'_\theta(X_1,\ldots,X_n)}{L_\theta(X_1,\ldots,X_n)}\Big]\Big)^2 \leq \mathrm{Var}_\theta[T]\,\mathrm{Var}_\theta\Big[\frac{L'_\theta(X_1,\ldots,X_n)}{L_\theta(X_1,\ldots,X_n)}\Big]\,.$$

Aus

$$\frac{L'_\theta(X_1,\dots,X_n)}{L_\theta(X_1,\dots,X_n)} = (\log L_\theta(X_1,\dots,X_n))' = \sum_{k=1}^{n} \frac{f'_\theta(X_k)}{f_\theta(X_k)}$$

und der Unabhängigkeit ist die letztere Varianz

$$\mathrm{Var}_\theta\Big[\frac{L'_\theta(X_1,\dots,X_n)}{L_\theta(X_1,\dots,X_n)}\Big] = n\,\mathrm{Var}_\theta\Big[\frac{f'_\theta(X_k)}{f_\theta(X_k)}\Big] = nI(\theta)\ .$$

Einsetzen ergibt die Behauptung. □

Betrachten wir nun die Normalverteilung mit Mittelwert θ und Varianz σ^2. Dann erhalten wir

$$\frac{f'_\theta(x)}{f_\theta(x)} = \frac{x-\theta}{\sigma^2}\ .$$

Die Fisher-Information wird also

$$I(\theta) = \mathrm{Var}_\theta\Big[\frac{X-\theta}{\sigma^2}\Big] = \frac{1}{\sigma^4}\sigma^2 = \frac{1}{\sigma^2}\ .$$

Der empirische Mittelwert hat Varianz

$$\mathrm{Var}_\theta[\hat\mu] = \frac{1}{n^2}\sum_{k=1}^{n}\mathrm{Var}_\theta[X_k] = \frac{\sigma^2}{n}\ .$$

Da $\mathrm{Var}_\theta[T] \geq (nI(\theta))^{-1} = \sigma^2/n$, ist $\hat\mu$ der beste unverfälschte Schätzer für den Mittelwert.

Ein analoges Resultat lässt sich für diskrete Verteilungen beweisen.

Wenn wir nun zwei unverfälschte Schätzer haben, so bevorzugen wir den Schätzer mit der kleineren Varianz. Da wir eine untere Schranke für die Varianz haben, definieren wir die **Effizienz** eines Schätzers als

$$e(T(\boldsymbol{X})) = \frac{1}{nI(\theta)\,\mathrm{Var}_\theta[T(\boldsymbol{X})]}\ .$$

Ein Schätzer ist dann also besser, falls er eine höhere Effizienz hat. Oft genügt es, wenn für grosse n die Varianz klein wird. Wir sagen ein Schätzer ist **asymptotisch effizient**, falls $\lim_{n\to\infty} e(T(\boldsymbol{X})) = 1$.

Falls der Schätzer nicht unverfälscht ist, kann man die Definition der Effizienz verallgemeinern zu

$$e(T(\boldsymbol{X})) = \frac{(1+B'(\theta))^2}{nI(\theta)\,\mathrm{Var}_\theta[T(\boldsymbol{X})]}\ .$$

Auch hier gilt allgemein $e(T(\boldsymbol{X})) \leq 1$. Da es möglicherweise keinen Schätzer gibt, für den $e(T(\boldsymbol{X})) = 1$ gilt, wird man meistens an asymptotisch effizienten Schätzern interessiert sein. Zum Beispiel folgt, dass unter schwachen Bedingungen der Maximum-Likelihood-Schätzer asymptotisch effizient ist.

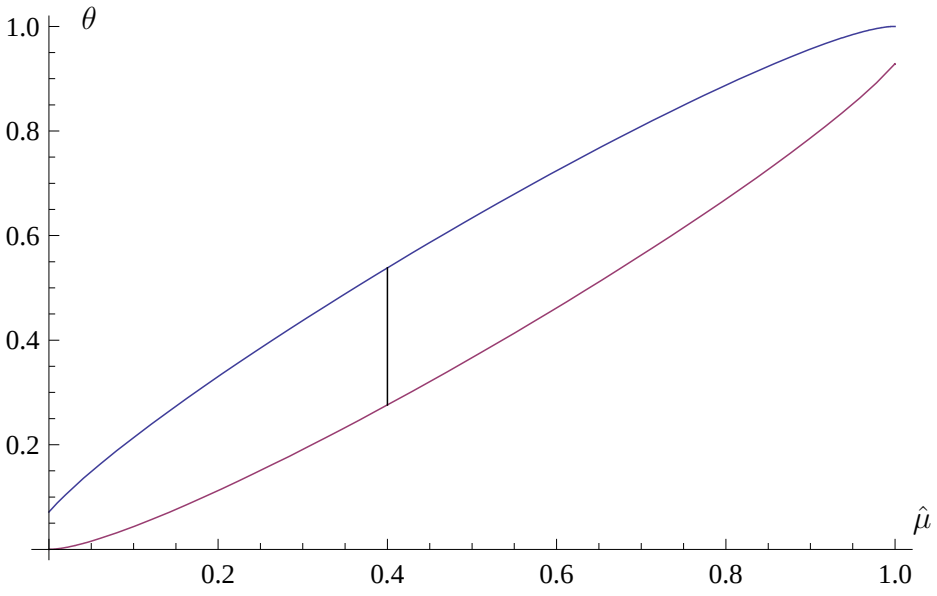

Abbildung 4.1: *Konfidenzintervalle für 0-1-Experimente*

4.7. Konfidenzintervalle

Wenn wir Parameter oder Kennzahlen schätzen, nützt es uns nicht viel, wenn wir wissen, dass der Schätzer T ist. Da die Daten zufällig sind, ist auch T zufällig, und man sollte wissen, wie weit T von θ_0 entfernt ist. Wenn wir also in der Zeitung lesen, dass das Wirtschaftswachstum auf 0.2% geschätzt wird, ist es ein Unterschied, ob die Varianz des Schätzers 0.01 oder 1 ist. Die Idee ist daher, nicht eine Zahl, sondern ein Intervall I von "wahrscheinlichen Werten" von θ anzugeben. Wir sagen $I = I(X_1, X_2, \ldots, X_n)$ ist ein **Konfidenzintervall zum Niveau** α, falls

$$\mathbb{P}_\theta[\theta \in I] \geq 1 - \alpha$$

für alle $\theta \in \Theta$.

Beispiele

- Um die Qualität von einer Serie von Produkten (z.B. Glühbirnen) zu testen, werden n Produkte untersucht. Wir setzen $X_k = 1$, falls der Test bestanden wird, und 0, falls er nicht bestanden wird. Wir nehmen an, dass die Produkte unabhängig sind, und den Test mit Wahrscheinlichkeit θ bestehen. Da $\hat{\mu}$ ein Schätzer für θ ist, setzen wir $I = [\hat{\mu} - \varepsilon(\theta), \hat{\mu} + \varepsilon(\theta)]$. Wir verlangen nun

$$\mathbb{P}_\theta[\theta \in I] = \mathbb{P}_\theta[|\hat{\mu} - \theta| \leq \varepsilon(\theta)] = \mathbb{P}_\theta\Big[\sqrt{n}\frac{|\hat{\mu} - \theta|}{\sqrt{\theta(1-\theta)}} \leq \frac{\varepsilon(\theta)\sqrt{n}}{\sqrt{\theta(1-\theta)}}\Big] \geq 1 - \alpha\,.$$

Wir können nun nach dem zentralen Grenzwertsatz die Verteilung mit einer Standardnormalverteilung approximieren. Also, wir suchen die Lösung zu

$$2\Phi\Big(\frac{\varepsilon(\theta)\sqrt{n}}{\sqrt{\theta(1-\theta)}}\Big) - 1 \geq 1-\alpha \;,$$

siehe auch (4.1). Dies ergibt

$$\varepsilon(\theta) = \Phi^{-1}(1-\alpha/2)\frac{\sqrt{\theta(1-\theta)}}{\sqrt{n}} \;.$$

Wir erhalten also,

$$|\hat{\mu}-\theta| \leq \Phi^{-1}(1-\alpha/2)\frac{\sqrt{\theta(1-\theta)}}{\sqrt{n}} = \sqrt{\theta(1-\theta)}\sqrt{\tau} \;,$$

wobei $\tau = n^{-1}(\Phi^{-1}(1-\alpha/2))^2$. Auflösung nach θ gibt

$$I = \frac{2\hat{\mu}+\tau}{2(1+\tau)} \pm \frac{\sqrt{\tau^2+4\tau\hat{\mu}(1-\hat{\mu})}}{2(1+\tau)} \;.$$

Die Konfidenzintervalle sind in Abbildung 4.1 illustriert.

- Seien $\{X_k\}$ normalverteilt mit Mittelwert θ und Varianz σ^2, wobei σ^2 bekannt sei. Zum Beispiel, man will sehen, ob eine Abfüllmaschine richtig eingestellt ist, das heisst X_i ist das gemessene Gewicht, wobei die Varianz aus Erfahrung bekannt ist. Wir verwenden ein Intervall der Form $I = [\hat{\mu}-\varepsilon, \hat{\mu}+\varepsilon]$. Also ist

$$\mathbb{P}_\theta[\theta \in I] = \mathbb{P}_\theta[|\hat{\mu}-\theta| \leq \varepsilon] = \mathbb{P}_\theta\Big[\sqrt{n}\frac{|\hat{\mu}-\theta|}{\sigma} \leq \frac{\sqrt{n}\varepsilon}{\sigma}\Big] \;.$$

  Dies ergibt $\varepsilon = \Phi^{-1}(1-\alpha/2)\sigma/\sqrt{n}$. Somit hängt in diesem Fall die Breite des Konfidenzintervall nicht von $\hat{\mu}$ ab.

- Seien $\{X_k\}$ normalverteilt mit Mittelwert θ und Varianz σ^2, wobei σ^2 nicht bekannt sei. Wir schätzen daher die Varianz $S^2 = (n-1)^{-1}\sum_{k=1}^n (X_k-\hat{\mu})^2$. Dadurch wird die Intervalllänge datenabhängig. Wir schreiben jetzt

$$\mathbb{P}_\theta[\theta \in I] = \mathbb{P}_\theta\Big[\sqrt{n}\frac{|\hat{\mu}-\theta|}{S} \leq \frac{\sqrt{n}\varepsilon}{S}\Big] \;.$$

  Das Problem ist, dass wir nun keine Normalverteilung mehr haben, da auch S zufällig ist. Die Verteilung von $\sqrt{n}(\hat{\mu}-\theta)/S$ kann explizit berechnet werden, und heisst **t-Verteilung mit $n-1$ Freiheitsgraden**. Sie hat die Dichtefunktion

$$t_{n-1}(x) = \frac{\Gamma(n/2)}{\sqrt{\pi(n-1)}\Gamma((n-1)/2)(1+x^2/(n-1))^{n/2}} \;.$$

Lässt man $n \to \infty$, so konvergiert diese Dichte zur Dichte der Normalverteilung. Die Verteilungsfunktion der t-Verteilung lässt sich in Tabellenbüchern finden, falls n nicht zu gross ist. Für grosse n lässt sich die Verteilung durch die Normalverteilung approximieren. Finden wir c, so dass

$$\mathbb{P}_\theta\Big[\sqrt{n}\frac{|\hat{\mu}-\theta|}{S} \le c\Big] = 1-\alpha \ ,$$

so haben wir $\varepsilon = Sc/\sqrt{n}$.

4.8. Lineare Regression

Nehmen wir an, dass wir Daten $\{(X_i, Y_i)\}$ betrachten. Die Vektoren $\{(X_i, Y_i)\}$ seien unabhängig, wobei X_i deterministisch oder zufällig sein kann. Zum Beispiel, wir messen die Länge Y_i eines Stabes, wenn die Temperatur X_i ist. Oder, X_i ist die Körperlänge des Vaters, Y_i die Körperlänge der Tochter.

Wir machen folgende Annahmen. Es besteht der Zusammenhang $Y_i = aX_i + b + \varepsilon_i$, wobei a und b unbekannte Zahlen sind, und die ε_i sind iid normalverteilte Zufallsvariablen mit Mittelwert 0 und Varianz σ^2. Die Variablen $\{\varepsilon_i\}$ sind unabhängig von $\{X_i\}$.

Das Problem ist nun, a und b zu schätzen. Wir haben die Log-Likelihoodfunktion

$$\log L = -\sum_{k=1}^{n} \frac{(Y_k - aX_k - b)^2}{2\sigma^2} - n\log\sigma - \frac{n}{2}\log(2\pi) \ .$$

Ableiten nach a und b ergibt die Schätzer

$$\hat{a} = \frac{n\sum_{k=1}^n X_k Y_k - (\sum_{k=1}^n X_k)(\sum_{k=1}^n Y_k)}{n\sum_{k=1}^n X_k^2 - (\sum_{k=1}^n X_k)^2}$$

und

$$\hat{b} = \frac{1}{n}\sum_{k=1}^{n} Y_k - \hat{a}\frac{1}{n}\sum_{k=1}^{n} X_k \ .$$

Aus unserer Konstruktion sehen wir auch das Folgende. Wir suchen die Gerade, so dass der quadratische Abstand der Y_i von der Gerade minimal wird. Wir minimieren also den quadratischen Fehler.

Setzen wir $\bar{X} = \frac{1}{n}\sum_{k=1}^n X_k$, $\bar{Y} = \frac{1}{n}\sum_{k=1}^n Y_k$, $\overline{XY} = \frac{1}{n}\sum_{k=1}^n X_k Y_k$ und $\overline{X^2} = \frac{1}{n}\sum_{k=1}^n X_k^2$. Dann können wir den Schätzer schreiben als

$$\hat{a} = \frac{\overline{XY} - \bar{X}\bar{Y}}{\overline{X^2} - \bar{X}^2} \ , \qquad \hat{b} = \bar{Y} - \hat{a}\bar{X} \ .$$

Da $\overline{XY} - \bar{X}\bar{Y}$ ein Schätzer für die Kovarianz ist, und $\overline{X^2} - \bar{X}^2$ ein Schätzer für die Varianz von X ist, haben wir die Regressionsformel aus Abschnitt 2.5, wobei wir die Grössen durch Schätzer ersetzt haben. Man könnte daher auch den Schätzer

$$\hat{a} = \frac{\overline{XY} - \bar{X}\bar{Y}}{(n-1)^{-1}\sum_{k=1}^{n}(X_k - \bar{X})^2}$$

verwenden, der die Varianz unverfälscht schätzt. Falls nun die Verteilung von ε_i verschieden von der Normalverteilung ist, erhält man somit trotzdem einen sinnvollen Schätzer, falls die zweiten Momente existieren.

Oft ist man auch an der Varianz σ^2 der Fehler interessiert. Zum Beispiel, falls man X_{n+1} beobachtet, und ein Konfidenzintervall für Y_{n+1} konstruieren will. Falls wir a und b kennen würden, könnten wir ε_i beobachten, und die Varianz einfach schätzen. Somit ersetzen wir a, und b durch die Schätzer und erhalten

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}(Y_k - \hat{a}X_k - \hat{b})^2 \ .$$

Man kann sofort sehen, dass dies der Maximum-Likelihood-Schätzer für σ^2 im normalverteilten Fall ist.

Wie man aus den Formeln für $\hat{a}$ und $\hat{b}$ sehen kann, sind die oben konstruierten Schätzer nichts anderes, als die Parameter einer Gerade, die die quadratischen Abstände zu den Punkten minimiert. Daher nennt man den Schätzer auch **kleinste-Quadrate-Schätzer**.

5. Testtheorie

5.1. Statistische Tests

In vielen praktischen Situationen ist man nicht direkt an den Parametern interessiert, sondern hat eine Vermutung, die man gerne beweisen möchte. Es gibt dann nur zwei mögliche Antworten: "Vermutung richtig", oder "Vermutung falsch". Zum Beispiel: Zwei Schulklassen schreiben die gleiche Klausur. Die eine Klasse hat Notendurchschnitt 3.1, die andere Notendurchschnitt 3.0 Kann man daraus schliessen, dass die zweite Klasse im Durschnitt besser ist, oder sind die Klassen gleich gut, und die zweite Klasse hatte nur zufällig den besseren Durchschnitt. Oder, ein neues Medikament wird getestet. Die Hälfte der Versuchspersonen bekommt das neue Medikament, die andere Hälfte nur ein Placebo. Kann man nun schliessen, dass das neue Medikament eine Wirkung hat?

Wir haben einen messbaren Raum $(\Omega, \mathcal{F})$ und eine Klasse von Wahrscheinlichkeitsmassen $\{\mathbb{P}_\theta : \theta \in \Theta\}$. Nun definieren wir eine Hypothese $\Theta_0 \subset \Theta$. Um den Test durchführen zu können, haben wir Daten $X_1, \ldots, X_n$. Ein Test ist nun eine Entscheidungsregel $\mathcal{E} : \mathbb{R}^n \to \{0, 1\}$. 0 bedeutet, die Hypothese wird beibehalten (dies bedeutet nicht, dass die Hypothese richtig ist), 1 bedeutet, die Hypothese wird verworfen. Den Test kann man auch durch den **kritischen Bereich**

$$K = \{\omega : \mathcal{E}(X_1, \ldots, X_n) = 1\}$$

charakterisieren.

Ein Test sollte nach Möglichkeit die richtige Entscheidung liefern. Das heisst, $\mathbb{P}_\theta[K]$ sollte auf $\Theta \setminus \Theta_0$ gross sein und auf Θ_0 klein.

Definition 5.1. *Wir sagen, ein Test hat das* **(Signifikanz-) Niveau** α*, falls* $\mathbb{P}_\theta[K] \leq \alpha$ *für alle* $\theta \in \Theta_0$*. Falls die Hypothese verworfen wird, obwohl sie richtig ist, sprechen wir vom* **Fehler erster Art**.

Beispiele

- Jemand hat gehört, dass die Ein-Euro-Münze nicht fair sei, und will dies nun untersuchen. Er hat die Hypothese "Münze fair" ($\Theta_0 = \{\frac{1}{2}\}$) und macht einen Test mit dem kritischen Bereich

$$K = \left\{ \left| \sum_{k=1}^{n} X_k - \frac{n}{2} \right| \geq c(n) \right\},$$

wobei $X_k = 1$, falls Kopf geworfen wird, und $X_k = 0$, falls Zahl geworfen wird. Mit der Normalapproximation erhält man

$$\mathbb{P}_{1/2}[K] = \mathbb{P}_{1/2}\Big[\frac{|\sum_{k=1}^n X_k - n/2|}{\sqrt{n/4}} \geq \frac{2c}{\sqrt{n}}\Big] \approx 2\Big(1 - \Phi\Big(\frac{2c}{\sqrt{n}}\Big)\Big) \leq \alpha \; .$$

Somit haben wir

$$c(n,\alpha) \geq \frac{\sqrt{n}}{2}\Phi^{-1}(1-\alpha/2) \; .$$

Wählt man also zum Beispiel $\alpha = 5\%$, ergibt sich $c(n,\alpha) \approx 0.98\sqrt{n}$.

Die obige Formel gibt nun zu jedem Niveau α und Anzahl Experimenten n eine Abweichung vom Mittelwert $\frac{1}{2}$, so dass der Fehler 1. Art das Niveau α einhält. Man möchte aber nach Möglichkeit auch entdecken, wenn $\theta = \frac{1}{2}$. Das lässt sich generell nicht erreichen, da θ beliebig nahe bei $\frac{1}{2}$ liegen kann, und daher auch $\hat{\mu}$ sehr nahe bei $\frac{1}{2}$ liegen wird. Wir können aber eine Abweichung Δ bestimmen, die wir als relevant bezeichnen wollen. Wir verlangen dann, dass

$$\mathbb{P}_\theta[K] \geq 1 - \beta \qquad \forall \theta : |\theta - \tfrac{1}{2}| \geq \Delta$$

mindestens mit Wahrscheinlichkeit $1-\beta$ erkannt wird. Die Menge $\Theta_1 = [0, \frac{1}{2} - \Delta] \cup [\frac{1}{2} + \Delta, 1]$ nennen wir die **Alternative** und die Wahrscheinlichkeit $\mathbb{P}_\theta[K^c]$ für $\theta \in \Theta_1$ nennen wir **Fehler zweiter Art**. Die Wahrscheinlichkeit einen Fehler zweiter Art zu machen, fällt mit dem Abstand zu $\frac{1}{2}$ und ist symmetrisch. Daher genügt es, $\mathbb{P}_{\theta_1}[K]$ für $\theta_1 = \frac{1}{2} + \Delta$ zu berechnen. Wir erhalten somit

$$\begin{aligned}
\mathbb{P}_\theta[K] \geq \mathbb{P}_{\theta_1}[K] &= 1 - \mathbb{P}_{\theta_1}\Big[\frac{n}{2} - c(n,\alpha) < \sum_{k=1}^n X_k < \frac{n}{2} + c(n,\alpha)\Big] \\
&= 1 - \mathbb{P}_{\theta_1}\Big[-\frac{c(n,\alpha) + n\Delta}{\sqrt{n\theta_1(1-\theta_1)}} < \frac{\sum_{k=1}^n X_k - n\theta_1}{\sqrt{n\theta_1(1-\theta_1)}} < \frac{c(n,\alpha) - n\Delta}{\sqrt{n\theta_1(1-\theta_1)}}\Big] \\
&\approx 2 - \Phi\Big(\frac{c(n,\alpha) + n\Delta}{\sqrt{n\theta_1(1-\theta_1)}}\Big) - \Phi\Big(\frac{c(n,\alpha) - n\Delta}{\sqrt{n\theta_1(1-\theta_1)}}\Big) \geq 1 - \beta \; .
\end{aligned}$$

Wählen wir zum Beispiel wieder $\alpha = \beta = 5\%$, so müssen wir n folgendermassen wählen

Δ	0.1	0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01
n	319	395	502	657	897	1294	2025	3604	8116	32482

- Bei einer Untersuchung im Jahre 1973, ob übersinnliche Begabungen existieren, wurden mit 15 Versuchspersonen je 500 Versuche gemacht, wobei eine Person

eine von vier Farben wählte, und die Versuchsperson die Farbe nennen sollte. Von den insgesamt 7500 Versuchen waren 2006 erfolgreich. Die Hypothese ist, dass dies reiner Zufall war. Hier ist $\Theta_0 = \{\frac{1}{4}\}$. Wir erhalten

$$\begin{aligned}\mathbb{P}_{1/4}\Big[\sum_{k=1}^{7500} X_k \geq 2006\Big] &= \mathbb{P}_{1/4}\Big[\frac{\sum_{k=1}^{7500} X_k - 1875}{\sqrt{1406.25}} \geq \frac{131}{\sqrt{1406.25}}\Big] \\ &\approx 1 - \Phi(3.4933) \approx 0.0002385\ .\end{aligned}$$

Die Hypothese wird also bereits auf sehr kleinem Niveau verworfen.

- **Vorzeichentest für den Median** Seien $\{Y_k : 1 \leq k \leq n\}$ Daten von einer Verteilung. Die Hypothese ist "Der Median ist M", wobei wir annehmen, dass $\mathbb{P}_\theta[Y_k = M] = 0$ für alle $\theta \in \Theta$. Das heisst, die Hypothese ist $\mathbb{P}_{\theta_0}[Y_k < M] = \mathbb{P}_{\theta_0}[Y_k > M] = \frac{1}{2}$. Ein einfacher Test ist der folgende. Wir setzen

$$X_k = \operatorname{sign}(Y_k - M) = \begin{cases} 1, & \text{falls } Y_k > M, \\ -1, & \text{falls } Y_k < M. \end{cases}$$

Unter der Hypothese haben wir somit $\mathbb{P}_\theta[X_k = 1] = \mathbb{P}[X_k = -1] = \frac{1}{2}$. Der Vorzeichentest ist dann der Test mit dem kritischen Bereich

$$K = \Big\{\Big|\sum_{k=1}^{n} \mathbb{1}_{X_k=1} - \frac{n}{2}\Big| \geq c\Big\}\ .$$

Diesen Test haben wir oben als 0-1-Experiment formuliert betrachtet.

Eine weitere hilfreiche statistische Grösse ist der p-Wert. Als p-Wert bezeichnen wir die Wahrscheinlichkeit unter der Hypothese, dass die Teststatistik einen mindestens so extremen Wert ergibt wie in den betrachteten Daten. In den meisten Fällen ist dies das selbe, wie das grösste Signifikanzniveau, bei dem die Hypothese verworfen wird.

Beispiel Nehmen wir an, wir werfen eine Ein-Euro-Münze 100 Mal. Wir erwarten 50 Mal Kopf. Nehmen wir an, wir erhalten 59 Mal Kopf. Die Wahrscheinlichkeit, unter der Hypothese mindestens 59 Mal Kopf zu erhalten ist $p = 0.044313$. Dies ist der p-Wert des Experiments. Das bedeutet, dass ein Test "Münze fair" gegen die Alternative "Kopf ist wahrscheinlicher" auf dem 5%-Niveau verwerfen würde, aber nicht mehr auf dem 4%-Niveau.

5.2. Der Likelihood-Quotienten-Test

Nehmen wir an, wir haben nur zwei mögliche Verteilungen, $\Theta = \{\theta_0, \theta_1\}$. Die Hypothese ist $\Theta_0 = \{\theta_0\}$. Die Daten sind $\{X_k : 1 \le k \le n\}$, mit Likelihoodfunktion $L_\theta(x_1, x_2, \ldots, x_n)$. Der Likelihood-Quotienten-Test hat den kritischen Bereich

$$K = \Big\{ \frac{L_{\theta_1}(x_1, x_2, \ldots, x_n)}{L_{\theta_0}(x_1, x_2, \ldots, x_n)} \ge c \Big\} ,$$

wobei wir $z/0 = \infty$ setzen. Der Quotient $L_{\theta_1}(x_1, \ldots, x_n)/L_{\theta_0}(x_1, \ldots, x_n)$ heisst **Likelihood-Quotient**. Die Konstante c muss dann so bestimmt werden, dass das Niveau α nicht überschritten wird: $\mathbb{P}_{\theta_0}[K] \le \alpha$.

Satz 5.2. (Neymann–Pearson-Lemma) *Der Likelihood-Quotienten-Test ist optimal. Das heisst, ist $\tilde{K}$ der kritische Bereich eines beliebigen Tests mit $\mathbb{P}_{\theta_0}[\tilde{K}] \le \mathbb{P}_{\theta_0}[K]$, so gilt $\mathbb{P}_{\theta_1}[\tilde{K}] \le \mathbb{P}_{\theta_1}[K]$.*

Beweis. Wir betrachten im Beweis nur den absolutstetigen Fall. Der Beweis im diskreten Fall folgt analog. Bezeichnen wir die Entscheidungsregeln mit $\mathcal{E} = \mathbb{1}_K$ und $\tilde{\mathcal{E}} = \mathbb{1}_{\tilde{K}}$. Die Voraussetzung lässt sich dann schreiben als $\int(\mathcal{E} - \tilde{\mathcal{E}})L_{\theta_0}\, \mathrm{d}\boldsymbol{x} \ge 0$, wobei wir die Argumente im Integral weglassen. Wir bemerken zuerst, dass $\mathcal{E} = 1$, falls $L_{\theta_0} = 0$ und $L_{\theta_1} > 0$, das heisst, $\mathcal{E} - \tilde{\mathcal{E}} \ge 0$. Weiter gilt,

$$(\mathcal{E} - \tilde{\mathcal{E}})\Big(\frac{L_{\theta_1}}{L_{\theta_0}} - c\Big) \ge 0 ,$$

da beide Terme das selbe Vorzeichen haben. Dann gilt

$$\begin{aligned}
\int(\mathcal{E} - \tilde{\mathcal{E}})L_{\theta_1}\, \mathrm{d}\boldsymbol{x} &= \int_{\{L_{\theta_0}=0\}}(\mathcal{E} - \tilde{\mathcal{E}})L_{\theta_1}\, \mathrm{d}\boldsymbol{x} + \int_{\{L_{\theta_0}>0\}}(\mathcal{E} - \tilde{\mathcal{E}})\frac{L_{\theta_1}}{L_{\theta_0}}L_{\theta_0}\, \mathrm{d}\boldsymbol{x} \\
&\ge \int_{\{L_{\theta_0}>0\}}(\mathcal{E} - \tilde{\mathcal{E}})\frac{L_{\theta_1}}{L_{\theta_0}}L_{\theta_0}\, \mathrm{d}\boldsymbol{x} = \int(\mathcal{E} - \tilde{\mathcal{E}})\frac{L_{\theta_1}}{L_{\theta_0}}L_{\theta_0}\, \mathrm{d}\boldsymbol{x} \\
&= \int(\mathcal{E} - \tilde{\mathcal{E}})\Big(\frac{L_{\theta_1}}{L_{\theta_0}} - c\Big)L_{\theta_0}\, \mathrm{d}\boldsymbol{x} + c\int(\mathcal{E} - \tilde{\mathcal{E}})L_{\theta_0}\, \mathrm{d}\boldsymbol{x} \ge 0 .
\end{aligned}$$

Dies ist die Behauptung. □

5.3. Parametertests für die Normalverteilung

5.3.1. Student's t-Test

Seien $X_1, X_2, \ldots, X_n$ normalverteilt mit Mittelwert m und Varianz σ^2. Wir wollen testen, ob $m = m_0$. Die Hypothese ist also $\Theta_0 = \{m = m_0\}$, die Alternative ist

$\Theta_1 = \{m \neq m_0\}$. Eine einfache Idee ist die Hypothese zu verwerfen, falls $|\hat{\mu} - m_0|$ gross ist. Wir wissen, dass unter der Hypothese $\hat{\mu}$ normalverteilt ist mit Mittelwert m_0 und Varianz σ^2/n. Also ist $\sqrt{n}(\hat{\mu} - m_0)/\sigma$ standard normalverteilt. Das Problem ist aber, dass wir σ^2 nicht kennen.

Wir müssen statt σ^2 den Schätzer

$$S^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \hat{\mu})^2$$

verwenden. Dies ergibt die Teststatistik

$$T = \sqrt{n}\frac{\hat{\mu} - m_0}{S} \ .$$

Wir haben früher bemerkt, dass T eine t-Verteilung mit $n-1$ Freiheitsgraden besitzt. Wir erhalten somit den Test

$$K = \{|T| \geq c(n)\} = \{\sqrt{n}|\hat{\mu} - m_0| \geq Sc(n)\} \ .$$

Falls n gross ist, kann man statt der t-Verteilung die Standard-Normalverteilung verwenden.

Dieser Test heisst Student's t-Test, da er von William S. Gosset unter dem Pseudonym "Student" veröffentlicht wurde, da sein Arbeitgeber Publikationen nicht erlaubte.

5.3.2. χ^2-Streuungstest

Nehmen wir an, wir kennen den Mittelwert und wollen die Hypothese $\{\sigma^2 = \sigma_0^2\}$ (oder $\{\sigma^2 \geq \sigma_0^2\}$) testen. Kennen wir den Mittelwert m_0, so wissen wir, dass

$$T = \sum_{k=1}^{n}\frac{(X_k - m_0)^2}{\sigma_0^2}$$

χ^2 mit n Freiheitsgraden verteilt ist. Das heisst, T hat die Verteilung mit der Dichte

$$f_n(x) = \frac{1}{2^{n/2}\Gamma(n/2)}x^{n/2-1}\mathrm{e}^{-x/2}\mathbb{1}_{x>0} \ .$$

Wir können somit einen zweiseitigen Test $K = \{T < c_1 \text{ oder } T > c_2\}$ durchführen. Testen wir $\{\sigma^2 \geq \sigma_0^2\}$, so muss man einen einseitigen Test $K = \{T < c\}$ verwenden, da T auch klein wird, wenn die richtige Varianz σ^2 klein wird.

Ist der Mittelwert nicht bekannt, so verwenden wir den Schätzer $\hat{\mu}$ und die Teststatistik

$$T = \sum_{k=1}^{n} \frac{(X_k - \hat{\mu})^2}{\sigma_0^2} \ .$$

Unter der Nullhypothese ist diese Statistik χ^2 verteilt mit $n-1$ Freiheitsgraden. Die Quantile der χ^2-Verteilung findet man in Tabellenbüchern.

5.4. Vergleich von zwei Verteilungen

In vielen Studien muss man zwei Verteilungen miteinander vergleichen. Zum Beispiel hat man eine Gruppe, die ein bestimmtes Medikament erhält, und eine Kontrollgruppe, die ein Placebo erhält. Oder man hat zwei Maschinen, die rote Blutkörperchen zählen. Man will dann wissen, ob die beiden Verteilungen gleich sind oder verschieden.

Seien $X_1, X_2, \ldots, X_n$ und $Y_1, Y_2, \ldots, Y_m$ unabhängige Zufallsvariablen. Die Variablen $\{X_k\}$ haben alle die Verteilung F_1 und die Variablen $\{Y_k\}$ haben alle die Verteilung F_2. Die Hypothese ist, dass die beiden Verteilungen identisch sind.

5.4.1. t-Test

Beginnen wir mit der Annahme, dass die X normalverteilt sind mit Mittelwert μ_1 und Varianz σ^2, die Y normalverteilt mit Mittelwert μ_2 und Varianz σ^2. Die Hypothese lautet also $\{\mu_1 = \mu_2\}$.

Bezeichnen wir mit $\overline{X} = n^{-1} \sum X_k$ und $\overline{Y} = m^{-1} \sum Y_k$ die empirischen Mittelwerte der beiden Gruppen. Da $\overline{X}$ und $\overline{Y}$ für grosse n und m nahe beim Mittelwert liegen, ist eine einfache Testidee, die Hypothese zu verwerfen, wenn $|\overline{X} - \overline{Y}|$ gross ist. Die Varianz lässt sich dann schätzen als

$$S^2 = \frac{1}{n+m-2} \Big(\sum_{k=1}^{n} (X_k - \overline{X})^2 + \sum_{k=1}^{m} (Y_k - \overline{Y})^2 \Big) \ .$$

Da $\overline{X} - \mu_1$ normalverteilt ist mit Mittelwert 0 und Varianz σ^2/n und $\overline{Y} - \mu_2$ normalverteilt ist mit Mittelwert 0 und Varianz σ^2/m, hat $\overline{X} - \overline{Y}$ die Varianz $(n^{-1} + m^{-1})\sigma^2$. Wir verwenden daher die Teststatistik

$$T = \frac{\overline{X} - \overline{Y}}{S\sqrt{n^{-1} + m^{-1}}} \ .$$

Es zeigt sich, dass T t-verteilt ist mit $n+m-2$ Freiheitsgraden. Der gesuchte Test ist dann

$$K = \{|T| \geq c(n+m)\} = \{\sqrt{nm}|\overline{X} - \overline{Y}| \geq c(n+m)S\sqrt{n+m}\} \ .$$

5.4.2. F-Test

Nehmen wir an, wir wollen testen, ob zwei unabhängige Stichproben unterschiedliche Varianzen haben. Nehmen wir zuerst an, dass wir die Mittelwerte μ_1 und μ_2 kennen. Wir schätzen die Varianzen als

$$S_1^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \mu_1)^2 \ , \qquad S_2^2 = \frac{1}{m}\sum_{k=1}^{m}(Y_k - \mu_2)^2$$

Die Teststatistik $T = S_1/S_2$ ist dann F-verteilt mit n, m Freiheitsgraden. Die F-Verteilung mit n, m Freiheitsgraden ist die Verteilung mit der Dichtefunktion

$$f_{n,m}(x) = n^{n/2}m^{m/2}\frac{\Gamma(n/2+m/2)}{\Gamma(n/2)\Gamma(m/2)}\frac{x^{n/2-1}}{(nx+m)^{(m+n)/2}} \ .$$

Mit einem zweiseitigen Test lässt sich dann der Test durchführen.

Wollen wir nur $\{\sigma_1^2 \leq \sigma_2^2\}$ testen, so können wir einen einseitigen Test wählen.

Sind die Mittelwerte nicht bekannt, so ersetzen wir die Mittelwerte durch die empirischen Mittelwerte, also

$$S_1^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X})^2 \ , \qquad S_2^2 = \frac{1}{m-1}\sum_{k=1}^{m}(Y_k - \bar{Y})^2$$

Die Teststatistik T ist dann F-verteilt mit $n-1, m-1$ Freiheitsgraden. Der Test lässt sich somit analog durchführen.

5.4.3. Wilcoxon-Test

Die Annahme der Normalverteilung ist aber relativ stark. Wir wollen nun keine Annahme über die Verteilung machen. Ordnen wir die Variablen X und Y nach ihrer Grösse, müssten unter der Hypothese die X und Y in zufälliger Reihenfolge angeordnet sein. Zählen wir für jedes X_k wieviele Y_j grösser sind, so müsste im Durchschnitt die Hälfte der Y grösser sein. Bilden wir also die Statistik

$$U = \sum_{i,j} \mathbb{1}_{X_i<Y_j} \ .$$

Wir nehmen nun an, dass unter der Hypothese $\mathbb{P}[X_i < Y_j] = \frac{1}{2}$ gilt. Wir erhalten $\mathbb{E}[U] = \frac{1}{2}nm$. Ein einfacher Test ist dann

$$K = \{|U - \tfrac{1}{2}nm| \geq c\} \ .$$

Sind beide n und m gross, dann ist $(U - \frac{1}{2}nm)/\sqrt{\mathrm{Var}[U]}$ annähernd standardnormalverteilt. Wir beweisen dies hier nicht. Wir müssen also die Varianz von U bestimmen. Die Formel ergibt

$$\mathrm{Var}[U] = \mathrm{Var}\Big[\sum_{i,j} \mathbb{1}_{X_i<Y_j}\Big] = \sum_{i,j} \mathrm{Var}[\mathbb{1}_{X_i<Y_j}] + \sum_{(i,j)\neq(k,\ell)} \mathrm{Cov}[\mathbb{1}_{X_i<Y_j}, \mathbb{1}_{X_k<Y_\ell}] \ .$$

Wir haben, dass $\mathrm{Var}[\mathbb{1}_{X_i<Y_j}] = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$. Dieser Term kommt nm mal vor. Ist $i \neq k$ und $j \neq \ell$, dann ist die Kovarianz 0. Ist $i \neq k$ und $j = \ell$, dann erhalten wir

$$\mathbb{E}[\mathbb{1}_{X_i<Y_j}\mathbb{1}_{X_k<Y_j}] = \mathbb{E}[\mathbb{1}_{X_i\vee X_k<Y_j}] = \tfrac{2}{6} = \tfrac{1}{3} \ .$$

Wir haben nämlich sechs verschiedene Anordnungen von X_i, X_k und Y_j, die alle die gleiche Wahrscheinlichkeit haben. Und bei zwei dieser Anordnungen ist Y_j an dritter Stelle. Somit haben wir $\mathrm{Cov}[\mathbb{1}_{X_i<Y_j}, \mathbb{1}_{X_k<Y_j}] = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$. Dieser Wert kommt $n(n-1)m$ vor. Analog haben wir $nm(m-1)$ mal den Ausdruck $\mathrm{Cov}[\mathbb{1}_{X_i<Y_j}, \mathbb{1}_{X_i<Y_\ell}] = \frac{1}{12}$. Also ergibt sich

$$\mathrm{Var}[U] = \frac{nm}{4} + \frac{n(n-1)m}{12} + \frac{nm(m-1)}{12} = \frac{nm(n+m+1)}{12} \ .$$

Der gesuchte Test ist dann

$$K = \Big\{\Big|U - \frac{nm}{2}\Big| \geq c\Big\} = \Big\{\frac{\sqrt{3}|2U - nm|}{\sqrt{nm(n+m+1)}} \geq \frac{c\sqrt{12}}{\sqrt{nm(n+m+1)}}\Big\} \ .$$

Die Konstante c kann also aus der Normalapproximation

$$c = \sqrt{\frac{nm(n+m+1)}{12}}\Phi^{-1}(1-\alpha/2)$$

bestimmt werden.

5.4.4. Rangsummentest

Äquivalent zum Wilcoxon-Test ist der Rangsummentest. Wir ordnen die Daten und bestimmen die Summe der Ränge der Y. Zum Beispiel bei der Anordnung

X	X	Y	Y	X	X	Y	Y	Y	X
1	2	3	4	5	6	7	8	9	10

erhalten wir die Rangsumme $3+4+7+8+9=31$. Ist $\{Y_{(i)}\}$ die Ordnungsstatistik von $\{Y_j\}$, das heisst $Y_{(1)} < Y_{(2)} < \cdots < Y_{(m)}$, dann erhalten wir für die Rangsumme

$$T_Y = \sum_{j=1}^{m} \text{Rang}\,(Y_j) = \sum_{i=1}^{m} \text{Rang}\,(Y_{(i)}) = \sum_{i=1}^{m} \Big(\sum_{k=1}^{n} \mathbb{1}_{X_k < Y_{(i)}} + i \Big) = U + \tfrac{1}{2} m(m+1)\ .$$

Somit lässt sich der Rangsummentest auf den Wilcoxon-Test zurückführen.

5.4.5. Verbundene Stichproben

Oft hat man die Situation, dass die beiden Datenreihen abhängig sind. Zum Beispiel, in einem Labor zählt man weisse Blutkörperchen mit einer Maschine. Als man eine neue Maschine anschafft, will man wissen, ob die neue und die alte Maschine gleich zählen. Man nimmt daher aus Blutproben je einen Tropfen, und zählt den einen Tropfen mit der alten Maschine (X_i), den anderen Tropfen mit der neuen Maschine (Y_i). Die beiden Werte X_i und Y_i werden im Normalfall verschieden sein, da die Menge der weissen Blutkörperchen in den Tropfen zufällig ist. Die Datenreihen werden aber abhängig sein, da, falls sich viele weisse Blutkörperchen in der i-ten Probe befinden, beide Werte X_i und Y_i gross sein werden.

Ein weiteres Beispiel ist der Test eines Medikamentes. Nehmen wir an, wir möchten ein Medikament gegen einen hohen Cholesterinspiegel testen. Wir erheben dann die Cholesterinwerte X_i vor der Einnahme des Medikamentes. Nach der Einnahme des Medikamentes über einen Monat, werden die Cholesterinwerte Y_i wieder erhoben. Diese Werte sind dann natürlich abhängig voneinander.

Wir wollen nun testen, ob die Randverteilungen einer verbundenen Stichprobe $\{(X_i, Y_i)\}$ identisch sind.

t-Test Nehmen wir an, (X_i, Y_i) sei zweidimensional normalverteilt, und die Randverteilungen seien identisch. Dann sind die $X_i - Y_i$ normalverteilt mit Mittelwert 0. Üblicherweise wird die Varianz nicht bekannt sein, und muss geschätzt werden. Wir schätzen somit den Mittelwert

$$\bar{d} = \frac{1}{n} \sum_{k=1}^{n} X_k - Y_k$$

und die Varianz

$$S^2 = \frac{1}{n-1} \sum_{k=1}^{n} (X_k - Y_k - \bar{d})^2\ .$$

Dann ist die Teststatistik unter der Hypothese

$$T = \sqrt{n}\frac{\bar{d}}{S}$$

t-verteilt mit $n-1$ Freiheitsgraden. Der t-Test sagt uns daher, ob Differenz der Mittelwerte vereinbar ist mit den Daten. Der Test kann auch einseitig gemacht werden. Im Beispiel mit den Cholesterinwerten will man effektiv zeigen, dass $Y_i < X_i$, und somit wird der kritische Bereich von der Form $\{T > c\}$ für ein $c > 0$ sein.

Vorzeichentest Eine einfache Testidee ist das Vorzeichen von $X_i - Y_i$ zu betrachten. Wir zählen, wie oft das Vorzeichen positiv ist, $V = \sum_{i=1}^n \mathbb{1}_{X_i > Y_i}$. Unter der Hypothese, dass beide Maschinen gleich zählen, haben wir dann $\mathbb{E}[\mathbb{1}_{X_i<Y_i}] = \mathbb{P}[X_i > Y_i] = \frac{1}{2}$ (wir nehmen an, dass $\mathbb{P}[X_i = Y_i] = 0$). Die Variablen $\mathbb{1}_{X_i>Y_i}$ sind dann unabhängige 0-1-Experimente mit Erfolgsparameter $\frac{1}{2}$. Für den Test

$$K = \{|V - n/2| \geq c\} = \Big\{\Big|\sum_{i=1}^n \mathbb{1}_{X_i>Y_i} - \frac{n}{2}\Big| \geq c\Big\}$$

lässt sich dann c aus der Binomialverteilung oder aus der Normalapproximation bestimmen.

5.4.6. χ^2-Unabhängigkeitstest

Nehmen wir an, dass wir zwei Merkmale haben, von denen wir vermuten, dass sie unabhängig sind. Zum Beispiel, aus einer Kfz-Unfallstatistik kennen wir die Merkmale männlich/weiblich und das Bundesland des Unfallfahrers. Wir teilen die Daten nun in Klassen mit den Merkmalen ein und zählen, wie viele der Daten in die entsprechende Klasse fallen. Sei also h_{ij} die Anzahl der Daten, die 1. Merkmal i und zweites Merkmal j haben, wobei $i \in \{1, 2, \ldots, n\}$ und $j \in \{1, 2, \ldots, m\}$. Dann bilden wir die Randhäufigkeiten

$$h_{i\cdot} = \sum_{j=1}^m h_{ij} \ , \qquad\qquad h_{\cdot j} = \sum_{i=1}^n h_{ij} \ .$$

Damit wir eine Normalapproximation machen dürfen, müssen wir die Klassen so wählen, dass $h_{ij} \geq 5$ für alle i, j. Ist dies nicht der Fall, müssen wir Klassen zusammenfassen. Die Gesamtanzahl Daten bezeichnen wir mit $h_{\cdot\cdot} = \sum_{i=1}^n \sum_{j=1}^m h_{ij}$.

Sind die Merkmale wirklich unabhängig, so erwarten wir

$$\hat{h}_{ij} = \frac{h_{i\cdot}}{h_{\cdot\cdot}} \frac{h_{\cdot j}}{h_{\cdot\cdot}} h_{\cdot\cdot}$$

Daten mit den Merkmalen i, j. Die Varianz ist

$$h_{..}\frac{h_{i.}h_{.j}}{h_{..}^2}\left(1-\frac{h_{i.}h_{.j}}{h_{..}^2}\right)\approx\frac{h_{i.}h_{.j}}{h_{..}}\ ,$$

vorausgesetzt, dass $\frac{h_{i.}h_{.j}}{h_{..}^2}$ klein ist (viele Klassen). Somit ist

$$\frac{h_{ij}-\hat{h}_{ij}}{\sqrt{\hat{h}_{ij}}}$$

näherungsweise standardnormalverteilt. Das heisst, die Teststatistik

$$T=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(h_{ij}-\hat{h}_{ij})^2}{\hat{h}_{ij}}$$

ist näherungsweise χ^2 verteilt mit $(n-1)(m-1)$ Freiheitsgraden. Kennen wir $h_{i.}$ und $h_{.j}$, so können wir nur $(n-1)(m-1)$ der Zellen frei wählen. Die restlichen $m+n-1$ Zellen sind durch die Randhäufigkeiten bestimmt.

5.5. Verteilungstests

Wenn wir die Parameter einer Verteilung aus Daten schätzen, wissen wir ja nicht, ob die angenommene Klasse von Verteilungen auch die korrekte Verteilung enthält. Man sollte daher testen, ob die gefundene Verteilung auch eine zu den Daten passende Verteilung ist. Wir betrachten nun zwei mögliche Testverfahren.

5.5.1. Der χ^2-Anpassungstest

Ein einfacher Test bedient sich der Ideen des χ^2-Unabhängigkeitstest (siehe Abschnitt 5.4.6). Wir teilen den Wertebereich der Daten in Zellen ein, so dass mindestens 5 Daten pro Zelle vorkommen. Nehmen wir an, dass wir m Zellen haben, die jeweils n_k Daten enthalten. Sei $n=\sum_{k=1}^{m}n_k$ die Anzahl der Daten. Wollen wir nun testen, ob $F_0(x)$ eine Verteilung ist, die mit den Daten vereinbar ist, so berechnen wir zuerst die erwartete Anzahl Daten $\bar{n}_k=F_{0,k}n$ pro Zelle, wobei $F_{0,k}$ die Wahrscheinlichkeit unter $F_0(x)$ ist, dass ein Datenpunkt in der k-ten Zelle liegt. Wie im χ^2-Unabhängigkeitstest ist die Teststatistik

$$T=\sum_{k=1}^{m}\frac{(n_k-\bar{n}_k)^2}{\bar{n}_k}$$

näherungsweise χ^2-verteilt mit $m-1$ Freiheitsgraden.

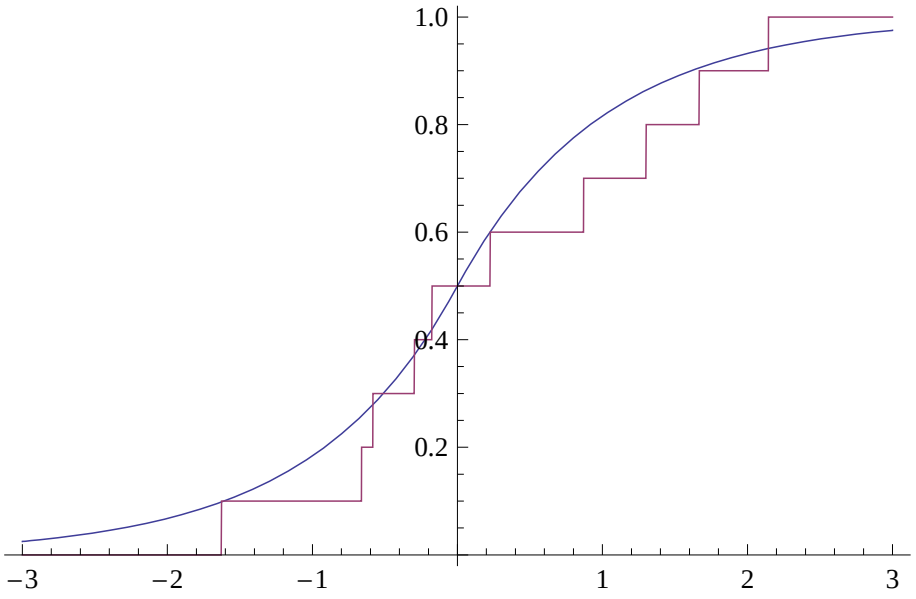

Abbildung 5.1: *Empirische und wirkliche Verteilung*

5.5.2. Der Kolmogorov–Smirnov-Test

Der Nachteil des χ^2-Anpassungstests ist, dass die Testvariablen normalverteilt sein sollten, oder dass man sehr viele Daten haben sollte. Der Kolmogorov–Smirnov-Test macht keine Annahmen diesbezüglich. Er basiert auf dem Satz von Glivenko–Cantelli.

Satz 5.3. *Sei $F_n(x)$ die empirische Verteilungsfunktion*

$$F_n(x) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{X_k \leq x} \ .$$

Dann konvergiert $F_n(x)$ fast sicher gleichmässig gegen $F_X(x)$, das heisst

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}} |F_n(x) - F_X(x)| = 0 \ .$$

Beweis. Nach dem Gesetz der grossen Zahl konvergiert $F_n(x)$ nach $F_X(x)$ für alle x. Weiter konvergiert $F_n(x-) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_{X_k<x}$ nach $F_X(x-)$. Sei $\varepsilon > 0$. Wir wählen $-\infty = t_0 < t_1 < \cdots < t_m = \infty$, so dass $F_X(t_k-) - F_X(t_{k-1}) \leq \frac{1}{2}\varepsilon$ für alle $1 \leq k \leq m$. Zum Beispiel können wir $t_k = \sup\{t : F_X(t) \leq F_X(t_{k-1}) + \frac{1}{2}\varepsilon$ wählen, da $F_X(t)$ rechtsstetig ist. Dann gibt es ein n_0, so dass $|F_n(t_k) - F_X(t_k)| < \frac{1}{2}\varepsilon$ und

$|F_n(t_k-) - F_X(t_k-)| < \frac{1}{2}\varepsilon$ für alle $n \geq n_0$ und alle $1 \leq k \leq m-1$ gilt. Sei $x \in \mathbb{R}$. Dann existiert ein k, so dass $t_{k-1} \leq x < t_k$. Wir haben dann

$$\begin{aligned} F_n(x) - F_X(x) &\leq F_n(t_k-) - F_X(t_{k-1}) \\ &= (F_n(t_k-) - F_X(t_k-)) + (F_X(t_k-) - F_X(t_{k-1})) < \tfrac{1}{2}\varepsilon + \tfrac{1}{2}\varepsilon = \varepsilon \ , \end{aligned}$$

und

$$\begin{aligned} F_n(x) - F_X(x) &\geq F_n(t_{k-1}) - F_X(t_k-) \\ &= (F_n(t_{k-1}) - F_X(t_{k-1})) - (F_X(t_k-) - F_X(t_{k-1})) > -\varepsilon \ , \end{aligned}$$

Also gilt $|F_n(x) - F_X(x)| < \varepsilon$. Dies zeigt die gleichmässige Konvergenz. □

Will man nun testen, ob die Daten die Verteilung $F_0(x)$ haben, bildet man die absoluten Differenzen

$$d_{0,k} = |F_n(X_{(k-1)}) - F_0(X_{(k)})| \ , \qquad d_{1,k} = |F_n(X_{(k)}) - F_0(X_{(k)})| \ ,$$

wobei $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ die Ordnungsstatistik bezeichnet. Hierbei benutzen wir $F_n(X_{(0)}) = 0$. Setzen wir $d = \max\{d_{i,k} : i \in \{0,1\}, 1 \leq k \leq n\}$, so können wir die Hypothese verwerfen, falls d einen kritischen Wert überschreitet. Die kritischen Werte kann man in Tabellen finden. Die Teststatistik ist unter der Hypothese unabhängig von der Verteilung F_0, falls $F_0(x)$ stetig ist, da unter der Nullhypothese $F_0(X_k)$ in $(0,1)$ gleichverteilt ist, und $F_n(X_k)$ eine Permutation der Werte i/n für $i \in \{1, 2, \ldots, n\}$ ist. In der Tat ist für eine stetige Verteilungsfunktion für $x \in (0,1)$

$$\mathbb{P}[F_0(X) \leq x] = \mathbb{P}[X \leq F_0^{-1}(x)] = F_0(F_0^{-1}(x)) = x \ .$$

5.6. Konfidenzintervalle und Tests

Vergleichen wir Konfidenzintervalle und Testtheorie, können wir erkennen, dass ein Zusammenhang besteht. Wir können also ein Konfidenzintervall zum Niveau α auch definieren als die Menge aller θ, für die ein Test zum Signifikanzniveau α die Hypothese $\Theta = \{\theta\}$ nicht verwirft. Dies ist die übliche Definition eines Konfidenzintervalls.

6. Simulation

6.1. Erzeugung von Zufallszahlen

Um auf dem Computer simulieren zu können, benötigt man einen Pseudo-Zufallszahlengenerator. Eine beliebte Methode ist die Folgende. Man wählt einen Startwert z_0 in $\{0, 1, \ldots, m-1\}$. Dann erzeugt man die Zahlen $z_{n+1} = (az_n + c) \mod m$. Die Zufallszahlen sind dann $u_n = z_n/m$. Wenn man die Zahlen a, c, m geschickt wählt, erhält man so eine Folge, die gleichverteilten Zufallsvariablen ähnlich sehen.

Da auf diese Weise nur m verschiedene Zahlen erreicht werden können, muss man m so gross wie möglich wählen. Weiter sollte man a und c so wählen, dass alle m Zahlen vorkommen. Und dann sollte es nicht möglich sein, gute Vorhersagen über die nächste Zahl treffen zu können, wenn man a, c und m nicht kennt. Zum Beispiel dürfen Teilintervalle von $[0, 1]$ nicht periodisch getroffen werden.

Ein in der Literatur verwendeter Generator ist $m = 2^{32}$ (Bitlänge), $a = 1\,664\,525$ und $c = 1\,013\,904\,223$. Die Wahl von m hat den Vorteil, dass der Computer automatisch modulo m rechnet. Das heisst, man benötigt weniger Rechenzeit, um die Zahlen zu erzeugen. Auch die Division z_n/m ist einfach. Eine Möglichkeit, die Periode zu vergrössern ist ein multilinearer Generator

$$z_{n+1} = (a_0 z_n + a_1 z_{n-1} + \cdots + a_k z_{n-k}) \mod m \,.$$

Dann benötigt man $k + 1$ Startwerte, die man z.B. durch einen einfachen Generator erzeugen könnte. Der Generator kann verbessert werden durch Vergrösserung der Bitlänge, durch Kombination von verschiedenen Zufallszahlengeneratoren, und durch Manipulation durch ein Schieberegister.

6.2. Inversionsverfahren

Definieren wir die Umkehrfunktion $F^{-1}(x) = \inf\{y : F(y) \geq x\}$. Ist U eine in $(0, 1)$ gleichverteilte Zufallsvariable, dann hat $X = F^{-1}(U)$ die Verteilungsfunktion F,

$$\mathbb{P}[X \leq x] = \mathbb{P}[F^{-1}(U) \leq x] = \mathbb{P}[F(F^{-1}(U)) \leq F(x)] = \mathbb{P}[U \leq F(x)] = F(x) \,.$$

Beispiele

- *Gleichverteilung auf* (a, b) Hier ist $F(x) = (x-a)/(b-a)$ auf (a, b). Also haben wir $F^{-1}(y) = a + (b - a)y$. Dies ergibt $X = a + (b - a)U$.

- *Diskrete Gleichverteilung* Ist $\mathbb{P}[X = k] = n^{-1}$ für $k = 1, 2, \ldots, n$, erhalten wir $X = \lfloor Un \rfloor + 1$. Hier bezeichnet $\lfloor x \rfloor = \sup\{n \in \mathbb{N} : n \leq x\}$ den Ganzzahlteil von x.

- 0-1 *Experiment* Ist $\mathbb{P}[X = 1] = p = 1 - \mathbb{P}[X = 0]$, so setzen wir $X = 0$ falls $U < 1 - p$, und $X = 1$ sonst. Dies können wir erreichen, indem wir $X = \lfloor U + p \rfloor$ setzen.

- *Exponentialverteilte Variablen* Ist $F(x) = 1 - \mathrm{e}^{-\alpha x}$ für $x > 0$, erhalten wir $F^{-1}(y) = -\alpha^{-1} \log(1 - y)$. Also gibt die Methode $X = -\alpha^{-1} \log(1 - U)$. Da $1 - U$ auch gleichverteilt auf $(0, 1)$ ist, können wir $X = -\alpha^{-1} \log U$ setzen.

- *Paretoverteilte Variablen* Für $F(x) = 1 - (1 + x/\beta)^{-\alpha}$ erhalten wir $F^{-1}(y) = \beta[(1 - y)^{-1/\alpha} - 1]$. Also erhalten wir $X = \beta[U^{-1/\alpha} - 1]$.

- *Geometrische Verteilung* Wir haben $F(x) = 1 - q^{\lfloor x \rfloor + 1}$. Also können wir die geometrische Verteilung als $X = \lfloor \log U / \log q \rfloor$ erzeugen. Dies ist also eine abgerundete Exponentialverteilung.

6.3. Simulation mit Hilfe anderer Variablen

In vielen Fällen ist die Funktion $F^{-1}(y)$ kompliziert oder lässt sich nicht in geschlossener Form darstellen. In manchen dieser Fälle kann man X aus anderen Zufallsvariablen erhalten.

Beispiele

- *Binomialverteilte Zufallsvariablen* Sind $\{Y_i\}$ unabhängige 0-1 Experimente mit Parameter p, dann ist $\sum_{j=1}^{n} Y_j$ binomialverteilt mit Parametern n und p. Wir können also eine binomialverteilte Variable mit $X = \sum_{j=1}^{n} \lfloor U_j + p \rfloor$ erzeugen, wobei U_j unabhängige gleichverteilte Variablen sind.

- *Poissonverteilte Variablen* Wir haben in (2.2) gesehen, dass falls T_i unabhängig exponentialverteilte Variablen mit Parameter 1 sind, und wir $N_t = \sup\{n : \sum_{i=1}^{n} T_i \leq t\}$ setzen, dann ist N_t Poissonverteilt mit Paramter t. Somit können wir mit $X = \inf\{n : -\sum_{i=1}^{n} \log U_i > \lambda\} - 1 = \inf\{n : \prod_{i=1}^{n} U_i < \mathrm{e}^{-\lambda}\} - 1$ eine poissonverteilte Zufallsvariable mit Parameter λ erzeugen.

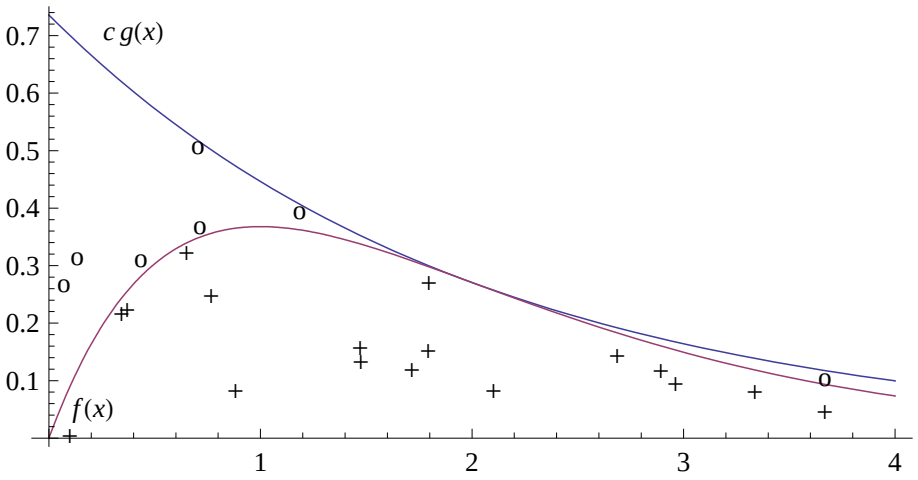

Abbildung 6.1: *Simulation mittels der Verwerfungsmethode*

- *Hypergeometrischverteilte Zufallsvariablen* Die hypergeometrische Verteilung gibt die Verteilung der Anzahl der roten Kugeln, wenn man aus einer Urne mit K roten und $N-K$ schwarzen Kugeln n Kugeln zieht. Seien daher U_i gleichverteilte Variablen. Wir setzen $Y_k = \lfloor U_k + (K - \sum_{j=1}^{k-1} Y_j)/(N-k+1) \rfloor$, da $(K - \sum_{j=1}^{k-1} Y_j)/(N-k+1)$ der Erfolgsparameter beim k-ten Ziehen ist. Dann hat $X = \sum_{j=1}^{n} Y_j$ eine hypergeometrische Verteilung.

6.4. Die Verwerfungsmethode

Sein nun $F(x)$ absolutstetig mit Dichte $f(x)$. Sei $g(x)$ die Dichte einer anderen Verteilungsfunktion $G(x)$, die einfach erzeugbar ist. Nehmen wir an, es gibt eine Konstante c, so dass $f(x) \leq cg(x)$ für alle x. Wir erzeugen dann unabhängige Zufallsvariablen $\{Y_k\}$ mit Verteilungsfunktion $G(x)$ und unabhängige gleichverteilte $\{U_k\}$. Wir gehen folgendermassen vor. Wir starten mit $n = 1$. Ist $U_n \leq f(Y_n)/(cg(Y_n))$, dann setzen wir $X = Y_n$, sonst erhöhen wir n um eins. Das heisst, wir simulieren die Variable, und akzeptieren sie dann mit Wahrscheinlichkeit $f(Y_n)/(cg(Y_n))$. Abbildung 6.1 zeigt eine Simulation der Gleichverteilung unter der Fläche begrenzt durch $cg(x)$. Die Punkte markiert mit + sind die akzeptierten Punkte unter der Fläche $f(x)$, die Simulationswerte sind die Werte auf der Abszisse. Die Punkte, die mit o markiert sind, werden verworfen.

Kennt man Y_k, wird die Variable mit Wahrscheinlichkeit $f(Y_1)/(cg(Y_1))$ akzeptiert. Die Variable wird also mit Wahrscheinlichkeit

$$\int \frac{f(x)}{cg(x)} g(x)\, \mathrm{d}x = \frac{1}{c} \int f(x)\, \mathrm{d}x = \frac{1}{c}$$

akzeptiert. Wir werden also eine geometrisch verteilte Anzahl Variablen mit Parameter $1 - 1/c$ verwerfen, bevor wir Y_k akzeptieren. Eine Variable wird akzeptiert und ist kleiner als x mit Wahrscheinlichkeit

$$\int_{-\infty}^{x} \frac{f(y)}{cg(y)} g(y)\, \mathrm{d}y = \frac{1}{c} \int_{-\infty}^{x} f(y)\, \mathrm{d}y = \frac{F(x)}{c} \; .$$

Bedingen wir also darauf, dass die Variable akzeptiert wird, haben wir die bedingte Verteilung $F(x)$. Daher hat X die Verteilungsfunktion $F(x)$.

Betrachten wir eine diskrete Verteilung und setzen $\mathbb{P}[X = n] = f(n)$ und $\mathbb{P}[Y = n] = g(n)$, dann funktioniert die Methode analog. Ist also c so gewählt, dass $cg(n) \geq f(n)$ für alle n, so akzeptieren wir die Variable Y_n mit Wahrscheinlichkeit $f(Y_n)/(cf(Y_n))$ und verwerfen sonst.

Beispiel Wir wollen eine Gammaverteilung mit Parametern $\gamma > 1$ und α erzeugen. Es genügt, den Fall $\alpha = 1$ zu betrachten, da X/α die gesuchte Verteilung besitzt, wenn X Gammaverteilt mit Parametern γ und 1 ist. Also haben wir $f(x) = x^{\gamma-1}\mathrm{e}^{-x}/\Gamma(\gamma)$. Sei $g(x) = \frac{1}{2}\mathrm{e}^{-x/2}$. Da $x^{\gamma-1}\mathrm{e}^{-x/2}$ das Maximum in $2(\gamma - 1)$ annimmt, erhalten wir mit Hilfe der Stirlingschen Formel

$$c = \frac{2^\gamma(\gamma-1)^{\gamma-1}\mathrm{e}^{-\gamma+1}}{\Gamma(\gamma)} \leq \frac{2^\gamma(\gamma-1)^{\gamma-1}\mathrm{e}^{-\gamma+1}}{\sqrt{2\pi}(\gamma-1)^{\gamma-1/2}\mathrm{e}^{-\gamma+1}} = \frac{2^\gamma}{\sqrt{2\pi(\gamma-1)}} \; .$$

Wir erzeugen daher $Y_k = -2\log\tilde{U}_k$ und U_k, wobei $\{\tilde{U}_k\}$ und $\{U_k\}$ unabhängig sind. Die oben beschriebene Methode liefert dann die gesuchten Variablen.

Ist $\gamma \in \mathbb{N}$, lassen sich die Γ verteilten Variablen einfacher als $-\sum_{k=1}^{\gamma} \log\tilde{U}_k$ erzeugen. Ist $\gamma \notin \mathbb{N}$, ist es effizienter, $g(x) = (2x)^{\lfloor\gamma\rfloor-1}\mathrm{e}^{-x/2}/\Gamma(\lfloor\gamma\rfloor)$ zu wählen, da dann c kleiner wird. Ist $\frac{1}{2} < \gamma < 1$, so kann man für g die χ^2-Verteilung wählen, die man aus der Normalverteilung erhält. Für $\gamma = \frac{1}{2}$ hat man eine χ^2-Verteilung.

6.5. Normalverteilte Variablen

Ist X standardnormalverteilt, so ist $\sigma X + \mu$ normalverteilt mit Mittelwert μ und Varianz σ^2. Also genügt es zu betrachten, wie man standardnormalverteilte Variablen erzeugt. Da die Verteilungsfunktion, beziehungsweise ihre Umkehrfunktion, sich nicht in geschlossener Form darstellen lässt, ist das Inversionsverfahren nicht geeignet, um die Variablen zu simulieren.

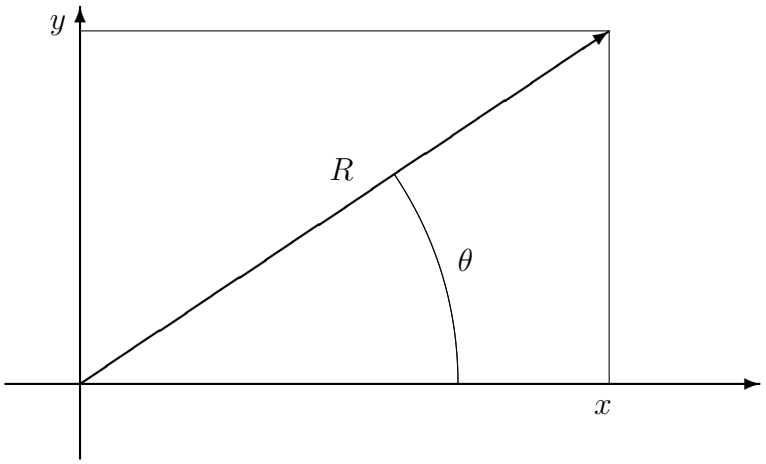

Abbildung 6.2: *Karthesische und Polarkoordinaten*

Benötigen wir abhängige normalverteilte Variablen, können wir dies durch eine Transformation von unabhängigen Variablen erreichen. Zum Beispiel, im zweidimensionalen Fall sind $(\sigma_1 X_1 + \mu_1, \sigma_2(\rho X_1 + \sqrt{1-\rho^2} X_2) + \mu_2)$ normalverteilt mit Korrelationskoeffizient ρ.

6.5.1. Die Box–Muller Methode

Seien X und Y zwei unabhängige standard-normalverteilte Zufallsvariablen, die wir als Punkt in $\mathbb{R}^2$ betrachten. Diesen Punkt können wir auch mit Polarkoordinaten (R, θ) darstellen, das heisst, $X = R\cos\theta$ und $Y = R\sin\theta$, wobei $\theta \in [0, 2\pi)$. Wir wollen nun die gemeinsame Verteilung von (R, θ) bestimmen. Sei A die Menge $\{(\rho\cos\varphi, \rho\sin\varphi) : \rho \le r, \varphi \in [0, \phi]\}$. Dann erhalten wir durch den Wechsel zu Polarkoordinaten

$$\begin{aligned}\mathbb{P}[R \le r, \theta \le \phi] &= \iint_A \frac{1}{2\pi} \mathrm{e}^{-(x^2+y^2)/2} \,\mathrm{d}y \,\mathrm{d}x = \int_0^r \int_0^\phi \frac{1}{2\pi} \mathrm{e}^{-\rho^2/2} \rho \,\mathrm{d}\varphi \,\mathrm{d}\rho \\ &= \frac{\phi}{2\pi}(1 - \mathrm{e}^{-r^2/2}) \,.\end{aligned}$$

Dies bedeutet, R und θ sind unabhängig, θ ist auf $(0, 2\pi)$ gleichverteilt, und R^2 ist exponentialverteilt mit Parameter $\frac{1}{2}$. Somit erhalten wir zwei unabhängige standardnormalverteilte Variablen

$$\begin{aligned}X &= \sqrt{-2\log U_1}\,\cos(2\pi U_2) \,, \\ Y &= \sqrt{-2\log U_1}\,\sin(2\pi U_2) \,.\end{aligned}$$

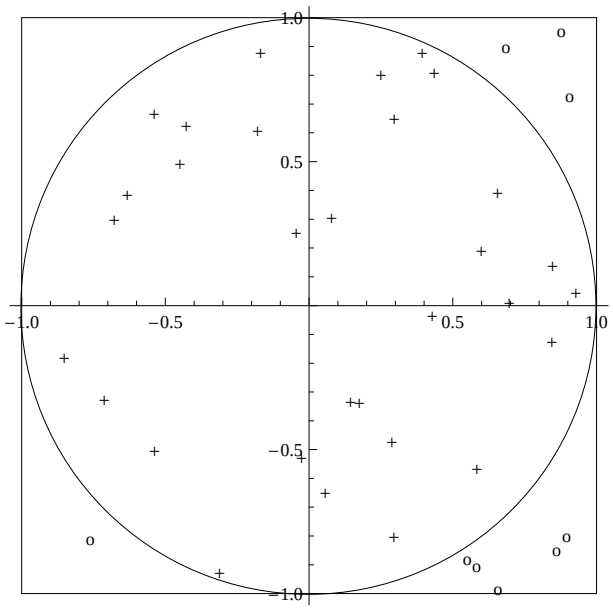

Abbildung 6.3: *Simulation der Gleichverteilung im Einheitskreis*

6.5.2. Die Polar Marsaglia Methode

Betrachten wir wieder Polarkoordinaten (R, θ) von zwei standardnormalverteilten Zufallsvariablen. Wir definieren nun die Transformation

$$(R, \theta) \mapsto (\mathrm{e}^{-R^2/4}, \theta) \; .$$

Radius und Winkel bleiben unabhängig. Für die Verteilung des Radius erhalten wir für $0 < r \leq 1$

$$\mathbb{P}[\mathrm{e}^{-R^2/4} \leq r] = \mathbb{P}[R^2 \geq -4 \log r] = r^2 \; .$$

Dies bedeutet, dass $(\mathrm{e}^{-R^2/4}, \theta)$ (bezüglich kartesischen Koordinaten) im Einheitskreis gleichverteilt ist. Da gleichverteilte Variablen einfach zu erzeugen sind, können wir die standardnormalverteilten Variablen so konstruieren. Wir setzen $V_i = 2U_i - 1$. Falls $W = V_1^2 + V_2^2 > 1$, verwerfen wir die Variablen, ansonsten setzen wir

$$X = V_1 \sqrt{\frac{-2 \log W}{W}} \; ,$$
$$Y = V_2 \sqrt{\frac{-2 \log W}{W}} \; .$$

Dies gilt, da $\sqrt{W} = \mathrm{e}^{-R^2/4}$ (also $R = \sqrt{-2 \log W}$) und $(V_1/\sqrt{W}, V_2/\sqrt{W}) = (\cos\theta, \sin\theta)$. Abbildung 6.3 zeigt eine Simulation von 40 Punkten, wobei 9 Punkte verworfen werden. Wir bemerken, dass die Wahrscheinlichkeit einen Punkt zu akzeptieren bei $\pi/4$ liegt, also von vier Simulationen etwa eine verworfen wird.

6.6. Monte-Carlo Simulation

6.6.1. Die Methode

Für viele Grössen, die man berechnen möchte, ist die exakte Berechnung sehr kompliziert. Zum Beispiel benötigen Banken und Versicherungen heute ein Kapital, das dem sogenannten Value-at-Risk zum Niveau 99.5% entspricht. Dies ist ein Eigenkapital, das mit Wahrscheinlichkeit 0.995 ausreicht. Die Portfolien sind normalerweise so komplex, dass es sehr schwer ist, diesen Wert exakt zu berechnen.

Die Idee ist, die entsprechenden Werte beziehungsweise Prozesse zu simulieren. Wenn wir die Grösse n mal unabhängig simuliert haben, so erhält man aus der entspechenden Variablen X_i

$$\frac{1}{n}\sum_{k=1}^{n} f(X_k) \approx \mathbb{E}[f(X)] \; .$$

Ist zum Beispiel X_k der Verlust, so ist

$$\frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_{X_k \leq x}$$

eine Näherung der Verteilungsfunktion.

Die Varianz des Schätzers

$$\operatorname{Var}\Big[\frac{1}{n}\sum_{k=1}^{n} f(X_k)\Big] = n^{-1}\operatorname{Var}[f(X_k)]$$

ist ein Mass für die Qualität der Schätzung. Da wir die Varianz $\operatorname{Var}[f(X_k)]$ nicht kennen, schätzen wir sie durch die empirische Varianz

$$\frac{1}{n-1}\sum_{k=1}^{n}\Big(f(X_k) - \overline{f(X)}\Big)^2 \; ,$$

wobei $\overline{f(X)}$ der geschätzte Mittelwert ist.

Da relativ grosse Genauigkeit benötigt, sind viele Simulationen nötig. Da die gesuchte Variable durch einen komplexen Prozess erhalten wird, braucht eine Simulaion meist viel Zeit. Daher ist es von Vorteil, wenn die Varianz kleiner ist.

Beispiel 6.1. Man könnte auf die Idee kommen, π mittels einer Simulation zu berechnen. Man simuliert zwei auf $(0,1)$ gleichverteilte Zufallsvariablen X und Y.

Dann testet man, ob $X^2+Y^2 \leq 1$. Das heisst, man testet, ob der simulierte Punkt in der Ebene im Einheitskreis liegt. Die Fläche des Kreisviertels ist $\pi/4$, die Fläche des Quadrates ist 1. Somit ist die Wahrscheinlichkeit, dass der Punkt in den Einheitskreis liegt, ist $\pi/4$. Ist $I_n = 1$, falls $X_n^2 + Y_n^2 \leq 1$ und $I_n = 0$ sonst, so hat man einen Schätzer für π

$$\frac{4}{N} \sum_{n=1}^{N} I_n \, .$$

Eine Simulation mit $N = 100\,000$ gibt den Wert 3.13688, mit $N = 1\,000\,000$ erhalten wir 3.14078. Mit $N = 10\,000\,000$ erhalten wir den ziemlich genauen Wert 3.14164. ■

Beispiel 6.2. In der Finanzmathematik sei $\{S_n\}$ der Preis eines Aktivs. Im Black–Scholes-Modell modelliert man

$$S_n = S_0 \exp\Big\{\sum_{k=1}^{n} X_k\Big\} \, ,$$

wobei $\{X_k\}$ unabhängige normalverteilte Variablen mit Mittelwert μ und Varianz σ^2 sind. Eine asiatische Option ist eine Option mit der Auszahlung

$$\Big(\frac{1}{N-n_0} \sum_{k=n_0+1}^{N} S_k - K\Big)^+ \, ,$$

wobei K der Ausübungspreis ist. Hier wird der Durchschnitt genommen, damit es unwahrscheinlich ist, dass ein einflussreicher Händler den Preis über eine längere Zeit ohne grosse Kosten manipulieren kann. Um den Preis zu bestimmen, müssen wir den Wert

$$\mathbb{E}\Big[\Big(\frac{1}{N-n_0} \sum_{k=n_0+1}^{N} S_k - K\Big)^+\Big]$$

bestimmen, wobei hier $\mathbb{E}$ das Mass ist, das für die Preisbestimmung verwendet wird. Das Problem ist nun, dass sich die Verteilung einer Summe von Lognormalverteilungen nicht effizient ausdrücken lässt. Um den Preis bestimmen zu können, werden daher oft Monte-Carlo Simulationen verwendet. ■

6.6.2. Varianzreduzierende Methoden

Die Zwei-Schätzer-Methode Nehmen wir an, dass wir den Erwartungswert einer Grösse $h(X_1, X_2, \ldots, X_n)$ berechnen wollen, wobei $\{X_k\}$ Zufallsvariablen sind.

Nehmen wir weiter an, dass wir eine Funktion $\tilde{h}$ kennen, so dass

$$\mathbb{E}[h(X_1, X_2, \ldots, X_n)] = \mathbb{E}[\tilde{h}(X_1, X_2, \ldots, X_n)] \; .$$

Die Varianzen sollen endlich sein. Wir definieren dann den Schätzer

$$h^*(X_1, X_2, \ldots, X_n) = \alpha h(X_1, X_2, \ldots, X_n) + (1-\alpha)\tilde{h}(X_1, X_2, \ldots, X_n) \; .$$

Dieser Schätzer hat die Varianz

$$\alpha^2 \operatorname{Var}[h] + (1-\alpha)^2 \operatorname{Var}[\tilde{h}] + 2\alpha(1-\alpha) \operatorname{Cov}[h, \tilde{h}] \; .$$

Die Varianz wird für

$$\alpha^* = \frac{\operatorname{Var}[\tilde{h}] - \operatorname{Cov}[h, \tilde{h}]}{\operatorname{Var}[h] + \operatorname{Var}[\tilde{h}] - 2\operatorname{Cov}[h, \tilde{h}]}$$

minimal. Ist $\alpha^* \neq 1$, also $\operatorname{Var}[h] \neq \operatorname{Cov}[h, \tilde{h}]$, so reduziert sich die Varianz des Schätzers. Die Methode hat weiter den Vorteil, dass die Zufallsvariablen $X_1, \ldots, X_n$ nur einmal erzeugt werden müssen, was Zeit im Programm spart. Wir bemerken noch, dass normalerweise α^* nicht berechnet werden kann. Man schätzt dann α^* und verwendet eine Approximation an die optimale Wahl.

Beispiel 6.2 (Fortsetzung). In der Finanzmathematik erzeugt man standard-normalverteilte Variablen Z_k. Dann ist auch $-Z_k$ standard-normalverteilt. In unserem Beispiel haben wir also $X_k = \mu + \sigma Z_k$, und setzt dann $\tilde{X}_k = \mu - \sigma Z_k$. Die gesuchte Funktion $h(Z_1, \ldots, Z_N)$ und $\tilde{h}(Z_1, \ldots, Z_N) = h(-Z_1, \ldots, -Z_N)$ haben dann den selben Erwartungswert und selbe Varianz. Damit erhalten wir $\alpha^* = \frac{1}{2}$. Da wir erwarten, dass $\operatorname{Cov}[h, \tilde{h}] < 0$, wird die Varianz des Schätzers halbiert. ■

Die Kontroll-Variablen-Methode Sei $\tilde{h}(X_1, \ldots, X_n)$ eine Variable, so dass der Erwartungswert $\mathbb{E}[\tilde{h}(X_1, \ldots, X_n)]$ einfach berechnet werden kann. Wir nehmen an, dass die Varianz $\operatorname{Var}[\tilde{h}(X_1, \ldots, X_n)]$ endlich sei. Der Schätzer

$$h^*(X_1, \ldots, X_n) = h(X_1, \ldots, X_n) - c(\tilde{h}(X_1, \ldots, X_n) - \mathbb{E}[\tilde{h}(X_1, \ldots, X_n)])$$

hat dann den selben Erwartungswert wie h und Varianz

$$\operatorname{Var}[h(X_1, \ldots, X_n)] + c^2 \operatorname{Var}[\tilde{h}(X_1, \ldots, X_n)] - 2c \operatorname{Cov}[h(X_1, \ldots, X_n), \tilde{h}(X_1, \ldots, X_n)] \, ;$$

Wählt man

$$c = \frac{\operatorname{Cov}[h(X_1, \ldots, X_n), \tilde{h}(X_1, \ldots, X_n)]}{\operatorname{Var}[\tilde{h}(X_1, \ldots, X_n)]}$$

wird die Varianz minimal. Die optimale Wahl ergibt die Varianz

$$\mathrm{Var}[h(X_1,\dots,X_n)] - \frac{(\mathrm{Cov}[h(X_1,\dots,X_n),\tilde{h}(X_1,\dots,X_n)])^2}{\mathrm{Var}[\tilde{h}(X_1,\dots,X_n)]}\,.$$

Um eine Varianz nahe bei Null zu erhalten, sucht man $\tilde{h}$, so dass

$$(\mathrm{Cov}[h(X_1,\dots,X_n),\tilde{h}(X_1,\dots,X_n)])^2 \approx \mathrm{Var}[h(X_1,\dots,X_n)]\,\mathrm{Var}[\tilde{h}(X_1,\dots,X_n)]\,.$$

Das heisst, die Korrelation sollte nahe bei 1 oder -1 liegen.

Das optimale c kann selten genau bestimmt werden. Sind Varianz von $\tilde{h}$ und h ungefähr gleich und die Korrelation nahe bei 1, so scheint $c = 1$ eine gute Wahl zu sein.

Beispiel 6.2 (Fortsetzung). Hätte man für die asiatische Option statt dem arithmetischen Mittel das geometrische Mittel verwendet, also die Auszahlung

$$\Big(\Big(\prod_{k=n_0+1}^{N} S_n\Big)^{1/(N-n_0)} - K\Big)^+\,,$$

so wäre es einfach den Preis der Option berechnen. Man kann für die Simulation nun verwenden, dass $(\prod_{k=n_0+1}^{N} S_n)^{1/(N-n_0)}$ und $(\frac{1}{N-n_0}\sum_{k=n_0+1}^{N} S_k - K)^+$ stark korreliert sind. Korrigiert man die Simulation entsprechend, lässt sich die Varianz stark vermindern, so dass die simulierten Preise genauer werden. ■

6.7. Importance Sampling

Nehmen wir an, dass wir den Erwartungswert $\mathbb{E}[h(X)]$ für eine Variable X mit Dichtefunktion $f(x)$ berechnen wollen. Sei $g(x)$ eine Dichtefunktion mit der Eigenschaft, dass $g(x) > 0$ für alle x mit $f(x) > 0$. Dann gilt

$$\mathbb{E}[h(X)] = \int h(x)f(x)\,\mathrm{d}x = \int h(x)\frac{f(x)}{g(x)}g(x)\,\mathrm{d}x\,.$$

Ist Y eine Variable mit der Dichtefunktion $g(x)$, so haben wir also

$$\mathbb{E}[h(X)] = \mathbb{E}\Big[h(Y)\frac{f(Y)}{g(Y)}\Big]\,.$$

Für die Monte-Carlo-Simulation sucht man nun die Dichtefunktion $g(x)$, für die die Varianz beziehungsweise das zweite Moment

$$\mathbb{E}\Big[\Big(h(Y)\frac{f(Y)}{g(Y)}\Big)^2\Big] = \int h^2(x)\frac{f^2(x)}{g(x)}\,\mathrm{d}x$$

minimal wird. Würden wir $g(x)$ kennen und durch eine Funktion $g(x) + \varepsilon v(x)$ ersetzen, dann müsste $\int g(x) + \varepsilon v(x)\,\mathrm{d}x = 1$ gelten, und die gesuchte Varianz müsste ein Minimum in $\varepsilon = 0$ haben. Wir müssen daher

$$\int h^2(x)\frac{f^2(x)}{g(x)+\varepsilon v(x)}\,\mathrm{d}x + \delta\Big[\int (g(x)+\varepsilon v(x))\,\mathrm{d}x - 1\Big]$$

minimieren. Leiten wir nach ε ab, erhalten wir

$$\int \Big[\delta - h^2(x)\frac{f^2(x)}{(g(x)+\varepsilon v(x))^2}\Big]v(x)\,\mathrm{d}x = 0\ .$$

In $\varepsilon = 0$, erhalten wir

$$\int \Big[\delta - h^2(x)\frac{f^2(x)}{g^2(x)}\,\mathrm{d}x\Big]v(x)\,\mathrm{d}x = 0\ .$$

Da dies für alle $v(x)$ gelten muss, erhalten wir $g(x) = h(x)f(x)/\sqrt{\delta}$. Da $\int g(x)\,\mathrm{d}x = 1$, wird dies für $g(x) = h(x)f(x)/\mathbb{E}[h(X)]$ erreicht. Das heisst, wir simulieren die Grösse $h(x)f(x)/g(x) = \mathbb{E}[h(X)]$. Dies ist eine deterministische Grösse. Damit wird die Varianz 0, und eine Simulation würde ausreichen. Natürlich ist dies keine mögliche Wahl, da man die gesuchte Grösse $\mathbb{E}[h(X)]$ kennen müsste. Man wird aber versuchen, g so zu wählen, dass g gross wird für Werte, an denen $h(x)f(x)$ gross wird. Dadurch wird das Gewicht der Verteilung an Stellen verlagert, die für den Erwartungswert gewichtig sind.

Die Methode funktioniert analog auch für diskrete Wahrscheinlichkeitsverteilungen. Meistens wird die Methode für stochastische Prozesse angewendet, indem die Gewichte geschickt verschoben werden, so dass die interessanten Szenarien wahrscheinlicher werden.

Beispiel 6.3. Betrachten wir die Irrfahrt $S_n = \sum_{k=1}^n X_k$, wobei X_k normalverteilt ist mit Mittelwert μ und Varianz 1. Die Variablen $\{X_k\}$ seien unabhängig. Wir haben $\mathbb{E}[X_k] = \mu$. Nach dem Gesetz der grossen Zahl konvergiert S_n/n nach μ. Ist $\mu > 0$, so konvergiert also S_n nach Unendlich, und $m = \inf_{n\in\mathbb{N}} S_n$ ist endlich. Starten wir mit einem Anfangskapital $a > 0$, so reicht das Kapital aus, falls $a + m \geq 0$. Wir können dann die Ruinwahrscheinlichkeit $\mathbb{P}[a+m < 0]$ durch Simulation bestimmen.

Es gibt zwei Probleme. Wir müssen die Simulation irgendwann abbrechen, zum Beispiel, wenn der Prozess einen grossen Wert k erreicht, von dem aus Ruin sehr unwahrscheinlich wird. Brechen wir ab, so berechnen wir nur eine Näherung des gesuchten Wertes. Ist a gross, so wird die Ruinwahrscheinlichkeit klein. Das heisst,

wir müssen sehr viele Pfade simulieren, um einen interessanten Pfad zu erhalten. Damit wird die Varianz des numerischen Schätzers im Vergleich zur gesuchten Grösse sehr gross. Die Simulation ist also sehr ineffizient.

Sei nun $g(x)$ die Dichte der Normalverteilung mit Mittelwert $-\mu$ und Varianz 1. Dann haben wir

$$\frac{f(x)}{g(x)} = \frac{\exp\{-\frac{1}{2}(x-\mu)^2\}}{\exp\{-\frac{1}{2}(x+\mu)^2\}} = \exp\{2\mu x\} \; .$$

Wir können dies für jede Variable X_k machen. Dann können wir ein Ereignis A, das durch $X_1, X_2, \ldots, X_n$ bestimmt ist, durch

$$\mathbb{P}[A] = \widetilde{\mathbb{E}}\Big[\mathbb{1}_A \prod_{k=1}^{n} \exp\{2\mu X_k\}\Big] = \widetilde{\mathbb{E}}[\mathbb{1}_A \exp\{2\mu S_n\}]$$

ausdrücken, wobei X_k unter dem Mass $\tilde{\mathbb{P}}$ normalverteilt mit Mittelwert $-\mu$ ist. Setzen wir $\tau = \inf\{n : S_n < -a\}$, so erhalten wir (wir beweisen dies hier nicht)

$$\mathbb{P}[a+m<0] = \mathbb{P}[\tau < \infty] = \widetilde{\mathbb{E}}[\mathbb{1}_{\tau<\infty} \exp\{2\mu S_\tau\}] \; .$$

Da der Wert $-a$ wegen dem negativen Trend sicher erreicht wird, haben wir also die Ruinwahrscheinlichkeit

$$\mathbb{P}[a+m<0] = \widetilde{\mathbb{E}}[\exp\{2\mu S_\tau\}] \; .$$

Wir haben nun, dass Ruin für jeden Pfad eintritt, das heisst, wir benötigen kein Abbruchkriterium und jeder Pfad ist zum numerischen Berechnen für unsere Grösse wichtig. Man kann zeigen, dass diese Methode asymptotisch (für $a \to \infty$) optimal ist. Insbesondere erhalten wir aus $S_\tau < -a$ die Abschätzung $\mathbb{P}[a+m<0] < \exp\{-2\mu a\}$. ■

Augen	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
geord.	1	1	2	3	4	5	6	6	6	6	5	4	3	2	1	1
ungeord.	1	3	6	10	15	21	25	27	27	25	21	15	10	6	3	1

Tabelle A.1: *Summe der Augen von drei Würfeln*

A. Geschichte der Stochastik

Die Entwicklung der Stochastik ist mit dem Glücksspiel verknüpft. Schon in vorhistorischer Zeit hat man offensichtlich gespielt. Man verwendete dazu einen kleinen Knochen des Schafes (Astragalus). Dieser Knochen diente als Würfel, und hat vier Seiten, auf die er fallen kann. Einen ersten wirklichen Nachweis, dass der Knochen zum Würfelspiel verwendet wurde, hat man aus einer ägyptischen Grabmalerei von 3500 vor Christus. Schon kurze Zeit später tauchten die ersten Würfel auf. Diese Würfel waren aber nicht fair. Das schien auch nicht nötig, da man glaubte, die Götter würden bestimmen, wie die Würfel fallen.

In der Antike wurde der Zufall immer wieder beschrieben, zum Beispiel durch Demokrit, Aristoteles und Epikur. Im Mittelalter schrieb Thomas von Aquino, dass der Zufall nur im Denken der Menschen existiere, da Gott zielstrebig handle. Auf die Idee, das nicht Fassbare mit Wahrscheinlichkeiten zu beschreiben, kamen diese Philosophen noch nicht.

Ein erstes Mal tauchen etwas wie Wahrscheinlichkeiten im 13. Jahrhundert auf. Der (unbekannte) Autor hat gezählt, auf wie viele Arten man die Summen 3 bis 18 mit drei Würfeln werfen kann. Es wird sowohl gezählt, auf wie viele Arten man die Summe geordnet schreiben kann, als auch, auf wie viele Arten man die Summen mit drei Würfen ungeordnet erhalten kann. Zum Beispiel, 5 kann man auf zwei Arten schreiben $(1+1+3,\ 1+2+2)$, und auf 6 Arten erhalten $((1,1,3),\ (1,2,2),\ (1,3,1),\ (2,1,2),\ (2,2,1),\ (3,1,1))$. Es gibt aber keinen Hinweis darauf, dass man die Häufigkeiten der Würfe damit verbunden hat.

Ein erstes Wahrscheinlichkeitsproblem taucht erstmals in einer Handschrift etwa um 1380 auf, die heute in der Nationalbibliothek von Florenz aufbewahrt wird. Dieses Problem wurde dann auch im Buch *Summa de Arithmetica Geometria Proportioni et Proportionalità* von Luca Pacioli im Jahr 1494 als Übung formuliert:

> Zwei Männer spielen “balla”. Beide haben 10 Goldmünzen. Der erste, der sechs Spiele gewonnen hat, bekommt alle 20 Münzen. Das Spiel muss abgebrochen werden, nachdem der erste Spieler fünf Spiele und der zweite drei

Spiele gewonnen hat. Beide Spieler sind gleich geschickt. Wie müssen die Spieler die zwanzig Münzen aufteilen?

Pacioli betrachtete dies nicht als Übung der Wahrscheinlichkeitsrechnung, die damals noch nicht existierte, sondern als Übung zur Proportionalitätstheorie. Der unbekannte Autor von 1380 hatte die richtige Lösung zufällig mit einem falschen Argument gefunden. Pacioli gab als Lösung 5:3 an. 1556 kam Niccolò Fontana Tartaglia auf die Antwort 2:1. 1558 schlug Gianbattista Francesco Peverone die Antwort 6:1 vor. Die richtige Antwort, 7:1, wurde schliesslich von Blaise Pascal 1654 gefunden. Man benötigte also etwa 270 Jahre, um ein Problem zu lösen, dass ein Student heute in weniger als fünf Minuten löst.

Um etwa 1550 schrieb Gerolamo Cardano ein Buch *Liber de Ludo Aleae* über das Würfelspiel. Er benutzte erstmals den Begriff, den wir heute Wahrscheinlichkeit nennen. Obwohl sein Buch Fehler enthielt, erkannte er zwei Grundregeln: die Wahrscheinlichkeit von zwei disjunkten Ereignissen ist die Summe der Wahrscheinlichkeiten und die Wahrscheinlichkeit, dass zwei unabhängige Ereignisse gleichzeitig eintreffen, ist gleich dem Produkt der beiden Wahrscheinlichkeiten. Cardanos Buch wurde aber nicht veröffentlicht, so dass es den Zeitgenossen von Cardano und seinen Nachkommen nicht bekannt war.

Die Wahrscheinlichkeitsrechnung, wie wir sie heute kennen, wurde daher im Sommer 1654 in einer Serie von Briefen von Pierre de Fermat und Blaise Pascal entwickelt. Dort wurde vor allem das Würfelspiel diskutiert. Insbesondere wird dort das obige Gewinnaufteilungsproblem und das Würfelproblem diskutiert. Beim Würfelproblem wirft man k Würfel und gewinnt, falls alle Würfel eine Sechs zeigen. Wie oft muss man würfeln können, damit man mit höherer Wahrscheinlichkeit gewinnt als verliert? Blaise Pascal wandte später die Theorie von Fermat auf eine Reihe von Problemen an.

Christiaan Huyghens hörte in Paris von Fermats und Pascals Theorie und schrieb ein Buch über Wahrscheinlichkeit, *De Ratiociniis in Ludo Aleae*, das in 1656 erschien. Er definierte den *Erwartungswert*, und benutzte diesen gekonnt, um Probleme zu lösen.

Die Entwicklung der *Statistik* geht schon ins 17. Jahrhundert zurück. Der Begriff bedeutet *Wissenschaft der Staaten*, und wurde ursprünglich verwendet, um Bevölkerungszahlen zu erheben. Der Begriff geht auf Aristoteles zurück, der in seinem Buch *Politeia* 158 Staaten beschrieb. Der erste, der Statistik ähnlich benutzte, wie wir es heute tun, war John Graunt, der 1661 aus Zahlen über Begräbnisse und

Geburten, und einer Schätzung der Anzahl Bewohner innerhalb der Mauern, die Bevölkerungszahl von London schätzte. Später haben Abraham de Moivre, Daniel Bernoulli und Edmond Halley die Ideen aufgenommen, und die Statistik auf ein stochastisches Fundament gestellt.

Ein wichtiges Anwendungsgebiet der Statistik war die Versicherung. Ursprünglich war eine Versicherung eine Wette, bei der ein Händler wettete, dass das Schiff mit seinen Waren nicht zurückkam. Auf diese Weise erhielt er eine Kompensation, falls das Schiff unterging. Erste dieser Art von Wetten wurden in Italien getätigt. 1608 gründete Edward Lloyd ein Kaffeehaus, in dem Händler und Seefahrer ihre Geschäfte beim Kaffee tätigten. Ab 1700 wurden dann die *Lloyd's News* veröffentlicht, die Informationen zu Schiffsbewegungen und anderen Neuigkeiten publizierte, die für die Versicherungshändler interessant waren. Daraus entstand dann Lloyds, eine der grössten Versicherungsgesellschaften der Welt.

Ähnlich begannen Lebensversicherungen. Zuerst wettete man auf Leben; und um gut kontrollieren zu können, ob die Person noch lebt, nicht auf das eigene, sondern auf das Leben eines Adligen. Um die Wetten fair zu gestalten, brauchte man zuverlässige Sterbetabellen. Eine der bekanntesten ersten Tabellen stellte der englische Astronom Edmond Halley her. 1671 verwendete Johan de Witt die Methoden von Christiaan Huygens und berechnete den Preis einer Witwen-Leibrente (*Waerdye van Lijf-renten naer Proportie van Los-renten*). Er führte das Äquivalenzprinzip ein, wonach der Wert der Prämien eines Versicherungsvertrages gleich dem Wert der Versicherungsleistungen sein sollte. Das erste Lehrbuch zur Versicherungsmathematik *Annuities upon Lives* wurde 1725 von Abraham de Moivre verfasst. Dieses Prinzip wurde erstmals von der *Society for Equitable Assurances on Lives and Survivorships* angewandt. Bis die Gesellschaft im Jahr 2000 das Neukundengeschäft einstellen musste, war es die älteste Lebensversicherungsgesellschaft; was die Wichtigkeit der stochastischen Grundlagen im Versicherungsgeschäft unterstreicht. Filip Lundberg modellierte 1900 in seiner Dissertation erstmals den Überschuss eines Versicherungsportfolios als stochastischen Prozess. Sein Modell ist heute noch das Basismodell der Schadenversicherungsmathematik. Mit einer geeigneten Umformulierung ist es auch das Basismodell der Warteschlangentheorie, die man benötigt um industrielle Prozesse sowie auch Prozesse im Computer zu optimieren.

Ein Meilenstein in der Geschichte der Stochastik war Jakob Bernoullis Buch *Ars Conjectandi*, die 'Vermutungs- oder Mutmassungskunst'. Daher kommt auch der Name *Stochastik*: vom griechischen Wort *stochazesthai*, und ist bekannt aus einem Buch von Platon. Bernoullis Buch enthielt unter anderem das *Schwache Gesetz*

der grossen Zahl, das besagt, dass der Durchschnitt von identisch verteilten und unabhängigen Zufallsvariablen sich immer mehr dem Mittelwert annähert, falls die Anzahl der Variablen steigt. Dieses Gesetz liegt im Prinzip der Statistik und der Versicherungsmathematik zugrunde. Das Buch wurde erst nach Bernoullis Tod 1713 durch seinen Bruder Johann fertiggestellt und herausgegeben.

Bernoullis Ideen wurden später von Abraham de Moivre wieder aufgenommen. Er veröffentlichte 1718 sein Buch *Doctrine of Chances*. In der zweiten Ausgabe von 1738 war der zweite fundamentale Satz veröffentlicht, der *zentrale Grenzwertsatz*, den de Moivre erstmals 1733 auf Latein publiziert hatte. Dieser besagt, dass sich die normierte Verteilung (Mittelwert 0 und Varianz 1) des Durchschnittes der Normalverteilung annähert. De Moivre fand zwar nicht die Dichte der Normalverteilung, aber er fand eine Reihe von Reihenentwicklungen der Normalverteilung. Diese Dichte wurde 1808 erstmals vom Amerikaner Robert Adrain gefunden. Ein Jahr später entdeckte Karl Friedrich Gauss das selbe Resultat. Da damals Adrains Resultat in Europa noch nicht bekannt war, nennt man die Normalverteilung heute auch *Gauss-Verteilung*. Der Zusammenhang zwischen dem zentralen Grenzwertsatz und der Normalverteilung wurde schliesslich in Pierre-Simon de Laplaces Buch *Théorie Analytique des Probabilités* gezeigt.

Das Problem der Stochastik war damals, dass ein mathematisches Fundament fehlte. Als 1900 David Hilbert am Mathematiker-Welt-Kongress eine Rede über die Mathematik des 20. Jahrhunderts gab, nannte er 23 ungelöste Probleme der Mathematik. Das sechste Problem war, Stochastik auf eine axiomatische Grundlage zu stellen. Diese Problem wurde erst 1933 gelöst, als Andrei Nikolaevich Kolmogorov seinen Artikel *Grundbegriffe der Wahrscheinlichkeitsrechnung* veröffentlichte. Er benutzte darin die mathematische Masstheorie, um Wahrscheinlichkeiten zu beschreiben.

Die moderne Wahrscheinlichkeitstheorie beschäftigt sich vor allem mit stochastischen Prozessen. Die Brownsche Bewegung tritt heute in vielen Problemen auf. Thorvald Nicolai Thiele benutzte schon 1880 die Brownsche Bewegung intuitiv, ohne sie sich als Prozess vorzustellen. Louis Bachelier benutzte 1900 in seiner Dissertation *Théorie de la spéculation* die Brownsche Bewegung, um einen Börsenkurs zu modellieren. Diese Idee wurde später von Paul Anthony Samuelson aufgegriffen. Die Arbeiten von Fischer Black, Myron S. Scholes, Robert Merton, Michael Harrison, David Kreps und Stan Pliska über das Samuelsonsche Modell ab 1972 waren der Anfang der sehr erfolgreichen modernen Finanzmathematik und wurden 1997 mit dem Nobelpreis ausgezeichnet. Der Siegeszug der Diffusionsprozesse in der Physik

begann mit einer Arbeit 1905 von Albert Einstein. Sie bilden heute die Basis der Quantentheorie und der stochastischen Physik. Obwohl Albert Einstein den stochastischen Grundstein gelegt hatte, akzeptierte er die Quantentheorie nicht ("Gott würfelt nicht").

In der Stochastik beschreibt man heute zwei Arten von Problemen: Die erste Art, sind "Experimente", die wir als zufällig auffassen, wie zum Beispiel das Glückspiel. Die zweite Art von Problemen sind komplexe Systeme. Oft kann man ein Experiment (z.B in der Physik) analytisch beschreiben. Durch die Komplexität des Systems ergibt aber eine sehr kleine, oft nicht messbare, Abweichung des Anfangszustandes einen total verschiedenen Zustand des Systems zu einem späteren Zeitpunkt. Man modelliert daher den Zustand des Systems als stochastisches Modell, um Aussagen über "wahrscheinliche" Zustände des Systems in der Zukunft zu erhalten. Zu dieser zweiten Art der Verwendung der Stochastik kann man auch die Quantenmechanik zählen.

B. Kombinatorik

In Laplace-Modellen muss man die Anzahl der Elementarereignisse zählen, die eine bestimmte Eigenschaft haben. Dies ist der Gegenstand der Kombinatorik. Wir wollen hier ein paar Beispiele betrachten.

Nehmen wir an, wir haben die Elemente $\{1, 2, 3, \ldots, n\}$. Dies können z.B. n Wettkämpfer sein. Wir wollen nun die Anzahl mögliche Ranglisten bestimmen, oder die Anzahl der Möglichkeiten, das Podest zu besetzen. Oder die besten 10 Wettkämpfer qualifizieren sich für den Final, und wir wollen die Anzahl der möglichen Startlisten für den Final bestimmen.

Nehmen wir an, wir ziehen k Mal eines der Elemente, und legen das gezogene Element wieder zurück. Wir können also ein Element mehrmals ziehen. Dann haben wir für das erste Element n Möglichkeiten, für das zweite n Möglichkeiten, usw. Also erhalten wir insgesamt n^k Möglichkeiten für das Resultat der Ziehung.

Ziehen wir nun k Mal eines der Elemente, legen aber das gezogene Element nicht mehr zurück, so haben wir für das erste Element n Möglichkeiten. Für das zweite Element können wir das erste Element nicht mehr ziehen. Also bleiben noch $n-1$ Möglichkeiten. Für das dritte Element bleiben dann noch $n-2$ Möglichkeiten, usw. Es ist klar, dass $k \leq n$ gelten muss, da nach n Ziehungen kein Element mehr vorhanden ist. Wenn wir das k-te Element ziehen, bleiben noch $n-k+1$ Möglichkeiten übrig. Somit haben wir

$$n(n-1)\cdots(n-k+1) = \frac{n(n-1)\cdots(n-k+1)(n-k)\cdots 1}{(n-k)\cdots 1} = \frac{n!}{(n-k)!}$$

Möglichkeiten.

Oft sind wir nicht an der Reihenfolge interessiert, mit der wir die Elemente ziehen, sondern nur welche Elemente zur Gruppe der k Elemente zählen. Wenn wir also die Elemente nicht zurücklegen und nicht auf die Reihenfolge der Ziehung achten, so besteht jede der Möglichkeiten aus k verschiedenen Elementen. Wenn wir nun die Elemente nachträglich noch irgendwie anordnen, so haben wir nach obiger Formel $k!/(k-k)! = k!$ Möglichkeiten, in welcher Reihenfolge die Elemente gezogen wurden. Das heisst, wenn wir zuerst die Elemente geordnet ziehen, und danach die Elemente der Grösse nach ordnen, kommt jede Anordnung $k!$ Mal vor. Somit haben wir

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

Möglichkeiten.

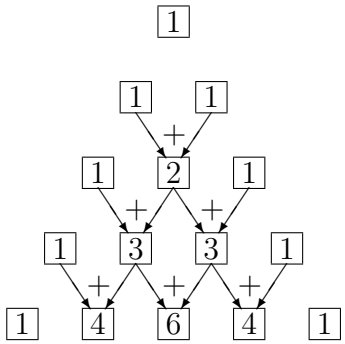

Abbildung B.1: *Pascalsches Dreieck*

Die Binomialkoeffizienten $\binom{n}{k}$ haben die folgenden Eigenschaften. Wir haben $\binom{n}{0} = \binom{n}{n} = 1$, was sofort durch Einsetzen in die Definition folgt. Weiter gilt für $k < n$

$$\begin{aligned}\binom{n}{k} + \binom{n}{k+1} &= \frac{n!}{k!(n-k)!} + \frac{n!}{(k+1)!(n-k-1)!} \\ &= \frac{n!}{(k+1)!(n-k)!}[(k+1)+(n-k)] = \frac{(n+1)!}{(k+1)!(n-k)!} \\ &= \binom{n+1}{k+1} . \end{aligned} \tag{B.1}$$

Diese Berechnungsmethode heisst *Pascalsches Dreieck*, siehe Abbildung B.1.

Es gilt weiter für $n \in \mathbb{N}$ die Formel

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} . \tag{B.2}$$

Dies können wir durch ein kombinatorisches Argument zeigen. Wir haben n Mal den Faktor $(x+y)$. Beim Ausmultiplizieren wählen wir aus jedem der Faktoren ein x oder y und multiplizieren diese. Das machen wir für jede mögliche Wahl. Für den Faktor $x^k y^{n-k}$ müssen wir also k Faktoren wählen, aus denen wir ein x nehmen. Dazu haben wir $\binom{n}{k}$ Möglichkeiten.

Man am kann die Formel auch mittels vollständiger Induktion zeigen. Ist $n = 0$, so ist

$$\sum_{k=0}^{0} \binom{0}{k} x^k y^{0-k} = \binom{0}{0} x^0 y^0 = 1 = (x+y)^0 .$$

Gilt die Formel für n, so erhalten wir

$$\begin{aligned}(x+y)^{n+1} &= (x+y)(x+y)^n = x\sum_{k=0}^{n}\binom{n}{k}x^k y^{n-k} + y\sum_{k=0}^{n}\binom{n}{k}x^k y^{n-k}\\ &= x^{n+1} + \sum_{k=0}^{n-1}\binom{n}{k}x^{k+1}y^{n-k} + \sum_{k=1}^{n}\binom{n}{k}x^k y^{n-k+1} + y^{n+1}\\ &= y^{n+1} + \sum_{k=1}^{n}\left(\binom{n}{k-1} + \binom{n}{k}\right)x^k y^{n+1-k} + x^{n+1}\\ &= \sum_{k=0}^{n+1}\binom{n+1}{k}x^k y^{n+1-k}\,,\end{aligned}$$

wobei wir (B.1) verwendet haben. Dies zeigt (B.2). Insbesondere erhalten wir

$$\sum_{k=0}^{n}\binom{n}{k} = (1+1)^n = 2^n\,,$$

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

und für $n \geq 1$

$$\sum_{k=0}^{n}(-1)^k\binom{n}{k} = (1-1)^n = 0\,.$$

Wir könnten nun auch die Elemente wieder zurücklegen, aber die Reihenfolge der Ziehung nicht beachten. Wir ordnen also die k gezogenen Elemente, so dass $\omega_1 \leq \omega_2 \leq \cdots \leq \omega_k$. Wir können dann zuerst die Gleichheitszeichen bestimmen. Wenn wir wissen, wieviele Gleichheitszeichen wir haben, ziehen wir die nötigen Zahlen, ordnen sie und setzen sie ein. Da es maximal $k-1$ Gleichheitszeichen gibt, könnten wir daher zusätzlich zu den n Zahlen noch $k-1$ nummerierte Gleichheitszeichen in die Urne legen, und dann ohne Zurücklegen ziehen. Die Gleichheitszeichen werden zuerst an der richtigen Stelle eingesetzt, dann die restlichen Zahlen geordnet und eingesetzt. Das ist als ob man so ordnet, dass die Gleichheitszeichen kleiner sind als die Zahlen und ensprechend geordnet werden. Somit ziehen wir geordnet k Elemente aus $n+k-1$ Elementen, das heisst $\binom{n+k-1}{k}$ Möglichkeiten.

Alternativ können wir diese Zahl auch so erhalten. m Gleichheitszeichen kann man auf $\binom{k-1}{m}$ Arten wählen. Danach können müssen wir noch $k-m$ Elemente ziehen, ordnen und einsetzen. Da haben wir $\binom{n}{k-m}$ Möglichkeiten. Somit erhalten wir

$$\sum_{m=0}^{k-1}\binom{k-1}{m}\binom{n}{k-m}$$

Möglichkeiten, um k Elemente mit Zurücklegen ohne Beachtung der Reihenfolge zu wählen. Betrachten wir nun das Produkt $(1+x)^{k-1}(1+x)^n$, insbesondere die Terme $x^m x^{k-m} = x^k$. Die Koeffizienten sind genau die Summanden der obigen Summe, siehe (B.2). Zählen wir also diese Summanden zusammen, müssen wir den Koeffizienten von x^k, also $\binom{n+k-1}{k}$ erhalten. Also gibt es $\binom{n+k-1}{k}$ Möglichkeiten, um k Elemente mit Zurücklegen ohne Beachtung der Reihenfolge zu wählen.

Folgendes Beispiel findet man in de Moivre [4, Problem III]. Auf wie viele Arten lassen sich p Punkte mit n (unterscheidbaren) Würfeln erreichen? Es ist klar, dass $n \leq p \leq 6n$. Sehen wir zunächst davon ab, dass die maximale Augenzahl 6 ist. Legen wir die Punkte nebeneinander, müssen wir die Punkte so in n Teile trennen, dass jedes Teil mindestens einen Punkt enthält. Wir müssen also $n-1$ Trennpunkte aus $p-1$ möglichen wählen. Dies sind dann $\binom{p-1}{n-1}$ Möglichkeiten. Eventuell haben wir aber nun Würfel mit mehr als 6 Punkten.

Ist $p-7 < n-1$, so kann es keine Trennung mit mit mehr als sechs Punkten geben, und wir sind fertig. Ist $p-7 \geq n-1$, und gibt es einen Würfel mit mehr als sechs Punkten, so können wir sechs Punkte wegnehmen. Dann bleiben $p-6$ Punkte zu trennen, und wir haben $\binom{p-7}{n-1}$ Möglichkeiten, diese Punkte auf die n Würfel zu verteilen. Wir haben $\binom{n}{1}$ Möglichkeiten, den Würfel zu wählen, bei dem wir die sechs Punkte entfernt haben. Nun haben wir aber die Möglichkeiten, bei denen mindestens k Würfel mehr als sechs Punkte haben, k Mal entfernt. Dies ist aber nur möglich, wenn $p-13 \geq n-1$. Wir haben hier $\binom{p-13}{n-1}\binom{n}{2}$ Möglichkeiten dazu. Jetzt haben wir also wieder zuviele Möglichkeiten, bei denen mindestens drei Würfel mehr als sechs Punkte haben. Gehen wir so weiter, siehe auch Hilfssatz 1.3 v), so erhalten wir die Formel

$$\sum_{k=0}^{n-1}(-1)^k\binom{p-6k-1}{n-1}\binom{n}{k}\,,$$

wobei wir $\binom{i}{j} = 0$ setzen, wenn $i < j$.

Würfeln wir also mit drei Würfeln und ist $p = 11$, so erhalten wir

$$\binom{10}{2}\binom{3}{0} - \binom{4}{2}\binom{3}{1} = 45 - 18 = 27$$

Möglichkeiten. Ist $p = 14$, so ergibt sich

$$\binom{13}{2}\binom{3}{0} - \binom{7}{2}\binom{3}{1} = 78 - 63 = 15\,.$$

Sind die Würfel nicht unterscheidbar, also wenn wir Punkte der Grösse nach ordnen, so ist das Problem viel komplizierter. Wir müssen dann unterscheiden, wie oft eine Punktzahl vorkommt.

Literatur

[1] **Feller, W.** (1968). *An Introduction to Probability Theorie and its Applications*. 3. Auflage, Band I. Wiley, New York.

[2] **Georgii, H.O.** (2009). *Stochastik. Einführung in die Wahrscheinlichkeitstheorie und Statistik*. 4. Auflage. De Gruyter, Berlin.

[3] **Henze, N.** (2017). *Stochastik für Einsteiger*. 11. Auflage, Springer Spektrum, Wiesbaden.

[4] **de Moivre, A.** (1967). *The doctrine of chances: A method of calculating the probabilities of events in play*. Chelsea Publishing Co., New York.

[5] **Rolski, T., Schmidli, H., Schmidt, V.** and **Teugels, J.L.** (1999). *Stochastic Processes for Insurance and Finance*. Wiley, Chichester.

Index