Stochastik für das Lehramt

Mathias Beiglböck Nathanaël Berestycki

11. November 2026

Inhaltsverzeichnis

1 Einführung **3**
1.1 Begriffe . . . 3
1.2 Gleichwahrscheinliche Ausgänge . . . 5
1.3 Wahrscheinlichkeit und Wirklichkeit . . . 7
1.4 Zusammenfassung des Kapitels . . . 11

2 Axiomatischer Zugang **12**
2.1 Axiome der Wahrscheinlichkeitstheorie . . . 12
2.2 Elementare Eigenschaften . . . 13
2.3 Bedingte Wahrscheinlichkeit . . . 14
2.4 Der Satz von Bayes . . . 16
2.5 Unabhängigkeit . . . 17
2.6 Unabhängigkeit von mehreren Ereignissen . . . 18
2.7 Zusammenfassung des Kapitels . . . 19

3 Zufallsvariablen **20**
3.1 Verteilungen auf $\mathbb{R}$ . . . 20
3.2 Zufallsvariablen und ihre Verteilungen . . . 23
3.3 Diskrete Zufallsvariablen; wichtige Beispiele . . . 25
3.4 Die Poissonapproximation der Binomialverteilung . . . 29
3.5 Erwartungswert . . . 30
3.6 Länge einer Zufallsvariable . . . 33
3.7 Varianz einer Zufallsvariable . . . 36
3.8 Unabhängige Zufallsvariablen . . . 38
3.9 Varianz einer Summe . . . 39
3.10 Kovarianz und Korrelation . . . 41
3.11 Zusammenfassung des Kapitels . . . 43

4 Ungleichungen und Abweichungen **44**
4.1 Jensen'sche Ungleichung . . . 44
4.2 Entropie einer Zufallsvariable . . . 46
4.3 Markov'sche Ungleichung . . . 49
4.4 Chebyshev'sche Ungleichung . . . 49
4.5 Schwaches Gesetz der großen Zahlen . . . 50

5 Bedingte Verteilung und bedingte Erwartung **52**
5.1 Bedingte Verteilung . . . 52
5.2 Bedingte Erwartung . . . 54
5.3 Turmeigenschaft . . . 56
5.4 Bedingte Erwartung bezüglich σ-Algebren . . . 58
5.5 Martingale . . . 60
5.6 Stoppzeiten . . . 63
5.7 Konvergenz von Martingalen . . . 66
5.8 Uniform intergrierbare Martingale . . . 67
5.9 Kolmogorovs 0-1-Gesetz . . . 69
5.10 Starkes Gesetz der großen Zahlen unter Momentenbedingungen . . . 70
5.11 Klassische Variante des starken Gesetzes der großen Zahlen . . . 71
5.12 Skizze eines weiteren Beweises des starken Gesetzes der großen Zahlen . . . 72
5.13 Maximalungleichungen . . . 73
5.14 Zusammenfassung . . . 74

6 Stetige Zufallsvariablen **75**
6.1 Dichtefunktion, Beispiele . . . 75
6.2 Erwartungswert einer stetigen Zufallsvariable. . . . 77
6.3 Gemeinsame Verteilung . . . 81
6.4 Unabhängigkeit von Zufallsvariablen . . . 82
6.5 Transformation von Zufallsvariablen . . . 87

7 Normalverteilung und Zentraler Grenzverteilungssatz **91**
7.1 Momentenerzeugende Funktion . . . 91
7.2 Die Normalverteilung . . . 93
7.3 Der zentrale Grenzverteilungssatz . . . 94
7.4 Beispiele und Anwendungen . . . 97

8 Einführung in die Statistik **99**
8.1 Was ist Statistik? . . . 99
8.2 Empirische Verteilungsfunktion . . . 100
8.3 Maximum-Likelihood Schätzung . . . 100
8.4 Abstrakter Blickwinkel auf Parameterschätzung . . . 103
8.5 Konfidenzintervalle . . . 106
8.6 Lineare Regression . . . 109

8.7 Hypothesentest und Neyman–Pearson lemma 113

1 Einführung

Diese Vorlesungsunterlagen basieren zu einem großen Teil auf den ausgezeichneten (englischsprachigen) Unterlagen von Richard Weber, die unter

`http://www.statslab.cam.ac.uk/~rrw1/prob/index`

zur Verfügung stehen.

Prüfung. Am Ende des Kurses findet eine schriftliche Prüfung statt. Weitere Details folgen.

Wir bitten um Verständnis, dass die Unterlagen erst im Laufe der Vorlesung entstehen. Insbesondere werden sich jene Teile des Skriptums, die in der Vorlesung noch nicht besprochen wurden, noch deutlich ändern.

1.1 Begriffe

In diesem Kurs führen wir in die Grundlagen der Wahrscheinlichkeitstheorie ein. Verglichen mit anderen Teilen der Mathematik ist sie ein junges Gebiet: Die ersten systematischen Überlegungen stammen von Pascal und Fermat aus dem 17. Jahrhundert, der Formalismus, den wir heute verwenden, geht auf Kolmogorov (1933) zurück.

Die Theorie benötigt drei Zutaten. Ausgangspunkt ist die Annahme, dass wir alle möglichen Ausgänge (auch: Fälle, Ergebnisse) eines Zufallsexperiments beschreiben können. Sie bilden eine Menge Ω, den **Grundraum**.

Beispiel 1.1. Wurf eines Würfels. Es gibt 6 Möglichkeiten, wir wählen daher $\Omega = \{1, \dots, 6\}$.

Dieses Zufallsexperiment ist leicht zu beschreiben. Oft ist der Grundraum Ω aber komplizierter.

Beispiel 1.2. Unendlich viele Würfe einer Münze: Wir können

$$\Omega = \{0,1\}^{\mathbb{N}} = \big\{\omega = (\omega_1, \omega_2, \dots) : \omega_i \in \{0,1\}\big\}$$

wählen, also die Menge aller 0-1-Folgen.

Oder, noch komplizierter:

Beispiel 1.3. Bewegung eines mikroskopisch kleinen Teilchens in einer Flüssigkeit. Hier wählen wir $\Omega = C(\mathbb{R}_+, \mathbb{R}^3)$, die Menge aller stetigen Bahnen im Raum.

Als zweite Zutat führen wir die Menge $\mathcal{A}$ der **beobachtbaren Ereignisse** ein. Sie besteht aus Teilmengen von Ω, d.h. $\mathcal{A} \subseteq \mathcal{P}(\Omega)$. Ein Element $A \in \mathcal{A}$ nennen wir ein **Ereignis**. Wir sagen, dass A *eintritt*, wenn der tatsächlich realisierte Ausgang ω in A liegt.

Ereignisse lassen sich oft auch in Worten beschreiben:

Beispiel 1.4. "Die Augenzahl ist gerade" $\leftrightarrow A = \{2, 4, 6\} \subseteq \Omega$.

Beispiel 1.5. "Der dritte Wurf ist Kopf" $\leftrightarrow A = \{\omega \in \Omega : \omega_3 = 1\}$.
"Mindestens die Hälfte der ersten 10 Würfe ist Kopf" $\leftrightarrow B = \{\omega \in \Omega : \sum_{i=1}^{10} \omega_i \geq 5\}$.

Beispiel 1.6. "Das Teilchen verlässt bis zur Zeit T die Kugel vom Radius R um 0 nicht" $\leftrightarrow A = \{\omega \in \Omega : \sup_{0 \leq t \leq T} \|\omega(t)\| \leq R\}$.

Mit den Operationen der Mengenlehre kann man aus gegebenen Ereignissen neue bilden. Seien A, B Ereignisse.

- $A^c = \Omega \setminus A$ ist das Komplement von A: A^c tritt genau dann ein, wenn A nicht eintritt.
- $A \cup B$ ist die Vereinigung von A und B: $A \cup B$ tritt genau dann ein, wenn A oder B (oder beide) eintreten.
- $A \cap B$ ist der Durchschnitt von A und B: $A \cap B$ tritt genau dann ein, wenn sowohl A als auch B eintreten.
- $A \setminus B = A \cap B^c$: $A \setminus B$ tritt genau dann ein, wenn A eintritt, B aber nicht.

Auch Beziehungen zwischen Mengen haben eine anschauliche Bedeutung:

- $A \subseteq B$: Wenn A eintritt, dann tritt auch B ein.
- $A \cap B = \emptyset$: A und B heißen *disjunkt*; sie können nicht gleichzeitig eintreten.

Beispiel 1.7. Seien A, B und C Ereignisse. Das Ereignis E, dass genau eines der drei eintritt, ist

$$E = (A \cap B^c \cap C^c) \cup (A^c \cap B \cap C^c) \cup (A^c \cap B^c \cap C).$$

Ist Ω endlich (wie in Beispiel 1.1) oder abzählbar unendlich (*diskreter Fall*), so wählen wir als $\mathcal{A}$ meist die Potenzmenge $\mathcal{P}(\Omega)$. Im Allgemeinen (wie in den Beispielen 1.2 und 1.3) besteht $\mathcal{A}$ aber nicht aus allen Teilmengen von Ω.

Die dritte und wichtigste Zutat ist das **Wahrscheinlichkeitsmaß** $\mathbb{P}$, eine Abbildung von $\mathcal{A}$ nach $[0, 1]$. Für $A \in \mathcal{A}$ ist $\mathbb{P}(A)$ die Wahrscheinlichkeit, dass A eintritt. Je größer $\mathbb{P}(A)$, desto eher rechnen wir damit, dass A eintritt.

Wie man $\mathbb{P}$ in einer konkreten Situation festlegt, ist für Anwendungen die entscheidende Frage. Die Mathematik allein beantwortet sie nicht; wir kommen in Abschnitt 1.3 ausführlich darauf zurück. Im Wahrscheinlichkeitsteil der Vorlesung wird $\mathbb{P}$ stets (zumindest implizit) vorgegeben sein. Im Statistikteil lernen wir Methoden kennen, $\mathbb{P}$ aus Daten zu schätzen.

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

1.2 Gleichwahrscheinliche Ausgänge

Bevor wir den axiomatischen Zugang entwickeln, betrachten wir einen einfachen, aber wichtigen Spezialfall: Der Grundraum Ω ist endlich und alle Ausgänge sind gleich wahrscheinlich. Dann ist $\mathcal{A} = \mathcal{P}(\Omega)$ und

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} \qquad (A \subseteq \Omega).$$

Insbesondere ist $\mathbb{P}(\emptyset) = 0$ und $\mathbb{P}(\Omega) = 1$.

Viele Standardaufgaben lassen sich so lösen: Man muss "nur" die *günstigen* Ausgänge zählen und durch die Anzahl der *möglichen* Ausgänge dividieren (Laplace). In der Praxis kann das Zählen allerdings sehr anspruchsvoll sein. Warum gerade diese Ausgänge gleich wahrscheinlich sein sollen, ist eine Frage an die reale Situation, nicht an die Mathematik; siehe Abschnitt 1.3.

Beispiel 1.8. Wir werfen zwei Würfel. Wie groß ist die Wahrscheinlichkeit, dass mindestens eine 6 fällt? Welche Augensumme ist am wahrscheinlichsten?

Zunächst wählen wir einen Grundraum. Zweckmäßig ist $\Omega = \{(i,j) : 1 \le i,j \le 6\}$: Ein Ausgang ist ein Paar von Augenzahlen, i für den ersten, j für den zweiten Würfel. Es ist $|\Omega| = 36$, und wir nehmen an, dass alle Paare gleich wahrscheinlich sind. Das Ereignis "mindestens eine 6" ist

$$A = \{(1,6), \ldots, (6,6), (6,1), \ldots, (6,5)\},$$

also $|A| = 11$ und $\mathbb{P}(A) = 11/36$.

Sei S_k das Ereignis, dass die Augensumme gleich k ist ($2 \le k \le 12$). Günstig sind die Ausgänge $(i, k-i)$ mit $1 \le i \le 6$ und $1 \le k - i \le 6$, d.h.

$$(k-6) \vee 1 \le i \le 6 \wedge (k-1),$$

wobei $a \wedge b = \min(a,b)$ und $a \vee b = \max(a,b)$. Daher ist

- $|S_k| = k - 1$ für $k \le 7$,
- $|S_k| = 13 - k$ für $k \ge 7$.

Am wahrscheinlichsten ist also die Augensumme 7, mit $\mathbb{P}(S_7) = 6/36 = 1/6$. □

Beispiel 1.9. Wir schreiben zufällig r Ziffern zwischen 0 und 9 hin. Wie groß ist die Wahrscheinlichkeit, dass (i) keine Ziffer größer als k ist, und (ii) k die größte vorkommende Ziffer ist ($0 \le k \le 9$)?

Wir wählen $\Omega = \{0, \ldots, 9\}^r = \{\omega = (\omega_1, \ldots, \omega_r) : 0 \le \omega_i \le 9\}$, also $|\Omega| = 10^r$.

(i) Sei A_k das Ereignis, dass keine Ziffer größer als k ist. Dann ist $A_k = \{\omega \in \Omega : \omega_i \le k \text{ für } 1 \le i \le r\}$, also $|A_k| = (k+1)^r$ und $\mathbb{P}(A_k) = (k+1)^r/10^r$.

(ii) Sei B_k das Ereignis, dass k die größte vorkommende Ziffer ist. Dann ist $B_k = A_k \setminus A_{k-1}$, und wegen $A_{k-1} \subseteq A_k$ gilt $|B_k| = |A_k| - |A_{k-1}|$. Somit

$$\mathbb{P}(B_k) = \frac{(k+1)^r - k^r}{10^r}. \qquad \square$$

Bei solchen Beispielen muss man stets darauf achten, ob *mit* oder *ohne* Zurücklegen gezogen wird.

Beispiel 1.10. Ich habe n Schlüssel in der Tasche, von denen genau einer sperrt. Ich wähle zufällig einen aus und probiere ihn. Wie groß ist die Wahrscheinlichkeit, dass ich genau beim r-ten Versuch (und nicht früher) Erfolg habe, wenn ich (i) falsche Schlüssel wieder zurück in die Tasche gebe, (ii) falsche Schlüssel beiseitelege?

Sei A_r das Ereignis, dass der r-te Versuch der erste erfolgreiche ist.
(i) Mit Zurücklegen gilt

$$\mathbb{P}(A_r) = \frac{(n-1)^{r-1} \cdot 1}{n^r}.$$

(ii) Ohne Zurücklegen gilt

$$\mathbb{P}(A_r) = \frac{(n-1)(n-2)\cdots(n-r+1)\cdot 1}{n(n-1)\cdots(n-r+1)} = \frac{1}{n}.$$

Das ist auch anschaulich klar: Der richtige Schlüssel steht in einer zufälligen Reihenfolge mit gleicher Wahrscheinlichkeit an jeder der n Stellen. $\square$

Beispiel 1.11 (Geburtstagsproblem)**.** Wie viele Personen müssen in einem Raum sein, damit mit Wahrscheinlichkeit größer als 1/2 mindestens zwei am selben Tag Geburtstag haben?

Der Einfachheit halber ignorieren wir Schaltjahre und nehmen an, dass alle 365 Tage gleich wahrscheinlich sind. Man könnte vermuten, dass man etwa halb so viele Personen braucht, also 183. Tatsächlich genügen aber 23. Sei $f(r)$ die Wahrscheinlichkeit, dass es unter r Personen zwei mit gleichem Geburtstag gibt. Dann ist

$$1 - f(r) = \mathbb{P}(\text{alle Geburtstage verschieden}) = \frac{364}{365} \cdot \frac{363}{365} \cdots \frac{365-(r-1)}{365}.$$

Auswerten ergibt $f(22) = 0.476\ldots$ und $f(23) = 0.507\ldots$.

Für die Intuition hilft es zu bedenken, dass es unter 23 Personen $\binom{23}{2} = 253$ Paare gibt, und jedes davon hat mit Wahrscheinlichkeit $\frac{1}{365}$ am selben Tag Geburtstag. $\square$

Beispiel 1.12 (Lotto)**.** Wie groß ist die Wahrscheinlichkeit, beim Lotto "6 aus 45" alle sechs Zahlen richtig zu tippen? Dazu überlegen wir, auf wie viele Arten 6 aus 45 Kugeln gezogen werden können.

Für die erste Kugel gibt es 45 Möglichkeiten, für die zweite 44, ..., für die sechste 40. Ziehungen, die sich nur in der Reihenfolge unterscheiden, sehen wir aber als gleich an, und es gibt $6! = 6 \cdot 5 \cdots 1$ Reihenfolgen. Insgesamt gibt es also

$$\frac{45 \cdot 44 \cdots 40}{6 \cdot 5 \cdots 1} = \binom{45}{6}$$

verschiedene Ziehungen, und die Wahrscheinlichkeit für einen Sechser ist $1/\binom{45}{6}$. Allgemein gibt es beim Lotto "k aus n" genau $\binom{n}{k}$ Möglichkeiten. □

Man könnte viele weitere Abzählaufgaben angeben und versuchen, sie zu systematisieren. In der Praxis ist es aber meist einfacher, sich die passende Formel am konkreten Beispiel selbst zu überlegen. Eine Variante des letzten Beispiels:

Beispiel 1.13. Beim Bridge werden 52 Karten auf 4 Spielerinnen verteilt, jede erhält 13 Karten. Analog zum Lotto gibt es

$$\frac{52!}{13! \cdot 13! \cdot 13! \cdot 13!}$$

Möglichkeiten, die Karten zu verteilen.

Dieses Beispiel motiviert den *Multinomialkoeffizienten*

$$\binom{n}{n_1, \ldots, n_k} := \frac{n!}{n_1! \cdots n_k!},$$

wobei $n_1, \ldots, n_k$ natürliche Zahlen mit $n_1 + \ldots + n_k = n$ sind.

Bemerkung 1.14 (Stirlingformel)**.** In Abzählaufgaben tritt häufig $n!$ auf. Die *Stirlingformel* besagt

$$n! \approx \sqrt{2\pi n} \left(\frac{n}{e}\right)^n \qquad (n \to \infty),$$

wobei $\approx$ bedeutet, dass der Quotient beider Seiten gegen 1 konvergiert. Die Näherung ist schon für kleine n erstaunlich gut: Der relative Fehler ist für $n = 1$ kleiner als 10%, für $n = 10$ kleiner als 1%.

Werfen wir etwa eine faire Münze $2n$-mal, so ist die Wahrscheinlichkeit, genau gleich oft Kopf wie Zahl zu werfen,

$$\binom{2n}{n} \frac{1}{2^{2n}} = \frac{(2n)!}{2^{2n}(n!)^2} \approx \frac{1}{\sqrt{\pi n}},$$

wie man durch Einsetzen der Stirlingformel sieht. Diese Wahrscheinlichkeit geht also gegen 0.

1.3 Wahrscheinlichkeit und Wirklichkeit

Die Rechenregeln für Wahrscheinlichkeiten, die wir in Kapitel 2 als Axiome festhalten werden, sagen uns, wie man mit Wahrscheinlichkeiten *rechnet*. Sie sagen aber nicht, was die Aussage "$\mathbb{P}(A) = 0.3$" über die Welt bedeutet, und auch nicht, wie man in einer konkreten Situation zu $\mathbb{P}$ kommt. Das ist keine Besonderheit der Stochastik: Auch die Axiome der euklidischen Geometrie sagen nicht, was ein Punkt "ist", und ob ein Feld tatsächlich rechteckig ist, entscheidet nicht die Mathematik, sondern das Maßband.

Wer Wahrscheinlichkeitsrechnung anwendet oder unterrichtet, durchläuft daher immer, meist stillschweigend, einen **Modellierungskreislauf**:

1. *Reale Situation*: ein Würfel, eine Wahl, ein medizinischer Test, das Wetter morgen.
2. *Modell*: Wahl von $(\Omega, \mathcal{A}, \mathbb{P})$. Hier fließen Annahmen ein, die *nicht* aus der Mathematik folgen, etwa "alle Seiten sind gleich wahrscheinlich" oder "die Würfe beeinflussen einander nicht".
3. *Mathematik*: Folgerungen innerhalb des Modells. (Das ist der Hauptinhalt dieser Vorlesung.)
4. *Interpretation und Überprüfung*: Was bedeutet das Ergebnis für die reale Situation, und passt es zu dem, was wir beobachten? Falls nicht, zurück zu Schritt 2.

Die Schritte 2 und 4 sind die Brücke zwischen Theorie und Wirklichkeit. Dazu gibt es verschiedene, historisch gewachsene Sichtweisen.

Drei klassische Interpretationen.

- **Laplace (Symmetrie).** Gibt es n Ausgänge, von denen "keiner bevorzugt ist", so hat jeder die Wahrscheinlichkeit $1/n$. Das funktioniert hervorragend bei Würfeln, Münzen, Karten und Lotto. Es versagt, wo keine Symmetrie in Sicht ist: Mit welcher Wahrscheinlichkeit landet ein Reißnagel auf der Spitze?
- **Frequentistisch (Häufigkeit).** Wiederholen wir ein Experiment N-mal "unabhängig voneinander" und ist $N(A)$ die Anzahl der Versuche, bei denen A eintritt, so erwarten wir
$$\mathbb{P}(A) = \lim_{N\to\infty} \frac{N(A)}{N}. \tag{1.1}$$
In diesem Sinn ist $\mathbb{P}(A)$ die relative Häufigkeit, mit der A auf lange Sicht eintritt. Das erklärt, warum Wahrscheinlichkeiten *messbar* sind, und liefert eine Antwort beim Reißnagel. Später werden wir Unabhängigkeit präzise definieren, und (1.1) wird sich als wichtiger Satz herausstellen: das **Gesetz der großen Zahlen** (Abschnitt 4.5). Als *Definition* von Wahrscheinlichkeit taugt (1.1) allerdings nicht: Viele Ereignisse sind nicht wiederholbar ("Morgen regnet es", "Partei X gewinnt die nächste Wahl"), und das Gesetz der großen Zahlen setzt ein Wahrscheinlichkeitsmaß bereits voraus.
- **Subjektiv (Grad der Überzeugung).** $\mathbb{P}(A)$ drückt aus, wie sehr eine Person mit dem Eintreten von A rechnet, etwa gemessen an den Wettquoten, die sie für fair hält (de Finetti). Damit kann man auch Einzelereignissen Wahrscheinlichkeiten geben, und der Satz von Bayes (Abschnitt 2.4) beschreibt, wie man solche Einschätzungen im Licht neuer Daten anpasst. Der Preis: Zwei Personen dürfen verschiedene Wahrscheinlichkeiten für dasselbe Ereignis haben, und die Theorie allein sagt nicht, wer "recht" hat.

Bemerkenswert ist, dass alle drei Sichtweisen zu *denselben* Rechenregeln führen, nämlich zu den Axiomen in Kapitel 2. Die Axiome sind gerade so gebaut, dass sie mit jeder dieser Interpretationen verträglich sind. Das ist ihre Stärke, erklärt aber auch, warum sie zur Frage "Was bedeutet $\mathbb{P}(A)$?" schweigen.

Symmetrie und Unabhängigkeit als Fundament (Burdzy). Der Probabilist Krzysztof Burdzy hat vorgeschlagen, die Brücke zur Wirklichkeit nicht über Häufigkeiten oder Überzeugungen zu schlagen, sondern über Eigenschaften der realen Situation, die man *vor* jedem Experiment erkennen kann.[1] Er formuliert fünf "Gesetze der Wahrscheinlichkeit", hier sinngemäß übersetzt:

(L1) Wahrscheinlichkeiten sind Zahlen zwischen 0 und 1, die Ereignissen mit ungewissem Ausgang zugeordnet werden.

(L2) Können A und B nicht gleichzeitig eintreten, so ist $\mathbb{P}(A \text{ oder } B) = \mathbb{P}(A) + \mathbb{P}(B)$.

(L3) Haben A und B physikalisch nichts miteinander zu tun, so sind sie im mathematischen Sinn unabhängig: $\mathbb{P}(A \text{ und } B) = \mathbb{P}(A)\mathbb{P}(B)$.

(L4) Gibt es eine Symmetrie der möglichen Ausgänge, die das Ereignis A auf das Ereignis B abbildet, so ist $\mathbb{P}(A) = \mathbb{P}(B)$.

(L5) Ein Ereignis hat Wahrscheinlichkeit 0 genau dann, wenn es nicht eintreten kann, und Wahrscheinlichkeit 1 genau dann, wenn es eintreten muss.

(L1) und (L2) entsprechen im Wesentlichen unseren Axiomen. Neu ist, dass (L3) und (L4) ausdrücklich sagen, *woher* in Anwendungen konkrete Wahrscheinlichkeiten kommen: aus Symmetrien und aus dem Fehlen eines physikalischen Zusammenhangs. Das sind Aussagen über die Welt, keine Mathematik; sie können sich im Einzelfall als falsch herausstellen. Burdzy betont, dass die Symmetrie in (L4) eine *physikalische* Symmetrie der Situation sein muss, nicht bloß eine Lücke in unserem Wissen. (L5) schließlich sagt, wie Wahrscheinlichkeitsaussagen mit der Beobachtung in Kontakt kommen; dazu gleich mehr.

Beispiel 1.15 (Würfel)**.** Ein Würfel ist ein Würfel: Zu je zwei Seiten gibt es eine Drehung, die den Würfel in sich überführt und die eine Seite auf die andere abbildet. Nach (L4) haben alle Seiten dieselbe Wahrscheinlichkeit, nach (L1) und (L2) also jeweils $1/6$. Das ist genau das Laplace-Modell, aber mit einer Begründung, die man am Objekt überprüfen kann. Ein gezinkter Würfel (etwa mit einem Bleigewicht an einer Seite) hat diese Symmetrie nicht, und das Laplace-Modell ist dann schlicht nicht anwendbar.

Beispiel 1.16 (Zwei Würfel und ein Irrtum von Leibniz)**.** Beim Werfen zweier Würfel ist nicht die Augensumme gleichverteilt, sondern das *Paar* (i, j): Die Symmetrien beider Würfel und (L3) ergeben $\mathbb{P}(\{(i,j)\}) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$. Daraus folgt $\mathbb{P}(\text{Summe } 11) = 2/36$, aber $\mathbb{P}(\text{Summe } 12) = 1/36$ (Beispiel 1.8). Leibniz hielt beide für gleich wahrscheinlich, weil jede Summe nur "auf eine Art" zustande komme. Welche Ausgänge gleich wahrscheinlich sind, muss also aus der physikalischen Situation begründet werden und kann nicht am Papier entschieden werden.

[1] K. Burdzy, *Probability is Symmetry: On Foundations of the Science of Probability*, Manuskript (2003); ausführlicher in *The Search for Certainty*, World Scientific (2009).

Beispiel 1.17 (Reißnagel)**.** Für einen einzelnen Wurf eines Reißnagels gibt es keine Symmetrie zwischen "Spitze oben" und "Spitze seitlich"; die Wahrscheinlichkeit p ist unbekannt. Trotzdem liefern (L3) und (L4) ein Modell: Die Würfe beeinflussen einander nicht (L3), und es ist egal, ob wir einen Wurf als ersten oder als siebenten durchführen, d.h. das Vertauschen der Würfe ist eine Symmetrie (L4). Damit sind die Würfe unabhängig und haben alle dieselbe unbekannte Erfolgswahrscheinlichkeit p. Erst *dann* kommen die Häufigkeiten ins Spiel: Nach dem Gesetz der großen Zahlen ist $N(A)/N$ für großes N nahe bei p, und wir können p schätzen (Kapitel 8). Die Häufigkeitsinterpretation ist in dieser Sichtweise also keine Definition, sondern ein *Satz*, der auf Symmetrie und Unabhängigkeit aufbaut.

Wie überprüft man eine Wahrscheinlichkeitsaussage? Die Aussage "Diese Münze ist fair" kann durch keine einzelne Beobachtung widerlegt werden: Jede endliche Folge von Würfen hat positive Wahrscheinlichkeit. Überprüfbar werden Wahrscheinlichkeitsmodelle erst, wenn sie Ereignisse mit Wahrscheinlichkeit *nahe* bei 0 oder 1 vorhersagen. Das ist die praktische Form von (L5):

Ein Ereignis mit sehr kleiner Wahrscheinlichkeit erwartet man in einem einzelnen Versuch nicht.

Werfen wir eine faire Münze 10 000-mal, so liegt der Anteil der Köpfe nach der Chebyshev'schen Ungleichung (Abschnitt 4.5) mit Wahrscheinlichkeit mindestens 0.99 zwischen 45% und 55%; mit dem zentralen Grenzverteilungssatz (Abschnitt 7.3) sieht man, dass diese Wahrscheinlichkeit sogar viel näher bei 1 liegt. Beobachten wir 60% Köpfe, so ist das kein Widerspruch zur Mathematik, aber ein sehr starkes Indiz, dass das Modell "faire Münze" für diese Münze nicht passt. Genau diese Logik steckt hinter Hypothesentests und Konfidenzintervallen (Kapitel 8). Burdzy geht so weit zu sagen, dass Wahrscheinlichkeiten strikt zwischen 0 und 1 vor allem als Zwischenschritte dienen, um zu solchen überprüfbaren Aussagen zu gelangen.

Einzelereignisse. "Morgen regnet es mit Wahrscheinlichkeit 30%" betrifft ein Ereignis, das es nur einmal gibt. Man kann eine solche Prognose trotzdem an der Wirklichkeit messen, nämlich über viele Tage hinweg: Ein Wetterdienst ist gut *kalibriert*, wenn es auf lange Sicht an etwa 30% jener Tage regnet, für die er 30% angesagt hat. Hier treffen sich subjektive und frequentistische Sichtweise: Die einzelne Prognose ist eine Einschätzung, ihre Qualität zeigt sich in Häufigkeiten.

Bemerkung 1.18 (Für den Unterricht)**.** Im Schulunterricht tauchen alle drei Interpretationen auf, oft ohne dass sie auseinandergehalten werden: Laplace-Wahrscheinlichkeiten beim Würfeln, relative Häufigkeiten bei Experimenten und Simulationen, subjektive Einschätzungen im Alltag ("Wie wahrscheinlich ist es, dass . . . ?"). Es lohnt sich, das explizit zu machen und bei jeder Aufgabe zu fragen: *Woher kommt hier die Wahrscheinlichkeit?* Zwei typische Fehlvorstellungen lassen sich mit (L3) und dem Gesetz der großen Zahlen gut besprechen:

- *Spielerfehlschluss*: "Nach fünfmal Rot beim Roulette ist Schwarz überfällig." Die Kugel hat kein Gedächtnis; die Drehungen sind physikalisch unabhängig (L3), die Wahrscheinlichkeit für Schwarz ist also unverändert.

- *"Das Gesetz der großen Zahlen gleicht aus"*: Bei $2n$ Münzwürfen konvergiert zwar der *Anteil* der Köpfe gegen $1/2$, die *Differenz* zwischen der Anzahl der Köpfe und der Anzahl der Zahlen wird aber typischerweise immer größer, nämlich von der Größenordnung $\sqrt{n}$. Die Wahrscheinlichkeit, genau gleich viele Köpfe wie Zahlen zu werfen, geht sogar gegen 0, wie $1/\sqrt{\pi n}$ (Bemerkung zur Stirlingformel). Ausgeglichen wird nicht, es wird verdünnt.

1.4 Zusammenfassung des Kapitels

Am Ende dieses Kapitels sollten Sie mit folgenden Begriffen und Techniken vertraut sein:

- Wahrscheinlichkeitsraum, (beobachtbare) Ereignisse;

- Berechnung von Wahrscheinlichkeiten mittels Kombinatorik, wenn alle Ausgänge gleich wahrscheinlich sind;

- Näherungen solcher Wahrscheinlichkeiten mit der Stirlingformel;

- der Modellierungskreislauf und die Frage, woher Wahrscheinlichkeiten in realen Situationen kommen (Symmetrie, Unabhängigkeit, Häufigkeit, subjektive Einschätzung).

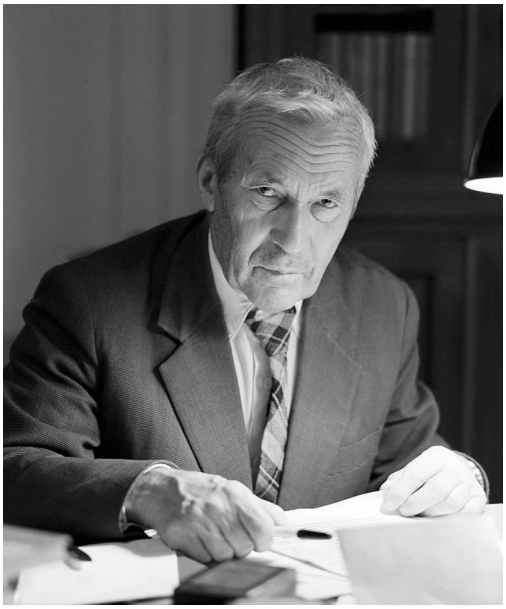


Abbildung 1: Andrey Kolmogorov (1903–1987), der im Jahr 1933 den axiomatischen Zugang zur Wahrscheinlichkeitstheorie einführte.

2 Axiomatischer Zugang

2.1 Axiome der Wahrscheinlichkeitstheorie

Ein Wahrscheinlichkeitsraum ist ein Tripel $(\Omega, \mathcal{F}, \mathbb{P})$, wobei $\Omega \neq \emptyset$, $\mathcal{F} \subseteq \mathcal{P}(\Omega)$, $\mathbb{P} : \mathcal{F} \to [0, 1]$. Von $\mathcal{F}$ verlangen wir, dass es eine σ-Algebra ist, d.h.

1. $\emptyset \in \mathcal{F}$, $\Omega \in \mathcal{F}$.
2. Falls $A \in \mathcal{F}$, dann auch $A^c \in \mathcal{F}$.
3. Falls $A_1, A_2, \ldots \in \mathcal{F}$, dann auch $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Damit $\mathbb{P}$ ein Wahrscheinlichkeitsmaß ist, muss es die folgenden Axiome erfüllen:

I. Für alle $A \in \mathcal{F}$ gilt $0 \leq \mathbb{P}(A) \leq 1$.

II. $\mathbb{P}(\Omega) = 1$.

III. Sind $A_1, A_2, \ldots \in \mathcal{F}$ paarweise disjunkt, so gilt

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

Axiom III. nennen wir die **σ-Additivität** von $\mathbb{P}$. Falls Ω höchstens abzählbar ist, nehmen wir in der Regel $\mathcal{F} = \mathcal{P}(\Omega)$.

Beispiel 2.1. Sei $\Omega = \{\omega_1, \omega_2, \ldots\}$. Gegeben $p_1, p_2, \ldots \geq 0$ mit $\sum_{i=1}^{\infty} p_i = 1$, definiert

$$\mathbb{P}(A) := \sum_{i:\omega_i \in A} p_i$$

ein Wahrscheinlichkeitsmaß auf $\mathcal{P}(\Omega)$. Tatsächlich ist für abzählbaren Grundraum Ω jedes Wahrscheinlichkeitsmaß von dieser Form.

Wie in Abschnitt 1.3 besprochen, ist es für die Intuition hilfreich, $\mathbb{P}(A)$ als relative Häufigkeit des Eintretens von A bei vielen unabhängigen Wiederholungen zu verstehen, siehe (1.1).

2.2 Elementare Eigenschaften

Wir beweisen ein paar elementare, aber nützliche Eigenschaften von Wahrscheinlichkeitsmaßen.

Proposition 2.2. *Sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. Dann gilt:*

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

(ii) $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ *für* $A \in \mathcal{F}$,

(iii) für $A, B \in \mathcal{F}$ *mit* $A \subset B$ *gilt* $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \geq \mathbb{P}(A)$ *(Monotonie),*

(iv) für alle $A, B \in \mathcal{F}$, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

Beweis. Sei $A \in \mathcal{F}$, und sei $A_1 = A, A_2 = A^c, A_3 = \emptyset, A_4 = \emptyset, \ldots$. Dann sind die A_n paarweise disjunkt, und mit der σ-Additivität (Axiom III.) folgt

$$1 = \mathbb{P}(\Omega) = \mathbb{P}(\cup_n A_n) = \sum_n \mathbb{P}(A_n).$$

Deshalb gilt $\mathbb{P}(\emptyset) = 0$, und $\mathbb{P}(A) + \mathbb{P}(A^c) = 1$. Das beweist (i) und (ii).

Für (iii) bemerken wir, dass $B = A \cup (B \setminus A)$ und dass diese Mengen disjunkt sind. Deswegen (σ-Additivität) gilt $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \geq \mathbb{P}(A)$.

Für (iv) schreiben wir $A \cup B$ als $[A \cap B] \cup [A \setminus (A \cap B)] \cup [B \setminus (A \cap B)]$, wobei die drei Ereignisse disjunkt sind. Deshalb gilt auch

$$\begin{aligned} \mathbb{P}(A \cup B) &= \mathbb{P}(A \cap B) + \mathbb{P}(A \setminus (A \cap B)) + \mathbb{P}(B \setminus (A \cap B)) \\ &= \mathbb{P}(A \cap B) + \mathbb{P}(A) - \mathbb{P}(A \cap B) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \quad \text{durch (iii)} \\ &= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \end{aligned}$$

wie gewünscht. □

Beispiel 2.3. Wir wählen eine Zahl N aus $\{1, \ldots, 500\}$, wobei jedes Ergebnis gleich wahrscheinlich sein soll. Wie groß ist die Wahrscheinlichkeit, dass N durch 5 oder 3 teilbar ist?

Sei A (beziehungsweise B) das Ereignis, dass N durch 5 (beziehungsweise 3) teilbar ist. Dann gilt (weil 5 und 3 teilerfremd sind)

$$\begin{aligned}\mathbb{P}(A) &= \frac{\lfloor 500/5 \rfloor}{500} = \frac{100}{500} \\ \mathbb{P}(B) &= \frac{\lfloor 500/3 \rfloor}{500} = \frac{166}{500} \\ \mathbb{P}(A \cap B) &= \frac{\lfloor 500/(3 \times 5) \rfloor}{500} = \frac{33}{500},\end{aligned}$$

deshalb erhalten wir

$$\mathbb{P}(A \cup B) = \frac{100 + 166 - 33}{500} = \frac{233}{500} = 0.466. \qquad \square$$

Proposition 2.4 (Stetigkeit des Wahrscheinlichkeitsmaßes)**.** *Für $A_1, A_2, \ldots \in \mathcal{F}$ mit $A_1 \subset A_2 \subset \ldots$ gilt*

$$\mathbb{P}(A) = \lim_{n \to \infty} \mathbb{P}(A_n)$$

wobei $A = \cup_{n=1}^{\infty} A_n$. Äquivalent dazu gilt für $B_1, B_2, \ldots \in \mathcal{F}$ mit $B_1 \supset B_2 \supset \ldots$, dass

$$\mathbb{P}(B) = \lim_{n \to \infty} \mathbb{P}(B_n)$$

wobei $B = \cap_{n=1}^{\infty} B_n$.

Beweis. Sei $A_1' = A_1$ und $A_n' = A_n \setminus A_{n-1}$ für $n \geq 2$. Dann gilt $A_n = \cup_{i=1}^{n} A_i'$, deshalb ist $A = \cup_{i=1}^{\infty} A_i'$, wobei die A_n' paarweise disjunkt sind. Deswegen gilt

$$\begin{aligned}\mathbb{P}(A) &= \sum_{i=1}^{\infty} \mathbb{P}(A_i') \quad \text{(wegen } \sigma\text{-Additivität)} \\ &= \lim_{n \to \infty} \sum_{i=1}^{n} \mathbb{P}(A_i') \quad \text{(Definition von unendlichen Reihen)} \\ &= \lim_{n \to \infty} \mathbb{P}(A_n) \quad \text{(wegen } \sigma\text{-Additivität).}\end{aligned}$$

Die zweite Aussage bekommen wir, indem wir Komplemente nehmen. $\square$

2.3 Bedingte Wahrscheinlichkeit

Seien A und B Ereignisse mit $\mathbb{P}(B) > 0$. Ein entscheidendes Konzept der Wahrscheinlichkeitstheorie ist die bedingte Wahrscheinlichkeit. Es erlaubt uns, Aussagen wie "wenn B eintritt, dann ist die Wahrscheinlichkeit von A gleich p" zu machen. Betrachten wir zum Beispiel $A =$ "der Bus wird zu spät kommen" und $B =$ "es wird regnen". Unter der Voraussetzung, dass B eintritt, ist es wahrscheinlicher, dass A eintritt.

Definition 2.5. *Seien A und B Ereignisse mit $\mathbb{P}(B) > 0$. Die bedingte Wahrscheinlichkeit von A, gegeben B, ist*

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

Diese Definition ist mit (1.1) intuitiv einfach zu rechtfertigen. Wenn wir das Experiment N-mal wiederholen, achten wir nur auf Fälle, in denen B eintritt. (Andere Experimente können ignoriert werden.) Unter jenen Experimenten, für die B eintritt, zählen wir, wie oft A eintritt. Deswegen sollte die bedingte Wahrscheinlichkeit gleich

$$\lim_{N\to\infty} \frac{N(A \cap B)}{N(B)} = \lim_{N\to\infty} \frac{N(A \cap B)/N}{N(B)/N} = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

sein.

Beispiel 2.6. Ein Würfel wird geworfen. Sei B das Ereignis, dass das Resultat eine gerade Zahl ist. Und sei A das Ereignis, dass wir eine 6 würfeln. Was ist $\mathbb{P}(A|B)$? Was ist $\mathbb{P}(B|A)$?

Die Rechnung ist nicht schwer: $\mathbb{P}(A) = 1/6$, $\mathbb{P}(B) = 1/2$ und $\mathbb{P}(A \cap B) = \mathbb{P}(A) = 1/6$. Deshalb gilt

$$\mathbb{P}(A|B) = 1/3, \mathbb{P}(B|A) = 1. \qquad \square$$

Beispiel 2.7. Eine Familie hat 2 Kinder. Was ist die Wahrscheinlichkeit, dass beide Jungs sind, gegeben dass mindestens einer ein Junge ist? (Wir nehmen an, dass jedes Kind entweder ein Junge oder ein Mädchen sein kann, und beide Möglichkeiten gleich wahrscheinlich sind).

Viele Leute antworten $1/2$, aber das ist nicht richtig. Warum? Es gibt für die 2 Kinder insgesamt 4 Möglichkeiten:

$$\Omega = \{JJ, JM, MJ, MM\}.$$

Das Ereignis $B =$"mindestens einer ist ein Junge" ist deswegen $B = \{JJ, JM, MJ\}$: die einzige Information, die uns gegeben wird, ist, dass MM nicht eintritt. Deswegen gilt $\mathbb{P}(A|B) = 1/3$, wobei A das Ereignis "Beide sind Jungs" ist. $\square$

Für ein ähnlich kontraintuitives, berühmtes Problem suchen Sie nach dem "**Monty Hall Problem**"!

Proposition 2.8 (Gesetz der vollständigen Wahrscheinlichkeit)**.** *Seien A, B Ereignisse mit $0 < \mathbb{P}(B) < 1$. Dann gilt*

$$\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c).$$

In Worten bedeutet dieser Satz, dass man über die zwei Möglichkeiten B und B^c zerlegen kann, um die Gesamtwahrscheinlichkeit von A zu berechnen.

Beweis. Es reicht, A in der Form $A = (A \cap B) \cup (A \cap B^c)$ zu zerlegen und zu bemerken, dass die zwei Ereignisse disjunkt sind. Deshalb gilt

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c) \\ &= \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}\mathbb{P}(B) + \frac{\mathbb{P}(A \cap B^c)}{\mathbb{P}(B^c)}\mathbb{P}(B^c).\end{aligned}$$

Daraus schließen wir sofort auf das Ergebnis. □

Beispiel 2.9 ("Ruin des Spielers", "Gambler's ruin"). Eine Münze wird wiederholt geworfen. Bei jedem Wurf bekommt der Spieler 1€ (wenn Kopf kommt) oder verliert 1€ (wenn Zahl kommt). Anfänglich hat er x Euro. Er spielt weiter, bis er entweder pleite ist oder N Euro hat $(0 < x < N)$.

Sei p_x die Wahrscheinlichkeit, dass er N Euro erreicht, bevor er pleite ist.

Nach dem Gesetz der vollständigen Wahrscheinlichkeit gilt

$$p_x = \frac{1}{2}p_{x+1} + \frac{1}{2}p_{x-1};$$

mit den Randbedingungen $p_0 = 0; p_N = 1$. (Die eindeutige Lösung dieser Gleichung ist $p_x = x/N$.) □

2.4 Der Satz von Bayes

Das folgende Zitat aus [3] soll den Satz von Bayes motivieren.

As you consider the next question, please assume that Steve was selected at random from a representative sample. An individual has been described by a neighbor as follows: "Steve is very shy and withdrawn, invariably helpful but with little interest in people or in the world of reality. A meek and tidy soul, he has a need for order and structure, and a passion for detail." Is Steve more likely to be a librarian or a farmer? [...]

Did it occur to you that there are more than 20 male farmers for each male librarian in the United States? Because there are so many more farmers, it is almost certain that more 'meek and tidy' souls will be found on tractors than at library information desks.

Satz 2.10 (Satz von Bayes). *Sei $\{B_i : i \in I\}, B_i \in \mathcal{F}, \mathbb{P}(B_i) > 0$ eine Partition des Wahrscheinlichkeitsraumes, wobei I eine höchstens abzählbare Indexmenge ist, und bezeichne A ein Ereignis mit $\mathbb{P}(A) > 0$. Dann gilt für $j \in I$*

$$\mathbb{P}(B_j|A) = \frac{\mathbb{P}(A|B_j)\mathbb{P}(B_j)}{\sum_{i\in I}\mathbb{P}(A|B_i)\mathbb{P}(B_i)}.$$

Beweis. Folgt direkt aus der Definition der bedingten Wahrscheinlichkeit. □

Beispiel 2.11. Ein medizinischer Test erkennt eine infizierte Person mit einer Wahrscheinlichkeit von 98%. Allerdings hat er eine 'false positive' Rate von 1%. Angenommen, 0.1% der Bevölkerung haben die Krankheit. Wenn eine zufällige Person positiv getestet wurde, wie groß ist die Wahrscheinlichkeit, dass sie tatsächlich infiziert ist? Wir schreiben + für das Ereignis, positiv getestet zu werden, und K für das Ereignis, erkrankt zu sein. Dann erhalten wir mit dem Satz von Bayes

$$\mathbb{P}(K|+) = \frac{\mathbb{P}(+|K)\mathbb{P}(K)}{\mathbb{P}(+|K)\mathbb{P}(K) + \mathbb{P}(+|K^c)\mathbb{P}(K^c)} \tag{2.1}$$

$$= \frac{0.98 \cdot 0.001}{0.98 \cdot 0.001 + 0.01 \cdot 0.999} \approx 0.09 \tag{2.2}$$

D.h. nur etwa 9% der Leute, die positiv getestet werden, sind tatsächlich infiziert.

2.5 Unabhängigkeit

Beispiel 2.12. Ein Würfel wird geworfen und wir betrachten die Ereignisse G "Augenzahl gerade" und D "Augenzahl durch drei teilbar". Dann gilt $\mathbb{P}(G|D) = \mathbb{P}(G)$. Anders gesagt: Zu wissen, dass D eintritt, verrät einem nichts darüber, ob G eintreten wird; "das Eintreten von D ist unabhängig vom Eintreten von G".

Definition 2.13. *Zwei Ereignisse A, B heißen* unabhängig*, falls*

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

Falls zusätzlich $\mathbb{P}(B) > 0$, dann sind A, B genau dann unabhängig, wenn

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

Falls A, B unabhängige Ereignisse sind, so auch A, B^c, wie man leicht nachrechnet:

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

Beispiel 2.14. Wir werfen zwei Würfel. A ist das Ereignis, dass der erste Würfel eine gerade Augenzahl zeigt. B ist das Ereignis, dass der zweite Würfel eine gerade Augenzahl zeigt. Es ist dann leicht einzusehen, dass A und B unabhängig sind.

Sei weiters C das Ereignis, dass die Summe der beiden Augenzahlen ungerade ist. Man kann sich dann auch überlegen, dass A und C unabhängig sind. (Am besten malt man dazu ein Bild.)

Unabhängige Experimente: Wenn wir über unabhängige Ereignisse reden, stellen wir uns typischerweise vor, dass zwei Experimente durchgeführt werden, die nichts miteinander zu tun haben. Gegeben seien $\Omega_1 = \{\alpha_1, \alpha_2, \ldots\}$ und $\Omega_2 = \{\beta_1, \beta_2, \ldots\}$ mit Wahrscheinlichkeitsverteilungen $p_1, p_2, \ldots$ und $q_1, q_2, \ldots$. Wenn wir in diesem Zusammenhang von zwei unabhängigen Experimenten reden, dann meinen wir damit, dass wir den Grundraum $\Omega_1 \times \Omega_2$ mit der durch

$$\mathbb{P}((\alpha_i, \beta_j)) = p_i q_j$$

definierten Wahrscheinlichkeitsverteilung betrachten.

Seien nun $A \subseteq \Omega_1, B \subseteq \Omega_2$. Wir können dann das Ereignis A auch als Ereignis auf $\Omega_1 \times \Omega_2$ auffassen, indem wir es mit der Menge $A \times \Omega_2$ identifizieren. Analog verfahren wir mit der Menge B. Dann gilt

$$\mathbb{P}(A \cap B) = \sum_{\alpha_i \in A, \beta_j \in B} p_i q_j = \sum_{\alpha_i \in A} p_i \sum_{\beta_j \in B} q_j = \mathbb{P}(A)\mathbb{P}(B).$$

Das erklärt, warum wir die beiden Experimente als unabhängig bezeichnen.

2.6 Unabhängigkeit von mehreren Ereignissen

Definition 2.15. *Ereignisse $A_1, A_2, \ldots$ heißen* unabhängig *oder* gemeinsam unabhängig*, falls für alle $i_1 < i_2 < \ldots < i_k$ gilt, dass*

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}) = \mathbb{P}(A_{i_1})\mathbb{P}(A_{i_2}) \ldots \mathbb{P}(A_{i_k}).$$

Wie zuvor überlegt man sich leicht, dass die Unabhängigkeit von Ereignissen erhalten bleibt, wenn man manche der Ereignisse A_i durch ihre Komplementärereignisse A_i^c ersetzt.

Analog zu oben erhalten wir gemeinsam unabhängige Ereignisse, indem wir Ereignisse auf unterschiedlichen Wahrscheinlichkeitsräumen stattdessen auf dem gemeinsamen Produktraum betrachten.

Es kann übrigens passieren, dass Ereignisse $A_1, A_2, \ldots$ paarweise unabhängig sind, aber nicht gemeinsam. Etwa gilt in Beispiel 2.14

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

Beispiel 2.16. Wir nehmen an, dass eine Münze einmal geworfen wird, mit möglichen Ausgängen $\{K, Z\} = \Omega$. Für $p \in [0, 1]$ ist die *Bernoulliverteilung* $B(1, p)$ gegeben durch

$$\mathbb{P}(K) = p, \quad \mathbb{P}(Z) = 1 - p =: q.$$

Beispiel 2.17. Wenn wir die obige Münze n-mal werfen, erhalten wir eine Folge von *Bernoulliexperimenten*. Die Anzahl der Köpfe, die man dabei erhält, ist eine Zahl in der Menge $\Omega = \{0, 1, \ldots, n\}$. Gegeben $k \in \Omega$, gibt es $\binom{n}{k}$ Wurffolgen, die zu k Köpfen führen, und jede dieser Folgen passiert mit Wahrscheinlichkeit $p^k(1-p)^{n-k}$. Insgesamt erhalten wir die *Binomialverteilung* $B(n, p)$ durch

$$\mathbb{P}(k \text{ Köpfe}) = p_k = \binom{n}{k} p^k (1-p)^{n-k}.$$

Wie viele Köpfe erwarten wir im Durchschnitt? Im nächsten Kapitel werden wir besprechen, wie man diese Frage formalisiert.

2.7 Zusammenfassung des Kapitels

Am Ende dieses Kapitels sollten Sie mit folgenden Begriffen und Resultaten vertraut sein:

- den Axiomen der Wahrscheinlichkeitstheorie;
- der Definition der bedingten Wahrscheinlichkeit, dem Gesetz der vollständigen Wahrscheinlichkeit und dem Satz von Bayes;
- dem Begriff der Unabhängigkeit (für zwei oder mehrere Ereignisse).

3 Zufallsvariablen

3.1 Verteilungen auf $\mathbb{R}$

Bis jetzt konnten wir alle Beispiele so verstehen, dass bestimmten reellen Zahlen Wahrscheinlichkeiten zugeordnet wurden: der Augenzahl eines Würfels, dem Ergebnis 0 oder 1 eines Münzwurfs, der Anzahl der Einsen bei n Münzwürfen. In all diesen Fällen können wir als Grundraum einfach $\Omega = \mathbb{R}$ mit $\mathcal{F} = \mathcal{P}(\mathbb{R})$ wählen. Die Wahrscheinlichkeit sitzt dabei auf einer höchstens abzählbaren Menge $\mathcal{W} \subseteq \mathbb{R}$ möglicher Werte.

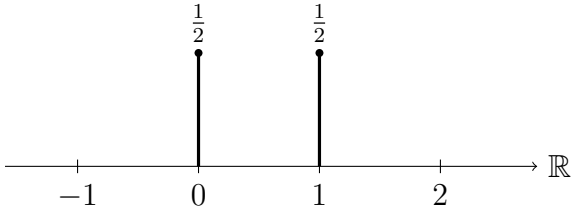
Beispiel 3.1 (Münze)**.** Wir modellieren den Wurf einer fairen Münze, die 0 oder 1 zeigt, durch den Wahrscheinlichkeitsraum $(\mathbb{R}, \mathcal{P}(\mathbb{R}), P)$ mit

$$P(A) := \sum_{x \in A} p(x), \qquad p(x) = \begin{cases} \frac{1}{2} & x = 0, \\ \frac{1}{2} & x = 1, \\ 0 & \text{sonst.} \end{cases}$$

Ein Ereignis $A \subseteq \mathbb{R}$ ist hier die Menge jener Ausgänge des Münzwurfs, die wir als "unser Ereignis" ansehen. Zum Beispiel ist

$$P(\{0\}) = P(\{1\}) = \tfrac{1}{2}, \quad P(\{2\}) = 0, \quad P(\{0,1\}) = 1, \quad P(\{1,2,3,4,5\}) = \tfrac{1}{2}, \quad P([0.3,7]) = \tfrac{1}{2}.$$

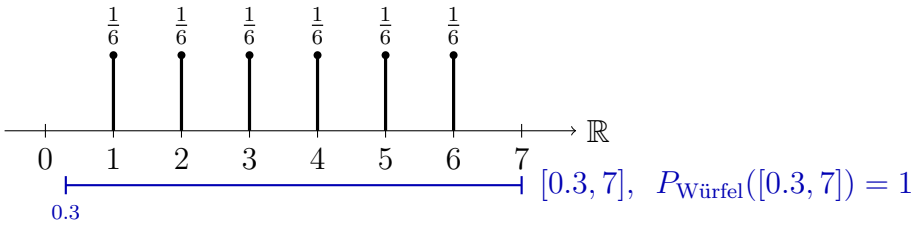
Man stellt sich p am besten als Stabdiagramm vor: Über jedem Punkt $x \in \mathbb{R}$ steht ein Stab der Höhe $p(x)$; fast überall ist die Höhe 0.



Beispiel 3.2 (Würfel)**.** Analog beschreiben wir einen Würfel durch $(\mathbb{R}, \mathcal{P}(\mathbb{R}), P_{\text{Würfel}})$ mit $P_{\text{Würfel}}(A) = \sum_{x \in A} p(x)$ und

$$p(x) = \begin{cases} \frac{1}{6} & x \in \{1, 2, \ldots, 6\}, \\ 0 & \text{sonst.} \end{cases}$$

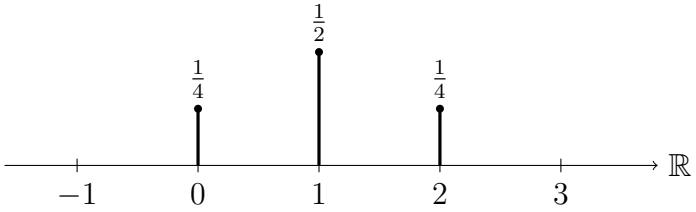
Dann ist etwa $P_{\text{Würfel}}(\{1\}) = \frac{1}{6}$, $P_{\text{Würfel}}(\{1,2,3\}) = \frac{1}{2}$ und $P_{\text{Würfel}}([0.3,7]) = 1$.



Beispiel 3.3 (Binomialexperiment)**.** Eine Münze, die mit Wahrscheinlichkeit $q \in [0,1]$ eine 1 und mit Wahrscheinlichkeit $1-q$ eine 0 zeigt, wird n-mal unabhängig geworfen. Wie in Beispiel 2.17 ist die Wahrscheinlichkeit, insgesamt k Einsen (und $n-k$ Nullen) zu sehen, gleich $\binom{n}{k} q^k (1-q)^{n-k}$. Die Anzahl der Einsen beschreiben wir daher durch ein Wahrscheinlichkeitsmaß P auf $\mathbb{R}$:

- $n = 2$, $q = \frac{1}{2}$: $\quad P(\{0\}) = \frac{1}{4}$, $P(\{1\}) = \frac{1}{2}$, $P(\{2\}) = \frac{1}{4}$;
- $n = 2$, q beliebig: $\quad P(\{0\}) = (1-q)^2$, $P(\{1\}) = 2q(1-q)$, $P(\{2\}) = q^2$;
- n beliebig: $\quad P(\{0\}) = (1-q)^n$, $P(\{1\}) = nq(1-q)^{n-1}$, $\ldots$, $P(\{k\}) = \binom{n}{k} q^k (1-q)^{n-k}$, $\ldots$, $P(\{n\}) = q^n$,

und wieder $P(A) = \sum_{x \in A} p(x)$ mit $p(k) = P(\{k\})$ für $k \in \{0, \ldots, n\}$ und $p(x) = 0$ sonst. Für $n = 2$, $q = \frac{1}{2}$ sieht das Stabdiagramm so aus:



In allen drei Beispielen legt eine Funktion $p : \mathbb{R} \to [0,1]$ fest, wie die Wahrscheinlichkeit auf $\mathbb{R}$ verteilt ist. Für $p : \mathbb{R} \to [0,\infty)$ und $A \subseteq \mathbb{R}$ verstehen wir dabei unter $\sum_{x \in A} p(x)$ das Supremum aller endlichen Teilsummen $\sum_{x \in E} p(x)$, $E \subseteq A$ endlich. (Ist A höchstens abzählbar, ist das die übliche Reihe; da alle Summanden ≥ 0 sind, spielt die Reihenfolge keine Rolle.)

Definition 3.4. *Eine Funktion $p : \mathbb{R} \to [0,1]$ heißt* Wahrscheinlichkeitsfunktion *(WF)*[2] *(oder* Zähldichte*), falls*

$$\sum_{x \in \mathbb{R}} p(x) = 1.$$

Wahrscheinlichkeitsfunktionen sind bei uns *immer auf ganz* $\mathbb{R}$ definiert. Wenn wir nur einige Werte angeben, ist stets $p(x) = 0$ für alle anderen $x \in \mathbb{R}$ gemeint.

Bemerkung 3.5. Ist p eine WF, so ist $\{x \in \mathbb{R} : p(x) > 0\}$ höchstens abzählbar: Die Menge $\{x : p(x) > \frac{1}{n}\}$ hat weniger als n Elemente (sonst wäre die Summe größer als 1), und $\{x : p(x) > 0\} = \bigcup_{n \geq 1} \{x : p(x) > \frac{1}{n}\}$ ist eine abzählbare Vereinigung endlicher Mengen.

Proposition 3.6. *(a) Ist p eine Wahrscheinlichkeitsfunktion, so ist*

$$P(A) := \sum_{x \in A} p(x), \qquad A \subseteq \mathbb{R},$$

ein Wahrscheinlichkeitsmaß auf $(\mathbb{R}, \mathcal{P}(\mathbb{R}))$.

[2] Im Gegensatz zum englischen Pendant "probability mass function" wird der Begriff "Wahrscheinlichkeitsfunktion" in der deutschen Literatur leider nicht ganz einheitlich verwendet.

(b) Ist umgekehrt P ein Wahrscheinlichkeitsmaß auf $(\mathbb{R}, \mathcal{P}(\mathbb{R}))$ und gibt es eine höchstens abzählbare Menge $\mathcal{W} \subseteq \mathbb{R}$ mit $P(\mathcal{W}) = 1$, so gibt es genau eine Wahrscheinlichkeitsfunktion p mit $P(A) = \sum_{x \in A} p(x)$ für alle $A \subseteq \mathbb{R}$, nämlich

$$p(y) = P(\{y\}), \qquad y \in \mathbb{R}.$$

Beweis. (a) Offenbar ist $P(A) \geq 0$ und $P(\mathbb{R}) = 1$. Sind $A_1, A_2, \ldots$ paarweise disjunkt, so gilt $\sum_{x \in \bigcup_n A_n} p(x) = \sum_n \sum_{x \in A_n} p(x)$, weil man Reihen mit nichtnegativen Gliedern beliebig umordnen und zusammenfassen darf.

(b) Gilt $P(A) = \sum_{x \in A} p(x)$ für alle A, so folgt mit $A = \{y\}$, dass $p(y) = P(\{y\})$; also gibt es höchstens ein solches p. Setzen wir umgekehrt $p(y) := P(\{y\})$, so ist $p(y) \leq P(\mathcal{W}^c) = 0$ für $y \notin \mathcal{W}$. Für $A \subseteq \mathbb{R}$ ist $P(A \setminus \mathcal{W}) \leq P(\mathcal{W}^c) = 0$, und $A \cap \mathcal{W}$ ist höchstens abzählbar, also nach der σ-Additivität

$$P(A) = P(A \cap \mathcal{W}) = \sum_{y \in A \cap \mathcal{W}} P(\{y\}) = \sum_{y \in A} p(y).$$

Mit $A = \mathbb{R}$ sieht man, dass p eine WF ist. □

Ein Wahrscheinlichkeitsmaß P auf $(\mathbb{R}, \mathcal{P}(\mathbb{R}))$ wie in (b) nennen wir eine *diskrete Verteilung* (auf $\mathbb{R}$). Nach Proposition 3.6 entsprechen diskrete Verteilungen genau den Wahrscheinlichkeitsfunktionen; wir sprechen daher auch von der *Verteilung mit WF p*.

Wichtige diskrete Verteilungen. Da der Buchstabe p für die WF reserviert ist, bezeichnen wir Erfolgswahrscheinlichkeiten in dieser Liste mit $q \in [0, 1]$. (Später schreiben wir, wie in der Literatur üblich, oft auch $B(n, p)$ usw., wenn keine Verwechslung droht.) Angegeben sind jeweils die Werte $p(x) > 0$; sonst ist $p(x) = 0$.

- **Gleichverteilung** auf einer endlichen Menge $M \subseteq \mathbb{R}$: $p(x) = \frac{1}{|M|}$ für $x \in M$. Beispiel 3.2 ist die Gleichverteilung auf $\{1, \ldots, 6\}$.
- **Bernoulliverteilung** $B(q) = B(1, q)$: $p(1) = q$, $p(0) = 1 - q$. Beispiel 3.1 ist $B(\frac{1}{2})$.
- **Binomialverteilung** $B(n, q)$: $p(k) = \binom{n}{k} q^k (1-q)^{n-k}$ für $k \in \{0, 1, \ldots, n\}$, siehe Beispiel 3.3. Nach dem binomischen Lehrsatz $(a+b)^n = \sum_{k=0}^n \binom{n}{k} a^k b^{n-k}$ ist $\sum_{k=0}^n p(k) = (q + 1 - q)^n = 1$.
- **Geometrische Verteilung** $G(q)$ mit Parameter $q \in (0, 1]$: $p(k) = (1-q)^{k-1} q$ für $k \in \{1, 2, \ldots\}$. Sie beschreibt, wann bei unabhängigen Münzwürfen zum ersten Mal eine 1 kommt (siehe Beispiel 3.21 unten). Wegen der geometrischen Reihe ist $\sum_{k \geq 1} (1-q)^{k-1} q = 1$.

Im nächsten Abschnitt sehen wir, wie solche Verteilungen ganz natürlich entstehen: als Verteilungen von *Zufallsvariablen*.

3.2 Zufallsvariablen und ihre Verteilungen

Oft interessieren wir uns nicht nur für ein Experiment und seinen Ausgang, sondern auch für numerische Größen, die diesem Experiment zugeordnet sind. Zum Beispiel interessiert sich ein Spieler nicht nur für den Ausgang eines Spiels, sondern auch dafür, wie viel er gewinnt oder verliert. Ein Netzbetreiber interessiert sich dafür, wie lang ein typischer Anruf dauert, und wie viele Leitungen ungefähr zu einem bestimmten Zeitpunkt belegt sind.

Aus diesem Grund ist das Konzept der *Zufallsvariable* sehr wichtig und hilfreich. Eine Zufallsvariable ist einfach eine numerische Größe, die einem zufälligen Experiment zugeordnet ist. D.h., gegeben ein Ausgang $\omega \in \Omega$, können wir ihm eine bestimmte Größe $X(\omega) \in \mathbb{R}$ zuordnen. Formal ist eine Zufallsvariable X nur eine Funktion von Ω nach $\mathbb{R}$.

Definition 3.7. *Sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. Eine Zufallsvariable X ist eine Funktion $X : \Omega \to \mathbb{R}$, sodass für alle $a, b \in \mathbb{R}$ gilt: $\{\omega \in \Omega : X(\omega) \in (a, b]\} \in \mathcal{F}$.*

Vorstellung: $\omega \in \Omega$ beschreibt den Zustand, den "die Welt" (zufällig) einnimmt, und $X(\omega)$ ist der Ausgang eines Experiments, wenn die Welt im Zustand ω ist.

Manchmal ist es auch praktisch, Zufallsvariablen zu betrachten, die Werte in einem anderen Raum annehmen, z.B. in der Menge $\{\text{Kopf}, \text{Zahl}\}$, wenn wir über Münzwürfe reden.

Notation. Wir kürzen den Begriff Zufallsvariable mit ZV ab. Das Ereignis $\{\omega \in \Omega : X(\omega) \in (a, b]\}$ wird IMMER wie folgt abgekürzt:

$$\{\omega \in \Omega : X(\omega) \in (a, b]\} = \{X \in (a, b]\}.$$

Das heißt, wir lassen ω aus der Beschreibung des Ereignisses weg. Allgemeiner schreiben wir für $C \subseteq \mathbb{R}$

$$\{X \in C\} = \{\omega \in \Omega : X(\omega) \in C\} = X^{-1}(C)$$

für das *Urbild* von C unter X, und $\{X = x\} = X^{-1}(\{x\})$. (Das Urbild ist auch dann definiert, wenn X nicht umkehrbar ist.)

Beispiel 3.8. Sei $\Omega = \{1, \dots, 6\}^2$ mit $\mathcal{F} = \mathcal{P}(\Omega)$ und $\mathbb{P}(A) = \frac{|A|}{36}$, und sei

$$X(\omega_1, \omega_2) := \omega_1, \qquad Y(\omega_1, \omega_2) := \omega_2.$$

(ω_1, ω_2 sind die Koordinaten von $\omega = (\omega_1, \omega_2)$.) Wir behaupten, dass sich X wie ein Würfel verhält, d.h. etwa den Wert 3 mit Wahrscheinlichkeit $\frac{1}{6}$ annimmt. Dass X den Wert 3 annimmt, entspricht dem Ereignis

$$A = \{X = 3\} = X^{-1}(\{3\}) = \{(\omega_1, \omega_2) \in \Omega : \omega_1 = 3\} = \{(3, 1), (3, 2), \dots, (3, 6)\}.$$

Also ist $\mathbb{P}(X = 3) = \frac{|A|}{36} = \frac{6}{36} = \frac{1}{6}$, und genauso $\mathbb{P}(X = k) = \frac{1}{6}$ für $k = 1, \dots, 6$. Ebenso ist $\{Y = 2\} = \{(\omega_1, \omega_2) : \omega_2 = 2\}$ und $\mathbb{P}(Y = 2) = \frac{1}{6}$; auch Y ist ein Würfel. Auf demselben Wahrscheinlichkeitsraum leben also zwei Würfel, und zum Beispiel ist

$$\mathbb{P}(X = 3,\ Y = 2) = \mathbb{P}(\{X = 3\} \cap \{Y = 2\}) = \mathbb{P}(\{(3, 2)\}) = \tfrac{1}{36}.$$

Bemerkung 3.9. In diesem Kurs achten wir nicht auf die Bedingung $\{X \in (a,b]\} \in \mathcal{F}$, die aus der Maßtheorie stammt. Diese und ähnliche Bedingungen werden im Folgenden ignoriert.

Konvention. Wir verwenden Großbuchstaben wie X, Y, Z, um Zufallsvariablen darzustellen. Die Kleinbuchstaben wie x, y, z werden verwendet, um mögliche Werte dieser Zufallsvariablen darzustellen. Diese Konvention ist sehr hilfreich.

Definition 3.10. *Sei X eine Zufallsvariable auf $(\Omega, \mathcal{F}, \mathbb{P})$. Die Abbildung*

$$P_X(S) := \mathbb{P}(X \in S) = \mathbb{P}(X^{-1}(S)), \qquad S \subseteq \mathbb{R}, \tag{3.1}$$

heißt ***Verteilung*** *von X.*[3]

Proposition 3.11. *Die Verteilung P_X einer Zufallsvariable ist ein Wahrscheinlichkeitsmaß auf $(\mathbb{R}, \mathcal{P}(\mathbb{R}))$.*

Beweis. Offenbar ist $P_X(S) \geq 0$, und $P_X(\mathbb{R}) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in \mathbb{R}\}) = \mathbb{P}(\Omega) = 1$, denn $X^{-1}(\mathbb{R})$ ist ganz Ω. Die σ-Additivität folgt daraus, dass Urbilder mit Mengenoperationen verträglich sind, etwa $\{X \in S_1\} \cap \{X \in S_2\} = \{X \in S_1 \cap S_2\}$ und $\{X \in \bigcup_n S_n\} = \bigcup_n \{X \in S_n\}$ (Übung). □

Eine Zufallsvariable *transformiert* also das Wahrscheinlichkeitsmaß $\mathbb{P}$ auf Ω in das Wahrscheinlichkeitsmaß P_X auf $\mathbb{R}$: Wir erhalten einen Wahrscheinlichkeitsraum $(\mathbb{R}, \mathcal{P}(\mathbb{R}), P_X)$ von genau der Art, wie wir sie in Abschnitt 3.1 betrachtet haben.

Ist P ein Wahrscheinlichkeitsmaß auf $(\mathbb{R}, \mathcal{P}(\mathbb{R}))$, so sagen wir, X *hat die Verteilung* P, i.Z. $X \sim P$, falls $P_X = P$. Zum Beispiel bedeutet $X \sim B(n,p)$, dass X binomialverteilt ist. Zwei Zufallsvariablen X, Y (die auch auf verschiedenen Wahrscheinlichkeitsräumen definiert sein dürfen) heißen *gleich in Verteilung* oder *identisch verteilt*, falls $P_X = P_Y$, i.Z. $X \sim Y$.

Beispiel 3.12. Sei X das Resultat eines Würfels. Dann ist die Verteilung von X

$$P_X(S) = \mathbb{P}(X \in S) = \frac{|S \cap \{1, \ldots, 6\}|}{6},$$

also genau die Gleichverteilung auf $\{1, \ldots, 6\}$ aus Beispiel 3.2. Wir sagen, dass X die **uniforme Verteilung** auf $\{1, \ldots, 6\}$ hat oder dass X auf $\{1, \ldots, 6\}$ *gleichverteilt* ist.

Bemerkung 3.13. Die Verteilung einer Zufallsvariable X beschreibt keine bestimmte Erkenntnis über den Ausgang von X, sondern ihre **statistischen Eigenschaften**, z.B.: Ist es wahrscheinlicher, dass X positiv oder dass X negativ ist? Genauer gesagt beschreibt die Verteilung von X *alles*, was statistisch mit X passieren kann.

Formal haben wir Zufallsvariablen als Funktionen definiert. Das geschieht allerdings in erster Linie, weil es technisch praktisch ist; für unser intuitives Verständnis ist es nicht so wichtig. Das kann man insbesondere an der Art von Fragen festmachen, die wir in der Wahrscheinlichkeitstheorie stellen. Beispielsweise fragen wir uns, ob eine Zufallsvariable mit positiver Wahrscheinlichkeit positiv ist. Wir fragen uns *nicht*, ob eine Zufallsvariable differenzierbar ist. Grob gesagt stellen wir in der Wahrscheinlichkeitstheorie nur Fragen, die sich schon beantworten lassen, wenn man die Verteilung einer Zufallsvariablen kennt.

[3] Genau genommen nur für Mengen S mit $\{X \in S\} \in \mathcal{F}$; wie oben ignorieren wir diese Bedingung.

3.3 Diskrete Zufallsvariablen; wichtige Beispiele

Sei X eine Zufallsvariable (implizit definiert auf einem bestimmten Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \mathbb{P})$).

Definition 3.14. *Wir sagen, dass X* ***diskret*** *ist, falls eine höchstens abzählbare Menge $\mathcal{W} \subseteq \mathbb{R}$ existiert, sodass $\mathbb{P}(X \in \mathcal{W}) = 1$. Wir sagen dann, dass X* fast sicher *Werte in $\mathcal{W}$ annimmt. Die Funktion*

$$p_X(x) := p(x) := \mathbb{P}(X = x), \quad x \in \mathbb{R},$$

heißt Wahrscheinlichkeitsfunktion (WF) *von X.*

X ist also genau dann diskret, wenn P_X eine diskrete Verteilung im Sinne von Abschnitt 3.1 ist. Wegen $p_X(x) = P_X(\{x\})$ ist p_X dann nach Proposition 3.6 genau die WF von P_X. Insbesondere gilt $p_X(x) \geq 0$, $\sum_{x \in \mathbb{R}} p_X(x) = 1$ und

$$\mathbb{P}(X \in S) = \sum_{x \in S} p_X(x), \qquad S \subseteq \mathbb{R}. \tag{3.2}$$

Die Verteilung einer diskreten Zufallsvariable ist also durch ihre WF vollständig beschrieben. Meist wählt man $\mathcal{W} = \{x : p_X(x) > 0\}$; für $x \notin \mathcal{W}$ ist dann $p_X(x) = 0$.

Bemerkung 3.15. Ist Ω höchstens abzählbar, so ist jede Zufallsvariable $X : \Omega \to \mathbb{R}$ diskret: Die Bildmenge $\mathcal{W} := \{X(\omega) : \omega \in \Omega\}$ ist höchstens abzählbar, und

$$P_X(\mathcal{W}) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in \mathcal{W}\}) = \mathbb{P}(\Omega) = 1.$$

Die WF von X erhält man dann, indem man die Wahrscheinlichkeiten aller ω aufsummiert, die auf denselben Wert abgebildet werden:

$$p_X(y) = P_X(\{y\}) = \mathbb{P}(\{\omega \in \Omega : X(\omega) = y\}) = \sum_{\omega:\, X(\omega) = y} \mathbb{P}(\{\omega\}).$$

Beispiel 3.16. Sei $\Omega = \{\omega_1, \ldots, \omega_6\}$ mit $\mathbb{P}(\{\omega_i\}) = \frac{1}{6}$ und

$$X(\omega_1) = \tfrac{1}{2}, \quad X(\omega_2) = X(\omega_3) = 1, \quad X(\omega_4) = X(\omega_6) = 2, \quad X(\omega_5) = \tfrac{3}{2}.$$

Dann ist zum Beispiel $p_X(1) = \mathbb{P}(X^{-1}(\{1\})) = \mathbb{P}(\{\omega_2, \omega_3\}) = \frac{2}{6}$; ebenso $p_X(\frac{1}{2}) = p_X(\frac{3}{2}) = \frac{1}{6}$, $p_X(2) = \frac{2}{6}$ und $p_X(x) = 0$ sonst. Abbildung 2 zeigt rechts den Graphen von X als Funktion auf Ω und links, auf derselben $\mathbb{R}$-Achse, die WF von X: Die Wahrscheinlichkeit aller ω mit $X(\omega) = x$ wird zum Wert x "transportiert".

Zum Beispiel ist das Resultat X beim Würfeln eine diskrete Zufallsvariable, mit $\mathcal{W} = \{1, \ldots, 6\}$. Ihre WF ist einfach

$$p(x) = \frac{1}{6}, x \in \mathcal{W}, \qquad p(x) = 0 \text{ sonst.}$$

Wir machen daraus eine Definition:

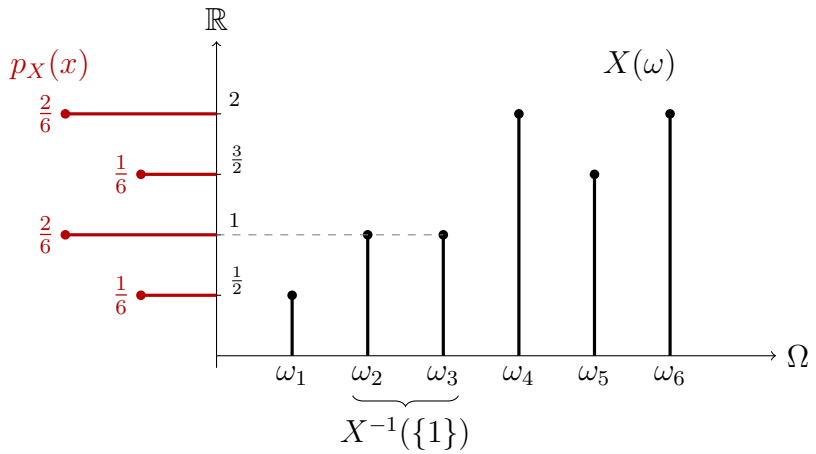


Abbildung 2: Rechts: die Zufallsvariable X als Funktion auf $\Omega = \{\omega_1, \ldots, \omega_6\}$. Links: ihre Wahrscheinlichkeitsfunktion p_X, die die Verteilung P_X auf $\mathbb{R}$ beschreibt. Die Zufallsvariable transformiert $\mathbb{P}$ auf Ω in P_X auf $\mathbb{R}$; so ist etwa $p_X(1) = \mathbb{P}(X^{-1}(\{1\})) = \mathbb{P}(\{\omega_2, \omega_3\}) = \frac{2}{6}$.

Definition 3.17. *Ein* Würfel *ist für uns eine Zufallsvariable X, die auf $\{1, \ldots, 6\}$ gleichverteilt ist, d.h. deren Wahrscheinlichkeitsfunktion Folgendes erfüllt:*

$$p_X(x) = \begin{cases} \frac{1}{6} & x \in \{1, 2, \ldots, 6\} \\ 0 & \textit{sonst} \end{cases}.$$

Eine Münze *ist für uns eine Zufallsvariable Y mit $Y \sim B(\beta)$ für ein $\beta \in [0, 1]$, d.h. deren Wahrscheinlichkeitsfunktion Folgendes erfüllt:*

$$p_Y(y) = \begin{cases} \beta & y = 1 \\ 1 - \beta & y = 0 \\ 0 & \textit{sonst} \end{cases}.$$

Im Fall $\beta = 1/2$ nennen wir Y eine faire Münze.

Wenn wir im Folgenden von einem *Würfel* oder einer *Münze* sprechen, meinen wir damit stets eine Zufallsvariable in diesem Sinn. Ob sie auf $\Omega = \{1, \ldots, 6\}$, auf $\Omega = \{1, \ldots, 6\}^2$ oder auf $\Omega = \mathbb{R}$ definiert ist, spielt für uns keine Rolle; es kommt nur auf die Verteilung an.

Beispiel 3.18. Zwei Würfel werden geworfen, wobei alle Möglichkeiten gleich wahrscheinlich sind. Seien X und Y die beiden Ergebnisse. Was ist die Wahrscheinlichkeitsfunktion p von $X + Y$?

Tatsächlich haben wir sie schon berechnet: das Ereignis $\{X + Y = k\}$ ist genau das

Ereignis S_k aus Beispiel 1.8. Deshalb gilt

$$p(k) = \begin{cases} (k-1)/36, & \text{falls } 2 \le k \le 7 \\ (13-k)/36 & \text{falls } 7 \le k \le 12 \\ 0 & \text{sonst.} \end{cases}$$

Die wichtigen Verteilungen aus Abschnitt 3.1 treten ganz natürlich als Verteilungen von Zufallsvariablen auf. Wie in der Literatur üblich, bezeichnen wir die Erfolgswahrscheinlichkeit ab jetzt wieder mit p; die WF schreiben wir dann als $\mathbb{P}(X = k)$ oder p_X.

Beispiel 3.19. (Fortsetzung von Beispiel 2.16.) Wir nehmen an, dass eine Münze einmal geworfen wird, mit möglichen Ausgängen $\{0, 1\}$. Sei X das Ergebnis, mit

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

Dann hat X die **Bernoulliverteilung**, $X \sim B(p) = B(1, p)$.

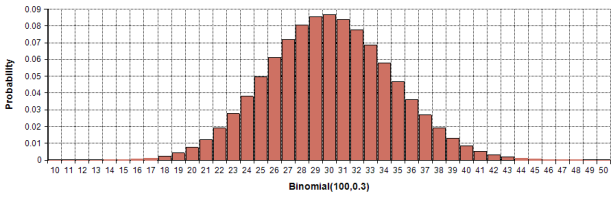


Abbildung 3: Die Binomialverteilung mit $n = 100$, $p = 0.3$.

Beispiel 3.20. (Fortsetzung von Beispiel 2.17). Wenn wir die obige Münze n-mal unabhängig werfen, erhalten wir eine Folge von *Bernoulliexperimenten*. Die Anzahl X der Köpfe, die man dabei erhält, ist eine Zahl in der Menge $\mathcal{W} = \{0, 1, \ldots, n\}$. Gegeben $k \in \mathcal{W}$, gibt es $\binom{n}{k}$ Wurffolgen, die zu k Köpfen führen, und jede dieser Folgen passiert mit Wahrscheinlichkeit $p^k(1-p)^{n-k}$. Insgesamt erhalten wir

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k \in \{0, \ldots, n\},$$

d.h. X hat die **Binomialverteilung**, $X \sim B(n, p)$ (vgl. Beispiel 3.3).

Beispiel 3.21. Wir betrachten eine unendliche Folge von unabhängigen Münzwürfen (Bernoulliexperimenten) und fragen uns, wie oft wir die Münze werfen müssen, bis wir das erste Mal Kopf bekommen; diese Anzahl nennen wir Z. Dann ist Z eine diskrete Zufallsvariable mit Werten in $\mathcal{W} = \{1, 2, \ldots\}$ (beachten Sie, dass der kleinste mögliche Wert von Z nach Definition $z = 1$ ist; d.h., der erste Wurf hat Nummer 1). Ihre Wahrscheinlichkeitsfunktion ist offenbar

$$\mathbb{P}(Z = k) = (1-p)^{k-1} p; \quad k \ge 1.$$

D.h. Z hat die **geometrische Verteilung** mit Parameter p, i.Z. $Z \sim G(p)$. Dass $\sum_{k=1}^{\infty} \mathbb{P}(Z = k) = 1$ gilt, bedeutet: Mit Wahrscheinlichkeit 1 kommt irgendwann ein Kopf.

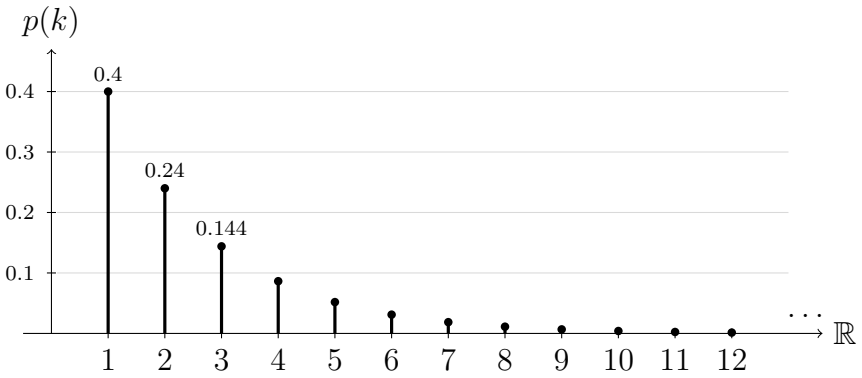


Abbildung 4: Die geometrische Verteilung mit $p = 0.4$: $\mathbb{P}(Z = k) = 0.6^{k-1} \cdot 0.4$, $k \geq 1$.

Umgekehrt ist jede diskrete Verteilung die Verteilung einer Zufallsvariable:

Lemma 3.22. *Sei P eine diskrete Verteilung auf $\mathbb{R}$. Dann gibt es eine Zufallsvariable X auf einem höchstens abzählbaren Wahrscheinlichkeitsraum mit $X \sim P$. Insbesondere gibt es zu jeder diskreten Zufallsvariable eine Zufallsvariable auf einem höchstens abzählbaren Wahrscheinlichkeitsraum mit derselben Verteilung.*

Beweis. Das könnte eine einfache Übungsaufgabe sein. □

Gemeinsame Verteilung. Angenommen, diskrete Zufallsvariablen X_1, X_2 sind auf demselben Wahrscheinlichkeitsraum definiert. Seien $\mathcal{W}_1, \mathcal{W}_2$ höchstens abzählbare Mengen mit $\mathbb{P}(X_i \in \mathcal{W}_i) = 1$. Dann können wir auch die Abbildung

$$X(\omega) := (X_1(\omega), X_2(\omega))$$

als Zufallsvariable (mit Werten in $\mathbb{R}^2$) interpretieren. Dann gilt

$$\mathbb{P}((X_1, X_2) \in \mathcal{W}_1 \times \mathcal{W}_2) = \mathbb{P}(X_1 \in \mathcal{W}_1, X_2 \in \mathcal{W}_2) = 1,$$

d.h. (X_1, X_2) liegt mit Wahrscheinlichkeit 1 in der abzählbaren Menge $\mathcal{W}_1 \times \mathcal{W}_2$ und ist daher auch eine diskrete Zufallsvariable. Ihre Verteilung

$$P_{(X_1,X_2)}(S) := \mathbb{P}((X_1, X_2) \in S), \qquad S \subseteq \mathbb{R}^2,$$

ist ein Wahrscheinlichkeitsmaß auf $(\mathbb{R}^2, \mathcal{P}(\mathbb{R}^2))$, die *gemeinsame Verteilung* von X_1 und X_2. Wie in Abschnitt 3.1 ist sie durch die *gemeinsame WF* $p(x_1, x_2) = \mathbb{P}(X_1 = x_1, X_2 = x_2)$ festgelegt. Wir sagen, dass Paare von Zufallsvariablen (X_1, X_2) und (Y_1, Y_2) dieselbe gemeinsame Verteilung haben, wenn $P_{(X_1,X_2)} = P_{(Y_1,Y_2)}$.

3.4 Die Poissonapproximation der Binomialverteilung

Beispiel 3.23. Die *Poissonverteilung* wird oft verwendet, um zu modellieren, wie viele Ereignisse innerhalb einer festen Zeitspanne stattfinden werden, zum Beispiel die Anzahl der Glühbirnen, die innerhalb eines Jahres ausfallen werden. Die Poissonverteilung $P(\lambda)$, $\lambda > 0$, ist die diskrete Verteilung auf $\mathbb{R}$ mit WF

$$p(k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, \dots$$

(und $p(x) = 0$ sonst). Heuristisch gibt der Parameter λ an, wie viele Ereignisse 'durchschnittlich' passieren werden; wir werden das später noch formalisieren.

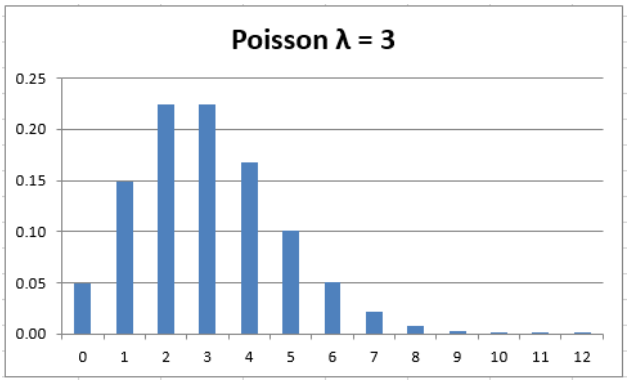


Abbildung 5: Die Poissonverteilung mit $\lambda = 3$.

Satz 3.24 (Poissonapproximation der Binomialverteilung)**.** *Sei $\lambda > 0$. Angenommen, $n \to \infty$ und $p = \lambda/n$ (eigentlich brauchen wir für p nur, dass $pn \to \lambda$). Dann gilt*

$$\mathbb{P}(X_n = k) \to \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, \dots,$$

wobei X_n die Binomialverteilung $B(n, p = \lambda/n)$ hat. Das heißt, für großes n und kleines p verhält sich $B(n, p)$ ungefähr wie $P(pn)$.

Beweis. Wir erinnern uns, dass $(1 - \frac{a}{n})^n \approx e^{-a}$. Dann gilt

$$\begin{aligned} \mathbb{P}(X_n = k) &= \binom{n}{k} p^k (1-p)^{n-k} \qquad (3.3) \\ &= \frac{1}{k!} \frac{n(n-1)\dots(n-k+1)}{n^k} (np)^k \left(1 - \frac{np}{n}\right)^{n-k} \to \frac{1}{k!} \lambda^k e^{-\lambda}. \end{aligned}$$

□

3.5 Erwartungswert

Sei X eine diskrete Zufallsvariable. Wie groß ist X *durchschnittlich*? Ist sie eher klein oder groß? Das ist eine statistische Frage über X, und natürlich können wir sie präzise mit Hilfe der Verteilung beantworten. Aber oft brauchen wir nur eine grobe Zusammenfassung oder eine Größenordnung. In diesem Fall ist der Erwartungswert von X sehr hilfreich.

Definition 3.25. *Der Erwartungswert einer diskreten Zufallsvariable X mit Wahrscheinlichkeitsfunktion $p(x)$ ist gleich*

$$\mathbb{E}[X] = \sum_{x:p(x)>0} xp(x) = \sum_x x\mathbb{P}(X = x)$$

falls $\sum_{x:p(x)>0} |x|p(x) < \infty$, andernfalls sagen wir, dass der Erwartungswert nicht existiert.

Offenbar hängt der Erwartungswert einer Zufallsvariable (und auch dessen Existenz) nur von der Verteilung der Zufallsvariable ab. Entsprechend nennen wir für eine diskrete Verteilung P auf $\mathbb{R}$ mit WF p die Zahl $\sum_x x\, p(x)$ (falls $\sum_x |x|\, p(x) < \infty$) den *Erwartungswert von P*. Der Erwartungswert von X ist dann gerade der Erwartungswert ihrer Verteilung P_X, und wir können etwa vom Erwartungswert der Binomialverteilung sprechen.

Beispiel 3.26. Sei X mit Wahrscheinlichkeit p gleich 1, sonst 0, d.h. $X \sim B(1, p)$. Dann ist $\mathbb{E}[X] = 1 \cdot p + 0 \cdot (1 - p) = p$.

Beispiel 3.27. Angenommen X ist ein Würfel. Dann gilt $\mathbb{E}[X] = \frac{1}{6}(1 + 2 + \ldots + 6) = 3.5$.

Beispiel 3.28. Sei X binomialverteilt, i.Z. $X \sim B(n, p)$. Dann gilt $\mathbb{E}[X] = np$.

Um das zu sehen, kann man direkt in die Definition des Erwartungswertes einsetzen, was

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

liefert. Ausgehend davon kommt man nach kurzer Rechnung zum Ergebnis. Wir werden unten aber gleich sehen, dass wir mit einem kleinen Trick das Ergebnis auch ohne Rechnung erhalten können.

Lemma 3.29. *Angenommen Ω ist höchstens abzählbar und der Erwartungswert von X existiert. Dann gilt*

$$\mathbb{E}[X] = \sum_{\omega\in\Omega} X(\omega)p_\omega,$$

wobei $p_\omega = \mathbb{P}(\{\omega\})$.

Beweis.

$$\mathbb{E}[X] = \sum_x xp(x) = \sum_x x\mathbb{P}(\{\omega : X(\omega) = x\}) = \sum_x \sum_{\omega\in\Omega:X(\omega)=x} x\mathbb{P}(\{\omega\}) = \tag{3.4}$$

$$\sum_x \sum_{\omega\in\Omega:X(\omega)=x} X(\omega)\mathbb{P}(\{\omega\}) = \sum_{\omega\in\Omega} X(\omega)\mathbb{P}(\{\omega\}). \tag{3.5}$$

□

Der folgende Satz sagt uns insbesondere, dass $\mathbb{E}$ ein *linearer Operator* ist.

Satz 3.30 (Eigenschaften des Erwartungswertes)**.** *Seien X, Y diskrete Zufallsvariablen, deren Erwartungswerte existieren.*

1. *Falls $X \geq 0$, dann gilt $\mathbb{E}[X] \geq 0$.*
2. *Falls $X \geq 0$ und $\mathbb{E}[X] = 0$, dann gilt $\mathbb{P}(X = 0) = 1$.*
3. *Für reelle Zahlen a, b gilt $\mathbb{E}[a + bX] = a + b\mathbb{E}[X]$.*
4. *Es gilt $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.*

Beweis. Wir dürfen annehmen, dass X, Y auf einem höchstens abzählbaren Wahrscheinlichkeitsraum definiert sind.

1. $X \geq 0$ impliziert, dass $\mathbb{E}[X] = \sum_{\omega\in\Omega} p_\omega X(\omega) \geq 0$.
2. Indirekt: Wenn es ein ω mit $p_\omega > 0$ und $X(\omega) > 0$ gibt, dann gilt auch $\mathbb{E}[X] = \sum_{\omega\in\Omega} p_\omega X(\omega) > 0$.
3. $\mathbb{E}[a + bX] = \sum_{\omega\in\Omega}(a + bX(\omega))p_\omega = a + b\sum_{\omega\in\Omega} X(\omega)p_\omega = a + b\mathbb{E}[X]$.
4. $\sum_{\omega\in\Omega} p_\omega(X(\omega) + Y(\omega)) = \sum_{\omega\in\Omega} p_\omega X(\omega) + \sum_{\omega\in\Omega} p_\omega Y(\omega)$.

□

Beispiel 3.31. Wir machen wieder das Binomialverteilungsexperiment, d.h., n Münzen werden unabhängig geworfen, wobei jede mit Wahrscheinlichkeit p Kopf zeigt und X bezeichne die Anzahl der Köpfe, die man insgesamt sieht, d.h. $X \sim B(n, p)$. Da $\mathbb{E}[X]$ nur von der Verteilung abhängt, berechnen wir damit den Erwartungswert der Binomialverteilung. Sei $X_i = 1$, falls die i-te Münze Kopf zeigt, und 0 sonst, also $X_i \sim B(1, p)$. Dann ist $X = X_1 + \ldots + X_n$ und somit gilt

$$\mathbb{E}[X] = \mathbb{E}[X_1] + \ldots + \mathbb{E}[X_n] = p + \ldots + p = np.$$

Beispiel 3.32. Wie oft müssen wir durchschnittlich eine Münze werfen, bis wir das erste Mal Kopf bekommen? Sei $X \sim G(p)$ (d.h., X ist eine Zufallsvariable mit geometrischer $G(p)$ Verteilung). Gesucht ist der Erwartungswert von X. Wir berechnen:

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{k\geq 1} k(1-p)^{k-1}p \\
&= p\sum_{k\geq 1} \frac{d}{dx}x^k|_{x=1-p} \\
&= p\frac{d}{dx}\Big(\sum_{k\geq 0} x^k\Big)\Big|_{x=1-p} \\
&= p\frac{d}{dx}\Big(\frac{1}{1-x}\Big)\Big|_{x=1-p} \\
&= 1/p.
\end{aligned}$$

Das ist konsistent mit unserer Intuition. Betrachten wir zum Beispiel den Fall $p = 1/100$. Dann werden wir den ersten Erfolg nach durchschnittlich $1/(1/100) = 100$ Würfen haben.

Der folgende Satz formalisiert, dass der Erwartungswert (in einem bestimmten Sinn) die beste Approximation einer Zufallsvariable durch eine Konstante ist.

Satz 3.33. *Sei X eine Zufallsvariable, sodass die Erwartungswerte von X und X^2 existieren. Dann ist $\mathbb{E}[X]$ die Konstante c, die*

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

minimiert.

Beweis. Wir schreiben zur Vereinfachung $m = \mathbb{E}(X)$.

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

Der letzte Ausdruck wird offenbar genau für $c = m$ minimal. □

Der Wert des Minimums wird später als *Varianz* bezeichnet.

Angenommen, X ist eine diskrete Zufallsvariable und $f : \mathbb{R} \to \mathbb{R}$ eine Funktion. Dann ist $f(X)$ ebenso eine Zufallsvariable. Um $\mathbb{E}[f(X)]$ zu berechnen, können wir theoretisch zuerst die Verteilung (oder genauer gesagt die WF, $p_{f(X)}$) von $f(X)$ beschreiben und schließlich den Erwartungswert durch Summierung von $xp_{f(X)}(x)$ berechnen. In der Praxis ist das jedoch meist nicht der geschickteste Weg. Viel einfacher ist es, direkt mit der Verteilung von X zu rechnen. Die folgende Proposition erlaubt uns das.

Proposition 3.34. *Sei X eine diskrete Zufallsvariable mit WF p, $\mathcal{W} = \{x : p(x) > 0\}$, und $f : \mathbb{R} \to \mathbb{R}$. Dann gilt*

$$\mathbb{E}(f(X)) = \sum_{x \in \mathcal{W}} f(x)p(x).$$

Insbesondere hängt auch $\mathbb{E}(f(X))$ nur von der Verteilung von X ab. Genauer gesagt konvergiert die Reihe auf der rechten Seite absolut genau dann, wenn die Reihe, die den Erwartungswert definiert, absolut konvergiert.

Beweis. Das ist eine einfache Folgerung aus Lemma 3.29. □

Beispiel 3.35. Sei X gleichverteilt auf $\{0, \dots, n-1\}$. Berechnen Sie $\mathbb{E}(\sin(2\pi X/n))$. Wir beobachten, dass

$$\begin{aligned}\mathbb{E}(\sin(2\pi X/n)) &= \frac{1}{n}\sum_{k=0}^{n-1}\sin(2k\pi/n)\\ &= \frac{1}{n}\Im\Big(\sum_{k=0}^{n-1} e^{2ik\pi/n}\Big)\\ &= \frac{1}{n}\Im\Big(\frac{1-e^{2i\pi}}{1-e^{2i\pi/n}}\Big) = 0.\end{aligned}$$

Finden Sie einen geometrischen Beweis!

3.6 Länge einer Zufallsvariable

Wie stark schwankt eine Zufallsvariable X um ihren Erwartungswert? Um diese Frage zu beantworten, ist eine geometrische Sichtweise sehr hilfreich: Wir fassen Zufallsvariablen als *Vektoren* auf und messen ihre Länge und ihren Abstand voneinander wie in der linearen Algebra.

Erinnerung: Skalarprodukt. Sei V ein Vektorraum über $\mathbb{R}$. Ein *Skalarprodukt* (oder *inneres Produkt*) ist eine Abbildung $\langle\cdot,\cdot\rangle : V \times V \to \mathbb{R}$ mit folgenden Eigenschaften:

1. $\langle x, y\rangle$ ist linear in x und linear in y (*bilinear*),
2. $\langle x, y\rangle = \langle y, x\rangle$ (*symmetrisch*),
3. $\langle x, x\rangle \geq 0$, und $\langle x, x\rangle = 0$ gilt nur für $x = 0$ (*positiv definit*).

Die zugehörige *Länge* (oder *Norm*) eines Vektors ist $\|x\|_2 := \sqrt{\langle x, x\rangle}$. Zwei Vektoren x, y heißen *orthogonal*, i.Z. $x \perp y$, falls $\langle x, y\rangle = 0$.

Beispiel 3.36. Auf $V = \mathbb{R}^n$ ist

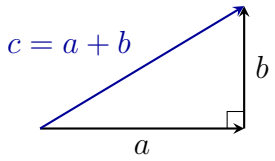
$$\langle x, y\rangle = \sum_{i=1}^{n} x_i y_i \qquad \text{für } x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix},\ y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

das übliche Skalarprodukt; $\|x\|_2$ ist dann die gewöhnliche (euklidische) Länge von x.

Die wichtigste geometrische Tatsache über orthogonale Vektoren gilt in jedem Vektorraum mit Skalarprodukt:

Satz 3.37 (Satz von Pythagoras)**.** *Seien $a, b, c \in V$ mit $a \perp b$ und $a + b = c$. Dann gilt*

$$\|a\|_2^2 + \|b\|_2^2 = \|c\|_2^2.$$



Beweis. Mit der Bilinearität und Symmetrie des Skalarprodukts erhalten wir

$$\|c\|_2^2 = \langle a+b, a+b\rangle = \langle a,a\rangle + 2\underbrace{\langle a,b\rangle}_{=0} + \langle b,b\rangle = \|a\|_2^2 + \|b\|_2^2.$$

□

Vektoren als Funktionen. Man kann einen Vektor $x \in \mathbb{R}^n$ auch als Funktion auffassen: Mit $\Omega = \{1, \dots, n\}$ entspricht x der Funktion

$$X : \Omega \to \mathbb{R}, \qquad X(i) := x_i.$$

Umgekehrt bestimmt jede solche Funktion einen Vektor. Zum Beispiel wird aus $x = (19, 7) \in \mathbb{R}^2$ die Funktion $X : \{1, 2\} \to \mathbb{R}$ mit $X(1) = 19$, $X(2) = 7$, siehe Abbildung 6. Addition und Vielfache von Vektoren werden dabei zur punktweisen Addition und Vervielfachung von Funktionen: $(X + Y)(\omega) = X(\omega) + Y(\omega)$, $(aX)(\omega) = a\,X(\omega)$.

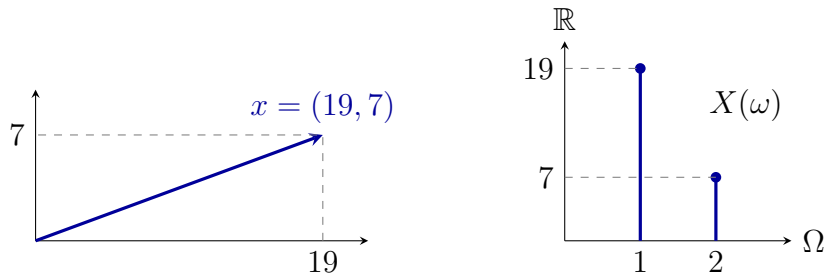


Abbildung 6: Links der Vektor $x = (19, 7) \in \mathbb{R}^2$, rechts dieselbe Information als Funktion $X : \{1, 2\} \to \mathbb{R}$ mit $X(1) = 19$, $X(2) = 7$.

Ganz allgemein bilden die Funktionen von einer Menge Ω nach $\mathbb{R}$ mit diesen Operationen einen Vektorraum über $\mathbb{R}$: Das Nullelement ist die Funktion X mit $X(\omega) = 0$ für alle $\omega \in \Omega$, das inverse Element zu X ist $-X$ mit $(-X)(\omega) = -X(\omega)$, und die Distributivgesetze $a(X + Y) = aX + aY$, $(a + b)X = aX + bX$ gelten, weil sie für jedes einzelne ω gelten.

Der Raum L^2. Zufallsvariablen sind Funktionen $\Omega \to \mathbb{R}$, also können wir sie wie Vektoren behandeln. Für das Skalarprodukt beschränken wir uns auf Zufallsvariablen, für die $\mathbb{E}[X^2]$ existiert. Wie in Lemma 3.29 sei Ω höchstens abzählbar.

Definition 3.38. *Sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum mit höchstens abzählbarem Ω. Wir setzen*

$$L^2(\Omega, \mathcal{F}, \mathbb{P}) := \{X : \Omega \to \mathbb{R} : \ \mathbb{E}[X^2] < \infty\}$$

und definieren für $X, Y \in L^2(\Omega, \mathcal{F}, \mathbb{P})$

$$\langle X, Y \rangle := \sum_{\omega \in \Omega} X(\omega) Y(\omega)\, \mathbb{P}(\{\omega\}) = \mathbb{E}[XY]$$

sowie die Länge *der Zufallsvariable X*

$$\|X\|_2 := \sqrt{\langle X, X \rangle} = \sqrt{\mathbb{E}[X^2]}.$$

Dabei ist XY die Zufallsvariable $(X \cdot Y)(\omega) := X(\omega) \cdot Y(\omega)$. Ist Ω endlich, so liegt jede Zufallsvariable in $L^2(\Omega, \mathcal{F}, \mathbb{P})$.

Proposition 3.39. *$L^2(\Omega, \mathcal{F}, \mathbb{P})$ ist ein Vektorraum, und $\langle \cdot, \cdot \rangle$ ist bilinear und symmetrisch mit $\langle X, X \rangle \geq 0$.*

Beweis. Seien $X, Y \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ und $a \in \mathbb{R}$. Wegen $(x - y)^2 \geq 0$ gilt $2|xy| \leq x^2 + y^2$ und damit $(x + y)^2 \leq 2x^2 + 2y^2$ für alle $x, y \in \mathbb{R}$. Also ist $\mathbb{E}[(X + Y)^2] \leq 2\mathbb{E}[X^2] + 2\mathbb{E}[Y^2] < \infty$, d.h. $X + Y \in L^2(\Omega, \mathcal{F}, \mathbb{P})$; ebenso ist $\mathbb{E}[(aX)^2] = a^2 \mathbb{E}[X^2] < \infty$. Außerdem existiert $\mathbb{E}[XY]$, denn $\mathbb{E}[|XY|] \leq \frac{1}{2}(\mathbb{E}[X^2] + \mathbb{E}[Y^2])$.

Die Bilinearität folgt aus der Linearität des Erwartungswertes: $\langle aX + bY, Z \rangle = \mathbb{E}[aXZ + bYZ] = a\langle X, Z \rangle + b\langle Y, Z \rangle$, und ebenso im zweiten Argument. Symmetrie ist klar, und $\langle X, X \rangle = \mathbb{E}[X^2] \geq 0$. □

Bemerkung 3.40. Aus $\langle X, X \rangle = \mathbb{E}[X^2] = 0$ folgt nur $\mathbb{P}(X = 0) = 1$, nicht $X(\omega) = 0$ für alle ω: Auf einem ω mit $\mathbb{P}(\{\omega\}) = 0$ kann X beliebig sein. Positiv definit ist $\langle \cdot, \cdot \rangle$ also erst, wenn wir Zufallsvariablen, die mit Wahrscheinlichkeit 1 übereinstimmen, als gleich ansehen (oder wenn $\mathbb{P}(\{\omega\}) > 0$ für alle ω gilt). Für unsere Zwecke ist das unerheblich.

Ist $\Omega = \{1, \ldots, n\}$ mit der Gleichverteilung, so ist $\langle X, Y \rangle = \frac{1}{n} \sum_{i=1}^n x_i y_i$, also bis auf den Faktor $\frac{1}{n}$ genau das übliche Skalarprodukt auf $\mathbb{R}^n$.

Der Erwartungswert als nächstgelegene Konstante. Eine Konstante $c \in \mathbb{R}$ können wir als die Zufallsvariable auffassen, die überall den Wert c annimmt. Ihr Abstand zu X ist $\|X - c\|_2$ (siehe Abbildung 7). Wir wollen X durch eine Konstante möglichst gut approximieren.

Proposition 3.41. *Sei $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$. Dann ist $\mathbb{E}[X]$ die konstante Zufallsvariable, die am nächsten an X liegt: Die Aufgabe*

$$\min_{c \in \mathbb{R}} \|X - c\|_2$$

wird gelöst, indem man $c = \mathbb{E}[X]$ wählt.

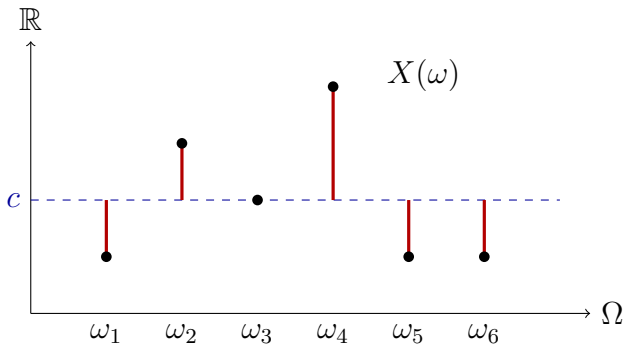


Abbildung 7: Eine Zufallsvariable X auf $\Omega = \{\omega_1, \dots, \omega_6\}$ mit Gleichverteilung und die konstante Zufallsvariable c (gestrichelt). Die roten Strecken sind die Abweichungen $X(\omega) - c$; es ist $\|X - c\|_2^2 = \sum_\omega (X(\omega) - c)^2 \, \mathbb{P}(\{\omega\})$. Hier ist $c = \mathbb{E}[X] = 2$ gewählt, die beste Konstante; dann ist $\|X - c\|_2^2 = \frac{1}{6}(1 + 1 + 0 + 4 + 1 + 1) = \frac{4}{3}$.

Beweis. Wir minimieren nicht $\|X - c\|_2$, sondern $\|X - c\|_2^2 = \mathbb{E}[(X - c)^2]$; da die Wurzel monoton ist, hat das dieselbe Lösung. Zur Vereinfachung setzen wir $m := \mathbb{E}[X]$. Dann gilt

$$\begin{aligned}
\|X - c\|_2^2 = \mathbb{E}[(X - c)^2] &= \mathbb{E}\big[\big((X - m) + (m - c)\big)^2\big] \\
&= \mathbb{E}\big[(X - m)^2 + 2(X - m)(m - c) + (m - c)^2\big] \\
&= \mathbb{E}[(X - m)^2] + 2(m - c)\underbrace{\mathbb{E}[X - m]}_{=\mathbb{E}[X] - m = 0} + (m - c)^2 \\
&= \mathbb{E}[(X - m)^2] + (m - c)^2.
\end{aligned}$$

Der letzte Ausdruck wird für $c = m$ minimal. □

Das ist dieselbe Rechnung wie im Beweis von Satz 3.33, jetzt geometrisch gelesen: Es ist $X - c = (X - m) + (m - c)$, und die beiden Summanden sind orthogonal, denn $\langle X - m, \, m - c \rangle = (m - c)\,\mathbb{E}[X - m] = 0$. Die Gleichung $\|X - c\|_2^2 = \|X - m\|_2^2 + (m - c)^2$ ist also der Satz von Pythagoras 3.37.

Damit zerlegen wir jede Zufallsvariable in zwei Teile:

$$X = \underbrace{\mathbb{E}[X]}_{\text{systematischer Anteil}} + \underbrace{(X - \mathbb{E}[X])}_{\text{unsystematischer Anteil}} .$$

Der systematische Anteil ist die beste deterministische Annäherung an X, der unsystematische Anteil ist der zufällige Rest. Wie groß dieser Rest ist, misst die Varianz.

3.7 Varianz einer Zufallsvariable

Der Erwartungswert einer Zufallsvariable ist eine gute Annäherung (in einem bestimmten Sinn die *beste* Annäherung, siehe Satz 3.33) einer Verteilung durch eine Konstante. Aber wie

gut ist diese Annäherung? Um dies zu messen, führen wir die *Varianz* einer Zufallsvariable ein, die die Ausbreitung der Verteilung misst.

Beispiel 3.42. Betrachten Sie zwei Spiele. Im ersten Spiel werfen wir eine faire Münze. Wenn Kopf kommt, bekomme ich 1\$, sonst verliere ich 1\$.

Im zweiten Spiel werfen wir eine Münze, für die $\mathbb{P}(\text{Kopf}) = 99.9\%$. Wenn man Kopf erhält, dann bekommt man 1\$. Sonst verliert man 999\$.

Welches Spiel möchten Sie spielen?

Tatsächlich sind beide Spiele *fair*: der erwartete Gewinn ist gleich 0 (überprüfen Sie es!). Trotzdem ist das Risiko im zweiten Spiel viel höher!

Obwohl die Erwartungswerte in diesen beiden Spielen gleich sind, sind die Varianzen sehr verschieden. Hier ist die Definition.

Definition 3.43. *Sei X eine Zufallsvariable mit Erwartungswert μ. Die Varianz von X ist*

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

wenn dieser Erwartungswert wohldefiniert ist. Dann ist die Standardabweichung

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

In der Sprache von Abschnitt 3.6 ist

$$\mathrm{Var}(X) = \|X - \mathbb{E}[X]\|_2^2, \qquad \sigma_X = \|X - \mathbb{E}[X]\|_2 :$$

Die Varianz ist die quadrierte Länge des unsystematischen Anteils, die Standardabweichung ist der Abstand von X zur besten Konstante $\mathbb{E}[X]$. Sowohl $\mathrm{Var}(X)$ als auch σ_X messen also, wie "zufällig" X ist. (Für $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ sind beide endlich.)

Nach Proposition 3.34 (mit $f(x) = (x-\mu)^2$) hängt auch die Varianz nur von der Verteilung von X ab; wir sprechen daher auch von der Varianz einer Verteilung.

Im obigen Beispiel gilt $\mu = 0$. Deshalb erhalten wir $\mathrm{Var}(X) = \mathbb{E}(X^2)$. Dies ist gleich 1 im ersten Spiel, und im zweiten Spiel gleich

$$\mathbb{E}(X^2) = p \cdot 1 + (1-p)999^2 \approx 10^3,$$

wobei $p = \mathbb{P}(\text{Kopf}) = 99.9\%$.

Um die Varianz zu berechnen, ist es oft praktisch, den folgenden Satz zu verwenden:

Proposition 3.44. *Sei X eine diskrete Zufallsvariable mit Erwartungswert μ. Dann gilt* $\mathrm{Var}(X) = \mathbb{E}(X^2) - \mu^2$.

Beweis. Das ist eine einfache Folge der Linearität des Erwartungswertes:

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

□

Beispiel 3.45. Sei $X \sim B(p)$ eine bernoulliverteilte Zufallsvariable. Dann gilt $\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$. Da $X \in \{0,1\}$, erhalten wir $X^2 = X$. Weiterhin ist $\mathbb{E}(X) = p$. Deswegen gilt

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

Beachten Sie, dass die größte Varianz (oder Standardabweichung) für $p = 1/2$ erreicht wird.

Beispiel 3.46. Sei $X \sim G(p)$ geometrisch verteilt. Zeigen Sie, dass die Varianz von X gleich $(1-p)/p^2$ ist.

Um $\mathrm{Var}(X)$ zu berechnen, erinnern wir uns, dass (wegen Beispiel 3.32) $\mathbb{E}(X) = 1/p$. Weiterhin finden wir nach Proposition 3.34

$$\mathbb{E}(X^2) = \sum_{k \geq 1} k^2 (1-p)^{k-1} p.$$

Den Rest überlassen wir als Übung. □

Was haben wir mit dieser Berechnung gelernt? Betrachten Sie zum Beispiel den Fall $p \to 0$. Dann haben wir $\mathbb{E}(X) = 1/p \gg 1$ und $\mathrm{Var}(X) \approx 1/p^2$. Deshalb erwarten wir, dass X typischerweise in der Größenordnung von $1/p$ ist, mit Abweichung (oder Fluktuation) auch in dieser Größenordnung. D.h., Werte wie $1/(2p)$ oder $2/p$ sind ganz typisch. Tatsächlich werden wir später sehen, dass die Verteilung von pX (die x-Achse schrumpft um den Faktor p) ungefähr wie eine Exponentialverteilung aussieht.

3.8 Unabhängige Zufallsvariablen

Definition 3.47. *Zufallsvariablen $X_1, \ldots, X_n$ (auf demselben Wahrscheinlichkeitsraum) heißen* unabhängig, *falls für alle $S_1, \ldots, S_n \subseteq \mathbb{R}$ die Ereignisse*

$$\{X_1 \in S_1\}, \ldots, \{X_n \in S_n\}$$

unabhängig im Sinne von Abschnitt 2.6 sind. Eine Folge $X_1, X_2, \ldots$ heißt unabhängig, falls $X_1, \ldots, X_n$ für jedes n unabhängig sind.

Bemerkung 3.48. Für diskrete Zufallsvariablen $X_1, \ldots, X_n$ ist das äquivalent dazu, dass für alle $x_1, \ldots, x_n \in \mathbb{R}$

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \mathbb{P}(X_1 = x_1) \ldots \mathbb{P}(X_n = x_n)$$

gilt. (Für die Rückrichtung summiert man über $x_1 \in S_1, \ldots, x_n \in S_n$.) Die gemeinsame WF unabhängiger Zufallsvariablen ist also das Produkt der einzelnen WFs.

Allgemeiner nennt man Familien von Mengen $\mathcal{F}_1, \mathcal{F}_2, \ldots$ unabhängig, falls $A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2, \ldots$ stets unabhängig sind. Die obige Definition besagt dann, dass die von $X_1, \ldots, X_n$ erzeugten σ-Algebren $\{\{X_i \in S\} : S \subseteq \mathbb{R}\}$ unabhängig sind.

Beispiel 3.49. In Beispiel 3.31 sind $X_1, \ldots, X_n$ unabhängige Zufallsvariablen. Mit $A_i = \{X_i = 1\}$ ist das genau die Unabhängigkeit der Münzwürfe aus Abschnitt 2.6.

Satz 3.50. *Seien $X_1, \ldots, X_n$ unabhängig und $f_1, \ldots, f_n : \mathbb{R} \to \mathbb{R}$ Funktionen. Dann sind auch $f_1(X_1), \ldots, f_n(X_n)$ unabhängige Zufallsvariablen.*

Beweis. Für $S_1, \ldots, S_n \subseteq \mathbb{R}$ ist $\{f_i(X_i) \in S_i\} = \{X_i \in f_i^{-1}(S_i)\}$. Diese Ereignisse sind nach Voraussetzung unabhängig. □

Satz 3.51. *Seien $X_1, \ldots, X_n$ unabhängige Zufallsvariablen, deren Erwartungswerte existieren. Dann gilt*

$$\mathbb{E}[X_1 \cdot \ldots \cdot X_n] = \mathbb{E}[X_1] \cdot \ldots \cdot \mathbb{E}[X_n].$$

Beweis. Wir bringen den Beweis für den Fall von diskreten Zufallsvariablen.

Wir schreiben $\mathcal{W}_i$ für den Wertebereich von X_i. Dann gilt

$$\mathbb{E}[X_1 \cdot \ldots \cdot X_n] = \sum_{x_1 \in \mathcal{W}_1} \ldots \sum_{x_n \in \mathcal{W}_n} x_1 \ldots x_n \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} \left(\sum_{x_i \in \mathcal{W}_i} x_i \mathbb{P}(X_i = x_i) \right) = \mathbb{E}[X_1] \cdot \ldots \cdot \mathbb{E}[X_n].$$

Der Beweis im allgemeinen Fall erfolgt über die Schritte 1) Beweis für Treppenfunktionen, 2) allgemeiner Fall über Grenzwertargument.

□

3.9 Varianz einer Summe

Die Identität in Satz 3.51 hat fundamentale Konsequenzen in der gesamten Wahrscheinlichkeitstheorie. Um sie zu verstehen, greifen wir die geometrische Sichtweise aus Abschnitt 3.6 wieder auf: Zufallsvariablen sind Vektoren mit dem inneren Produkt $\langle X, Y \rangle = \mathbb{E}[XY]$ und der Länge $\|X\|_2 = \sqrt{\mathbb{E}[X^2]}$. (Diese Intuition trägt auch für allgemeine Zufallsvariablen; man braucht dann etwas Maßtheorie, und statt von linearer Algebra spricht man von Funktionalanalysis.) Wir zerlegen X in den systematischen Anteil $\mu = \mathbb{E}[X]$ und den unsystematischen, stochastischen Fehler $X - \mathbb{E}[X]$; die Standardabweichung $\sigma_X = \|X - \mathbb{E}[X]\|_2$ ist die Länge des unsystematischen Anteils.

Die fundamentale Folgerung aus Satz 3.51 ist nun, dass Unabhängigkeit von Zufallsvariablen die Orthogonalität der unsystematischen Anteile impliziert:

Korollar 3.52. *Seien X, Y unabhängig, $\mathbb{E}[X^2], \mathbb{E}[Y^2] < \infty$. Dann gilt*

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

Beweis. $\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y] = 0.$ □

Sobald man auf eine Orthogonalitätsrelation trifft, sollte man an den Satz von Pythagoras denken:

Satz 3.53. *Seien X, Y unabhängige Zufallsvariable. Dann gilt*

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

Beweis. Seien $X' = X - \mathbb{E}[X], Y' = Y - \mathbb{E}[Y]$ die unsystematischen Anteile. Dann erhalten wir

$$\mathbb{E}[(X'+Y')^2] = \mathbb{E}[(X')^2 + 2X'Y' + Y'^2] = \mathbb{E}[(X')^2] + \mathbb{E}[(Y')^2].$$

□

Natürlich gilt ganz analog auch:

Satz 3.54. *Seien $X_1, \ldots, X_n$ unabhängige Zufallsvariablen. Dann gilt*

$$\mathrm{Var}(X_1 + \ldots + X_n) = \mathrm{Var}(X_1) + \ldots + \mathrm{Var}(X_n).$$

Beispiel 3.55. Seien $X_1, \ldots, X_n$ unabhängig mit $X_i \sim B(1,p)$, und $Z = X_1 + \ldots + X_n \sim B(n,p)$. Dann gilt ja (wie schon oben bemerkt) $\mathbb{E}[X_i] = \mathbb{E}[X_i^2] = p$ und damit $\mathrm{Var}[X_i] = p - p^2 = p(1-p)$. Wir folgern nun $\mathrm{Var}[Z] = np(1-p)$.

Korollar 3.56. *Seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte (d.h. $X_1 \sim X_2 \sim \ldots \sim X_n$) Zufallsvariablen mit Erwartungswert μ und Standardabweichung σ. Dann gilt*

$$\mathrm{Var}\left(\frac{X_1 + \ldots + X_n}{n}\right) = \frac{1}{n}\mathrm{Var}(X_1).$$

Das heißt, die Zufallsvariable

$$\bar{X}_n := \frac{X_1 + \ldots + X_n}{n}$$

hat Erwartungswert μ und Standardabweichung $\frac{\sigma}{\sqrt{n}}$.

Beweis. Wir beobachten, dass für eine Zufallsvariable Y und eine reelle Zahl $\alpha \geq 0$ gilt dass

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

Daraus folgt

$$\mathrm{Var}(\bar{X}_n) = \mathrm{Var}\left(\frac{X_1 + \ldots + X_n}{n}\right) = \frac{1}{n^2}(\mathrm{Var}(X_1 + \ldots + X_n)) = \frac{n\,\mathrm{Var}(X_1)}{n^2}.$$

□

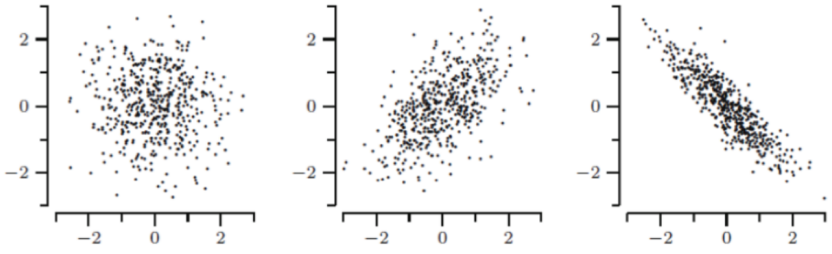


Abbildung 8: Typische Realisierungen von unkorrelierten, positiv korrelierten, und negativ korrelierten Zufallsvariablen.

3.10 Kovarianz und Korrelation

Eine Verallgemeinerung des Satzes von Pythagoras ist der Kosinussatz. Er entspricht dem Sachverhalt, dass der Winkel α zwischen zwei Vektoren $\vec{x}, \vec{y}$ durch

$$\cos(\alpha) = \frac{\langle \vec{x}, \vec{y} \rangle}{|\vec{x}| \cdot |\vec{y}|}$$

gegeben ist.

In der Wahrscheinlichkeitstheorie interessieren wir uns insbesondere für den Winkel zwischen den unsystematischen Anteilen von Zufallsvariablen. Dementsprechend definieren wir

Definition 3.57. *Seien X, Y Zufallsvariablen mit $\mathbb{E}[X^2], \mathbb{E}[Y^2] < \infty$. Die* Kovarianz *von X, Y ist gegeben durch*

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

Falls $\mathrm{Var}[X], \mathrm{Var}[Y] > 0$, ist die Korrelation von X, Y definiert als

$$\mathrm{cor}(X, Y) = \frac{\mathrm{Cov}[X, Y]}{\sqrt{\mathrm{Var}[X]\,\mathrm{Var}[Y]}}.$$

Die Kovarianz steht also für das innere Produkt zwischen den unsystematischen Anteilen von X und Y, die Korrelation von X und Y repräsentiert den Winkel zwischen X und Y.

Sind X, Y unabhängig, so gilt $\mathrm{Cov}(X, Y) = 0$, man sagt X, Y sind *unkorreliert*. Die Umkehrung gilt natürlich nicht. Überzeugen Sie sich selbst davon, indem Sie wenigstens zwei Beispiele von unkorrelierten Zufallsvariabeln finden, die nicht unabhängig sind.

Wenn $\mathrm{Cov}(X, Y) > 0$ sagt man dass X, Y *positiv korreliert* sind. Weil $\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$ wissen wir, dass dies der Fall ist, wenn X und Y eher zusammen größer als ihre Erwartungswerte, und eher zusammen kleiner als ihre Erwartungswerte sind.

Wir zählen ein paar einfache Eigenschaften der Covarianz auf:

1. Für $c \in \mathbb{R}$ gilt
$$\mathrm{Cov}(X, c) = 0, \quad \mathrm{Cov}(X + c, Y) = \mathrm{Cov}(X, Y).$$

2. $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$.

3. $\mathrm{Cov}(X + Y, Z) = \mathrm{Cov}(X, Z) + \mathrm{Cov}(Y, Z)$.

4. $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$.

5. $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + 2\,\mathrm{Cov}(X, Y) + \mathrm{Var}(Y)$.

In der Geometrie ist der Kosinus jedes Winkels betragsmäßig höchstens 1, bzw. ist das innere Produkt von zwei Vektoren (betragsmäßig) höchstens gleich dem Produkt der Längen dieser Vektoren, d.h. $|\langle \vec{x}, \vec{y} \rangle| \leq |\vec{x}| \cdot |\vec{y}|$. Die Entsprechung in unserem Kontext ist die *Cauchy-Schwarz Ungleichung*:

Satz 3.58. *Seien X, Y Zufallsvariablen. Dann gilt*

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

Gleichheit gilt genau dann wenn X, Y kolinear sind.

Beweis. Sei oBdA $\mathbb{E}[Y^2] > 0$. Wir setzen

$$W = X - Y\frac{\mathbb{E}[XY]}{\mathbb{E}[Y^2]}.$$

Dann gilt

$$0 \leq \mathbb{E}[W^2] = \mathbb{E}[X^2] - 2\frac{\mathbb{E}[XY]^2}{\mathbb{E}[Y^2]} + \frac{\mathbb{E}[XY]^2}{\mathbb{E}[Y^2]}.$$

Daraus folgt die Aussage unmittelbar. □

Als Konsequenz erhalten wir unmittelbar

Korollar 3.59. *Seien X, Y Zufallsvariablen. Dann gilt*

$$\begin{aligned} \mathrm{Cov}(X, Y)^2 &\leq \mathrm{Var}(X)\,\mathrm{Var}(Y) \\ |\,\mathrm{cor}(X, Y)| &\leq 1. \end{aligned}$$

Gleichheit gilt jeweils genau dann wenn $X - \mathbb{E}(X)$ und $Y - \mathbb{E}(Y)$ kolinear sind.

Wir schließen diesen Abschnitt mit zwei Bemerkungen.

1. Die Sätze des vorigen Kapitels 3.9 gelten natürlich nicht nur für unabhängige Zufallsvariablen, sondern allgemeiner auch für unkorrelierte Zufallsvariablen.

2. Unabhängigkeit ist in der Wahrscheinlichkeitstheorie ein omnipräsentes Konzept. Dementsprechend oft treten auch unkorrelierte Größen auf.

3.11 Zusammenfassung des Kapitels

Am Ende des Kapitels sollten Sie vertraut sein mit:

- Wahrscheinlichkeitsfunktionen und diskrete Verteilungen auf $\mathbb{R}$;
- Verteilung P_X einer Zufallsvariable als Wahrscheinlichkeitsmaß auf $\mathbb{R}$, Schreibweise $X \sim P$;
- Wahrscheinlichkeitsfunktion (WF) einer diskreten Zufallsvariablen;
- wichtige Verteilungen: Gleich-, Bernoulli-, Binomial-, geometrische Verteilung;
- Erwartungswert einer diskreten Zufallsvariablen (er hängt nur von ihrer Verteilung ab), und wie man ihn über die WF berechnen kann;
- Varianz einer Zufallsvariable;
- Unabhängigkeit von Zufallsvariablen, Additivität der Varianz einer Summe von unabhängige Zufallsvariablen;
- Cauchy–Schwarz Ungleichung.

4 Ungleichungen und Abweichungen

In diesem Kapitel sehen wir einige wichtige Ungleichungen über Erwartungswerte einer Funktion einer Zufallsvariablen. Wir verwenden diese Ungleichungen um:

- Die Abweichungen der Zufallsvariable zu kontrollieren (Markov-, Chebyshev-Ungleichungen);
- den Begriff der Entropie zu formulieren und sich diesem Begriff zu nähern (durch die Jensen'sche Ungleichung)
- um das (schwache) Gesetz der Großen Zahlen zu beweisen.

4.1 Jensen'sche Ungleichung

Wir beginnen mit der Jensen'schen Ungleichung, in der es um konvexe Funktionen geht. Erinnern Sie sich, dass eine Funktion $f : (a, b) \to \mathbb{R}$ konvex ist, wenn für alle $x, y \in (a, b)$ es gilt:

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y); \quad \lambda \in [0,1]. \tag{4.1}$$

Das heißt, der Funktionswert des Durchschnittes von x, y ist kleiner gleich der Durchschnitt der Funktionswerte von x und y. Wenn f von Klasse C^2 über (a, b) ist, dann ist (4.1) äquivalent zu $f'' \geq 0$. Zum Beispiel sind $f(x) = x^2, x^4, e^x$ konvexe Funktionen.

Wir können (4.1) auch als eine Ungleichung zwischen Erwartungswerten interpretieren. Sei X eine Zufallsvariable gleich x mit Wahrscheinlichkeit λ und gleich y mit Wahrscheinlichkeit $1 - \lambda$. Dann ist (4.1) äquivalent zu

$$f(\mathbb{E}(X)) \leq \mathbb{E}(f(X))$$

(wegen Proposition 3.34).

Diese Tatsache gilt in großer Allgemeinheit:

Satz 4.1 (Jensen'sche Ungleichung.)**.** *Sei f eine konvexe Funktion und sei X eine diskrete Zufallsvariable. Dann gilt*

$$f(\mathbb{E}(X)) \leq \mathbb{E}(f(X)). \tag{4.2}$$

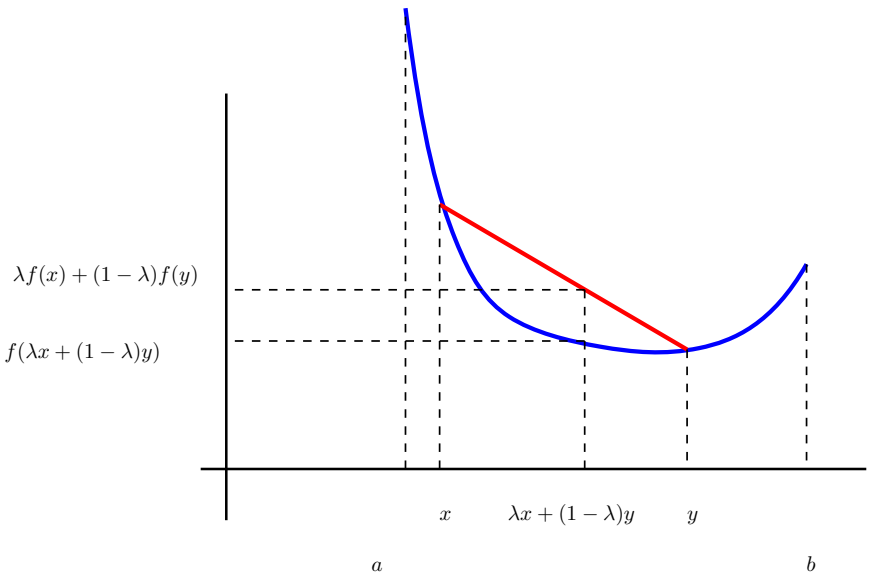
Gleichheit gilt genau dann wenn f fast sicher (bez. der Verteilung von X) affin ist.

Ist f strikt konvex, so gilt die Gleichheit in der Jensen'schen Ungleichung nur für konstantes X.

Beweis von Satz 4.1. Wir schreiben f' für die Rechtsableitung der Funktion f. Dann gilt für alle $\omega \in \Omega$.

$$f(\mathbb{E}X) + f'(\mathbb{E}X)(X(\omega) - \mathbb{E}X) \leq f(X(\omega)).$$

Indem wir auf beiden Seiten den Erwartungswert anwenden erhalten wir die Ungleichung. □



Beispiel 4.2. Für eine diskrete Zufallsvariable X gilt:

$$\mathbb{E}(X)^2 \leq \mathbb{E}(X^2). \tag{4.3}$$

Wir erhalten (4.3) direkt aus Satz 4.1, wobei $f(x) = x^2$ (konvex) setze. (4.3) folgt auch aus folgender Bemerkung:

$$\mathbb{E}(X^2) - \mathbb{E}(X)^2 = \mathrm{Var}(X) = \mathbb{E}((X-\mu)^2) \geq 0$$

wobei $\mu = \mathbb{E}(X)$. (Siehe Proposition 3.44).

Allgemeiner erhalten wir:

Korollar 4.3. *Sei $X \geq 0$ eine diskrete Zufallsvariable, und seien $0 < p < q$. Dann gilt*

$$\mathbb{E}(X^p) \leq \mathbb{E}(X^q)^{p/q}.$$

Insbesondere:

$$\mathbb{E}(X^q) < \infty \implies \mathbb{E}(X^p) < \infty.$$

Beweis. Wir verwenden Satz 4.1 mit $f(x) = x^\alpha$ wobei $\alpha = q/p > 1$ für $x \geq 0$: dann gilt $f''(x) = \alpha(\alpha-1)x^{\alpha-2} \geq 0$ für $x > 0$, d.h. f ist konvex. □

Wir sagen, dass $X \in L^p = L^p(\mathbb{P})$ wenn $\mathbb{E}(|X|^p) < \infty$. Korollar 4.3 kann auch wie folgt ausgedrückt werden: $L^p \supset L^q$ wenn $p < q$.

Beispiel 4.4. Sei X eine Zufallsvariable mit Wahrscheinlichkeitsfunktion

$$\mathbb{P}(X = k) = \frac{1}{\zeta(s) k^s}; \quad k \geq 1$$

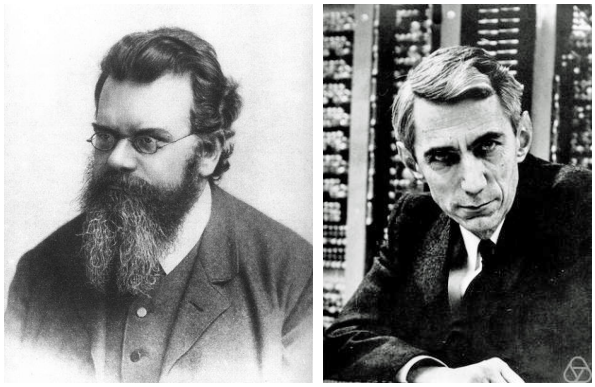


Abbildung 9: Ludwig Boltzmann und Claude Shannon

wobei $s > 1$ und $\zeta(s)$ die Riemann'sche Funktion ist. Für welches $p > 0$ gilt $X \in L^p$?

Bemerken Sie, dass

$$\mathbb{E}(X^p) = \frac{1}{\zeta(s)} \sum_{k \geq 1} k^p \frac{1}{k^s}$$

Die Reihe ist endlich genau dann, wenn $p - s < -1$. Das heißt, $X \in L^p \iff p < s - 1$. □

Sei

$$p_{\max} = \sup\{p > 0 : X \in L^p(\mathbb{P})\}. \tag{4.4}$$

Allgemein werden wir später sehen, dass $p_{\max}$ der Geschwindigkeit, mit der die Wahrscheinlichkeitfunktion im Unendlichen abfällt entspricht.

4.2 Entropie einer Zufallsvariable

Eine guter Artikel zur Bedeutung der Entropie in der Wahrscheinlichkeitstheorie ist zum Beispiel [2].

Der Begriff der Entropie wurde von Ludwig Boltzmann (1844–1906) eingeführt. Heuristisch beschreibt die Entropie, wie ungeordnet ein physikalisches System ist. Die Entropie eines Systems ist eine messbare Große, und ist zentral für den zweiten Hauptsatz der Thermodynamik. Dieser besagt, dass die Entropie von isolierten Systemen, die der spontanen Evolution überlassen werden, nicht mit der Zeit abnehmen kann. Dies ist ein Grundprinzip der **statistischen Mechanik**, heute ein wichtiger Zweig der Wahrscheinlichkeitsrechnung.

Im Jahr 1948 entwickelte der Bell-Labs-Wissenschaftler Claude Shannon ein ähnliches statistisches Konzept. Dies misst die Unsicherheit, die mit einer Zufallsvariable verbunden ist. Der Ursprung des Namens scheint von John von Neumann vorgeschlagen worden zu sein und wird durch die folgende wunderbare Anekdote erklärt:

> *My greatest concern was what to call it. I thought of calling it 'information', but the word was overly used, so I decided to call it 'uncertainty'. When I discussed it with John von*

Neumann, he had a better idea. Von Neumann told me, 'You should call it entropy, for two reasons: In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, nobody knows what entropy really is, so in a debate you will always have the advantage.

Um die Unsicherheit einer Zufallsvariable zu messen ist es zuerst notwendig, die folgende einfachere Situation zu überlegen. Sei A ein Ereignis von dem wir wissen, dass $\mathbb{P}(A) = p$ (z.B. die $\mathbb{P}($ Würfel zeigt 6 $) = 1/6$). Wie überrascht sind wir, wenn wir feststellen, dass A eintritt? (Also, wie überrascht sind wir, wenn eine 6 geworfen wird.)

Wir führen eine Funktion $S = S(p)$ ein (vielleicht kommt S vom englischen Wort "surprise"). Wir möchten, dass die Funktion die folgenden Anforderungen erfüllt:

- Da ein sicheres Ereignis nicht überraschend ist, $S(1) = 0$
- $S(p)$ ist eine fallende Funktion (je kleiner p, desto überraschender ist das Eintreten von A)
- Wenn A zweimal unabhängig eintritt, dann sind wir zweimal überrascht. Genauer gesagt, $S(pq) = S(p) + S(q)$.

Es gibt eine eindeutige Lösung S die diesen Anforderungen genügt, $S(p) = -c \log p$. Als Konvention nehmen wir $c = 1/\log 2$, sodass

$$S(p) = -\log_2 p.$$

$S(p)$ ist also auch die Anzahl der Binärziffern die wir brauchen um p zu approximieren (die Anzahl der 0en am Anfang der Binärentwicklung von p).

Definition 4.5. *Sei X eine diskrete Zufallsvariable mit WF $p(x), x \in \mathcal{W}$. Die* Entropie *oder* Shannon-Entropie *von X, bez. $\mathcal{L}_X$, ist die Größe*

$$H(X) = H(\mathcal{L}_X) = -\sum_x p(x) \log_2 p(x).$$

Beobachten Sie, dass $H(X) = \mathbb{E}(S(p(X)))$, sodass $H(X)$ misst, wie überraschend das Eintreten von X durchschnittlich ist. Das heißt, $H(X)$ ist eher groß, wenn die Werte die X annimmt eher unwahrscheinlich sind.

Beispiel 4.6. Sei X eine diskrete Zufallsvariable und $f : \mathbb{R} \to \mathbb{R}$ eine beliebige Funktion. Sei $Y = f(X)$. Dann $H(Y) \leq H(X)$. (Übung!)

Proposition 4.7. *Sei X eine diskrete Zufallsvariable mit Werten in $\mathcal{W} = \{x_1, \ldots, x_n\}$. Dann gilt*

$$H(X) \leq \log_2 n.$$

Diese obere Schranke wird genau dann erreicht, wenn X gleichverteilt auf $\mathcal{W}$ ist.

Beweis. Sei $p_i = \mathbb{P}(X = x_i)$, und ohne Beschränkung der Allgemeinheit nehmen wir $p_i > 0$ an. Weil $h(x) = x \lg_2 x$ konvex ist, gilt mit Jensen

$$\sum_{i \le n} h(p_i) \frac{1}{n} \ge h\left(\sum_{i \le n} p_i \frac{1}{n} \right).$$

Daraus folgt die Behauptung unmittelbar. □

Bemerkung 4.8 (Entropie als Maß für Informationsgehalt.)**.** Angenommen eine (faire) Münze X wird einmal pro Sekunde geworfen. Um diese Information von einem Sender zu einem Empfänger zu übertragen, Benötigt man einen Kanal der mit einer Kapazität von *einem bit pro Sekunde* übertragen kann.

Falls eine sehr unfaire Münze X (sagen wir $p = 0.01$) geworfen wird, können wir die zugehörige Information effizienter übertragen. Zum Beispiel könnten wir stets nur übertragen wieviele 0en zwischen je zwei 1ern geworfen werden. Wenn man bereit ist, eine bestimmte Verzögerung zu akzeptieren, genügt also ein Kanal mit deutlich geringerer Kapazität.

Shannon's "fundamental theorem for a noiseless channel" besagt, dass die nötige Kapazität des Kanals gerade $H(X)$ ist. Dieses Resultat gilt allgemeiner für diskrete Zufallsvariablen. Auch die Unabhängigkeit der einzelnen Würfel kann durch allgemeiner Voraussetzungen ersetzt werden.

Neben der Shannon-Entropie, betrachtet man in der Wahrscheinlichkeitstheorie auch oft:

Definition 4.9. *Seien $\mathbb{P}, \mathbb{Q}$ Wahrscheinlichkeitsmaße. Dann ist die* relative Entropie *oder* Kullback-Leibler-Divergenz *definiert durch*

$$H(\mathbb{Q}|\mathbb{P}) := \begin{cases} \infty & \mathbb{Q} \not\ll \mathbb{P} \\ \int \frac{d\mathbb{Q}}{d\mathbb{P}} \log \frac{d\mathbb{Q}}{d\mathbb{P}} & \textit{else.} \end{cases}$$

1. Es gilt stets $H(\mathbb{Q}|\mathbb{P}) \ge 0$, wobei Gleichheit genau für $\mathbb{P} = \mathbb{Q}$ gilt. Der Beweis is analog zu Proposition 4.7 (Übung). Die relative Entropie wird daher oft als Maß für die Abweichung von $\mathbb{Q}$ zu $\mathbb{P}$ interpretiert. (Es ist wichtig im Auge zu behalten, dass $H(\mathbb{Q}|\mathbb{P})$ im allgemeinen nicht symmetrisch in $\mathbb{P}, \mathbb{Q}$ ist, insbesondere ist H keine Metrik.)

2. Falls $\mathbb{P} = \sum_{i \le n} p_i \delta_{x_i}, \mathbb{Q} = \sum_{i \le n} q_i \delta_{x_i}$, so gilt

$$H(\mathbb{Q}|\mathbb{P}) = \sum_{i \le n} q_i \log(q_i/p_i) = \sum_{i \le n} q_i \log(1/p_i) - \sum_{i \le n} q_i \log(1/q_i) = H(\mathbb{Q}, \mathbb{P}) - H(\mathbb{Q}).$$

   Hier bezeichnet $H(\mathbb{Q}, \mathbb{P}) := \sum_{i \le n} q_i \log(1/p_i)$ die *Kreuzentropie*.

   Wir haben oben schon angemerkt, das die Shannon-entropie $H(\mathbb{Q})$ die notwendige Kapazität ist um ein $\mathbb{Q}$-Signal mit einem $\mathbb{Q}$-optimalen Code zu übertragen. Im Gegensatz dazu, ist $H(\mathbb{Q}, \mathbb{P})$ die notwendige Kapazität, die man braucht um ein $\mathbb{Q}$-Signal mit einem auf $\mathbb{P}$ optimierten Code zu übertragen. (Das folgt aus dem Satz von Kraft-McMillan.)

   Die relative Entropie entspricht daher die extra Kapazität, die ein Kanal benötigt, falls der Code auf ein anderes Signal optimiert wurde.

4.3 Markov'sche Ungleichung

Satz 4.10. *Sei X eine diskrete Zufallsvariable mit $\mathbb{E}[|X|] < \infty$ und $a > 0$. Dann gilt*

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

Beweis. Offenbar gilt $I_{\{|X|\geq a\}} \leq |X|/a$ und daraus folgt

$$\mathbb{P}(|X| \geq a) = \mathbb{E}[I_{\{|X|\geq a\}}] \leq \mathbb{E}[|X|/a] = \frac{\mathbb{E}[|X|]}{a}.$$

□

Zum Beispiel, wenn $X \geq 0$, $\mathbb{P}(X \geq 100\mu) \leq 1\%$.

Beispiel 4.11. Nehmen wir an, dass $\mathbb{E}[|X|^p] < \infty$. Dann gilt für jedes $t > 0$

$$\mathbb{P}(|X| > t) \leq \mathbb{E}(|X|^p)/t^p.$$

Deshalb fällt $\mathbb{P}(X > t)$ mindestens so schnell ab wie Konst. $\times\, t^{-p}$. Das ist ein Grund, warum das $p_{\max}$ von (4.4) etwas mit der Geschwindgkeit, mit der die Wahrscheinlichkeitfunktion im Unendlichen abfällt, zu tun hat.

4.4 Chebyshev'sche Ungleichung

Satz 4.12. *Sei X eine diskrete Zufallsvariable und $\varepsilon > 0$. Dann gilt*

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

Beweis. Genau wie bei der Markov'schen Ungleichung. (Alternativ kann man die Markov'sche Ungleichung direkt verwenden auf X^2, weil $\mathbb{P}(|X| > \varepsilon) = \mathbb{P}(X^2 > \varepsilon^2)$.) □

Es lohnt sich eine Folgerung im Kopf zu behalten: Falls $\mu = \mathbb{E}[X]$, dann gilt

$$\mathbb{P}(|X - \mu| \geq \varepsilon) \leq \frac{\mathrm{Var}(X)}{\varepsilon^2}.$$

Die folgende Form dieser Ungleichung ist besonders beachtenswert:

$$\mathbb{P}(|X - \mu| > k\sigma) \leq 1/k^2. \tag{4.5}$$

Mit anderem Worte, die Wahrscheinlichkeit, dass X mehr als k Standardbweichungen von seinem Erwartungswert ernfernt ist, ist kleiner als $1/k^2$. Zum Beispiel ist X mit Wahrscheinlichkeit mindestens 96% näher van μ als 5σ – dass ist immer wahr!

Manchmal ist uns die Verteilung von X bekannt; in diesem Fall können wir oft präzisere Schätzungen erhalten. Aber (4.5) gilt in jedem Fall (unter der Annahme, dass $\mathrm{Var}(X) < \infty$ i.e. $X \in L^2$). Obwohl einfach, ist das eine mächtige Schranke, wie wir sofort sehen werden.

4.5 Schwaches Gesetz der großen Zahlen

Satz 4.13 (L^2-Gesetz der großen Zahlen)**.** *Sei $X_1, X_2, \ldots$ eine Folge von unabhängig und identisch verteilten Zufallsvariablen mit Mittel μ und Varianz σ^2. Setze*

$$\bar{X}_n := \frac{X_1 + \ldots + X_n}{n}.$$

Dann gilt

$$\mathbb{E}[(\bar{X}_n - \mu)^2] = \sigma^2/n$$

und insbesondere

$$\lim_{n\to\infty} \mathbb{E}[(\bar{X}_n - \mu)^2] = 0.$$

Beweis. Es gilt (nach Pythagoras)

$$\mathbb{E}[(\bar{X}_n - \mu)^2] = \mathrm{Var}(\bar{X}_n) = \mathrm{Var}\left(\frac{X_1 + \ldots + X_n}{n}\right) = \frac{\mathrm{Var}(X_1 + \ldots + X_n)}{n^2} = \frac{n\,\mathrm{Var}(X_1)}{n^2}.$$

□

Mittels der Chebyshev'schen Ungleichung erhalten wir daraus unmittelbar:

Satz 4.14 (Schwaches Gesetz großen Zahlen)**.** *Sei $X_1, X_2, \ldots$ eine Folge von unabhängig und identisch verteilten diskreten Zufallsvariablen mit Mittel μ und Varianz σ^2. Setze wieder*

$$\bar{X}_n := \frac{X_1 + \ldots + X_n}{n}$$

Dann gilt für alle $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}[|\bar{X}_n - \mu| \geq \varepsilon] = 0. \tag{4.6}$$

Beweis. Nach der Chebyshev'schen Ungleichung gilt für $\varepsilon > 0$

$$\mathbb{P}[|\bar{X}_n - \mu| \geq \varepsilon] \leq \sigma^2/(n\varepsilon^2) \to 0.$$

□

Beispiele: Wenn wir einen Würfel immer wieder werfen, sehen wir im Schnitt 3.5 Augen. Wenn wir eine faire Münze immer wieder werfen kommen ungefähr 50% Köpfe.

Bemerkungen:

1. Wie schon früher angemerkt, liefert uns das (schwache) Gesetz großer Zahlen den Zusammenhang von unserem axiomatischen Zugang zu einer frequentistischen Sichtweise auf die Wahrscheinlichkeitstheorie: Sei A ein Ereignis mit $p = \mathbb{P}(A)$, wir stellen uns vor dass ein Experiment einen bestimmten positiven Ausgang hat (die Münze zeigt

Kopf oder so). Wir bezeichnen weiter mit $A_1, A_2, \ldots$ unabhängige Wiederholungen des Experiments und schreiben

$$N(A) := I_{A_1} + \ldots + I_{A_N}$$

für die Anzahl der positiven Ausgänge innerhalb der ersten N Experimente.

Nach dem Gesetz großer Zahlen gilt dann

$$\mathbb{P}\left(\left|\frac{N(A)}{N} - p\right| \geq \varepsilon\right) \to 0, \tag{4.7}$$

d.h. der Anteil positiven Ausgänge nähert sich genau der Wahrscheinlichkeit von A an.

Beweis von (4.7). Es gilt $p = \mathbb{E}[I_A]$ und daher

$$\mathbb{P}\left(\left|\frac{N(A)}{N} - p\right| \geq \varepsilon\right) = \mathbb{P}\left(\left|\frac{I_{A_1} + \ldots + I_{A_N}}{N} - p\right| \geq \varepsilon\right) \to 0.$$

□

2. Anstelle von Unabhängigkeit könnten wir auch Unkorreliertheit fordern. Es ist auch nicht wesentlich, dass alle Zufallsvariablen die selbe Verteilung haben, es reicht, dass die Erwartungswerte und Varianzen übereinstimmen (oder zumindest konvergieren).

3. Komplett weglassen können wir die Voraussetzung der Unabhängigkeit nicht: Das sieht man leicht indem man z.B. den Fall $X_1 = X_2 = \ldots$ betrachtet.

4. Die Konvergenzart in (4.6) heisst 'Konvergenz in Wahrscheinlichkeit', man spricht auch davon, dass die Verteilungen von $X_n, n \geq 1$ 'schwach' gegen den Mittelwert konvergieren.

   Im Gegensatz zum 'schwachen' Gesetz großer Zahlen spricht man von einem 'starken' Gesetz großer Zahlen wenn die Konklusion statt (4.6), ist dass

   $$\lim_{n\to\infty} \bar{X}_n(\omega) = \mu \tag{4.8}$$

   für alle ω aus einer Menge $\tilde{\Omega}$ mit $\mathbb{P}(\tilde{\Omega} = 1)$. Starke Konvergenz im Sinne von (4.8) impliziert schwache Konvergenz im Sinne von (4.6). (Übung?) Wenn man sich etwas mehr bemüht als wir das getan haben, kann man in der Situation von (4.14) auch starke Konvergenz zeigen. Dafür reicht es sogar vorauszusetzen, dass $\mathbb{E}[X_i]$ existiert, die Varianz braucht nicht endlich zu sein. (*Das* 'starke Gesetz großer Zahlen'.)

5. Ein grundlegender Satz in der Theorie dynamischer Systeme ist der *Ergodensatz* der eine Verallgmeinerung des starken Gesetzes großer Zahlen darstellt, er wird dort im Sinne von *Raummittel = Zeitmittel* interpretiert.

5 Bedingte Verteilung und bedingte Erwartung

5.1 Bedingte Verteilung

Seien X, Y diskrete Zufallsvariablen die auf dem selben Wahrscheinlichkeitsraum definiert sind. Wir denken insbesondere an den Fall, dass X und Y nicht unabhängig sind. Die *gemeinsame Verteilung*, genauer die *gemeinsame Wahrscheinlichkeitsfunktion* ist dann gegeben durch

$$p(x,y) = p_{(X,Y)}(x,y) = \mathbb{P}(X = x, Y = y)$$

wobei x, y in den jeweiligen Wertebereichen $\mathcal{W}_X$ und $\mathcal{W}_Y$ liegen. Diese Verteilung ist nichts anderes als die Verteilung des Vektors (X, Y), der auch eine diskrete Zufallsvariable (mit Werten in $\mathcal{W}_X \times \mathcal{W}_Y$) ist.

In diesem Kontext bezeichnet man die Verteilung / Wahrscheinlichkeitsfunktion von X

$$p_X(x) = \mathbb{P}(X = x) = \sum_{y \in \mathcal{W}_Y} p(x,y)$$

auch als *Randverteilung* oder *Marginalverteilung* von X.

Wenn wir $y \in \mathcal{W}_Y$ fixieren, betrachten wir auch die *bedingte Verteilung von X gegeben $Y = y$*, i.Z. $\mathbb{P}(X = \cdot | Y = y) = p_{\cdot|y}$:

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

Bemerkungen:

1. Offenbar erhält man die Randverteilung als entsprechend gewichtete Summe der bedingten Verteilungen:
$$p_X(x) = \sum_{y \in \mathcal{W}_Y} p_{x|y} p_Y(y).$$

2. Wenn X, Y unabhängig sind, dann ist für jedes $y \in \mathcal{W}_Y$ die bedingte Verteilung von X gegeben $Y = y$ gleich der Randverteilung von X. Die Umkehrung gilt auch, wie man sich leicht überlegt:

Proposition 5.1. *X und Y sind unabhängig genau dann, wenn $p(x,y) = p_X(x)p_Y(y)$, und auch genau dann, wenn $p_{x|y} = p_X(x)$ für jedes $x \in \mathcal{W}_X, y \in \mathcal{W}_Y$.*

Beispiel 5.2. Seien $X_1, \ldots, X_n$ unabhängig $B(1,p)$ verteilte Zufallsvariablen, und sei $Y = X_1 + \ldots + X_n$. Was ist die bedingte Verteilung von X_1, gegeben $Y = k$ (wobei $0 \leq k \leq n$)?

Natürlich kann X_1 nur gleich 0 oder 1 sein, auch gegeben $Y = k$. Deshalb ist die bedingte Verteilung auch Bernoulli. Aber was ist der Parameter? Ist er noch p oder ist er etwas anderes

unter der Annahme, dass $Y = k$? Wir berechnen:

$$\begin{aligned}\mathbb{P}(X_1 = 1|Y = k) &= \frac{\mathbb{P}(X_1 = 1; X_2 + \ldots + X_n = k-1)}{\mathbb{P}(Y = k)} \\ &= \frac{\mathbb{P}(X_1 = 1)\mathbb{P}(X_2 + \ldots + X_n = k-1)}{\mathbb{P}(Y = k)} \\ &= \frac{p\binom{n-1}{k-1}p^{k-1}(1-p)^{n-1-(k-1)}}{\binom{n}{k}p^k(1-p)^{n-k}} \\ &= \frac{\binom{n-1}{k-1}}{\binom{n}{k}} = \frac{k}{n}.\end{aligned}$$

Das heißt, gegeben $Y = k$ ist die bedingte Verteilung von X_1 eine Bernoulli Verteilung mit Parameter $p' = k/n$.

Beispiel 5.3. Nehmen wir an, dass die Anzahl Y der Tippfehler auf einer Seite Poisson(λ) verteilt ist, und Tippfehler unabhängig von einander mit Wahrscheinlichkeit p gefunden werden. Was ist die bedingte Verteilung von X gegeben Y? Was ist die gemeinsame Verteilung von X und Y? Was ist die gemeinsame Verteilung von X und $Y - X$ (die Anzahl der Tippfehler die nicht gefunden werden)? Was sind die Randverteilungen von X und Y?

In diesem Beispiel ist tatsächlich die bedingte Verteilung von X gegeben Y geben (und natürlich auch die Verteilung von Y). Das heißt, die Annahme ist

$$\begin{aligned}\mathbb{P}(Y = n) &= e^{-\lambda}\lambda^n/n! \\ \mathbb{P}(X = k|Y = n) &= \binom{n}{k}p^k(1-p)^{n-k}.\end{aligned}$$

Wir können von dieser Annahme die gemeinsame Verteilung von X und Y einfach berechnen:

$$\begin{aligned}\mathbb{P}(X = k, Y = n) &= \mathbb{P}(Y = n)\mathbb{P}(X = k|Y = n) \\ &= e^{-\lambda}\frac{\lambda^n}{n!}\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k} \\ &= e^{-\lambda}\frac{\lambda^n p^k(1-p)^{n-k}}{k!(n-k)!}; 0 \le k \le n.\end{aligned}$$

Die gemeinsame Verteilung von X und $Z = Y - X$ folgt sofort:

$$\begin{aligned}\mathbb{P}(X = k, Z = m) &= \mathbb{P}(X = k, Y = k + m) \\ &= e^{-\lambda}\frac{\lambda^{k+m}p^k(1-p)^m}{k!m!} \\ &= e^{-\lambda p}\frac{(\lambda p)^k}{k!} \times e^{-\lambda(1-p)}\frac{(\lambda(1-p))^m}{m!}\end{aligned}$$

und deshalb sind X und Z unabhängig und Poisson verteilt mit Parametern λp und $\lambda(1-p)$!

Um dieses Ergebnis zu würdigen, betrachten Sie $\lambda = 100, p = 1/2$ (sodass ich ungefähr die Hälfte aller Fehler finde). Falls ich 200 Fehler auf einer Seite finde, wie viele Fehler sind eigentlich insgesamt auf der Seite? Vielleicht 400? Aus der vorherigen Berechnung erhalten wir, dass die unentdeckten Tippfehler unabhängig von denen sind, die wir gefunden haben. Deshalb erwarten wir insgesamt ungefähr 250.

5.2 Bedingte Erwartung

Die *bedingte Erwartung von X gegeben $Y = y$* ist definiert als der Erwartungswert der bedingten Verteilung $\mathbb{P}(X = \cdot|Y = y)$, d.h.

$$\mathbb{E}[X|Y = y] = \sum_{x \in \mathcal{W}_X} x p_{x|y} = \frac{\mathbb{E}[X 1_{\{Y=y\}}]}{\mathbb{P}(Y = y)}.$$

Beispiel 5.4. Im Beispiel 5.3 gilt $\mathbb{E}(X|Y = n) = np$, $\mathbb{E}(Z|X = k) = \mathbb{E}(Z) = \lambda(1 - p)$, $\mathbb{E}(Y|X = k) = k + \lambda(1 - p)$.

Im Fall, dass die Zufallsvariablen unabhängig sind, erhalten wir:

Proposition 5.5. *Wenn X und Y unabhängig sind, dann gilt $\mathbb{E}(X|Y = y) = \mathbb{E}(X)$ und $\mathbb{E}(Y|X = x) = \mathbb{E}(Y)$.*

Normalerweise ist der Erwartungswert von X eine bestimmte (d.h., nicht zufällige) Größe. Das gilt auch wenn wir auf $Y = y$ bedingen. Aber hier kommt eine grundlegender (obwohl am Anfang sehr verwirrender) Gedanke der Wahrscheinlichkeitstheorie ins Spiel: Nehmen wir an, dass ein Beobachter den Wert von $Y = Y(\omega)$ feststellt und nichts anderes. Was ist seine Vermutung für X? Natürlich ist sie $\mathbb{E}(X|Y = y)$ für den beobachten Wert von $Y(\omega) = y$. Aber dieser Wert hängt vom Ereignis ω ab. Das heißt, wir sollten $\mathbb{E}(X|Y = y)$ als eine Zufallsvariable sehen, wobei wir $y = Y(\omega)$ einsetzen!

Definition 5.6. *Der bedingte Erwartungswert $\mathbb{E}(X|Y)$ ist die Zufallsvariable, die den Wert $\mathbb{E}(X|Y = y)$ nimmt, wenn $Y(\omega) = y$. Das heißt,*

$$\mathbb{E}(X|Y) = \sum_y \mathbb{E}(X|Y = y) 1_{\{Y=y\}}.$$

Wenn wir $\mathbb{E}(X|Y)$ berechnen, nehmen wir einen Durchschnitt nur über die Zufälligkeit von X, während wir so tun, als ob Y bekannt und fest wäre. Das ist eine sehr praktische und natürliche Sache, weil der Zufall oft "in Scheibchen" kommt (z.B. im Beispiel 5.3, gibt es den Zufall in der Anzahl der Tippfehler, und der Zufall in Bezug auf das Entdecken dieser Tippfehler). In diesem Fall ist es natürlich, die Scheibchen zu trennen.

Beispiel 5.7. Im Beispiel 5.3 ist $\mathbb{E}(X|Y) = Yp$, und $\mathbb{E}(Y|X) = X + \lambda(1 - p)$. Beachten Sie, dass beides Zufallsvariablen sind!

Beispiel 5.8. Im Beispiel 5.2 gilt $\mathbb{E}(X_1|Y) = Y/n$.

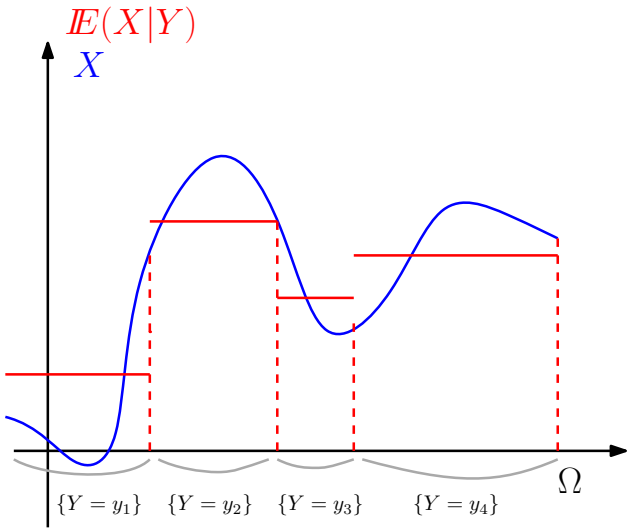


Abbildung 10: In diesem Bild sehen wir eine Zufallsvariable X als eine Funktion auf dem Grundraum Ω. Die bedingte Erwartung $\mathbb{E}(X|Y)$ ist der Mittelwert von X über die jeweilige Menge $\{Y = y\}$. Die **Turmeigenschaft** (Satz 5.11) sagt, dass der Mittelwert von X auch der Mittelwert von $\mathbb{E}(X|Y)$ ist.

Bemerkenswerter Weise erbt die bedingte Erwartung viele Eigenschaften des gewöhnlichen Erwartungswerts, die Beweise sind jeweils nicht schwer und verwenden ähnlich Argumente, wie wir sie aus dem Fall des gewöhnlichen Erwartungswertes kennen. Insbesondere haben wir in Satz 3.33 gesehen, dass der Erwartungswert $\mathbb{E}[X]$ die beste konstante Approximation der Zufallsvariable X ist. Im selben Sinne ist $\mathbb{E}[X|Z]$ die beste Approximation an die Funktion X gebeben die Information aus Z:

Satz 5.9. *Seien X, Z Zufallsvariablen und $\mathbb{E}X^2 < \infty$. Dann gibt es eine Funktion f die*

$$\mathbb{E}[(X - f(Z))^2]$$

minimiert und für jedes solche f gilt

$$\mathbb{E}[X|Z] = f(Z).$$

Weitere Eigenschaften der bedingten Erwartung:

Satz 5.10. *Seien X, Y, Z Zufallsvariablen. Dann gilt z.B.*

1. *Linearität: $\mathbb{E}[aX + bY|Z] = a\mathbb{E}[X|Z] + b\mathbb{E}[Y|Z]$.*
2. *Wenn Y weniger Information als Z enthält, d.h. wenn es eine Funktion mit $Y = f(Z)$ gibt, dann gilt*
$$\mathbb{E}[XY|Z] = Y\mathbb{E}[X|Z].$$

3. *Sind X, Z unabhängig, dann gilt $\mathbb{E}[X|Z] = \mathbb{E}[X]$.*

4. *Jensen'sche Ungleichung: Sei f eine konvexe Funktion. Dann gilt*

$$f(\mathbb{E}[X|Z]) \leq \mathbb{E}[f(X)|Z].$$

5.3 Turmeigenschaft

In 99% der Fälle verwenden wir den bedingten Erwartungswert, um einen (unbedingten) Erwartungswert zu berechnen. Der folgende wichtige Satz (die sogenannte **Turmeigenschaft**) erlaubt uns dies zu tun.

Satz 5.11 (Turmeigenschaft)**.** *Seien X, Y (diskrete) Zufallsvariablen. Dann gilt*

$$\mathbb{E}(X) = \mathbb{E}[\mathbb{E}(X|Y)].$$

Beweis. Wir berechnen einfach

$$\begin{aligned}
\mathbb{E}[\mathbb{E}(X|Y)] &= \sum_y \mathbb{P}(Y = y)\mathbb{E}(X|Y = y) \\
&= \sum_y \mathbb{P}(Y = y) \sum_x x\mathbb{P}(X = x|Y = y) \\
&= \sum_{x,y} x\mathbb{P}(X = x, Y = y) \\
&= \sum_x x\mathbb{P}(X = x) \\
&= \mathbb{E}(X),
\end{aligned}$$

wie gewünscht. □

Um diesen Satz zu verstehen, machen wir ein paar Beispiele.

Beispiel 5.12. Fortsetzung des Beispiels 5.2. $\mathbb{E}(X_1) = \mathbb{E}[\mathbb{E}(X_1|Y)] = \mathbb{E}[Y/n] = np/n = p$ (das war offensichtlich!)

Beispiel 5.13. Berechnen Sie $\mathbb{E}(X)$ und $\mathbb{E}(Y)$ mit der Hilfe von Satz 5.11. Überprüfen Sie, dass es mit dem übereinstimmt, was wir schon über die Verteilung von X und Y wissen.

Ein wichtiges Anwendungsbeispiel zur Turmeigenschaft ist:

Satz 5.14 (Wald'scher Satz)**.** *Nehmen wir an, dass $X_1, X_2, \ldots,$ identisch verteilt sind, und sei $S_n = \sum_{i=1}^n X_i$. Sei N eine unabhängige und ganzzahlige Zufallsvariable. Dann gilt*

$$\mathbb{E}(S_N) = \mathbb{E}(N)m,$$

wobei $m = \mathbb{E}(X_1)$.

Beweis. Wir bedingen auf $N = n$. Dann ist $\mathbb{E}(S_N|N = n) = \sum_{i=1}^{n} \mathbb{E}(X_i|N = n) = nm$, weil X_i und N unabhängig sind. Deshalb folgt

$$\mathbb{E}(S_N) = \mathbb{E}[\mathbb{E}(S_N|N)] = \sum_{k=0}^{\infty} \mathbb{P}(N = k)\mathbb{E}(S_N|N = k) = \sum_{k=0}^{\infty} \mathbb{P}(N = k)km = m\mathbb{E}(N),$$

wie gewünscht. □

Beispiel 5.15. N Personen steigen in einen Aufzug, sei X_i ihr jeweiliges Gewicht. Dann ist das erwartete Gesamtgewicht $\mathbb{E}(T) = \mathbb{E}(N)\mathbb{E}(X_i)$.

Beispiel 5.16. Eine Person A hat eine Krankheit und steckt in der ersten Periode X_A Leute an. Jede Person B die angesteckt wurde, steckt in der nächsten Periode wiederum X_B neue Personen an. Wir nehmen an, dass diese Ansteckungen unabhängig und identisch verteilt sind. Sei $R = \mathbb{E}[X_A]$. Dann werden in der zweiten Periode im Schnitt R^2 Personen infiziert. (usw)

Beispiel 5.17. Hier ist eine etwas aufwändigere Anwendung. Wir werfen eine Münze mit $\mathbb{P}(\text{Kopf}) = p$. Sei N die Anzahl der Würfe bis wir zwei Köpfe in Folge bekommen. (z.B., in dem Fall, dass die Reihenfolge der Würfe $ZKZZKZKZZKK$ ist, erhalten wir $N = 11$). Dann beobachten Sie, dass

$$N = \sum_{i=1}^{M} X_i + 1 \tag{5.1}$$

wobei M und X_i unabhängige Zufallsvariablen sind. Dabei sind M und X_1 geometrisch verteilte Zufallsvariable, während die Verteilung von $X_i, i \geq 2$ eine geometrische Verteilung, bedingt auf $X_i > 1$, ist.

Um (5.1) zu sehen, zerlegen wir die Reihenfolge der Würfe in Stücke, die mit K enden: z.B.,

$$(ZK)(ZZK)(ZK)(ZZK)(K).$$

Dabei ist M die Anzahl der Stücke (hier $M = 4$) und sind X_i ihre jeweilige Länge (hier $X_1 = 2, X_2 = 3, X_3 = 2, X_4 = 3$).

Wir schließen daraus und aus dem Wald'schen Satz, dass

$$\mathbb{E}(N) = (\mathbb{E}(M) - 1)\mathbb{E}(X_2) + \mathbb{E}(X_1) + 1.$$

Jetzt bemerken Sie, dass $\mathbb{E}(M) = \mathbb{E}(X_1) = 1/p$ (Erwartungswert einer geometrischen Zufallsvariable), und $\mathbb{E}(X_2) = 1 + 1/p$ (Übung!) Deshalb erhalten wir insgesamt

$$\begin{aligned}\mathbb{E}(N) &= (1/p - 1)(1/p + 1) + 1/p + 1 \\ &= 1/p^2 - 1 + 1/p + 1 \\ &= 1/p^2 + 1/p.\end{aligned}$$

Bemerkung 5.18. Allgemeiner ist es möglich zu beweisen, dass wenn wir k Köpfe in Folge möchten, dann müssen wir durchschnittlich $(1/p + 1/p^2 + \ldots + 1/p^k)$ warten.

5.4 Bedingte Erwartung bezüglich σ-Algebren

Definition 5.19. *Sei X eine Zufallsvariable auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \mathbb{P})$ sodass $\mathbb{E}|X| < \infty$ und sei $\mathcal{G} \subseteq \mathcal{F}$ eine σ-Algebra. Dann heißt eine Zufallsvariable Z die* bedingte Erwartung *von X gegeben $\mathcal{G}$, in Zeichen,*

$$Z = \mathbb{E}[X|\mathcal{G}]$$

falls Z $\mathcal{G}$-messbar ist und für alle $G \in \mathcal{G}$ gilt, dass

$$\mathbb{E}[Z1_G] = \mathbb{E}[X1_G]. \tag{5.2}$$

Satz 5.20. *Unter den Voraussetzung der obigen Definition existiert die bedingte Erwartung und ist f.s. eindeutig bestimmt.*

Beweis. Wir zeigen zunächst die Eindeutigkeit: Angenommen Z_1, Z_2 sind Versionen von $\mathbb{E}[X|\mathcal{G}]$. Dann sind beide $\mathcal{G}$-messbar und erfüllen $\mathbb{E}[(Z_1 - Z_2)1_G] = 0$ für $G \in \mathcal{G}$. Falls sie nicht gleich sind, gilt oBdA $\mathbb{P}(Z_1 < Z_2) > 0$. Dann gibt es auch $n \in \mathbb{N}$ für dass die Menge $G = \{Z_2 - Z_1 \geq n\} \in \mathcal{G}$ erfüllt dass $\mathbb{P}(G) > 0$. Aber dann folgt

$$\mathbb{E}(Z_2 - Z_1)1_G \geq \mathbb{P}(G)/n > 0$$

also können die beiden Zufallsvariablen nicht beide Versionen der bedingten Erwartung sein.

Die Existenz der bedingten Erwartung folgt schnell aus dem *Satz von Radon-Nikodym*: Wir definieren auf $\mathcal{G}$ ein *signiertes Maß* ν durch

$$\nu(G) := \mathbb{E}X1_G.$$

Falls $\mathbb{P}(G) = 0$ gilt auch $\nu(G) = \mathbb{E}X1_G = 0$, also ist ν absolut stetig bezüglich $\mathbb{P}$. Weiters gilt für alle $G \in \mathcal{G}$ dass $|\nu(G)| \leq \mathbb{E}|X|$, daher ist ν ein endliches signiertes Maß. Der Satz von Radon-Nikodym besagt daher dass es eine $\mathcal{G}$-messbare Funktion Z gibt die eine Dichte von ν bezüglich $\mathbb{P}$ ist, d.h. $Z = d\nu/d\mathbb{P}$. Das bedeutet aber gerade dass

$$\nu(G) = \mathbb{E}ZI_G$$

für $G \in \mathcal{G}$. □

Bemerkungen:

1. Wenn die Bedingung (5.2) für einen durchschnittsstabilen Erzeuger gilt, dann auch schon für die ganze σ-Algebra. Insbesondere passt unsere neue Definition der bedingten Erwartung zu der die wir vorher im diskreten Fall betrachtet haben.

2. Man kann wiederum zeigen, dass $\mathbb{E}[X|\mathcal{G}]$ genau jene $\mathcal{G}$-messbare Zufallsvariable Z ist, die
$$\mathbb{E}[(X - Z)^2]$$
minimiert. Umgekehrt kann man diese Eigenschaft nutzen um ohne Verwendung von Radon-Nikodym zu zeigen, dass die bedingte Erwartung existiert.

Satz 5.21. *Eigenschaften der bedingten Erwartung (wir betrachten dabei jeweils Zufallsvariablen mit endlichem Erwartungswert).*

1. *Für die triviale σ-Algebra $\mathcal{G} = \{\emptyset, \Omega\}$ gilt $\mathbb{E}[X|\mathcal{G}] = \mathbb{E}[X]$.*
2. *(Turmeigenschaft 1) $\mathbb{E}[\mathbb{E}[X|\mathcal{G}]] = \mathbb{E}[X]$.*
3. *Falls X $\mathcal{G}$-messbar ist, so folgt $\mathbb{E}[X|\mathcal{G}] = X$.*
4. *(Linearität) $\mathbb{E}[aX + bY|\mathcal{G}] = a\mathbb{E}[X|G] + b\mathbb{E}[Y|\mathcal{G}]$.*
5. *(Positivität) Falls $X \geq 0$ so gilt auch $\mathbb{E}[X|\mathcal{G}] \geq 0$.*
6. *Es gelten die drei Konvergenzsätze wenn man $\mathbb{E}[.]$ durch $\mathbb{E}[.|\mathcal{G}]$ ersetzt (Lemma von Fatou, Satz von der monotonen Konvergenz, Satz von der dominierten Konvergenz).*
7. *Es gilt die bedingte Version der Jensensch'en Ungleichung: $f(\mathbb{E}[X|\mathcal{G}]) \leq \mathbb{E}[f(X)|\mathcal{G}]$ für eine konvexe Funktion f. Insbesondere gilt daher $\|\mathbb{E}[X|G]\|_2 \leq \|X\|_2$.*
8. *(Turmeigenschaft 2) Falls $\mathcal{G} \subseteq \mathcal{H}$ so gilt*
$$\mathbb{E}[\mathbb{E}[X|\mathcal{H}]|\mathcal{G}] = \mathbb{E}[X|\mathcal{G}].$$
9. *Falls Y beschränkt und $\mathcal{G}$-messbar ist, so gilt*
$$\mathbb{E}[XY|\mathcal{G}] = \mathbb{E}[X|\mathcal{G}]Y.$$
10. *Falls X unabhängig von $\mathcal{G}$ ist, so gilt*
$$\mathbb{E}[X|\mathcal{G}] = \mathbb{E}[X].$$

Beweis. Die meisten Eigenschaften sind relativ einfach zu zeigen und wir lassen die Beweise aus. Interessant ist vielleicht:

9. Dieser Beweis folgt mit dem Standardtrick der Maßtheorie: für Indikatorfunktionen $Y = 1_G$, $g \in \mathcal{G}$ folgt der Satz direkt aus der Definition. Wegen Linearität gilt er dann auch für $\mathcal{G}$-messbare Treppenfunktionen. Dann schließt man (zB) mit dem Satz von dominierter Konvergenz.

□

5.5 Martingale

Definition 5.22. *Gegeben sei Wahrscheinlichkeitsraum. Eine* Filtration *ist eine aufsteigende Folge von σ-Algebren $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \ldots \subseteq \mathcal{F}$. Ein stochastischer Prozess $X = (X_n)_{n=0}^{\infty}$ heisst* adaptiert *falls X_n für alle n $\mathcal{F}_n$-messbar ist.*

Ein Prozeß M ist ein Martingal *wenn für alle $n \in \mathbb{N}$ gilt, dass*

$$\mathbb{E}[M_{n+1}|\mathcal{F}_n] = M_n.$$

Ein Prozeß M ist ein Submartingal *wenn für alle $n \in \mathbb{N}$ gilt, dass*

$$\mathbb{E}[M_{n+1}|\mathcal{F}_n] \geq M_n.$$

Ein Prozeß M ist ein Supermartingal *wenn für alle $n \in \mathbb{N}$ gilt, dass*

$$\mathbb{E}[M_{n+1}|\mathcal{F}_n] \leq M_n.$$

Wenn man von stochastischen Prozessen spricht, geht man fast immer stillschweigend davon aus, dass diese adaptiert sind. Falls man keine Filtration spezifiziert, geht man oft davon aus, dass die vom Prozess X *erzeugte Filtration* $\mathcal{F}_n := \sigma(X_0, \ldots, X_n), n \geq 0$ gemeint ist.

Mittels der Turmeigenschaft zeigt man leicht:

Lemma 5.23. *Ein Prozeß M integrierbarer Zufallsvariablen ist ein Martingal genau dann wenn $\mathbb{E}[M_n|\mathcal{F}_k] = M_k$ für $n > k \geq 0$.*

Ist M ein Martingal, so gilt $\mathbb{E}M_n = \mathbb{E}M_0$ für all $n \in \mathbb{N}$.

Analoge Eigenschaften gelten natürlich für Sub- und Supermartingale.

Eine weitere Konsequenz der Turmeigenschaft ist:

Lemma 5.24. *Sei M ein Martingal bezüglich einer Filtration $(F_n)_n$. Dann ist M auch ein Martingal bezüglich der von M erzeugten Filtration.*

Wir betrachten einige Beispiele.

Beispiel 5.25. Als zentrales Beispiel betrachten wir die Irrfahrt auf $\mathbb{Z}$:

Seien $X_1, X_2, \ldots$ unabhängig und $\mathbb{P}(X_i = 1) = p, \mathbb{P}(X_i = -1) = 1 - p =: q$. Sei weiters

$$S_n := S_0 + X_1 + \ldots + X_n,$$

wobei $S_0 \in \mathbb{R}$ (und meistens $S_0 = 0$). Als Filtration betrachten wir $\mathcal{F}_n := \sigma(S_0, S_1, \ldots, S_n)$.

Dann heißt der Prozess $(S_n)_{n=0}$ *einfache Irrfahrt* oder *random walk* mit Start in S_0. Im Fall $p = q = 1/2$ sprechen wir von der *symmetrischen Irrfahrt*. (Diese Bezeichnungen sind allerdings nicht ganz einheitlich. Oft meint man die symmetrische Irrfahrt auch wenn man es nicht extra dazusagt.)

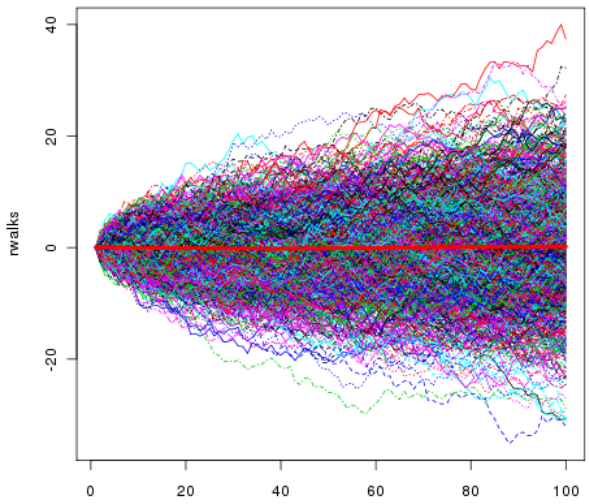


Abbildung 11: Viele Simulationen der Symmetrischen Irrfahrt

Konzeptionell interessiert uns oft das Verhalten der Pfade im Lauf der Zeit. Um eine erste Idee zu bekommen, berechnen wir Erwartungswert und Varianz. Offenbar gilt $\mathbb{E}X_i = p - q = 2p - 1$ und $\mathrm{Var}(X_i) = 1 - (p-q)^2 = (p+q)^2 - (p-q)^2 = 4pq$ und daher

$$\mathbb{E}[S_n] = S_0 + n(p-q) \tag{5.3}$$
$$\mathrm{Var}[S_n] = n4pq. \tag{5.4}$$

Die Varianz der mittleren Position hängt damit linear von n ab während die Standardabweichung in der Größenordnung von $\sqrt{n}$ liegt.

Insbesondere gilt für die symmetrische Irrfahrt

$$\mathbb{E}[S_n] = S_0 \tag{5.5}$$
$$\mathrm{Var}[S_n] = n. \tag{5.6}$$

Der Einfachheit halber betrachten wir den Fall $S_0 = 0$. Typische Fragen die man sich stellt sind: Gegeben $a, b \in \mathbb{N}$, erreicht die Irrfahrt $-a$ bevor sie b erreicht? Wie lange dauert es bis das passiert? Wird die Irrfahrt stets zu ihrem Ausgangspunkt zurückkehren? Wir kommen später auf diese Fragen zurück.

Eine besonders wichtige Eigenschaft der symmetrischen Irrfahrt ist, ein Martingal zu sein;

$$\mathbb{E}[S_{n+1}|\mathcal{F}_n] = S_n + \mathbb{E}[X_{n+1}|\mathcal{F}_n] = S_n + \mathbb{E}[X_{n+1}] = S_n. \tag{5.7}$$

Offenbar war hier für die Martingaleigenschaft nicht die gesamte Verteilung der X_n wichtig, sondern nur dass $\mathbb{E}[X_n] = 0$.

Im Fall der nicht symmetrischen Irrfahrt ist $\mathbb{E}S_n = n(p-q)$ und entsprechend ist S_n selbst kein Martingal. Allerdings erhalten wir, dass $M_n = S_n - (p-q)$ sehr wohl ein Martingal ist.

Beispiel 5.26 (Doob'sches Martingal)**.** Sei X eine integrierbare Zufallsvariable und $(\mathcal{F}_n)_n$ eine Filtration. Dann sieht man aus der Turmeigenschaft leicht, dass

$$M_n := \mathbb{E}[X|\mathcal{F}_n], \quad n \geq 1$$

ein Maringal ist.

Eine Operation die uns oft zu neuen Sub- bzw. Supermartingalen führt ist das anwenden von konvexen oder konkaven Funktionen. Mittels der bedingten Version der Jensen'schen Ungleichung sieht man nämlich leicht:

Lemma 5.27. *Sei M ein Martingal. Ist $f : \mathbb{R} \to \mathbb{R}$ konvex und $f(M_n), n \geq 1$ integrierbar, so ist $(f(M_n))_n$ ein Submartingal. Ist M selbst nur ein Submartingal, so ist $(f(M_n))_n$ jedenfalls ein Submartingal falls f monoton wachsend und konvex ist.*

Ist S_n wiederum die symmetrische Irrfahrt, so folgt, dass $Y_n := S_n^2$ ein Submartingal definiert welches wegen $\mathbb{E}Y_n = n$ offenbar kein Martingal ist. Wir erhalten jedoch wieder ein Martingal wenn wir den Prozess

$$M_n := S_n^2 - n$$

betrachten. Es gilt nämlich

$$\mathbb{E}[M_{n+1}|F_n] = \tag{5.8}$$
$$\mathbb{E}[(S_n + X_{n+1})^2 - n - 1|F_n] = \tag{5.9}$$
$$\mathbb{E}[S_n^2 + 2S_nX_{n+1} + X_{n+1}^2 - n - 1|F_n] = M_n. \tag{5.10}$$

Martingale spielen in vielen Anwendungen der Wahrscheinlichkeitstheorie eine große Rolle, insbesondere weil sie unser Modell für faire Spiele sind. Um das zu formalisieren führen wir den Begriff der Spielstrategie ein:

Definition 5.28. *Ein Prozeß beschränkter Zufallsvariabler $(H_n)_{n=1}^{\infty}$ heißt* Strategie / Handelsstrategie / previsibel *falls H_n für jedes n aus der Information von $\mathcal{F}_{n-1}$ bestimmt werden kann, d.h. falls H_n messbar bezüglich $\mathcal{F}_{n-1}$ ist.*

Sei $(S_n)_{n=0}^{\infty}$ ein weiterer Prozess. Die der Strategie $(H_n)_{n=1}^{\infty}$ zugehörige Auszahlung ist gegeben durch

$$(H \cdot S)_n := \sum_{k=1}^{n} H_k(S_k - S_{k-1}). \tag{5.11}$$

Wir können uns vorstellen, dass $(H \cdot S)_n$ den Gewinn oder Verlust beschreibt, den man macht, wenn man an jedem Tag $k-1$ genau H_k Aktien zum Preis S_{k-1} kauft, um sie am Tag k dann zum Preis S_k zu verkaufen.

Ist $(M_n)_{n=0}^{\infty}$ ein Martingal (und jedes H_n beschränkt und previsibel), so gilt

$$\mathbb{E}[H_k(M_{k+1} - M_k)|\mathcal{F}_k] = \tag{5.12}$$
$$\mathbb{E}[H_k(M_{k+1} - M_k)|\mathcal{F}_k] = \tag{5.13}$$
$$H_k\mathbb{E}[M_{k+1} - M_k|\mathcal{F}_k] = 0, \tag{5.14}$$

also auch $\mathbb{E}[H_k(M_{k+1} - M_k)] = 0$.

Insbesondere gilt daher für jedes Martingal

$$\mathbb{E}[(H \cdot M)_n] = 0. \tag{5.15}$$

Das bedeutet, egal mit welcher Strategie wir handeln, im Durchschnitt werden wir weder Gewinn noch Verlust machen. In diesem Sinn ist ein Martingal ein *faires Spiel*.

Tatsächlich kann man auch (unschwer) zeigen, dass (5.15) charakteristisch für Martingale ist, d.h. wenn (5.15) für jede (beschränkte) Strategie $(H_n)_{n=0}^{\infty}$ erfüllt ist, dann ist $(M_n)_{n=0}^{\infty}$ ein Martingal (Übung).

5.6 Stoppzeiten

Definition 5.29. *Eine Stoppzeit ist eine Zufallsvariable $\tau : \Omega \to \mathbb{N} \cup \{\infty\}$ sodass für alle $n \in \mathbb{N}$*

$$\{\tau \leq n\} \in \mathcal{F}_n.$$

Folgendes Lemma ist leicht zu sehen:

Lemma 5.30. *Eine Zufallsvariable $\tau : \Omega \to \mathbb{N} \cup \{\infty\}$ ist genau dann eine Stoppzeit, wenn für alle $n \in \mathbb{N}$*

$$\{\tau = n\} \in \mathcal{F}_n.$$

Beispiel 5.31. Falls $\tau = k$, wobei k eine feste natürliche Zahl ist, so ist τ offenbar eine Stoppzeit, unabhängig von der betrachteten Filtration.

Gegeben eine Filtration $(\mathcal{F}_n)$, $m < k$ und $A \in \mathcal{F}_k$, ist

$$\tau(\omega) := \begin{cases} m & \text{falls } \omega \in A \\ k & \text{falls } \omega \in A^c \end{cases}$$

eine Stoppzeit.

Gegeben ein adaptierter Prozess (X_n) und eine Borelmenge $B \subseteq \mathbb{R}$, so ist

$$\tau_B := \inf\{n : X_n \in B\}$$

eine Stoppzeit. (Übung)

Definition 5.32 (gestoppter Prozeß)**.** *Sei $X = (X_n)_{n=0}^{\infty}$ ein adaptierter Prozess und τ eine Stoppzeit. Dann ist der gestoppte Prozess $X^\tau = (X_n^\tau)_{n=0}^{\infty}$ definiert durch*

$$X_n^\tau(\omega) := X_{n\wedge\tau(\omega)}(\omega).$$

Falls die Stoppzeit τ f.s. endlich ist (also $\mathbb{P}(\tau < \infty) = 1$, so definieren wir die Zufallsvariable X_τ durch

$$X_\tau(\omega) := \begin{cases} X_{\tau(\omega)}(\omega) & \text{falls } \tau(\omega) \in \mathbb{N} \\ 0 & \text{falls } \tau(\omega) = \infty \end{cases}.$$

Falls τ f.s. endlich ist, gilt offenbar

$$X_\tau = \lim_{n\to\infty} X_n^\tau.$$

Wir betrachten die *buy-and-hold* Strategie

$$H_k^\tau := \begin{cases} 1 & \text{falls } k \le \tau \\ 0 & \text{sonst.} \end{cases}$$

Für eine Stoppzeit τ gilt dann

$$\{H_k^\tau = 0\} = \{\tau = 0\} \cup \ldots \cup \{\tau = k-1\} \in \mathcal{F}_k,$$

d.h. H^τ ist previsibel. Weiters gilt offenbar für einen stochastischen Prozeß X dass

$$X_n^\tau = X_0 + (H^\tau \cdot X)_n.$$

Indem wir auf beiden Seiten den Erwartungswert anwenden erhalten wir damit:

Lemma 5.33. *Sei X ein Martingal und τ eine Stoppzeit. Dann gilt für alle $n \in \mathbb{N}$*

$$\mathbb{E}X_n^\tau = \mathbb{E}X_0.$$

Eine wichtige Frage ist, wann wir bei obiger Gleichung n gegen ∞ gehen lassen dürfen. Der nächste Satz nennt hinreichende Bedingungen dafür:

Satz 5.34 (Doob's Optional Stopping Theorem)**.** *Sei τ eine Stoppzeit und X ein Martingal. Dann ist X_τ integrierbar und $\mathbb{E}X_\tau = \mathbb{E}X_0$ in jeder der folgenden Situationen:*

1. *τ ist beschränkt, d.h. es gibt ein $c \in \mathbb{R}_+$ mit $\tau \le c$.*
2. *X ist beschränkt, d.h. es gibt ein $c \in \mathbb{R}_+$ sodas $\sup_n |X_n| \le c$ und τ ist fast sicher endlich.*
3. *$\mathbb{E}\tau < \infty$ und X hat beschränkte Zuwächse, d.h. es gibt ein $c \in \mathbb{R}_+$ mit $|X_{n+1} - X_n| \le c$ für alle n.*

Beweis. Unter den beiden ersten Bedingungen ist der Beweis jeweils recht einfach.

Beweis von (3): Die Folge X_n^τ erfüllt

$$|X_n^\tau| \leq |X_0| + c\tau$$

und die Zufallsvariable auf der rechten Seite ist integrierbar, also ist der Satz von der dominierten Konvergenz anwendbar. □

Satz 5.35. *Seien $a, b \in \mathbb{N}$ und S_n die symmetrische Irrfahrt mit Start in 0. Die Irrfahrt trifft fast sicher irgendwann auf $-a$ oder b. Für die Zeit $\tau_{a,b}$ die bis dahin vergeht gilt*

$$\mathbb{E}\tau_{a,b} = ab.$$

Die Irrfahrt trifft mit Wahrscheinlichkeit $\frac{b}{a+b}$ auf $-a$ bevor sie auf b trifft.

Beweis. Dass

$$\tau_{a,b} := \min\{n : S_n \in \{-a, b\}$$

endlich ist, könnte eine Übungsaufgabe sein.

Dann gilt

$$0 = \mathbb{E}S_0 = \mathbb{E}S_\tau = (-a)\mathbb{P}(\text{Irrfahrt auf } a \text{ getroffen}) + b\mathbb{P}(\text{Irrfahrt auf } b \text{ getroffen}).$$

Schreiben wir p_a für die Wahrscheinlichkeit zuerst auf $-a$ zu treffen und p_b für die Wahrscheinlichkeit zuerst auf b zu treffen so gilt daher

$$0 = p_a * (-a) + p_b * b,$$

woraus (wegen $p_a + p_b = 1$) $p_a = b/(a+b)$ folgt.

Wir haben oben schon gesehen, dass auch $M_n = S_n^2 - n$ ein Martingal ist und es gilt wiederum

$$0 = \mathbb{E}M_0 = \mathbb{E}[M_{n\wedge\tau_{a,b}}] = \mathbb{E}[S^2_{n\wedge\tau_{a,b}}] - \mathbb{E}[n \wedge \tau_{a,b}] \tag{5.16}$$

$$\to a^2\mathbb{P}(\text{Irrfahrt auf } -a \text{ getroffen}) + b^2\mathbb{P}(\text{Irrfahrt auf } b \text{ getroffen}) - \mathbb{E}[\tau_{a,b}], \tag{5.17}$$

wenn wir n gegen ∞ gehen lassen (und einmal dominierte und einmal monotone Konvergenz verwenden). Wir erhalten also $\mathbb{E}[\tau_{a,b}] = a^2b/(a+b) + ab^2/(a+b) = ab$. □

Indem wir b gegen unendlich gehen lassen, erhalten wir aus obigem Satz leicht, dass $\mathbb{P}[\tau_{a,\infty} < \infty] = 1$, aber $\mathbb{E}[\tau_{a,\infty}] = \infty$.

5.7 Konvergenz von Martingalen

Gegeben ein stochastischer Prozess $(M_n)_n$ betrachten wir eine "buy low, sell high" Strategie: Wir stellen uns vor, dass M_n den Preis einer Aktie am Tag n beschreibt (und erlauben auch negative Preise). Immer wenn M_n unter die Null gefallen ist, kaufen wir eine Aktie. Sobald M_n auf Eins oder darüber gestiegen ist, verkaufen wir die Aktie wieder. Sei $(H_n)_n$ die entsprechende Strategie. Bei jedem *upcrossing* des Intervalls $[0,1]$ machen wir dann einen Euro Gewinn. Die Anzahl aller upcrossings bis zum Zeitpunkt n bezeichnen wir mir $U_n^{[0,1]}$. Tatsächlich ist unser Gewinn oder Verlust $(H^{[0,1]} \cdot M)_n$ bis zum Zeitpunkt n nicht unbedingt größer als $U_n^{[0,1]}$, da es auch passieren kann, dass M_n nachdem wir es unterhalb der 0 gekauft haben weiter an Wert verliert. Was aber jedenfalls richtig ist, ist die Ungleichung

$$U_n^{[0,1]} \leq (H^{[0,1]} \cdot M_n) + (M_n)_-, \tag{5.18}$$

wobei $(M_n)_-$ den Negativteil von M_n bezeichnet.

Wenn man das Argument mit dem Intervall $[a,b]$ wiederholt, erhält man

$$(b-a)U_n^{[0,1]} \leq (H^{[a,b]} \cdot M_n) + (M_n - a)_-,$$

Indem wir auf beiden Seiten Erwartungswerte nehmen, sehen wir.

Lemma 5.36 (Doob's upcrossing lemma)**.** *Sei M ein Supermartingal. Dann gilt*

$$(b-a)\mathbb{E}U_n^{[a,b]} \leq \mathbb{E}(M_n - a)_- \leq \mathbb{E}[|M_n| + |a|]].$$

Insbesondere impliziert $\sup \mathbb{E}[|M_n|] < \infty$ *dass*

$$\mathbb{E}[U_\infty^{[a,b]}] < \infty, \mathbb{P}[U_\infty^{[a,b]} = \infty] = 0.$$

Damit zeigen wir:

Satz 5.37 (Doob'scher Konvergenzsatz)**.** *Sei M ein Supermartingal mit* $\sup \mathbb{E}|M_n| < \infty$*. Dann existiert* $\lim_{n\to\infty} M_n$ *fast sicher.*

Beweis. Die Menge von $\omega \in \Omega$ auf der der Grenzwert nicht existiert können wir schreiben als

$$\{\liminf_n M_n < \limsup_n M_n\} = \bigcup_{a<b,a,b\in\mathbb{Q}} \{\liminf_n M_n < a < b < \limsup_n M_n\} \tag{5.19}$$

$$\subseteq \bigcup_{a<b,a,b\in\mathbb{Q}} \{U_\infty^{[a,b]} = \infty\}. \tag{5.20}$$

Nachdem rechts eine Vereinigung von abzählbar vielen Nullmengen steht, sind wir fertig. □

Falls M ein Submartingal ist, erhalten wir natürlich das selbe Resultat indem wir den obigen Satz auf $-M$ anwenden.

Wenn wir ein Supermartingal betrachten das von unten beschränkt ist, ist natürlich auch die Folge der absoluten Erwartungswerte beschränkt. Also erhalten wir:

Korollar 5.38. *Sei M ein Supermartingal dass von unten beschränkt ist. Dann existiert $M_\infty = \lim_{n\to\infty} M_n$ fast sicher.*

Wir haben oben schon besprochen, dass für die einfache symmetrische Irrfahrt S_n die Treffzeit von (-1) fast sicher endlich ist. Der Prozeß $Y_n := S_n^\tau, n \geq 1$ ist ein von unten beschränktes Martingal und $S_\tau = \lim_{n\to\infty} S_\tau^n = -1$ bzw. $Y_\infty = \lim_{n\to\infty} Y_n$ fast sicher. Allerdings gilt in dieser Situation *nicht* $\mathbb{E}Y_0 = \mathbb{E}Y_\infty$.

Im nächsten Kapitel wollen wir noch ein bisschen besser verstehen, unter welchen Umständen wir sicherstellen können, dass diese Gleichung erfüllt ist.

5.8 Uniform intergrierbare Martingale

Ist eine Zufallsvariable X integrierbar, so sieht man leicht, dass

$$\lim_{k\to\infty} \int_{|X|\geq k} |X|\, d\mathbb{P} \to 0.$$

Eine Folge von Zufallsvariablen (X_n) heißt uniform integrierbar falls

$$\lim_{k\to\infty} \sup_n \int_{|X_n|\geq k} |X_n|\, d\mathbb{P} \to 0. \tag{5.21}$$

Bemerkung 5.39. Aus der uniformen Integrierbarkeit folgt, dass $\sup_n \mathbb{E}[|X_n|] < \infty$, aber diese Bedingung ist nicht hinreichend.

Eine Folge von Zufallsvariablen (X_n), ist uniform integrierbar genau dann wenn $\sup_n \mathbb{E}[|X_n|] < \infty$ und

$$\lim_{\varepsilon\to 0} \sup_n \sup_{B:\mathbb{P}(B)\leq\varepsilon} \int_B |X_n|\, d\mathbb{P} \to 0. \tag{5.22}$$

Das könnte eine Übungsaufgabe sein.

Eine Folge von Zufallsvariablen ist jedenfalls dann uniform integrierbar, wenn eine der folgenden Bedingungen erfüllt ist:

1. Es gibt eine integrierbare dominierende Funktion, d.h. es gibt eine Zufallsvariable Y mit $\mathbb{E}Y < \infty$ und $Y \geq |X_n|$ für alle $n \in \mathbb{N}$.

2. Die Folge is in L_2 beschränkt. Dann gilt nämlich für $k \to \infty$ dass

$$\sup_n \int_{|X_n|>k} |X_n|\, d\mathbb{P} \leq 1/k \sup_n \int X^2\, d\mathbb{P} \to 0.$$

   Allgemeiner sieht man so, dass Beschränktheit in L_p für ein $p > 1$ ausreicht um uniforme Integrierbarkeit zu zeigen.

3. Es gibt eine integrierbare Zufallsvariable X sodass $X_n = \mathbb{E}[X|\mathcal{F}_n]$ für eine Folge von σ-Algebren $\mathcal{F}_1, \mathcal{F}_2, \ldots$ Dann gilt nämlich nach der Definition der bedingten Erwartung, der bedingten Jensenungleichung und der Markovungleichung

$$\int_{|X_n|\geq k} |X_n|\, d\mathbb{P} = \int_{|X_n|\geq k} |\mathbb{E}[X|F_n]|\, d\mathbb{P} = \int_{|X_n|\geq k} |X|\, d\mathbb{P} \leq \sup_{B:\mathbb{P}(B)<\mathbb{E}|X|/k} \int_B |X| d\mathbb{P}.$$

Die rechte Seite konvergiert für $k \to \infty$ gegen 0, natürlich gleichmäßig in n.

Mit dem Konzept der uniformen Integrierbarkeit können wir den Satz von der dominierten Konvergenz verallgemeinern:

Satz 5.40. *Angenommen $X_n \to X$ fast sicher und die Folge $(X_n)_n$ ist uniform integrierbar. Dann gilt auch*

$$\lim_n \mathbb{E}|X_n - X| = 0,$$

(d.h. $X_n \to X$ in L^1) und natürlich $\lim_n \mathbb{E}X_n = \mathbb{E}X$.

Beweis. Wegen dem Lemma von Fatou und $\sup_n \mathbb{E}|X_n| < \infty$ ist X integrierbar. Daraus folgert man unschwer, dass auch $Y_n := |X_n - X|, n \in \mathbb{N}$ uniform integrierbar ist.

Wir müssen noch zeigen, dass $\lim_n \mathbb{E}Y_n = 0$. Wegen der uniformen Integrierbarkeit der $Y_n, n \in \mathbb{N}$, wird $\sup_n \int_{Y_n \geq k} Y_n\, d\mathbb{P}$ für große k beliebig klein. Andererseits gilt $\lim_{n\to\infty} \mathbb{E}Y_n \wedge k = 0$ für jedes k nach dem Satz über dominierte Konvergenz. □

Satz 5.41. *Sei $(M_n)_n$ ein uniform integrierbares Martingal. Dann existiert*

$$M_\infty := \lim_n M_n$$

fast sicher und in L^1. Außerdem gilt

$$M_n = \mathbb{E}[M_\infty|F_n], n \geq 1]. \tag{5.23}$$

Umgekehrt ist jedes Martingal mit einer ("Doob'schen") Darstellung $M_n = \mathbb{E}[X|\mathcal{F}_n], n \geq 1$ uniform integrierbar und es gilt $M_n \to \mathbb{E}[X|\mathcal{F}_\infty]$ für $\mathcal{F}_\infty = \sigma\left(\bigcup_n \mathcal{F}_n\right)$.

Das heißt wir können jedes uniform integrierbare Martingal "abschließen" indem wir ihm M_∞ als letztes Element hinzufügen.

Beweis. Wegen der uniformen Integrierbarkeit ist $\sup_n \mathbb{E}|M_n| < \infty$ und die fast sichere Konvergenz $M_\infty = \lim_n M_n$ folgt aus dem Doob'schen Konvergenzsatz. Im letzten Satz haben wir auch gesehen, dass wegen uniformer Integrierbarkeit dann auch L^1-Konvergenz gilt.

Wir zeigen noch, dass $M_n = \mathbb{E}[M_\infty|\mathcal{F}_n]$ fast sicher. Für $r \geq n$ und $B \in \mathcal{F}_n$ liefert die Martingaleigenschaft dass

$$\int_B M_r\, d\mathbb{P} = \int_B M_n\, d\mathbb{P}.$$

Andererseits gilt

$$\left| \int_B M_r \, d\mathbb{P} - \int_B M_\infty \, d\mathbb{P} \right| \leq \int_B M_r - M_\infty \, d\mathbb{P} \leq \mathbb{E}|M_r - M_\infty| \to 0$$

für $r \to \infty$. Daher muß $\int_B M_n \, d\mathbb{P} = \int_B M_\infty \, d\mathbb{P}$ gelten.

Dass die Darstellung $M_n = \mathbb{E}[X|\mathcal{F}_n], n \geq 1$ uniforme Integrierbarkeit impliziert, haben wir oben schon gesehen. Setze $M_\infty := \lim_n M_n$. (Wegen uniformer Integrierbarkeit ist der punktweise Grenzwert natürlich der L^1-Grenzwert.) Wir müssen noch sehen überlegen, dass

$$M_\infty = \mathbb{E}[X|\mathcal{F}_n].$$

Falls $A \in \mathcal{F}_n$ für ein $n \in \mathbb{N}$, gilt natürlich $\mathbb{E}[I_A M_\infty] = \mathbb{E}[I_A X]$. Weil $\bigcup_n \mathcal{F}_n$ ein Durchschnittsstabiler Erzeuger von $\mathcal{F}_\infty = \sigma\left(\bigcup_n \mathcal{F}_n\right)$ ist, gilt die Gleichheit auch auf $\mathcal{F}_\infty$. □

Besonders einfach lässt sich ein Martingal verstehen falls (M_n) in L^2 beschränkt bleibt. Grund dafür ist die fundamentale Orthogonalitätsrelation

$$\mathbb{E}[(M_m - M_n)Y] = 0$$

für $k \leq n \leq m$ und $Y \in L^2(\Omega, \mathcal{F}_k, \mathbb{P})$. Daraus folgt nämlich mit Pythagoras

$$\|M_n - M_0\|_2^2 = \sum_{k=1}^{n} \|M_n - M_0\|_2^2.$$

Damit sehen wir:

Satz 5.42. *Sei $(M_n)_n$ ein Martingale. Falls $\sup_n \mathbb{E}M_n^2 < \infty$, existiert*

$$M_\infty := \lim_{n \to \infty} M_n$$

fast sicher und in L^2 (und es gilt $M_n := \mathbb{E}[M_\infty|\mathcal{F}_n]$).

Beweis. Die L^2-Konvergenz erhält man leicht aus der Vollständigkeit von L^2. Nur für die punktweise Konvergenz brauchen wir den Doob'schen Konvergenzsatz. □

5.9 Kolmogorovs 0-1-Gesetz

Satz 5.43. *Seien $(X_n)_{n \geq 1}$ unabhängige Zufallsvariablen. Definieren wir $\mathcal{G}_n = \sigma(X_n, X_{n+1}, \ldots)$. Für jedes Ereignis A in der terminalen σ-Algebra*

$$\mathcal{G}_\infty := \cap_{n \geq 1} \mathcal{G}_n$$

gilt: Die Wahrscheinlichkeit $P(A)$ ist entweder 0 oder 1.

Beweis. Sei $(\mathcal{F}_n)_n$ die von $X_1, X_2, \ldots$ erzeugte Filtration Wir betrachten die Zufallsvariablen $Y_n = \mathbb{E}[I_A|\mathcal{F}n]$. Nach dem Martingalkonvergenzsatz gibt es ein Y sodass $Y_n \to Y$ fast sicher und in L^1 und es ist $Y = \mathbb{E}[I_A|\mathcal{F}_\infty] = I_A$.

Für jedes $n \in N$ gilt wegen $A \in \sigma(X_{n+1}, X_{n+2}, \ldots)$ und der Unabhängigkeit der $X_1, X_2, \ldots$

$$Y_n = \mathbb{E}[I_A|\mathcal{F}_n] = \mathbb{E}[I_A] = \mathbb{P}(A).$$

Daraus folgt $I_A = \mathbb{P}(A)$ fast sicher, also $\mathbb{P}(A) \in \{0, 1\}$. □

Wenn alle Elemente einer σ-Algebra Wahrscheinlichkeit 0 oder 1 habe, so sind auch alle bezüglich dieser σ-Algebra meßbaren Funktionen konstant (Übung). In der Situation des 0-1-Gesetzes erhalten wir daraus z.B. dass $\limsup_n X_n$ fast sicher konstant ist.

5.10 Starkes Gesetz der großen Zahlen unter Momentenbedingungen

Mittels der Martingalkonvergenzsätze von oben ist es reicht leicht eine Version des Gesetzes der großen Zahlen zu beweisen.

Wir brauchen dazu noch zwei elementare Lemmata aus der Analysis.

Lemma 5.44 (Cesaro)**.** *Angenommen eine Folge reeller Zahlen erfüllt $v_n \to v_\infty$, dann gilt auch*

$$\frac{v_1 + \ldots + v_n}{n} \to v_\infty.$$

Beweis. Sei $\varepsilon > 0$. Wenn wir N groß genug wählen ist $v_n \geq v_\infty - \varepsilon$ für alle $n \geq N$. Daher gilt

$$\liminf_n \frac{v_1 + \ldots + v_n}{n} \geq \liminf_n \frac{(n-N)(v_\infty - \varepsilon)}{n} = v_\infty - \varepsilon.$$

Analog erhält man $\limsup_n v_n \leq v_\infty + \varepsilon$. □

Lemma 5.45 (Kronecker)**.** *Seien $x_n, n \geq 1$ reelle Zahlen sodass*

$$u_n := \sum_{k=1}^{n} \frac{x_k}{k} \to u_\infty.$$

Dann gilt

$$\frac{s_n}{n} \to 0, \quad \textit{wobei } s_n := x_1 + \ldots + x_n.$$

Beweis. Wegen $u_n - u_{n-1} = \frac{x_n}{n}$ haben wir

$$s_n = \sum_{k=1}^{n} k(u_k - u_{k-1}) = nu_n - \sum_{k=1}^{n} u_k.$$

Wenn wir beide Seiten durch n dividieren können wir nach dem vorigen Lemma (Cesaro) schließen, dass $\frac{s_n}{n} \to u_\infty - u_\infty = 0$. □

Aus dem Lemma von Kronecker und den Martingalkonvergenzsätzen aus dem letzten Kapitel erhalten wir nun leicht die folgende Version des Gesetzes der großen Zahlen:

Satz 5.46 (Starkes Gesetz der großen Zahlen unter Momentenbedingungen)**.** *Sei $W_1, W_2, \ldots$ eine Folge von unabhängigen Zufallsvariablen mit $\mathbb{E}W_i = 0, i \in \mathbb{N}$ und*

$$\sum_{n=1}^{\infty} \frac{\operatorname{Var} W_n}{n^2} < \infty.$$

Dann gilt fast sicher

$$\lim_{n\to\infty} \frac{W_1 + \ldots + W_n}{n} = 0.$$

Beweis. Unter den Voraussetzungen des Satzes definiert $M_n := W_1/1 + \ldots + W_n/n, n \geq 1$ ein Martingal dass in L^2 beschränkt ist. Insbesondere existiert

$$\lim_n (W_1/1 + \ldots + W_n/n)(\omega)$$

für fast alle $\omega \in \Omega$. Aus dem Lemma von Kronecker folgt dann das Gewünschte. □

5.11 Klassische Variante des starken Gesetzes der großen Zahlen

Die klassische Variante des Gesetzes der großen Zahlen besagt:

Satz 5.47. *Sei $X_1, X_2, \ldots$ eine i.i.d. Folge von integrierbaren Zufallsvariablen mit $\mathbb{E}X_i = \mu$. Dann gilt fast sicher*

$$\lim_n \frac{S_n}{n} = \mu, \quad \text{wobei } S_n := X_1 + \ldots + X_n.$$

Wir bringen hier das Argument von [1]. Zur Vorbereitung benötigen wir ein paar Lemmata. Insbesondere das erste ist auch für sich genommen interessant:

Lemma 5.48. *Seien $Y_1, Y_2, \ldots$ i.i.d. mit $\mathbb{E}Y_i > 0$ und $Y_i \leq C$ für ein $C \in \mathbb{R}$ und $S_n := Y_1 + \ldots + Y_n$. Sei $b \geq 0$. Dann erfüllt $\tau := \inf\{n : S_n > b\}$ dass $\mathbb{E}\tau < \infty$.*

Beweis. Um das zu zeigen, betrachten wir das Martingal

$$M_n := S_n - \mathbb{E}[Y_1]n.$$

Dann gilt natürlich $0 = \mathbb{E}[M_{n\wedge\tau}]$ also

$$\mathbb{E}[Y_1]\mathbb{E}[n \wedge \tau] = \mathbb{E}[S_{n\wedge\tau}] \leq C + b.$$

Weil die rechte Seite beschränkt bleibt, muß auch die linke Seite für $n \to \infty$ beschränkt bleiben, daher gilt jedenfalls $\mathbb{P}(\tau < \infty) = 1$. Mit dem Satz von der monotonen Konvergenz folgt nun $\mathbb{E}[n \wedge \tau] \leq (C + b)/\mathbb{E}[Y_1]$. □

Lemma 5.49. *Seien $Y_1, Y_2, \dots$ i.i.d. mit $\mathbb{E}Y_i > 0$. Dann gilt fast sicher*

$$\inf_n Y_1 + \ldots + Y_n > -\infty$$

Beweis. Gegeben potentiell unbeschränkte Y_i mit $\mathbb{E}Y_i > 0$, so gibt es nach dem Satz von der monotonen Konvergenz ein $C > 0$, sodass auch $\mathbb{E}Y_i \wedge C > 0$. Natürlich folgt aus $\inf(Y_1 \wedge C) + \ldots + (Y_n \wedge C) > 0$ dann auch $\inf Y_1 + \ldots + Y_n > -\infty$. Daher können wir oBdA annehmen, dass $Y_i \leq C$.

Wir setzen $S_n := Y_1 + \ldots + Y_n$. Für jedes n gilt

$$(0 = S_0, S_1, \dots, S_n) \sim (S_n - S_n, S_n - S_{n-1}, \dots, S_n - S_0),$$

d.h. beide Zufallsvektoren haben die selbe Verteilung.

Sei wieder $\tau := \inf\{n : S_n > 0\}$. Weiters nennen wir n einen (schwachen) Negativrekord der Folge $S_0, S_1, \dots, S_n$ falls

$$S_n = S_0 \wedge S_1 \wedge \ldots \wedge S_n.$$

Anders ausgedrückt: n ist ein (schwacher) Negativrekord, wenn die Werte $S_n - S_n, S_n - S_{n-1}, \dots, S_n - S_0$ alle ≤ 0 erfüllen. Daher erhalten wir

$$\mathbb{P}(\tau > n) = \mathbb{P}(n \text{ ist ein (schwacher) Negativrekord}).$$

Indem wir über alle n summieren, erhalten wir

$$\mathbb{E}[\tau] = \sum_n \mathbb{P}(\tau > n) = \mathbb{E}|\{n : n \text{ ist ein (schwacher) Negativrekord}\}|.$$

$\inf_n S_n = -\infty$ kann natürlich nur gelten, wenn es unendlich viele Negativrekorde gibt. Daher reicht es zu zeigen, dass $\mathbb{E}[\tau] < \infty$. □

Beweis von Satz 5.47. Sei $\varepsilon > 0$ und $Y_n := X_n - \mu + \varepsilon$ für $n \in \mathbb{N}$ sodass $\mathbb{E}Y_n > 0$. Aus

$$\inf_n Y_1 + \ldots + Y_n > 0,$$

folgt dann

$$\liminf_n (Y_1 + \ldots + Y_n)/n \geq 0$$

fast sicher, also auch $\liminf_n (X_1 + \ldots + X_n)/n \geq \mu - \varepsilon$.

Analog erhalten wir $\limsup_n (X_1 + \ldots + X_n)/n \leq \mu + \varepsilon$. □

5.12 Skizze eines weiteren Beweises des starken Gesetzes der großen Zahlen

Wir skizzieren einen weiteren Beweis von Satz 5.47.

Wir betrachten für $n \in \mathbb{N}$ die σ-Algebra

$$\mathcal{G}_{-n} := \sigma(S_n, S_{n+1}, S_{n+2}, \dots) = \sigma(S_n, X_{n+1}, X_{n+2}).$$

und setzen

$$Y_{-n} := \mathbb{E}[X_1|\mathcal{G}_{-n}].$$

Dann ist

$$\ldots, Y_{-3}, Y_{-2}, Y_{-1} \text{ ein Martingal bezüglich der Filtration } \ldots \subseteq \mathcal{G}_{-3} \subseteq \mathcal{G}_{-2} \subseteq \mathcal{G}_{-1}.$$

Die Familie von Zufallsvariablen $Y_{-n}, n \in \mathbb{N}$ ist uniform integrierbar und mit dem selben Beweis wie für den Doob'schen Konvergenzsatz erhält man, dass

$$Y_{-\infty} := \lim_{n\to\infty} Y_{-n}$$

fast sicher und in L^1 existiert. ("Doob'scher Konvergenzsatz für Rückwärtsmartingale")

Weiters gilt $\mathbb{E}[X_1|\mathcal{G}_{-n}] = \mathbb{E}[X_1|S_n]$, weil $X_{n+1}, X_{n+2}, \ldots$ keine weitere Information über X_1 liefern. Nach einem Satz aus der Maßtheorie gibt es in diesem Fall eine P_{S_n} eindeutig bestimmte Funktion $f : \mathbb{R} \to \mathbb{R}$ sodass

$$\mathbb{E}[X_1|S_n] = f(S_n).$$

Aus Symmetriegründen gilt dann natürlich auch $\mathbb{E}[X_i|S_n] = f(S_n)$ für $i = 2, \ldots, n$. Damit folgt

$$nf(S_n) = \sum_{i=1}^{n} \mathbb{E}[X_i|S_n] = \mathbb{E}[S_n|S_n] = S_n,$$

also $f(x) = x/n$ und somit

$$Y_{-n} = \mathbb{E}[X_1|\mathcal{G}_{-n}] = \frac{S_n}{n}.$$

Zusammen erhalten wir also, dass

$$\lim_{n\to\infty} \frac{S_n}{n} = Y_{-\infty}$$

fast sicher und in L^1. Weil $\lim_{n\to\infty} \frac{S_n}{n} \in \sigma(X_k, X_{k+1}, \ldots)$ für alle $k \in \mathbb{N}$ erhalten wir schließlich aus dem 0-1-Gesetz von Kolmogorov, das $Y_{-\infty}$ fast sicher konstant ist.

5.13 Maximalungleichungen

Wir haben oben gesehen, dass Martingale eine Reihe von nützlichen Eigenschaften haben, sie sind *faire Spiele* (d.h. $\mathbb{E}[(H \cdot X)_n] = 0$ für alle previsiblen (H)), sie erfüllen das optional stopping theorem $\mathbb{E}X_{\tau\wedge n} = \mathbb{E}X_0$ und haben starke Konvergenzeigenschaften.

Zum Abschluss des Kapitels erwähnen wir noch knapp *Martingalungleichungen*. Grob gesagt, erlauben Sie, das Verhalten eines Martingals durch dessen Werte zu Endzeitpunkt zu kontrollieren.

Exemplarische für diese große Klasse von Ungleichungen erwähnen wir die Doob'sche L^2-Ungleichung:

Satz 5.50. *Sei $(S_n)_{n=0}^N, S_0 = 0$ ein nicht-negatives Submartingal und $\bar{S}_k := \max_{i\le k} S_i$ das* laufende Maximum. *Dann gilt*

$$\mathbb{E}\bar{S}_N^2 \le 4\mathbb{E}S_N^2.$$

Beweisskizze. Gegeben reelle Zahlen $s_0, \dots, s_N$ gilt

$$\bar{s}_N^2 + 4\sum_{n=1}^{N} \bar{s}_{n-1}(s_n - s_{n-1}) \le 4s_N^2.$$

Wir beweisen diese Ungleichung nicht, aber wir betonen, dass das Argument elementar ist: es genügt die Terme umzuordnen und Quadrate zu ergänzen. Die Bedeutung der Ungleichung liegt darin, dass sie uns direkt die gewünschte Martingalungleichung liefert: indem wir die Ungleichung auf die Pfade des Martingals anwenden und Erwartungswerte nehmen erhalten wir nämlich

$$\mathbb{E}\bar{S}_N^2 + 4 \underbrace{\mathbb{E}\sum_{n=1}^{N} \bar{S}_{n-1}(S_n - S_{n-1})}_{\ge 0 \text{ wegen Submartingaleigenschaft}} \le 4\mathbb{E}\bar{S}_N^2.$$

□

5.14 Zusammenfassung

In diesem Kapitel haben wir den Begriff der bedingten Erwartung kennengelernt. Er ermöglicht uns eine zentrale Klasse stochastischer Prozesses zu definieren, die der *Martingale*. Schon mit einfachen Mitteln kann man Eigenschaften von Martingalen beweisen, die weitreichende Konsequenzen für die Stochastik haben. Exemplarisch haben wir das 0-1-Gesetz von Kolmogorov und verschiedene Beweise des Gesetzes der großen Zahlen kennengelernt.

6 Stetige Zufallsvariablen

6.1 Dichtefunktion, Beispiele

Bisher haben wir nur Zufallsvariablen betrachtet, die endlich oder abzählbar unendlich viele Werte annehmen. Offensichtlich ist das nicht genug für viele Zufallsexperimenten: betrachten Sie z.B. eine Person, die Dartpfeile wirft, oder die Bewegung eines mikroskopischen Teilchens in einer Flüssigkeit. In diesem Kapitel entwickeln wir die Theorie, sodass Zufallsvariablen überabzählbar unendlich viele Werte annehmen können. Zur Vereinfachung beschränken wir uns auf den Fall von Zufallsvariablen, die Werte in $\mathbb{R}^d$ annehmen.

Definition 6.1. *Sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. Wir sagen dass $X : \Omega \to \mathbb{R}^d$ (wobei $d \geq 1$) eine **stetige Zufallsvariable** ist, wenn es eine (stückweise stetige) Funktion $f : \mathbb{R}^d \to [0, \infty)$ gibt, sodass*

$$\mathbb{P}(X \in A) = \int_A f(x)dx = \int_{\mathbb{R}^d} f(x)1_A(x)dx$$

für jedes $A \subset \mathbb{R}^d$ sodass das Integral wohldefiniert ist.
*Die Funktion f bezeichnen wir als die **Dichtefunktion** der Verteilung von X.*

Meistens haben wir $d = 1$, dann gilt insbesondere für alle $-\infty \leq a \leq b \leq \infty$

$$\mathbb{P}(X \in [a, b]) = \mathbb{P}(X \in (a, b)) = \int_a^b f(x)\, dx.$$

Bemerkung 6.2. Wir machen ein paar Bemerkungen.

- Bemerken Sie, dass für jedes x, $\mathbb{P}(X = x) = 0$ und deshalb für jede abzählbare Menge $A \subset \mathbb{R}^d$, $\mathbb{P}(X \in A) = 0$. Aber

  $$\mathbb{P}(X \in \mathbb{R}^d) = \mathbb{P}(\cup_{x \in \mathbb{R}^d}\{X = x\}) = 1 \neq 0,$$

  weil die Vereinigung $\bigcup_{x \in \mathbb{R}^d}\{X = x\}$, zwar disjunkt, nicht aber abzählbar ist.

- Die Dichtefunktion (auf Englisch "probability density function") ist das stetige Analogon zur Wahrscheinlichkeitsfunktion die wir im diskreten Fall betrachtet haben:

  $$\mathbb{P}(X \in [x, x + dx]) = f(x)dx + o(dx),$$

  während die Wahrscheinlichkeitsfunktion $p(x)$ einer diskreten Zufallsvariable Y

  $$\mathbb{P}(Y = x) = p(x)$$

  erfüllt.

  Bitte beachten Sie jedoch dass man für eine stetige Zufallsvariable nicht "$\mathbb{P}(X = x) = f(x)$" schreiben kann, das wäre ein Unfug.

- Beachten Sie auch folgendes: die Wahrscheinlichkeitsfunktion $p(x)$ einer diskreten Zufallsvariable Y induziert ein Wahrscheinlichkeitsmass auf der Menge $\mathcal{W}$, während die Dichtefunktion nur eine Funktion f ist, die folgendes erfüllt:

$$f(x) \geq 0 \quad ; \quad \int_{\mathbb{R}^d} f(x)dx = 1.$$

  Insbesondere ist es möglich dass $f(x) > 1$: was zählt ist dass $\int_A f(x)dx \leq 1$ für jede Menge $A \subset \mathbb{R}^d$ gilt.

Beispiel 6.3. Seien $a, b \in \mathbb{R}$ mit $a < b$. Wir sagen dass X **gleichverteilt** auf (a, b) ist, wenn

$$\mathbb{P}(X \in A) = \int_a^b 1_A(x)\frac{dx}{b-a}.$$

D.h., X ist stetig mit Dichtefunktion f:

$$f(x) = \begin{cases} \frac{1}{b-a}; & \text{wenn } x \in (a, b) \\ 0 & \text{sonst.} \end{cases}$$

Bemerken Sie, dass f stückweise stetige ist, wie gewünscht. Allgemeiner gilt: falls $D \subset \mathbb{R}^d$ endliches Volumen hat, können wir eine Zufallsvariable betrachten, die gleichverteilt auf D ist: in diesem Fall ist die Dichtefunktion $f(x) = 1/|D|$ auf D und 0 sonst. Das wäre viellecht geeignet um einE Spieler*in, die Dart spielt, zu modellieren.

Definition 6.4. *Die Verteilungsfunktion einer stetigen Zufallsvariable mit Werten in* $\mathbb{R}$ *ist die Funktion*

$$F : x \in (-\infty, \infty) \mapsto \mathbb{P}(X \leq x) = \int_{-\infty}^x f(t)dt.$$

Man kann leicht überprüfen, dass die Verteilungsfunktion F die folgenden Eigenschaften erfüllt:

- F ist wachsend
- $\lim_{x\to-\infty} F(x) = 0; \lim_{x\to\infty} F(x) = 1.$
- F ist rechtsstetig.

Die Verteilungsfunktion einer Zufallsvariable können wir immer definieren (sowohl im diskreten als auch im stetigen Fall). Im stetigem Fall ist F differenzierbar an jedem Stetigkeitspunkt von f und

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

Der Begriff der Verteilungsfunktion ist vor allem aus theoretischen Gesichtspunkten relevant. Man kann zum Beispiel beweisen (mit Hilfe von Masstheorie), dass jede Funktion F die die obigen Eigenschaften erfüllt, die Verteilungsfunktion einer Zufallsvariablen ist: das ist der sogenannte Lebesgue–Stieltjès Satz.

In dieser Vorlesung wird die Verteilungsfunktion keine große Rolle spielen. Äquivalent zu Verteilungsfunktion und manchmal ein bisschen nützlicher ist die "Tailfunktion". Sie ist definiert durch $\bar{F}(x) = \mathbb{P}(X > x) = 1 - F(x)$.

Beispiel 6.5. Wir sagen, dass X **exponentialverteilt** mit Parameter $\lambda \geq 0$ ist, wenn

$$\bar{F}(x) = e^{-\lambda x}; x \geq 0 \text{ (und 1 sonst).}$$

D.h., X ist stetig mit Dichtefunktion

$$f(x) = \lambda e^{-\lambda x}; x \geq 0 \text{ (und 0 sonst).}$$

Die Exponentialverteilung ist das stetige Analogon zur Geometrischen Verteilung bei der

$$\mathbb{P}(X > n) = (1-p)^n.$$

Man kann sich auch in diesem Fall vorstellen, dass X die Zeit bis zu einem ersten Erfolg darstellt, wobei es in jeder Zeiteinheit dt es eine Wahrscheinlichkeit ungefähr gleich λdt gibt, dass ein Erfolg eintritt. Das ist der Inhalt des nächsten elementaren Satzes:

Proposition 6.6. *Sei X exponentialverteilt mit Parameter $\lambda \geq 0$. Dann gilt*

$$\mathbb{P}(X > t + h | X > t) = e^{-\lambda h}$$

*(Gegeben $X > t$, ist $X - t$ auch exponentialverteild). Wir bezeichnen diese Eigenschaft als die **Gedächtnislosigkeit** von X. Insbesondere gilt*

$$\mathbb{P}(X \in [t, t+h] | X > t) = \lambda h + o(h).$$

Der Parameter λ sollte daher als **Erfolgsrate** gedacht werden.

6.2 Erwartungswert einer stetigen Zufallsvariable.

Unsere erste Aufgabe ist, den Erwartungswert einer stetigen Zufallsvariable zu definieren und zu berechnen. Um dies zu tun, verwenden wir ein Grenzwertargument. Deshalb müssen wir zuerst diskrete Annäherungen definieren.

Definition 6.7. *Sei X eine (stetige) Zufallsvariable, und sei $n \geq 1$. Wir definieren*

$$\underline{X}_n = 2^{-n} \lfloor 2^n X \rfloor = \sum_{k=-\infty}^{\infty} k 2^{-n} 1_{\{k2^{-n} \leq X < (k+1)2^{-n}\}}.$$

Ebenso definieren wir

$$\bar{X}_n = 2^{-n} \lceil 2^n X \rceil = \sum_{k=-\infty}^{\infty} (k+1) 2^{-n} 1_{\{k2^{-n} \leq X < (k+1)2^{-n}\}}.$$

In Worten ist $\underline{X}_n$ der nächstgelegene Wert links von X innerhalb eines diskreten Gitters der Maschenweite 2^{-n} (und natürlich $\bar{X}_n$ entsprechend von rechts).

Bemerken Sie, dass $\bar{X}_n, \underline{X}_n$ diskrete Zufallsvariablen sind. Deshalb können wir über $\mathbb{E}(\underline{X}_n)$ und $\mathbb{E}(\bar{X}_n)$ reden. Die Frage ist, ob diese Größen konvergieren, wenn $n \to \infty$, und ob die Grenzwerte gleich sind.

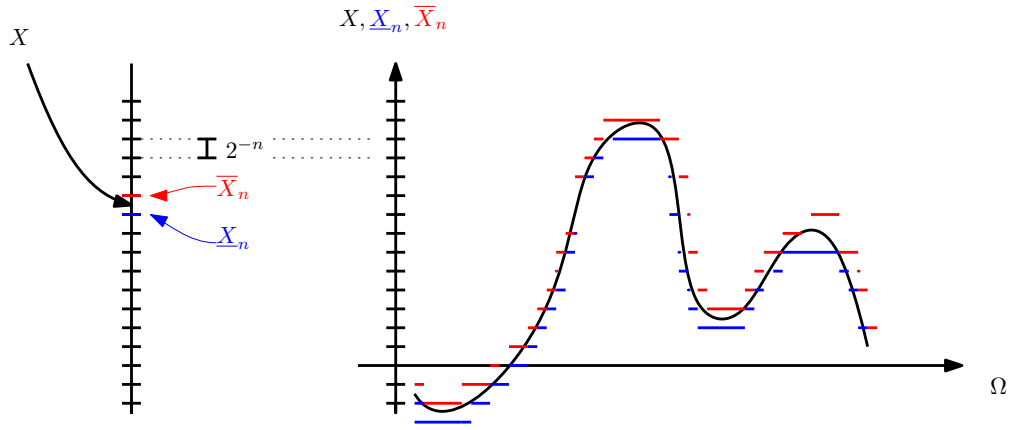


Abbildung 12: Eine stetige Zufallsvariable X, und ihre diskrete Annäherungen $\underline{X}_n \leq X \leq \overline{X}_n$.

Satz 6.8. *Sei X eine Zufallsvariable mit Werten in $\mathbb{R}$ sodass $\mathbb{E}\bar{X}_1$ existiert. Die obigen Grenzwerte existieren, und sind gleich: d.h.,*

$$\lim_{n\to\infty} \mathbb{E}(\bar{X}_n) = \lim_{n\to\infty} \mathbb{E}(\underline{X}_n).$$

*Wir nennen den gemeinsamen Grenzwert **Erwartungswert** von X und bezeichnen ihn mit $\mathbb{E}(X)$. Außerdem gilt*

$$\mathbb{E}(X) = \int_{\mathbb{R}} x f(x) dx,$$

wenn X eine stetige Zufallsvariable mit Dichte f ist und das Integral auf der rechten Seite endlich ist.

Beweis. Bemerken Sie, dass

$$\underline{X}_n \leq X \leq \bar{X}_n,$$

und $\underline{X}_n$ mit n wächst, während $\bar{X}_n$ mit n abfällt. Das heißt,

$$\underline{X}_n \leq \underline{X}_{n+1}; \quad \text{und } \bar{X}_n \geq \bar{X}_{n+1}.$$

Deshalb gilt

$$\mathbb{E}(\underline{X}_n) \leq \mathbb{E}(\underline{X}_{n+1}) \tag{6.1}$$

und der Grenzwert (wenn $n \to \infty$) existiert. Ebenso ist $\lim_{n\to\infty} \mathbb{E}(\bar{X}_n)$ wohldefiniert. Außerdem gilt

$$|\bar{X}_n - \underline{X}_n| \leq 2^{-n}.$$

Deshalb folgt

$$|\mathbb{E}(\bar{X}_n) - \mathbb{E}(\underline{X}_n)| = \mathbb{E}(\bar{X}_n - \underline{X}_n) \leq 2^{-n} \to 0$$

und wir erhalten

$$\lim_{n\to\infty} \mathbb{E}(\bar{X}_n) = \lim_{n\to\infty} \mathbb{E}(\underline{X}_n).$$

Mit Hilfe der Linearität des Erwartungswertes (und einem Grenzwertargument) folgt

$$\begin{aligned}
\mathbb{E}(\underline{X}_n) &= \sum_{k=-\infty}^{\infty} k2^{-n}\mathbb{P}(X \in [k2^{-n}, (k+1)2^{-n})) \qquad (6.2)\\
&= \sum_{k\in\mathbb{Z}} (k2^{-n}) \int_{k2^{-n}}^{(k+1)2^{-n}} f(x)dx \\
&= \sum_{k\in\mathbb{Z}} 2^{-n} \int_{k2^{-n}}^{(k+1)2^{-n}} \lfloor 2^n x \rfloor f(x)dx \\
&= \int_{\mathbb{R}} 2^{-n} \lfloor 2^n x \rfloor f(x)dx.
\end{aligned}$$

Wegen $x - 2^{-n} \leq 2^{-n}\lfloor 2^n x \rfloor \leq x$, erhalten wir daraus

$$\int_{\mathbb{R}} 2^{-n} \lfloor 2^n x \rfloor f(x)dx \to \int_{\mathbb{R}} x f(x)dx$$

wie gewünscht. □

Bemerkung 6.9. Streng genommen, gilt der obige Beweis nur für den Fall $\mathbb{E}(|\bar{X}_1|), \mathbb{E}(|\underline{X}_1|) < \infty$. (Dies ist erforderlich, um (6.1) und (6.2) zu rechtfertigen.) Man kann jedoch zeigen, dass diese Bedingung äquivalent zur Wohldefiniertheit von $\int_{\mathbb{R}} x f(x)dx$ ist.

Der Erwartungswert hat auch in der allgemeinen Situation noch die Eigenschaften, die wir aus dem diskreten Fall gewohnt sind:

Proposition 6.10. *Eigenschaften des Erwartungswertes. Es gilt*

1. $|\mathbb{E}X - \mathbb{E}\bar{X}_n|, |\mathbb{E}X - \mathbb{E}\underline{X}_n| \leq 1/2^n$.
2. $X \leq Y$ *impliziert* $\mathbb{E}X \leq \mathbb{E}Y$. *Insbesondere folgt aus* $|X - Y| \leq \eta$, *dass* $\mathbb{E}|X - Y| \leq \eta$.
3. $\mathbb{E}[aX + bY] = a\mathbb{E}X + b\mathbb{E}Y$.
4. *Ist* $\phi : \mathbb{R} \to \mathbb{R}$ *eine konvexe Funktion so gilt*

$$\phi(\mathbb{E}[X]) \leq \mathbb{E}[\phi(X)].$$

Beweis. 1. Folgt direkt aus dem Beweis des vorigen Satzes.

2. Das gilt offenbar für diskrete Zufallsvariablen. Für allgemeine folgt es wegen Eigenschaft 1. mit Approximation.

3. Dass der Erwartungswert für diskrete Zufallsvariablen linear ist, wissen wir schon und der allgemeine Fall folgt mit Approximation.

4. Da ϕ konvex ist, liegt die Tangente g von ϕ im Punkt $(\mathbb{E}X, \phi(\mathbb{E}X))$ unter ϕ sodass
$$\mathbb{E}[\phi(X)] \geq \mathbb{E}[g(X)] = g(\mathbb{E}[X]) = \phi(\mathbb{E}[X]).$$

□

Wir bringen zwei weitere Darstellungen des Erwartungswertes, die nützlich sind:

Proposition 6.11. *Angenommen $\mathbb{E}X$ existiert, dann gilt*
$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X \geq x)\, dx - \int_{-\infty}^0 \mathbb{P}(X \leq x)\, dx.$$

Beweis. Das könnte eine Übungsaufgabe sein. □

Proposition 6.12. *Sei X eine stetige Zufallsvariable auf $\mathbb{R}^d$ mit Dichte $f : \mathbb{R}^d \to [0, \infty)$ und sei $g : \mathbb{R}^d \to \mathbb{R}$ eine Funktion sodass $\mathbb{E}[g(X)] < \infty$. Dann gilt*
$$\mathbb{E}[g(X)] = \int g(x) f(x)\, dx. \tag{6.3}$$

Beweis. Wir werden nur eine Beweisskizze geben.

Sei zunächst $g(x) = I_B(x)$ für $B \subset \mathbb{R}^d$. Dann gilt nach Definition der Dichte
$$\mathbb{E}[g(X)] = \mathbb{E}[I_B(X)] = \mathbb{P}(X \in B) = \int I_B(x) f(x)\, dx = \int g(x) f(x)\, dx.$$
Als nächstes bemerken wir, dass sowohl die rechte als auch die linke Seite dieser Gleichung linear in g sind. Daraus folgt, dass (6.3) für alle g der Form
$$g(x) = \sum_{i=1}^n \beta_i I_{B_i} \tag{6.4}$$
mit $\beta_1, \ldots, \beta_n \in \mathbb{R}, B_1, \ldots, B_n \subset \mathbb{R}^d$ erfüllt ist.

Für allgemeines g beweist man die Aussage, indem man durch Funktionen der Gestalt (6.4) approximiert. □

Genau wie für diskrete Zufallsvariablen, betrachten wir auch im allgemeinen Fall die Varianz
$$\mathrm{Var}(X) := \mathbb{E}[(X - \mathbb{E}X)^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$$
Aus Proposition 6.12 folgt, dass wir für eine Zufallsvariable X mit Dichte g auch
$$\mathrm{Var}(X) = \int x^2 f(x)\, dx - \left(\int x f(x)\, dx \right)^2$$
schreiben können.

Die Standardabweichung wird wieder als die Wurzel der Varianz definiert.

Beispiel 6.13. Sei X exponentialverteilt mit Parameter λ. Dann gilt $\mathbb{E}X = 1/\lambda, \mathrm{Var}(X) = 1/\lambda^2$.

6.3 Gemeinsame Verteilung

Wir beschränken uns der Einfachheit halber meist auf den Fall von zwei Zufallsvariablen X und Y. Das macht die Notation einfacher, die Definitionen und Resultate gelten aber natürlich genauso auch im allgemeinen Fall.

Gegeben zwei Zufallsvariablen X und Y ist die *gemeinsame / multivariate Verteilungsfunktion* durch

$$F(x,y) = \mathbb{P}(X \leq x, Y \leq y), \quad F : \mathbb{R}^2 \to [0,1] \tag{6.5}$$

gegeben. Genau wie im eindimensionalen Fall kann man Verteilungsfunktionen leicht durch Monotonie und Rechtsstetigkeit charakterisieren. Die **Marginalverteilung** (oder **Randverteilung**) von X ist gegeben durch die Verteilungsfunktion

$$F_X(x) = \mathbb{P}(X \leq x) = \lim_{y \to \infty} F(x,y).$$

(Und analog für Y.)

Definition 6.14. *Wir sagen dass X, Y* gemeinsam stetig verteilt *sind wenn es eine Funktion $f : \mathbb{R} \times \mathbb{R} \to [0, \infty)$ gibt (die **gemeinsame Dichtefunktion**) mit*

$$\mathbb{P}((X,Y) \ \in A) = \int_A f(x,y)\, dxdy.$$

Das heißt, X und Y sind gemeinsam stetig verteilt, wenn der Vektor (X,Y) eine stetige Zufallsvariable auf $\mathbb{R}^2$ ist. Wenn X und Y gemeinsam stetig sind, kann $\mathbb{P}((X,Y) \in A) > 0$ nur gelten, wenn die Fläche von A positiv ist.

Beispiel 6.15. Bemerken Sie dass X und Y beide stetig sein können, ohne dass X und Y gemeinsam stetig sind. Sei z.B. X stetig und $Y = X$. Dann sind X und Y nicht gemeinsam stetig: die Verteilung von (X,Y) ist auf $\Delta = \{(x,y) : y = x\}$ konzentriert (bemerken Sie, dass die Fläche von Δ null ist).

Seien X und Y gemeinsam stetig verteilt. In diesem Fall erhält man für die gemeinsame Verteilungsfunktion

$$F(x,y) = \int_{u \leq x} \int_{v \leq y} f(u,v)\, dudv.$$

Umgekehrt gilt natürlich

$$f(x,y) = \frac{\partial^2 F(x,y)}{\partial x \partial y},$$

wenn die Verteilungsfunktion hinreichend differenzierbar ist. Weiters gilt

Proposition 6.16. *Seien X und Y gemeinsam stetig verteilt mit gemeinsamer Dichtefunktion f. Dann ist X stetig mit Dichtefunktion*

$$f_X(x) = \int_{y \in \mathbb{R}} f(x,y) dy.$$

Beweis. Bemerken Sie, dass

$$\mathbb{P}(X \in A) = \mathbb{P}(X \in A, Y \in \mathbb{R}) = \int_{x \in A} \underbrace{\int_{y \in \mathbb{R}} f(x,y)\, dy}_{=: f_X(x)} dx = \int_A f_X(x)\, dx, \tag{6.6}$$

wie gewünscht. □

Beispiel 6.17. Sei (X, Y) gleichverteilt auf der Scheibe $D = \{(x, y) : x^2 + y^2 = 1\}$. Dann ist X stetig mit Dichtefunktion

$$f_X(x) = \frac{2}{\pi}\sqrt{1 - x^2}; \quad -1 \leq x \leq 1$$

und 0 sonst. Zum Beispiel, kann man für $R = \sqrt{X^2 + Y^2}$ überprüfen, dass

$$\mathbb{E}(R^2) = 2\mathbb{E}(X^2) = 2\int_{-1}^{1} x^2 \frac{2}{\pi}\sqrt{1 - x^2} dx = 1/2.$$

Das obige Integral ist nicht ganz leicht zu berechnen. Allerdings erhalten wir nach Proposition 6.12 mithilfe von Polarkoordinaten:

$$\begin{aligned}
\mathbb{E}(R^2) &= \int_D (x^2 + y^2) f(x,y) dx dy \\
&= \frac{1}{\pi}\int_0^1 r^2 (2\pi r) dr \\
&= 2\int_0^1 r^3 dr = 1/2.
\end{aligned}$$

6.4 Unabhängigkeit von Zufallsvariablen

Im diskreten Fall, haben wir Unabhängigkeit durch die Gleichung $\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^n \mathbb{P}(X_i = x_i)$ definiert. Im stetigen Fall müssen wir diese Definition anpassen:

Definition 6.18. *Wir sagen, dass die Zufallsvariablen $X_1, \ldots, X_n$ mit Werten in $\mathbb{R}$ unabhängig sind, wenn für alle $A_1, \ldots, A_n \subset \mathbb{R}$ gilt, dass*

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \mathbb{P}(X_1 \in A_1) \ldots \mathbb{P}(X_n \in A_n).$$

Bemerken Sie, dass dies den diskreten Fall beinhaltet: wenn $X_1, \ldots, X_n$ diskret sind, sind die zwei Definitionen äquivalent zu einander.

Begriffe	**Diskreter Fall**	**Stetiger Fall**
Funktion	Wahrscheinlichkeitsfunktion $p(x)$	Dichtefunktion $f(x)$
Wahrscheinlichkeit $\mathbb{P}(X \in A)$	$\mathbb{P}(X \in A) = \sum_{x \in A} p(x)$	$\mathbb{P}(X \in A) = \int_A f(x)dx$
Erwartungswert	$\mathbb{E}(X) = \sum_x xp(x)$	$\mathbb{E}(X) = \int xf(x)dx$
Erwartunsgwert einer Funktion der ZV	$\mathbb{E}(g(X)) = \sum_x g(x)p(x)$	$\mathbb{E}(X) = \int g(x)f(x)dx$
Gemeinsame Verteilung	Gemeinsame WF $p(x, y)$	Gemeinsame Dichtefunktion $f(x, y)$
Randverteilung	$p_X(x) = \sum_y p(x, y)$	Wenn X und Y gemeinsam stetig sind, dann ist auch X stetig mit Dichtefunktion $f_X(x) = \int_y f(x, y)dy$
Bedingte Verteilung	$p_{X\|Y=y}(x) = \frac{p(x,y)}{p_Y(y)} = \frac{p(x,y)}{\sum_x p(x,y)}$	$f_{X\|Y=y}(x) = \frac{f(x,y)}{f_Y(y)} = \frac{f(x,y)}{\int_{\mathbb{R}} f(x,y)\,dx}$
Bedingte Erwartung	$\mathbb{E}[X\|Y = y] = \sum_x xp_{X\|Y=y}(x)$	$\mathbb{E}[X\|Y = y] = \int_{\mathbb{R}} xf_{X\|Y=y}(x)\,dx$

Abbildung 13: Eine Tabelle mit unseren Definitionen und Begriffen zur Erinnerung.

Satz 6.19. *Seien $X_1, \ldots, X_n$ stetige Zufallsvariablen mit Dichtefunktionen $f_1, \ldots, f_n$. Dann sind $X_1, \ldots, X_n$ unabhängig genau dann wenn $X_1, \ldots, X_n$ gemeinsam stetig sind und die gemeinsame Dichtefunktion durch*

$$f(x_1, \ldots, x_n) = f_1(x_1) \ldots f_n(x_n)$$

gegeben ist.

Beweis. Nehmen wir an, dass $(X_1, \ldots, X_n)$ stetig ist mit Dichtefunktion $f(x_1, \ldots, x_n) = f_1(x_1) \ldots f_n(x_n)$. Dann gilt

$$\begin{aligned}
\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) &= \mathbb{P}((X_1, \ldots, X_n) \in A) \\
&= \int_A f(x_1, \ldots, x_n) dx_1 \ldots dx_n \\
&= \int_{A_1 \times \ldots \times A_n} f_1(x_1) \ldots f_n(x_n) dx_1 \ldots dx_n \\
&= \left(\int_{A_1} f_1(x_1) dx_1 \right) \ldots \left(\int_{A_n} f_n(x_n) dx_n \right) \\
&= \mathbb{P}(X_1 \in A_1) \ldots \mathbb{P}(X_n \in A_n).
\end{aligned}$$

Deshalb sind $X_1, \ldots, X_n$ unabhängig.

Nehmen wir jetzt an, dass $X_1, \ldots, X_n$ unabhängig sind. Zu zeigen ist, dass

$$\mathbb{P}((X_1, \ldots, X_n) \in A) = \int \ldots \int_A f(x_1, \ldots, x_n) dx_1 \ldots dx_n$$

für jedes $A \subset \mathbb{R}^n$, wobei $f(x_1, \ldots, x_n) = f_1(x_1) \ldots f_n(dx_n)$, Wir zeigen es nur für den Fall, indem $A = A_1 \times \ldots \times A_n$ (für den allgemeinen Fall braucht man ein Approximationsargument das wir nicht angeben). In diesem Fall gilt

$$\begin{aligned}
\mathbb{P}((X_1, \ldots, X_n) \in A) &= \mathbb{P}(X_1 \in A_1, \ldots X_n \in A_n) \\
&= \mathbb{P}(X_1 \in A_1) \ldots \mathbb{P}(X_n \in A_n) \\
&= \int_{A_1} f_1(x_1) dx_1 \ldots \int_{A_n} f_n(x_n) dx_n \\
&= \int \ldots \int_A f_1(x_1) \ldots f_n(x_n) dx_1 \ldots dx_n \\
&= \int \ldots \int_A f(x_1, \ldots, x_n) dx_1 \ldots dx_n
\end{aligned}$$

wie gewünscht. □

Beispiel 6.20. Sei (X, Y) gleichverteilt auf dem Quadrat $D_1 = (-1, 1)^2$. Dann sind X und Y gleichverteilt. Sind andererseits (X, Y) gleichverteilt auf $D_2 = \{(x, y) : x^2 + y^2 < 1\}$, dann sind X und Y nicht unabhängig.

Beispiel 6.21. Seien X, Y unabhängige, gleichverteilte Zufallsvariablen auf $[0,1]$. Dann gilt für $0 \leq z \leq 1$:

$$\mathbb{P}(X+Y \leq z) = \int_0^1 \int_0^1 1_{\{x+y\leq z\}} dxdy = \text{Fläche}(\Delta) = z^2/2,$$

wobei Δ das Dreieck mit den Eckpunkten $(0,0), (0,z)$ und $(z,0)$ ist.

Den nächsten Satz schreiben wir der Einfachheit halber nur für zwei Zufallsvariablen X und Y, obwohl er natürlich auch für mehrere Zufallsvariablen gilt.

Satz 6.22. *Seien X und Y unabhängige, stetige Zufallsvariablen mit Dichtefunktionen f und g. Dann gilt $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$. Allgemeiner gilt*

$$\mathbb{E}(u(X)v(Y)) = \mathbb{E}(u(X))\mathbb{E}(v(Y))$$

für Funktionen $u, v : \mathbb{R} \to \mathbb{R}$ für die die zwei Erwartungswerte auf der rechten Seite wohldefiniert sind.

Beweis. Wir bemerken, dass $u(X)$ und $v(Y)$ unabhängig sind:

$$\begin{aligned} \mathbb{P}(u(X) \in A, v(Y) \in B) &= \mathbb{P}(X \in u^{-1}(A), Y \in v^{-1}(B)) \\ &= \mathbb{P}(X \in u^{-1}(A))\mathbb{P}(Y \in v^{-1}(B)) \\ &= \mathbb{P}(u(X) \in A)\mathbb{P}(v(Y) \in B). \end{aligned} \tag{6.7}$$

Deshalb erhalten wir aus Satz 6.19

$$\begin{aligned} \mathbb{E}(u(X)v(Y)) &= \int_{\mathbb{R}^2} u(x)v(y)f(x)g(y)dxdy \\ &= \left(\int_{\mathbb{R}} u(x)f(x)dx\right)\left(\int_{\mathbb{R}} v(y)g(y)dy\right) \\ &= \mathbb{E}(u(X))\mathbb{E}(v(Y)) \end{aligned}$$

wobei wir Proposition 6.12 (zweimal) angewendet haben. □

Satz 6.22 erlaubt uns im stetigen Fall viele Eigenschaften zu beweisen, die sind uns schon aus dem diskreten Fall bekannt sind.

Korollar 6.23. *Seien $X_1, \ldots, X_n$ unabhängig mit $\mathbb{E}(X_i^2) < \infty$. Dann gilt $\mathrm{Cov}(X_i, X_j) = 0$ wenn $i \neq j$, und*

$$\mathrm{Var}(X_1 + \ldots + X_n) = \mathrm{Var}(X_1) + \ldots + \mathrm{Var}(X_n).$$

Beweis. Sei $m_i = \mathbb{E}(X_i)$. Wir bemerken, dass $X_i - m_i$ und $X_j - m_j$ unabhängig sind (sh. Satz 3.50). Deshalb folgt

$$\mathrm{Cov}(X_i, X_j) = \mathbb{E}((X_i - m_i)(X_j - m_j)) = \mathbb{E}(X_i - m_i)\mathbb{E}(X_j - m_j) = 0.$$

(Erinnern Sie sich: das bedeutet, dass die unsystematischen, stochastischen Anteile von X_i und X_j orthogonal sind.) Außerdem, gilt

$$\begin{aligned}\mathrm{Var}(X_1 + \ldots + X_n) &= \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2 \sum_{1 \le i < j \le n} \mathrm{Cov}(X_i, X_j) \\ &= \sum_{i=1}^{n} \mathrm{Var}(X_i)\end{aligned}$$

(Im Wesentlichen ist die obige Formel der Inhalt des Satzes von Pythagoras). □

Insbesondere erhält man das (schwache) Gesetz der großen Zahlen für stetige Zufallsvariablen.

Satz 6.24. *Sei $X_1, X_2, \ldots$ eine Folge unabhängig identisch verteilter Zufallsvariablen, wobei $m = \mathbb{E}(X_i)$ und $\sigma^2 = \mathrm{Var}(X_i) < \infty$. Sei*

$$\bar{X}_n = \frac{X_1 + \ldots + X_n}{n}$$

der Stichprobenmittelwert. Dann gilt für jedes $\varepsilon > 0$,

$$\mathbb{P}(|\bar{X}_n - m| > \varepsilon) \to 0$$

für $n \to \infty$.

Beweis. Der Beweis ist gleich wie im diskreten Fall, weil $\mathrm{Var}(\bar{X}_n)) = \sigma^2/n \to 0$ und (wie schon bemerkt) die Chebyshev'sche Ungleichung auch im stetigen Fall gilt. □

Das Summieren von unabhängigen Zufallsvariablen kann man auch auf der Ebene der entsprechenden Dichten gut abbilden.

Proposition 6.25. *Seinen X, Y unabhängige stetige Zufallsvariablen mit Dichten f, g. Dann ist die Dichte von $Z := X + Y$ gegeben durch die* Faltung

$$(f * g)(z) := \int_{-\infty}^{\infty} f(x) g(z - x)\, dx.$$

Beweis. Es gilt

$$\mathbb{P}(Z \le u) = \int_{\{(x,y): x+y \le u\}} f(x) g(y)\, dxdy = \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{u-x} f(x) g(y)\, dydx.$$

Indem wir $z = x + y$ substituieren, erhalten wir

$$\mathbb{P}(Z \le u) = \int_{x=-\infty}^{\infty} \int_{z=-\infty}^{u} f(x) g(z - x)\, dx\, dz = \int_{z=-\infty}^{u} \int_{x=-\infty}^{\infty} f(x) g(z - x)\, dx\, dz,$$

d.h. $\int_{-\infty}^{\infty} f(x) g(z - x)\, dx$ ist Dichte von Z. □

Zum Abschluss des Abschnitts betrachten wir noch den Begriff der *bedingten Dichte*. Zur Motivation erinnern wir uns an die in Abschnitt 5.1 für diskrete Zufallsvariablen besprochene bedingte Verteilung: Gegeben diskrete Zufallsvariablen X, Y, haben wir dort die bedingte Verteilung $\mathbb{P}(X = \cdot|Y = y) = p_{\cdot|Y=y}$ durch

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

definiert.

Seien nun X, Y Zufallsvariablen mit gemeinsamer Dichte $f = f(x, y)$ und sei f_Y die Randdichte von Y. Für y mit $f_Y(y) \neq 0$ definiert man die bedingte Dichte von X gegeben $Y = y$ durch

$$f_{X|Y=y}(x) := \frac{f(x, y)}{f_Y(y)}.$$

Beispiel 6.26. Der Zufallsvektor (X, Y) sei gleichverteilt auf der Einheitskreisscheibe. Gegeben $Y = 0$ ist X gleichverteilt auf $[-1, 1]$. Gegeben $Y = 1/2$ ist X gleichverteilt auf $[-\sqrt{3}/2, \sqrt{3}/2]$.

Wie schon im diskreten Fall kann man auch bedinge Erwartung und bedingte Verteilung sehr viel abstrakter betrachten. Wir werden das hier jedoch nicht weiter verfolgen.

Bemerkung 6.27. In der Definition von $f_{X|Y=y}$ bemerken Sie, dass das Ereignis $\{Y = y\}$ null Wahrscheinlichkeit hat! Trotzdem macht diese Definition Sinn. Wir können in der Tat alle Eigenschafte die wir für eine bedingte Verteilung erwarten auch formal beweisen. Zum Beispel können wir die bedinge Erwartung durch $\mathbb{E}(X|Y = y) = \int_x x f_{X|Y=y}(x)dx$, definieren und erhalten dann auch die *Turmeigenschaft* $\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|Y))$.

6.5 Transformation von Zufallsvariablen

Sei X gleichverteilt auf einem Intervall $[a, b]$. Für reelle Zahlen α, β ist die Zufallsvariable

$$Y := r(X) := \alpha X + \beta$$

gleichverteilt auf dem Intervall $[r(a), r(b)]$.

Die Dichte von X auf $[a, b]$ ist per Definition $\frac{1}{b-a}$. Beim Übergang zu $r(X)$ transformiert sich die Dichte um den Faktor $\frac{1}{|\alpha|}$ und beträgt auf dem Intervall $[r(a), r(b)]$ nun $\frac{1}{(b-a)|\alpha|} = \frac{1}{|r(b)-r(a)|}$.

Sei nun allgemeiner r eine Bijektion und X eine Zufallsvariable mit Dichte f. Ist X stückweise gleichverteilt und r injektiv und stückweise affin, so erhalten wir mit obiger Überlegung für die Dichte g von $r(X)$ auf dem Bild von r.

$$g(y) = \frac{f(r^{-1}(y))}{|r'(r^{-1}(y))|}. \tag{6.8}$$

Wenn wir die Umkehrabbildung von r mit s bezeichnen, können wir das auch als

$$g(y) = f(s(y))|s'(y)| \tag{6.9}$$

schreiben.

Mittels geschickter Approximationsargumente, kann man diese Formeln auch sehr viel allgemeiner beweisen, z.B. gilt:

Satz 6.28. *Sei $r : \mathbb{R} \to \mathbb{R}$ lokal Lipschitz und bis auf eine Menge mit Länge 0 injektiv. Bezeichne $s : \mathbb{R} \to \mathbb{R}$ eine Umkehrabbildung am Bild B von r. Sei X eine stetige Zufallsvariable mit Dichte f. Dann ist $r(X)$ eine stetige Zufallsvariable. Die Dichte g von $Y = r(X)$ erfüllt*

$$g(y) = f(s(y))|s'(y)|$$

auf der Menge B und $g = 0$ sonst.

Die Gleichung (6.9) entspricht natürlich genau der aus der Analysis bekannten Substitutionsregel: Für $D \subset B$ gilt

$$\mathbb{P}(Y \in D) = \mathbb{P}(X \in s(D)) = \int_{s(D)} f(x)\,dx = \left| \begin{array}{c} x = s(y) \\ dx = |s'(y)|\,dy \end{array} \right| = \int_D f(s(y))|s'(y)|\,dy,$$

d.h. wir sehen wieder, dass $g(y) = f(s(y))|s'(y)|$ die Dichte von Y ist.

Wir betrachten nun den mehrdimensionalen Fall: Sei $X = (X_1, \ldots, X_n)$ ein Zufallsvektor der auf dem Einheitswürfel gleichverteilt ist. Sei weiters $A \in \mathbb{R}^{n\times n}$ eine Matrix mit Inverser I. Das Bild des Einheitswürfels unter der Abbildung A ist ein Paralellepiped mit Volumen $|\det(A)|$. Dementsprechend ist

$$r(X) = Ax$$

gleichverteilt mit Dichte $\frac{1}{|\det(A)|} = |\det(I)|$ (auf diesem Paralellepiped).

Im allgemeinen können wir hinreichend reguläre Abbildungen lokal durch affine Abbildungen annähern, wobei der lineare Anteil durch die Jacobimatrix der partiellen Ableitungen gegeben ist. Sei r injektiv und lokal Lipschitz mit inverser Abbildung s. Bezeichne weiters

$$J(y) := \begin{pmatrix} \frac{\partial s_1}{\partial y_1} & \cdots & \frac{\partial s_1}{\partial y_n} \\ \vdots & & \vdots \\ \frac{\partial s_n}{\partial y_1} & \cdots & \frac{\partial s_n}{\partial y_n} \end{pmatrix}$$

die Jacobimatrix der Abbildung s. Dann ist die Dichte von $r(X)$ gegeben durch

$$g(y) = f(s(y))|\det J(y)| \tag{6.10}$$

am Bild von r und 0 sonst (unter den selben Voraussetzungen wie in Satz 6.28).

Als nächstes machen wir einige Beispiele, zuerst in Dimension 1.

Beispiel 6.29. Sei X gleichverteilt auf $(0,1)$. Was ist die Dichtefunktion von $Y = X^2$? Wir können zum Beispiel Satz 6.28 mit $r(x) = x^2$ anwenden. Bemerken Sie, dass $r(x)$ lokal Lipschitz ist und bijektiv auf $(0,1)$ mit Umkehrabbildung $s(y) = \sqrt{y}$. Deshalb ist die Dichte von $Y = f(X)$ durch

$$g(y) = |s'(y)| 1_{s(y)\in(0,1)} = \frac{1}{2\sqrt{y}} 1_{y\in(0,1)}$$

gegeben.

Eine andere Art wäre wie folgt:

$$\begin{aligned}\mathbb{P}(Y \le y) &= \mathbb{P}(X^2 \le y) \\ &= \mathbb{P}(X \le \sqrt{y}) = \sqrt{y}\end{aligned}$$

für $0 \le y \le 1$. Mittels differenzieren erhalten wir

$$g(y) = \frac{d}{dy}\mathbb{P}(Y \le y) = 1/(2\sqrt{y})$$

für $0 \le y \le 1$ (und 0 sonst), wie gewünscht.

Beispiel 6.30. Sei X exponentialverteilt mit Parameter λ. Dann ist $Y = \alpha X$ auch exponentialverteilt mit Parameter λ/α.

Beispiel 6.31. Sei U gleichverteilt auf $(0,1)$. Dann ist $X = -\log(U)/\lambda$ exponentialverteilt mit Parameter λ.

Als nächstes machen wir ein paar multidimensionale Beispiele.

Beispiel 6.32. (Proposition 6.25 überarbeitet.) Seien X und Y unabhängig mit Dichten f und g. Was ist die Dichte von $X+Y$? In Proposition 6.25 haben wir sie schon als die **Faltung** von f und g berechnet. Hier ist ein anderer Beweis mit Hilfe des Transformationsatzes (6.10). Das Problem ist, dass $(x,y) \mapsto x+y$ keine Bijektion ist. Wir betrachten

$$(x,y) \mapsto r(x,y) = (u,v) = \left(\frac{x}{x+y}, x+y\right)$$

Dann ist r offensichtlich eine Bijektion mit Umkehrabbildung

$$s(u,v) = (uv, v-uv) = (uv, v(1-u)) = (s_1(u,v), s_2(u,v)).$$

Deshalb hat $(U,V) = r(X,Y)$ die (gemeinsame) Dichtefunktion

$$\phi(u,v) = f(uv)g(v(1-u))|\det J(u,v)|$$

wobei

$$\begin{aligned}J(u,v) &= \begin{pmatrix} \frac{\partial s_1}{\partial u} & \frac{\partial s_1}{\partial v} \\ \frac{\partial s_2}{\partial u} & \frac{\partial s_2}{\partial v} \end{pmatrix} \\ &= \det \begin{pmatrix} v & u \\ -v & 1-u \end{pmatrix} \\ &= v(1-u) - (-v)u = v\end{aligned}$$

die Jacobimatrix von s ist. Deshalb erhalten wir

$$\phi(u,v) = vf(uv)g(v(1-u)).$$

Um die Dichtefunktion von $X+Y=V$ zu berechnen, können wir jetzt die Randverteilung betrachten (wie in Proposition 6.16):

$$\begin{aligned} f_V(v) &= \int_u \phi(u,v)du \\ &= \int_u vf(uv)g(v(1-u))du \\ &= \int_x f(x)g(v-x)dx \end{aligned}$$

(wobei wir die Substitution $x = uv$, $dx = vdu$ betrachtet haben). Wir erhalten also wieder die Formel für die Faltung von f und g.

7 Normalverteilung und Zentraler Grenzverteilungssatz

7.1 Momentenerzeugende Funktion

Es kann oft sehr nützlich sein, die Verteilung eine Zufallsvariable über bestimmte, aus ihr abgeleitete Funktionen darzustellen. Insbesondere interessieren wir uns für:

Definition 7.1. *Sei X eine Zufallsvariable mit Werten in $\mathbb{R}$. Dann ist die* Momentenerzeugende Funktion *definiert durch*

$$G(z) = \mathbb{E}(\exp(zX)) = \int_x e^{zx} f(x)dx; z \in \mathbb{R}.$$

Zufallsvariablen mit der selben Momentenerzeugenden Funktion haben auch die selbe Verteilung (zumindest unter bestimmten Voraussetzungen). Der Beweis ist allerdings relativ technisch, weswegen wir ihn weglassen.

Satz 7.2. *Seien $X, \tilde{X}$ zwei Zufallsvariablen mit $G(z) = \tilde{G}(z)$ für alle $z \in (-\theta, \theta)$. Dann sind die Verteilungen von X und $\tilde{X}$ gleich: d.h., $\mathbb{P}(X \in A) = \mathbb{P}(\tilde{X} \in A)$ für jede Menge $A \subset \mathbb{R}$.*

Ein wesentlicher Grund dafür, dass wir uns für Momentenerzeugende Funktionen interessieren ist, dass sie uns erlauben die Summe von unabhängigen Zufallsvariablen zu verstehen: Auf der Ebene der Momentenerzeugenden Funktionen kann man nämlich einfach multiplizieren, während man auf der Ebene von Dichten über mit der (komplizierteren) Faltungsoperation arbeiten müsste.

Satz 7.3. *Seien X, Y unabhängig mit Momentenerzeugenden Funktionen G_X, G_Y. Dann gilt*

$$G_{X+Y}(z) = G_X(z)G_Y(z).$$

Beweis. $\mathbb{E}[e^{zX+zY}] = \mathbb{E}[e^{zX}e^{zY}] = \mathbb{E}[e^{zX}]\mathbb{E}[e^{zY}]$. □

Beispiel 7.4. Sei X gleichverteilt auf $[0, a]$. Dann ist die momentenerzeugende Funktion durch

$$G(z) = \frac{1}{a}\int_0^a e^{zx}dx = \begin{cases} \frac{e^{za}-1}{az} & (z \neq 0) \\ 1 & (z = 0) \end{cases}$$

gegeben.

Beispiel 7.5. Sei X exponentialverteilt mit Parameter $\lambda \geq 0$. Dann ist die momentenerzeugende Funktion durch

$$G(z) = \int_0^\infty e^{zx}\lambda e^{-\lambda x}dx = \begin{cases} \frac{\lambda}{\lambda - z} & z < \lambda \\ \infty & z \geq \lambda \end{cases}$$

gegeben.

Bemerkung 7.6. Es kann passieren, dass $G(z) = \infty$ wie im obigen Beispiel. Dann sagt $G(z)$ nichts über X aus. Tatsächlich ist $G(z) < \infty$ nur wenn die Tailfunktion $\bar{F}(x)$ schneller als e^{-zx} abfällt (für $z > 0$, und ähnlich auf der anderen Seite für $z < 0$). Die Bedingung $G(z) < \infty$ ist deshalb ziemlich restriktiv. Um diese einschränkende Bedingung zu vermeiden kann man sonst immer die **charakteristische Funktion** betrachten

$$\phi(t) = \mathbb{E}(e^{itX}) = \int_x e^{itx} f(x)dx = \int_x \cos(tx)f(x)dx + i\int_x \sin(tx)f(x)dx;$$

das heißt, ϕ ist die Fourier Transformation von f (abgesehen vom Faktor π). Dies ist eine bessere Wahl im Vergleich zur momentenerzeugenden Funktion, weil es zu einer allgemeineren und saubereren Theorie führt. Die Kehrseite ist dass, dieser Zugang etwas komplexe Analysis erfordert. Auf diesem Grund beschränken wir uns in diesem Kapitel meistens nur auf den Fall in dem $G(z) < \infty$. Dann sind alle Berechnungen reellwertig und deshalb einfacher zu verstehen und zu rechtfertigen. Der Großteil der Ergebnisse, die in diesem Kaptel bewiesen werden, gilt jedoch auch ohne diese Beschränkung. Die Beweise benötigen jeweils nur eine geringfügige Modifikation (im wesentlichen muß man z durch it ersetzen).

Wir weisen auf einige elementare Eigenschaften hin.

Proposition 7.7. *Sei X eine Zufallsvariable mit $G(z) < \infty$ für alle $z \in (-\theta, \theta)$, wobei $\theta > 0$. Dann ist $G(0) = 1$. Außerdem gilt $\mathbb{E}(X^n) < \infty$ für jedes $n = 0, 1, \ldots$ und*

$$G(z) = 1 + z\mathbb{E}(X) + \frac{z^2}{2!}\mathbb{E}(X^2) + \ldots.$$

Insbesondere gilt

$$\mathbb{E}(X^n) = G^{(n)}(0)$$

für jedes $n \geq 0$.

Beweis. (Skizze). Wir können die Taylorentwicklung der Exponentialfunktion verwenden, um zu erhalten, dass

$$e^{zX} = 1 + zX + \frac{z^2}{2!}X^2 + \ldots$$

Wir nehmen auf die beiden Seite Erwartungswerte und erhalten

$$G(z) = 1 + z\mathbb{E}(X) + \frac{z^2}{2!}\mathbb{E}(X^2) + \ldots$$

wie gewünscht. Aufgrund unserer Annahme $G(z) < \infty$ für $z \in (-\theta, \theta)$ können wir Erwartungswert und Summation vertauschen und erhalten $\mathbb{E}[\sum_n z^n X^n/n!] = \sum_n z^n \mathbb{E}(X^n)/n!$. □

Beobachten Sie, dass das obige Ableitungsargument nur das Folgende ergibt. Seien $X, \tilde{X}$ sodass $G(z) = \tilde{G}(z)$, für $z \in (-\theta, \theta)$. Dann gilt $\mathbb{E}(X^n) = \mathbb{E}(\tilde{X}^n)$ für jedes $n \geq 1$. Es kann jedoch passieren, dass $\mathbb{E}(X^n) = \mathbb{E}(\tilde{X}^n)$ für jedes $n \geq 0$, ohne dass X und $\tilde{X}$ die selbe Verteilung haben.

Bemerkung 7.8. Mit charakteristischen Funktionen formuliert besagt dieser Satz dass

$$\hat{f}(t) = \hat{g}(t); t \in (-\theta, \theta) \implies f = g, \text{ “fast überall”},$$

wobei $\hat{f}(t)$ und $\hat{g}(t)$ die Fourier Transformationen von f und g sind.

7.2 Die Normalverteilung

Definition 7.9. *Eine Zufallsvariable X heißt* normalverteilt *wenn die Dichte von X gegeben ist durch*

$$\phi_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{7.1}$$

In diesem Fall ist der Erwartungswert von X gleich μ und die Varianz gleich σ^2 (Übung). Wir schreiben dann $X \sim N(\mu, \sigma^2)$.

Um zu sehen, dass (7.1) tatsächlich eine Dichte definiert, erinnern wir uns zunächst an folgendes Integral:

Beispiel 7.10.

$$\left(\int_{-\infty}^{\infty} e^{-x^2/2}\,dx\right)^2 = \int_{-\infty}^{\infty} e^{-x^2/2}\,dx \int_{-\infty}^{\infty} e^{-y^2/2}\,dy \tag{7.2}$$

$$= \int_{\mathbb{R}^2} e^{-(x^2+y^2)/2}\,dxdy \tag{7.3}$$

$$= \left| \begin{array}{c} (x,y) = J(r,\phi) = (r\sin\alpha, r\cos\alpha) \\ \det(J) = r \end{array} \right| \tag{7.4}$$

$$= \int_0^{\infty} \int_0^{2\pi} re^{-r^2/2}\,drd\alpha = 2\pi. \tag{7.5}$$

Mittels Substitution erhält man daraus leicht, das (7.1) eine Dichte definiert.

Falls $X \sim N(0,1)$, heißt X *standardnormalverteilt*. Oft schreibt man ϕ für die Dichte der Standardnormalverteilung und Φ für die entsprechende Verteilungsfunktion. Durch Substitution erhält man leicht: $X \sim N(\mu, \sigma^2) \Rightarrow \frac{X-\mu}{\sigma} \sim N(0,1)$ (Übung). Dadurch kann man Fragen über die Normalverteilung leicht auf Fragen über die Standardnormalverteilung zurückführen.

Es gibt keine explizite Formel für die Verteilungsfunktion / Tailfunktion der (Standard-) Normalverteilung. Folgende Abschätzung für die Tailfunktion vermittelt jedoch einen guten Eindruck über deren Abfallen: Es gilt für $X \sim N(0,1)$ und $x > 0$

$$\frac{1}{x + 1/x}\phi(x) < \mathbb{P}(x \le X) = 1 - \Phi(x) < \frac{1}{x}\phi(x). \tag{7.6}$$

Um das rasche Abfallen der Tailfunktion einer normalverteilten Zufallsvariablen $Z \sim N(\mu, \sigma^2)$ zu illustrieren, betrachten wir auch folgende Tabelle die Werte für die Normalverteilung mit der Abschätzung aus der Chebychev'schen Ungleichung ($\mathbb{P}(|Z-\mu| > k\sigma) \le 1/k^2$) vergleicht:

$$\begin{array}{ll} \mathbb{P}(|Z-\mu| \ge \sigma) \le 1 & \mathbb{P}(|Z-\mu| \ge \sigma) = 0.32 \\ \mathbb{P}(|Z-\mu| \ge 2\sigma) \le 0.25 & \mathbb{P}(|Z-\mu| \ge 2\sigma) = 0.046 \\ \mathbb{P}(|Z-\mu| \ge 3\sigma) \le 0.11 & \mathbb{P}(|Z-\mu| \ge 3\sigma) = 0.0027 \\ \mathbb{P}(|Z-\mu| \ge 4\sigma) \le 0.06 & \mathbb{P}(|Z-\mu| \ge 4\sigma) = 0.00006. \end{array} \tag{7.7}$$

Grob wird dieses starke Abfallen auch wie folgt ausgedrückt: Ist eine Größe normalverteilt, so liegen etwa 70% innerhalb einer Standardabweichung, etwa 95% innerhalb von zwei Standardabweichungen und etwa 99.7% innerhalb von drei Standardabweichungen um den Mittelwert.

Beispiel 7.11. Angenommen zwei Rasensorten A und B werden zu gleichen Anteilen gesäht wobei die Halme von $A \sim N(10, 1.0^2)$ und die Halme von $B \sim N(10, 1.5^2)$. Von den Halmen die kürzer als 7 cm sind stammen (nach Bayes)

$$\frac{0.046}{0.046 + 0.0027} \approx 95\%$$

von Sorte B. (Wird die Länge 5.5 cm betrachtet, sind mehr als 99% von Sorte B, etc.)

Das Modell ist insofern nicht realistisch, als dass es suggeriert, dass es Halme negativer Länge gäbe. Andererseits betrifft das auch bei Sorte B weniger als einen von 10^{10} Halmen.

Beispiel 7.12. Sei X *normalverteilt* $X \sim N(\mu, \sigma^2)$. Dann ist die momentenerzeugende Funktion von X gegeben durch

$$G_X(z) = e^{\mu z + \frac{1}{2} z^2 \sigma^2}.$$

Um das zu sehen, rechnen wir

$$\begin{aligned}
\mathbb{E}[e^{zX}] &= \int_{-\infty}^{\infty} e^{zx} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx \\
&= \left| \begin{array}{l} y = (x-\mu)/\sigma \\ dy = dx/\sigma \end{array} \right| \\
&= \int_{-\infty}^{\infty} e^{z(\mu+\sigma y)} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}\, dy \\
&= e^{z\mu + \frac{1}{2}\sigma^2 z^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-z\sigma)^2}{2}}\, dy = e^{\mu z + \frac{1}{2} z^2 \sigma^2}.
\end{aligned}$$

Wir erhalten unmittelbar:

Satz 7.13. *Seien X, Y unabhängig, $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$. Dann gilt $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.*

Sei $Z \sim N(\mu, \sigma^2)$ und $\alpha \in \mathbb{R}_+$. Dann gilt $\alpha Z \sim N(\alpha\mu, (\alpha\sigma)^2)$.

7.3 Der zentrale Grenzverteilungssatz

Seien $X_1, \ldots, X_n$ unabhängig identisch verteilt mit Mittel 0 (oBdA) und Varianz σ^2. Dann hat die Zufallsvariable

$$\frac{X_1 + \ldots + X_n}{\sqrt{n}} =: S_n/\sqrt{n} \tag{7.8}$$

wieder Mittel 0 und Varianz σ^2. Aus Satz 7.1 wissen wir: Falls zusätzlich $X_1, \ldots, X_n \sim N(0, \sigma^2)$, dann ist auch $S_n/\sqrt{n}$ normalverteilt und mithin gilt

$$S_n/\sqrt{n} \sim N(0, \sigma^2). \tag{7.9}$$

Der zentrale Grenzverteilungssatz besagt, dass (7.9) ohne Normalverteilungsannahme immerhin noch approximativ stimmt. Genauer gilt folgendes:

Satz 7.14 (zentraler Grenzverteilungssatz)**.** *Sei $X_1, X_2, \ldots$ eine Folge unabhängig identisch verteilter Zufallsvariablen mit Mittel 0 und endlicher Varianz σ^2. Bezeichne $S_n := X_1 + \ldots + X_n$. Dann gilt für alle $x \in \mathbb{R}$*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n}{\sigma\sqrt{n}} \leq x\right) = \Phi(x), \tag{7.10}$$

wobei Φ die Verteilungsfunktion der Standardnormalverteilung bezeichnet.

Wir machen einige Bemerkungen zur Aussage des Grenzverteilungssatzes vor dem Beweis:

1. Falls die X_i nicht *zentriert* sind, d.h. wir $\mathbb{E}X_i \neq 0$ zulassen, wird die Aussage des Grenzverteilungssatzes zu

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x\right) = \Phi(x) \tag{7.11}$$

2. Die Gleichung (7.11) sagt aus, dass die Verteilung von S_n ungefähr $N(n\mu, n\sigma^2)$ ist. Zum Beispiel ist (7.10) offenbar äquivalent zu

$$\lim_{n\to\infty} \mathbb{P}\left(a < \frac{S_n}{\sigma\sqrt{n}} \leq b\right) = \Phi(b) - \Phi(a),$$

   und es kommt nicht darauf an, ob man hier mir den Zeichen $<$ oder $\leq$ arbeitet. (Das folgt aus der Stetigkeit von Φ.)

3. Die Konvergenz in (7.10) besagt gerade, dass die Verteilungsfunktionen F_n von $\frac{S_n - n\mu}{\sigma\sqrt{n}}$ *punktweise* gegen die Verteilungsfunktion Φ der Standardnormalverteilung konvergieren.

   Sind allgemeiner Zufallsvariablen $Y, Y_1, Y_2, \ldots$ mit Verteilungsfunktionen $F, F_1, F_2, \ldots$ gegeben, so sagen wir, dass die Folge $Y_1, Y_2, \ldots$ *in Verteilung* gegen Y konvergiert, falls

$$\lim_{n\to\infty} F_n(x) = F(x)$$

   für alle x in denen F stetig ist.

   Man schreibt dafür manchmal auch

$$Y_n \quad \Rightarrow_{\mathrm{D}} Y,$$

   wobei das D für "distribution" steht.

4. In der Situation des zentralen Grenzwertsatzes ist es natürlich zu fragen, ob es nicht auch eine standardnormalverteilte Zufallsvariable Z gibt, sodass

$$\lim_{n\to\infty} \frac{S_n - n\mu}{\sigma\sqrt{n}} = Z$$

(wie es ja beim Gesetz der großen Zahlen der Fall ist). Tatsächlich kann man zeigen, dass es keine solche Zufallsgröße geben kann: das Ergebnis im Satz 7.14 handelt sich nur um *die Verteilung* von S_n. Es sagt nichts aus über den Grenzwert von $(S_n - n\mu)/(\sigma\sqrt{n})$ wenn $n \to \infty$.

Der Beweis des zentralen Grenzwertsatzes basiert auf folgendem Stetigkeitssatz, den wir ohne Beweis angeben:

Satz 7.15 (Lévys Stetigkeitssatz)**.** *Seien $X, X_1, X_2, \ldots$ Zufallsvariablen deren momentenerzeugende Funktionen G, $G_1, G_2, \ldots$*

$$\lim_{n\to\infty} G_n(z) = G(z)$$

für alle $z \in (-\theta, \theta)$ erfüllen. Dann konvergiert $X_1, X_2, \ldots$ in Verteilung gegen X.

Beweiskizze des zentralen Grenzwertsatzes. Die momentenerzeugende Funktion G von X_i erfüllt

$$\begin{aligned} G(z) &= \mathbb{E}[e^{zX_i}] \\ &= 1 + z\mathbb{E}[X_i] + \frac{1}{2!}z^2\mathbb{E}[X_i^2] + \frac{1}{3!}z^3\mathbb{E}[X_i^3] + \ldots \\ &= 1 + \frac{\sigma^2}{2!}z^2 + o(z^2). \end{aligned}$$

Damit erhalten wir für die momentenerzeugende Funktion von $S_n/(\sigma\sqrt{n})$

$$\begin{aligned} \mathbb{E}\left[e^{z\frac{S_n}{\sigma\sqrt{n}}}\right] &= \mathbb{E}\left[e^{z\frac{X_1+\ldots+X_n}{\sigma\sqrt{n}}}\right] \\ &= \left(\mathbb{E}\left[e^{\frac{z}{\sigma\sqrt{n}}X_1}\right]\right)^n \\ &= \left(G(\tfrac{z}{\sigma\sqrt{n}})\right)^n \\ &= \left(1 + \frac{\sigma^2}{2!}(\tfrac{z}{\sigma\sqrt{n}})^2 + o(1/n)\right)^n \\ &= \left(1 + \tfrac{z^2}{2n} + o(1/n)\right)^n \to e^{z^2/2}, \end{aligned}$$

was gerade die momentenerzeugende Funktion der Standardnormalverteilung ist. Mittels des Satzes 7.15 ist der Beweis komplett.

(Falls die momentenerzeugenden Funktionen nicht existieren, argumentiert man ganz analog mit charakteristischen Funktionen.) □

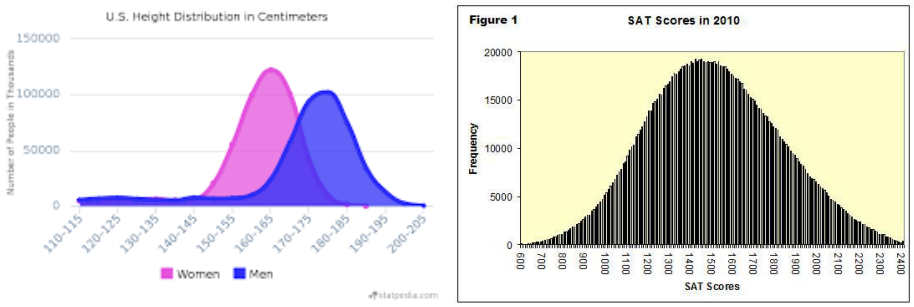


Abbildung 14: Größenverteilung für Frauen und Männer in den USA, und Verteilung der Punktzahlen in SAT.

7.4 Beispiele und Anwendungen

Wenn eine Größe die Summe von vielen kleinen Zufallseffekten ist, erwarten wir aus dem Satz 7.14, dass sie ungefähr normalverteilt ist. Betrachten Sie zum Beispiel die Verteilung der Körpergröße innerhalb einer Population, oder die Verteilung der Punktzahlen in der SAT Prüfung.

Wir bringen noch einige weitere Beispiele:

Beispiel 7.16. Seien $X_1, \ldots, X_n$ unabhängig Bernoulli (p) verteilt (d.h. $\mathbb{P}(X_i = 1) = p = 1 - \mathbb{P}(X_i = 0)$.) Wir möchten p schätzen. Sei $\hat{p}_n = (X_1 + \ldots + X_n)/n$. Wie groß sollten wir n wählen, sodass $|\hat{p}_n - p| \leq \varepsilon := 3\%$ mit zumindest 95.4% Wahrscheinlichkeit?

(Dies ist die Situation, die wir für Umfragen analysieren sollten.)

Nach dem zentralen Grenzverteilungsatz erhalten wir

$$\begin{aligned}\mathbb{P}(|\hat{p}_n - p| \geq \varepsilon) &\approx \mathbb{P}(|N(0, \sigma^2/n)| \geq \varepsilon) \\ &= \mathbb{P}(|N(0,1)| \geq \tfrac{\varepsilon\sqrt{n}}{\sigma})\end{aligned}$$

wobei $\sigma^2 = \mathrm{Var}(X_1) = p(1-p) \leq 1/4$. Deshalb gilt

$$\mathbb{P}(|\hat{p}_n - p| \geq \varepsilon) \leq \mathbb{P}(|N(0,1)| \geq 2\varepsilon\sqrt{n})$$

Die Größe auf der rechten Seite ist kleiner gleich 4.6% wenn $2\varepsilon\sqrt{n} \geq 2$ (siehe (7.7)). Äquivalent dazu erhalten wir

$$n \geq \lceil \tfrac{1}{\varepsilon^2} \rceil = 1,112.$$

Wenn wir die *Toleranz* auf 5% setzen (statt 4.6%), dann ergibt sich $2\varepsilon\sqrt{n} \geq q$, wobei

$$\int_{-q}^{q} e^{-\frac{x^2}{2}} \frac{dx}{\sqrt{2\pi}} = 5\%; \qquad q \approx 1.96.$$

Diese Bedingung ist äquivalent zu $n \geq 1,068$. Dies ist typisch für die Stichprobengröße und Fehlerspanne, die in kommerziellen Umfragen verwendet wird. Um die Frage komplett rigoros

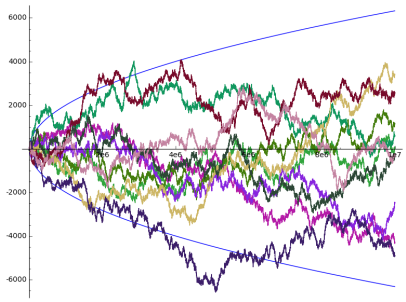


Abbildung 15: Einige Stichproben der symmetrischen Irrfahrt, im Vergleich zu $n \mapsto \pm\sqrt{n}$.

zu beantworten müssten wir auch die Fehler bei der Anwendung des zentralen Grenzverteilungsatzes schätzen. In der *wirklichen* Wirklichkeit gibt es viele anderen Sorten von Fehlern wie z.B. Stichprobenverzerrung.

Beispiel 7.17. Irrfahrt auf $\mathbb{Z}$. Sei $S_n = \sum_{i=1}^{n} X_i$, wobei $X_i = \pm 1$ mit Wahrscheinlichkeit $1/2$ und unabhängig sind. S_n ist die Position zum Zeitpunkt n der (symmetrischer) Irrfahrt auf $\mathbb{Z}$. Nach dem zentralen Grenzverteilungssatz erhalten wir

$$\frac{S_n}{\sqrt{n}} \Rightarrow_D N(0,1)$$

wenn $n \to \infty$. Das heißt, nach n Schritten ist die Position des Teilchens ungefähr $\pm\sqrt{n}$ (Größenordnung), und ihre Verteilung ungefähr die Standartnormalverteilung. Wir sagen, dass die Irrfahrt sich *diffus* ausbreitet.

Diese Tatsäche ist ein grundlegendes Naturgesetz (das sogenannte Ficks'sche Gesetz in der Physik). Sie ist tief verbunden mit Eigenschaften der Wärmeleitungsgleichung (heat equation)

$$\frac{\partial u}{\partial t} = \frac{1}{2}\Delta u$$

die sich auch diffus ausbreitet. Um mehr über dieses Thema zu erfahren sollten Sie mehr über Stochastische Prozessen und insbesondere die *Brownsche Bewegung* lernen.

8 Einführung in die Statistik

Ein gute Einführung in die mathematische Statistik bietet [4].

8.1 Was ist Statistik?

Die Statistik befasst sich mit der *Datenanalyse:* der Verwendung von Daten, um Schlussfolgerungen zu ziehen. Es geht um Fragen wie "Was sagen mir diese Daten?". Wir untersuchen in diesem Kurs mathematische Techniken zur Schätzung von Parametern, Anpassung von Modellen und dem Testen von Hypothesen.

Beispiel 8.1. Eine berühmte Studie untersuchte die Auswirkungen auf Herzinfarkte durch die Einnahme eines Aspirins jeden zweiten Tag. Die Ergebnisse nach 5 Jahren waren

Gruppe	Herzinfarkte	Keine Herzinfarkte	Herzinfarkte pro 1000
Aspirin	104	10,933	9.42
Placebo	189	10,845	17.13

Was können wir daraus schließen? Wie sicher sind wir uns bei unserer Schlussfolgerung?

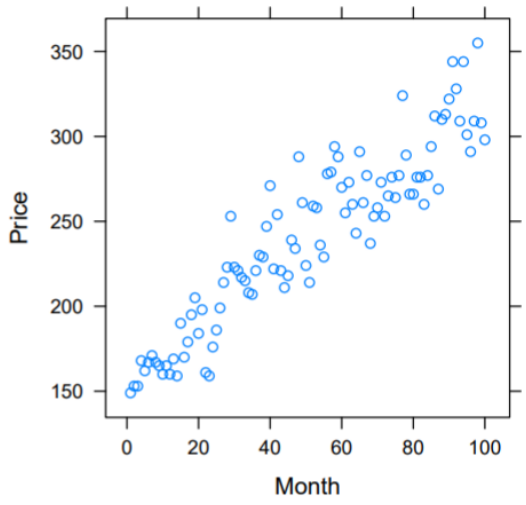


Abbildung 16: Scatterplot einiger Hauspreise in Oxford

Beispiel 8.2. Steigen die Hauspreise linear mit der Zeit? Nehmen wir an, dass wir einige Hauspreise $(x_i, y_i)_{i=1}^n$ beobachten, wobei y_i die Preise bezeichnen und x_i den entsprechenden Monat.

Ist es sinnvoll dass

$$y_i = \alpha + \beta x_i + \text{Fehler?}$$

Wenn ja, wie sollen wir α und β schätzen?

Unterschied zwischen Wahrscheinlichkeitstheorie und Statistik. Für die meisten Leute sind die zwei Begriffe fast Synonyme. Es gibt jedoch einen großen Unterschied. In der Wahrscheinlichkeitstheorie nehmen wir an, dass die Verteilungen der Zufallsvariablen gegeben sind. Wir arbeiten dann ihre Eigenschaften aus. Im Gegensatz dazu, haben wir in der Statistik *Daten*, von denen wir annehmen, dass sie aus einem unbekannten Wahrscheinlichkeitsmodell $\mathbb{P}$ generiert wurden. Wir möchten in der Lage sein, einige nützliche Dinge über $\mathbb{P}$ zu sagen; etwa welches Modell $\mathbb{P}$ die Daten am besten beschreibt.

In der Praxis geht Statistik über Mathematik hinaus. Um komplexe Daten zu verstehen braucht man nicht nur mathematische Kenntnisse, sondern auch gesunden Menschenverstand zusammen mit guter Intuition: Datenanalyse kann eine subtile Kunst sein, und kann sogar philosophische Fragen aufwerfen. Diese Aspekte werden uns jedoch natürlich nicht betreffen, und wir werden nur bei der Mathematik bleiben.

8.2 Empirische Verteilungsfunktion

Vor allem von theoretischem Interesse ist die Frage ob man die Verteilung einer Zufallsvariable schätzen kann, falls man beliebig viele Realisierungen beobachten kann.

Gegeben sei eine Folge $X_1, X_2, \ldots$ von unabhängig verteilten Zufallsvariablen mit Verteilungsfunktion F. Wir definieren dann die *empirische Verteilungsfunktion*

$$\hat{F}_n(t) = \frac{\text{Anzahl der Elemente im Sample } \leq t}{n} = \frac{1}{n}\sum_{i\leq n} I_{X_i\leq t}.$$

Dann gilt $\mathbb{E}\hat{F}_n(t) = \mathbb{P}(X_1 \leq t) = F(t)$ und $\operatorname{Var}(\hat{F}_n(t)) = \operatorname{Var} I_{X_1\leq t}/n$. Weil $I_{X_1\leq t}, I_{X_2\leq t}, \ldots$ unabhängig sind, gilt nach dem starken Gesetz der großen Zahlen, dass

$$F_n(t) \to F(t).$$

Der sogenannte Satz von Glivenko-Cantelli besagt, dass diese Konvergenz sogar gleichmäßig ist.

Da alle Eigenschaften, die wir in der Wahrscheinlichkeitstheorie betrachten nur von der Verteilung (bzw. Verteilungsfunktion) abhängen ist es, vom theoretischen Standpunkt, befriedigend, dass wir die Verteilung beliebig genau annähern können, wenn wir ein Experiment nur oft genug wiederholen.

In der Praxis nimmt man jedoch meist an, dass man bestimmte Informationen über die grundsätzliche Form der Verteilung hat und die eigentliche Verteilung dann schon über bestimmte Parameterwerte festgelegt wir. Entsprechend ist es dann ausreichend, diese Parameter zu schätzen. Das wichtigste Verfahren dafür wird im nächsten Kapitel diskutiert.

8.3 Maximum-Likelihood Schätzung

Rahmen Eine Zufallsvariable $X \in \mathbb{R}^d$ hat Verteilung $\mathbb{P}_\theta$, wobei $\theta \in \Theta$ ein uns unbekannter Paramater ist. Hier ist Θ die Menge aller möglichen Parameter (die wir normalerweise als

bekannt annehmen). Die Verteilung $\mathbb{P}_\theta$ kann diskret oder stetig (oder etwas exotischer) sein. Unser Ziel ist θ zu schätzen gegeben, dass wir unabhängig, identisch verteilte Zufallsvariablen $X_1, \ldots, X_n$ beobachtet haben, die jeweils die Verteilung $\mathbb{P}_\theta$ haben. Wenn der Parameter aus dem Kontext klar ist, werden wir ihn weglassen. Andererseits wollen wir manchmal betonen, dass ein Erwartungswert bezüglich $\mathbb{P}_\theta$ berechnet wird und schreiben dann $\mathbb{E}_\theta$.

Beispiel 8.3 (Umfrage)**.** $\mathbb{P}_\theta$ könnte die Bernoulli (θ) Verteilung sein. Hier ist $\Theta = [0, 1]$. Gegeben ist $(X_1, \ldots, X_n) \in \{0, 1\}^n$ (wobei $x = 1$ 'ja' und $x = 0$ 'nein' entspricht) wie sollen wir $\theta = \mathbb{P}_\theta(X_i = 1)$ schätzen?

Beispiel 8.4 (Glühbirne)**.** $\mathbb{P}_\theta$ könnte die Exponentialverteilung mit Parameter $\theta > 0$ sein (hier ist $\Theta = (0, \infty)$). Zum Beispiel sind $X_1, \ldots, X_n$ die Zeiten zwischen aufeinanderfolgenden Erneuerungen einer Glühbirne und unser Ziel ist θ (oder $1/\theta = \mathbb{E}_\theta(X_i)$) zu schätzen.

Beispiel 8.5. Die Verteilung könnte $\mathcal{N}(m, \sigma^2)$ sein. Hier ist $\Theta = \{(m \in \mathbb{R}, \sigma \in \mathbb{R})\}$. Bemerken Sie dass $\mathbb{P}_{(m,\sigma)} = \mathbb{P}_{(m,-\sigma)}$ es kann also sein, dass θ nicht eindeutig ist.

Seien $(X_1, \ldots, X_n)$ unsere Beobachtungen. Wir halten $\mathbf{x} = (X_1, \ldots, X_n)$ fest. Sei

$$f(\mathbf{x}; \theta)$$

die gemeinsame Dichtefunktion/Wahrscheinlichkeitsfunktion von $\mathbf{X} = (X_1, \ldots, X_n)$ für den Parameter θ. Manchmal schreibt man auch $f(\mathbf{x}|\theta)$ statt $f(\mathbf{x}; \theta)$. Wichtig ist, dass wir $f(\mathbf{x}; \theta)$ als Funktion von $\theta \in \Theta$ auffassen.

Definition 8.6. *Die Likelihood Funktion ist die Funktion*

$$\theta \in \Theta \mapsto L(\theta) = f(\mathbf{x}; \theta).$$

Die log Likelihood ist die Funktion $\ell(\theta) = \log L(\theta)$*.*

Die Likelihood und log-Likelihood Funktionen hängen natürlich von unseren Beobachtungen $\mathbf{x}$ ab. Aber wie gesagt halten wir sie fest, während θ variert.

Definition 8.7. *Der Maximum-Likelihood-Schätzer (MLS) ist der Wert* $\hat{\theta}$*, der* $L(\theta)$ *maximiert (sofern er wohldefiniert ist):*

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta).$$

Das heißt, $\hat{\theta}$ ist der Wert, für den die Beobachtungen $\mathbf{x} = (X_1, \ldots, X_n)$ am wahrscheinlichsten sind.

Beispiel 8.8. Im Umfragebeispiel hat man

$$L(\theta) = \theta^s (1 - \theta)^{n-s}$$

wobei $s = s(\mathbf{x}) = \sum_{i=1}^{n} x_i$. D.h.,

$$\ell(\theta) = s \log \theta + (n - s) \log(1 - \theta).$$

In dem man differenziert, erhält man

$$\frac{s}{\hat{\theta}} = \frac{n - s}{1 - \hat{\theta}},$$

oder

$$s(1 - \hat{\theta}) = (n - s)\hat{\theta};$$

d.h.

$$\hat{\theta} = s/n = \bar{x}_n.$$

Beispiel 8.9. Seien $X_1, \ldots, X_n$ unabhängig nach $N(\mu, \sigma^2)$ verteilt, wobei die Parameter μ, σ^2 unbekannt sind. Die Likelihood ist

$$L(\mu, \sigma^2) = \prod_{i \leq n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} \tag{8.1}$$

$$= (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2} \sum_{i \leq n} (x_i - \mu)^2}. \tag{8.2}$$

Die entsprechende log-Likelihood ist dann

$$l(\mu, \sigma^2) = \text{const.} - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i \leq n} (x_i - \mu)^2. \tag{8.3}$$

Für die Ableitungen erhalten wir

$$\frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i \leq n} (x_i - \mu) \tag{8.4}$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i \leq n} (x_i - \mu)^2. \tag{8.5}$$

Nullsetzen liefert uns

$$\hat{\mu} = \frac{1}{n} \sum_{i \leq n} x_i = \bar{x}_n \tag{8.6}$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i \leq n} (x_i - \hat{\mu})^2 = \frac{1}{n} \sum_{i \leq n} (x_i - \bar{x}_n)^2. \tag{8.7}$$

Ist σ^2, bekannt ist $\hat{\mu} = \frac{1}{n} \sum_{i \leq n} x_i = \bar{x}_n$ ebenfalls der MLE. Ist μ bekannt, so ist $\hat{\sigma}^2 = \frac{1}{n} \sum_{i \leq n} (x_i - \mu)^2$ der MLE.

8.4 Abstrakter Blickwinkel auf Parameterschätzung

Definition 8.10. *Formal ist eine* Statistik *eine Funktion T auf $\mathbb{R}^n$.*

Typischer Weise denken wir wieder, dass $(X_1, \dots, X_n)$ ein Zufallsvektor ist, den wir beobachten. Wenn wir von einer Statistik reden, stellen wir uns typischer Weise vor, dass wir eine Zuordnung haben, die einem Ausgang $\mathbf{x}$ von $(X_1, \dots, X_n)$ den Wert $T(\mathbf{x})$ zuordnet.

Ein typisches Beispiel für eine Statistik ist das Stichprobenmittel $T(x_1, \dots, x_n) = \bar{x}_n$.

Definition 8.11. *Seien wieder $X_1, \dots, X_n$ Zufallsvariablen, deren Verteilung durch einen Parameter θ beschrieben wird. Ein* Schätzer *für θ ist dann eine beliebige Statistik T die wir verwenden um θ zu schätzen.*

Das für uns wichtigste Beispiel eines Schätzers ist der MLE den wir im letzten Kapitel kennengelernt haben.

Heuristisch ist T ein guter Schätzer, wenn er Werte nahe am tatsächlichen Parameter annimmt. Wir werden uns bald damit beschäftigen, dass zu formalisieren.

Definition 8.12. *Seien $X_1, \dots$ Zufallsvariablen deren Verteilung durch $\theta \in \Theta$ parametrisiert wird. Eine Folge von Schätzern $T_n : \mathbb{R}^n \to \Theta$ heißt* konsistent *falls*

$$\lim_{n\to\infty} T_n(X_1, \dots, X_n) = \theta \tag{8.8}$$

in Wahrscheinlichkeit.

Meist ist aus dem Kontext klar, wie T_n von n abhängt und man spricht nur von der Konsistenz von T.

In der Situation von Beispiel 8.9 gilt nach dem schwachen Gesetz der großen Zahlen

$$T(X_1, \dots, X_n) = \bar{X}_n \to \mu.$$

Der MLE ist in vielen Fällen konsistent.

Definition 8.13. *Gilt $\mathbb{E}T(X_1, \dots, X_n) = \theta$, so nennen wir den Schätzer* erwartungstreu / unverzerrt / unbiased *andernfalls* verzerrt / biased

In Beispiel (8.9) gilt $\mathbb{E}\hat{\mu}(X_1, \dots, X_n) = \mu$ aber $\mathbb{E}\hat{\sigma}^2(X_1, \dots, X_n) = (n-1)\sigma^2/n$. Der letzte Schätzer ist also biased aber immer noch *asymptotisch unbiased.*

Definition 8.14. *Der* bias *eines Schätzers ist gegeben durch*

$$b(T) = \mathbb{E}_\theta T(X_1, \dots, X_n) - \theta,$$

die mittlere quadratische Abweichung *MQA (englisch MSE) durch*

$$MQA(T) = \mathbb{E}_\theta(T(X_1, \dots, X_n) - \theta)^2.$$

Systematische Fehler sind offenbar eher unerwünscht. MQA ist ein einfaches Maß für die Qualität eines Schätzers.

Satz 8.15. $MQA(T) = \mathrm{Var}(T) + (b(T))^2$.

Beweis. Wir schreiben $\mu = \mathbb{E}T$. Dann gilt

$$\begin{aligned} MQA(T) &= \mathbb{E}((T-\mu) + (\mu - \theta))^2 && (8.9)\\ &= \mathbb{E}((T-\mu)^2 + 2(\mu-\theta)(T-\mu) + (\mu-\theta)^2) && (8.10)\\ &= \mathbb{E}(T-\mu)^2 + (\mu-\theta)^2) = \mathrm{Var}(T) + (b(T))^2. && (8.11) \end{aligned}$$

□

In Beispiel 8.3 haben wir uns mit Bernoulli(θ) verteilten Zufallsvariablen $X_1, X_2, \ldots$ beschäftig. Wir haben schon festgestellt, dass der MLS gerade das Stichprobenmittel war, d.h.

$$\hat{\theta}(x_1, \ldots, x_n) = \bar{x}_n.$$

Ein anderer möglicher Schätzer ist

$$\hat{\theta}'(x_1, \ldots, x_n) := x_1/3 + 2x_2/3.$$

Beide Schätzer sind unbiased. Jedoch ist $\hat{\theta}$ konsistent, während $\hat{\theta}'$ das offenbar nicht ist.

Da beide Schätzer unbiased sind, ist die MQA jeweils gerade die Varianz:

$$MQA(\hat{\theta}) = \mathrm{Var}(\bar{X}_n) = \mathrm{Var}(X_1)/n = p(1-p)/n \quad (8.12)$$

$$MQA(\hat{\theta}') = \mathrm{Var}(X_1/3 + 2X_2/3) = p(1-p)(1/9 + 4/9). \quad (8.13)$$

(Was uns wiederum nicht sehr überrascht.)

Gilt für eine Folge von Schätzern $MQA(\hat{\theta}_n) \to 0$, so ist sie offenbar konsistent.

Natürliche Fragen an dieser Stelle sind:

1. Wie schnell konvergieren gute Schätzer typischer Weise, d.h. wie schnell fällt $MQA(\hat{\theta}_n)$ ab?

2. Wie findet man einen guten Schätzer, gibt es 'optimale' Schätzer. Ist der MLS ein guter Schätzer?

Wir geben nur ohne Beweis einige Resultate an, die Einblick bezüglich der obigen Fragen geben.

Ein wichtiges Resultat ist die sogenannte Cramér-Rao Schranke, die uns eine untere Schranke für die mögliche Güte eines Schätzers darstellt.

Satz 8.16 (Cramér-Rao)**.** *Seien $X_1, \ldots, X_n$ unabhängig identisch verteilt mit Dichte $f(x;\theta)$, $\theta \in \Theta$ mit $\Theta \in \mathbb{R}$ offen. Bezeichne weiters*

$$I(\theta) := \mathbb{E}_\theta \left(\frac{\partial}{\partial \theta} \log f(X_1;\theta) \right)^2 = \int \left(\frac{\partial}{\partial \theta} \log f(x;\theta) \right)^2 f(x,\theta)\, dx$$

die zu X_1 (bzw. zu $f(.;\theta)$) gehörige Fisher-Information.

Sei weiters $T : \mathbb{R}^n \to \Theta$ ein unverzerrter *Schätzer. Dann gilt (unter milden Regularitätsvoraussetzungen an f)*

$$\mathrm{Var}_\theta(T) \geq \frac{1}{nI(\theta)}. \tag{8.14}$$

Beweis. Wir betrachten zuerst den Fall $n = 1$ und definieren die Zufallsvariable

$$U_1(\theta) = \frac{\partial}{\partial\theta} \log f(X_1, \theta).$$

Falls (f_θ) hinreichend regulär ist, dass wir Integration und Differentiation vertauschen dürfen, erhalten wir

$$\mathbb{E}[U_1(\theta)] = \int \frac{\partial}{\partial\theta} \log f(x,\theta) f(x,\theta)\, dx = \int \frac{\partial}{\partial\theta} f(x,\theta)\, dx = \frac{\partial}{\partial\theta} \int f(x,\theta)\, dx = 0. \tag{8.15}$$

Damit folgt dann

$$\mathrm{Var}[U_1(\theta)] = \mathbb{E}[U_1(\theta)^2] = \int \left(\frac{\partial}{\partial\theta} \log f(x,\theta)\right)^2 f(x,\theta)\, dx = I(\theta). \tag{8.16}$$

Weil T unverzerrt ist, erhalten wir

$$\begin{aligned} 1 &= \frac{\partial}{\partial\theta} \int T(x) f(x,\theta)\, dx = \int T(x) \frac{\partial}{\partial\theta} f(x,\theta)\, dx = \int T(x) \left(\frac{\partial}{\partial\theta} \log f(x,\theta)\right) f(x,\theta)\, dx \\ &= \mathbb{E}[TU_1(\theta)] = \mathrm{Cov}[TU_1(\theta)] \leq \sqrt{\mathrm{Var}\, T\, \mathrm{Var}\, U_1(\theta)}, \end{aligned}$$

wobei wir im letzten Schritt die Cauchy-Schwarz Ungleichung verwendet haben. Es folgt $\mathrm{Var}\, T \geq 1/I(\theta)$.

Für allgemeines n setzen wir für $k \leq n$

$$U_k(\theta) = \frac{\partial}{\partial\theta} \log f(X_k, \theta)$$

und $U(\theta) = U_1(\theta) + \ldots + U_n(\theta)$.

Dann gilt

$$\begin{aligned} \mathbb{E}[U(\theta)] &= 0 \\ \mathrm{Var}[U(\theta)] &= n\, \mathrm{Var}[U_1(\theta)] = nI(\theta). \\ 1 &= \int T(x_1, \ldots, x_n) \frac{\partial}{\partial\theta} \prod_{i\leq n} f(x_i,\theta)\, dx_1 \ldots dx_n \\ &= \int T(x_1, \ldots, x_n) \left(\sum_{i\leq n} \frac{\partial}{\partial\theta} \log f(x_i,\theta)\right) \prod_{i\leq n} f(x_i,\theta)\, dx_1 \ldots dx_n = \mathrm{Cov}[U(\theta)T] \end{aligned}$$

Die gewünschte Schranke folgt dann wieder wie oben. □

Um so etwas wie die 'Norm' eines Fehlers zu erhalten, sollten wir natürlich die Wurzel aus $MQA(T)$ bzw. $\mathrm{Var}(T)$ nehmen. In diesem Sinn besagt (8.14), dass ein optimaler (unverzerrter) Schätzer stets einen Fehler der Größenordnung $\frac{1}{\sqrt{n}}$ aufweisen wird.

Falls wir eine Schätzer finden, sodass in der Cramer-Rao Schranke Gleichheit gilt, haben wir offenbar einen bestmöglichen Schätzer identifiziert.

Wenn man sich auf geeignete Klassen von Dichten ('exponentielle Familien') einschränkt, kann man zeigen, dass der MLS asymptotisch unverzerrt ist und auch die Cramér-Rao Schranke asymptotisch annimmt.

8.5 Konfidenzintervalle

Beispiel 8.17. Prognose nach einer Wahl (exit poll): Aufgrund einer repräsentativen Stichprobe am Ausgang von Wahllokalen wird angegeben, dass etwa 32% aller Wähler*innen für Partei A gestimmt haben. Weiters wird angegeben, dass das 95% Konfidenzintervall $(29\%, 35\%)$ beträgt. *Wie ist das genau zu interpretieren?*

Ein wesentliches Ziel dieses Abschnitts ist es die obige Frage zu beantworten. Wir gehen zunächst von einem mathematisch einfacheren Beispiel aus:

Beispiel 8.18. Seien $X_1, \dots, X_n$ standardnormalverteilt nach $N(\mu, \sigma^2)$ wobei σ bekannt ist und μ aus einem sample geschätzt werden soll. Wie wir oben schon besprochen haben ist der MLS gerade

$$\hat{\mu}(x_1, \dots, x_n) = \bar{x}_n.$$

Das Stichprobenmittel $\bar{X}_n$ erfüllt

$$\bar{X}_n \sim N(\mu, \sigma^2/n) \tag{8.17}$$
$$\iff \bar{X}_n - \mu \sim N(0, \sigma^2/n) \tag{8.18}$$
$$\iff \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1). \tag{8.19}$$

Ist $Z \sim N(0,1)$, so gilt $\mathbb{P}(-1.96 < Z < 1.96) = 0.95$. Damit erhalten wir

$$\mathbb{P}\left(-1.96 < \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.95 \tag{8.20}$$
$$\iff \mathbb{P}\left(\bar{X}_n - 1.96\sigma/\sqrt{n} < \mu < \bar{X}_n + 1.96\sigma/\sqrt{n}\right) = 0.95 \tag{8.21}$$
$$\iff \mathbb{P}\left(\text{das Intervall } \left(\bar{X}_n \pm 1.96\sigma/\sqrt{n}\right) \text{ enthält } \mu\right) = 0.95. \tag{8.22}$$

D.h. wir erhalten ein *zufälliges* Intervall dessen Endpunkte aus der Zufallsvariable $\bar{X}$ berechnet werden, sodass der Parameter μ mit hoher Wahrscheinlichkeit im Intervall liegt.

Definition 8.19. *Seien $X_1, \dots, X_n$ nach $\mathbb{P}_\theta$ verteilt. Seien zwei Statistiken $a(X_1, \dots, X_n)$, $b(X_1, \dots, X_n)$ gegeben. Dann ist*

$$(a(X_1, \dots, X_n), b(X_1, \dots, X_n))$$

ein Konfidenzintervall *zum Konfidenzlevel* 95% *falls für alle* θ

$$\mathbb{P}_\theta(a(X_1,\dots,X_n) < \theta < b(X_1,\dots,X_n)) \geq 95\%. \tag{8.23}$$

Man nennt $(a(X_1,\dots,X_n), b(X_1,\dots,X_n))$ dann ein 95% Konfidenzintervall. (Natürlich werden werden Konfidenzintervalle für andere Prozentsätze analog definiert.)

Wir bemerken, dass $a(X_1,\dots,X_n), b(X_1,\dots,X_n)$ natürlich nicht vom Parameter θ abhängen dürfen.

Im allgemeinen möchte man natürlich $a(X_1,\dots,X_n), b(X_1,\dots,X_n)$ so konstruieren, dass

1. die Länge von $(a(X_1,\dots,X_n), b(X_1,\dots,X_n))$ klein ist, während
2. die Wahrscheinlichkeit $\mathbb{P}_\theta(a(X_1,\dots,X_n) < \theta < b(X_1,\dots,X_n))$ groß ist.

Ist eine der Intervallgrenzen eines Konfidenzintervals gleich $\pm\infty$, spricht man von einem *einseitigen* Konfidenzintervall.

Beispiel 8.20 (Normalverteilungsapproximation des Binomialkonfidenzintervalls)**.** Wir gehen wieder zurück zum Beispiel mit den exit polls. Angenommen insgesamt jede μ-te Wähler*in wählt Partei A. Wir modellieren das indem wir annehmen, dass $X_1, X_2, \dots$ unabhängig Bernoulli-verteilt mit Parameter μ sind. Der MLS für μ ist

$$\hat{\mu}(X_1,\dots,X_n) = \bar{X}_n$$

und wir erhalten $\mathrm{Var}_\mu(\bar{X}_n) = \mu(1-\mu)/n$. Wie schon in Beispiel 7.16 approximieren wir die Binomialverteilung mit der Normalverteilung und erhalten für das 95% Konfidenzintervall

$$\begin{cases} 0.95 = \mathbb{P}(|\hat{\mu} - \mu| \leq \varepsilon) \approx & \mathbb{P}(|N(0, \mu(1-\mu)/n)| \leq \varepsilon) \\ = & \mathbb{P}\left(|N(0,1)| \leq \varepsilon/\sqrt{\mu(1-\mu)/n}\right). \end{cases} \tag{8.24}$$

Indem wir wieder $0.95 = \mathbb{P}(|N(0,1)| \geq 1.96$ beachten, erhalten wir aus (8.24)

$$\varepsilon/\sqrt{\mu(1-\mu)/n} \approx 1.96. \tag{8.25}$$

Das heißt, wir erhalten, dass mit 95% Wahrscheinlichkeit

$$\mu \in \left(\hat{\mu} - 1.96 * \sqrt{\mu(1-\mu)/n}, \hat{\mu} + 1.96 * \sqrt{\mu(1-\mu)/n}\right).$$

Es handelt sich hier um *kein* Konfidenzintervall, da die beiden Intervallgrenzen noch vom Parameter μ abhängen.

Um doch zu einem Intervall zu gelangen, kann man hier μ durch den Wert Schätzwert $\hat{\mu}$ ersetzen und bekommt das folgende ungefähre Konfidenzintervall

$$\left(\hat{\mu} - 1.96 * \sqrt{\hat{\mu}(1-\hat{\mu})/n}, \hat{\mu} + 1.96 * \sqrt{\hat{\mu}(1-\hat{\mu})/n}\right),$$

welches auch Standardintervall (zum level 95%) genannt wird.

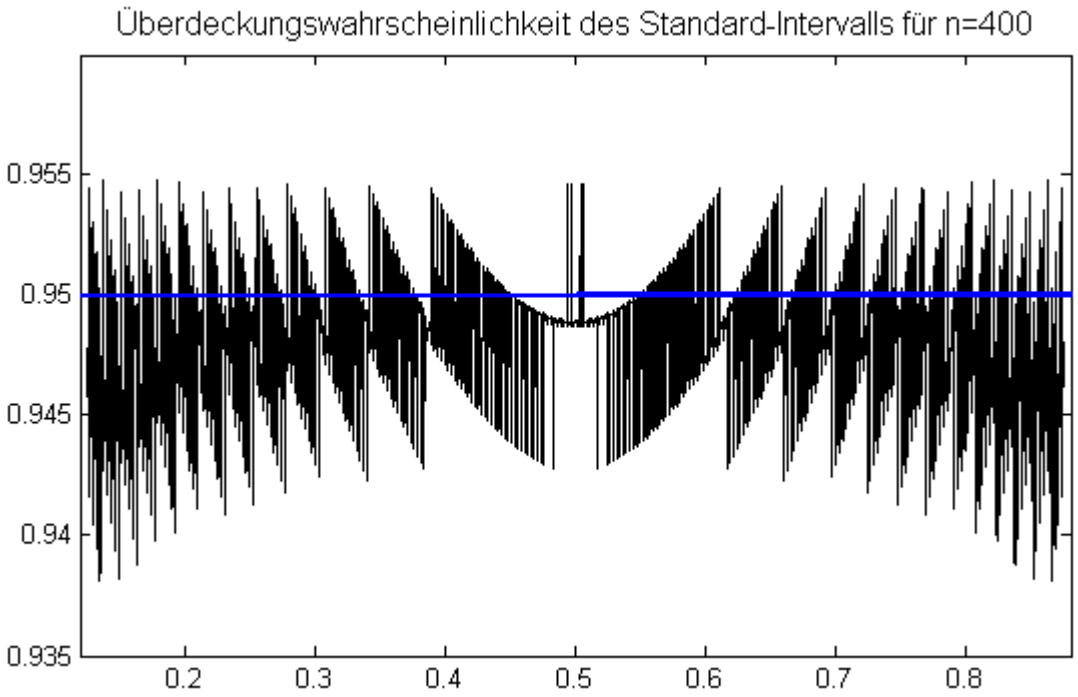


Abbildung 17: Die Verwendung des Standard-Intervalls kann problematisch sein.

Fragt man z.B. Beispiel $n = 1003$ Menschen und erhält dass 321 Menschen Partei A gewählt haben, ergibt sich $\hat{\mu} = 321/1003 = 32\%$ und ein Standardintervall von $(29\%, 35\%)$.

In der Abbildung ist als Beispiel die Überdeckungswahrscheinlichkeit für $n = 400$ und $p \in [0.125, 0.875]$ illustriert.

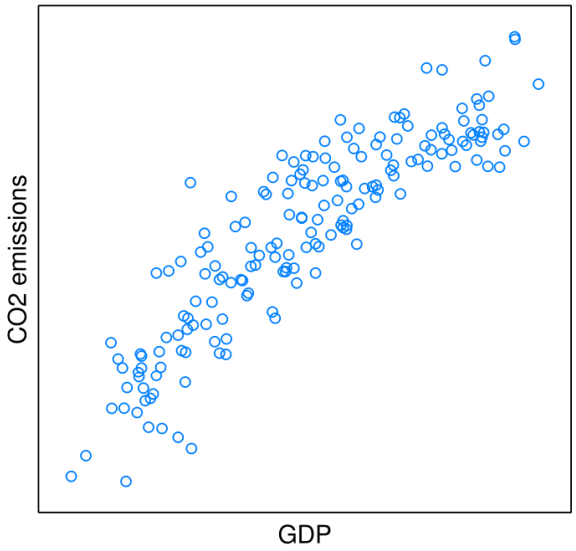
Bei gegebenem n hängt die Breite des Standardintervalls von $\hat{\mu}$ ab und ist für $\hat{\mu} = 1/2$ am größten. Wenn man garantieren will, dass die Schwankung höchstens $\pm 3\%$ beträgt muß man also fordern, dass

$$1.96\sqrt{1/2(1-1/4)/n} \approx 0.03 \iff n \approx 1068.$$

Um eine 'Schwankungsbreite' von $\pm 3\%$ zu erreichen, ist es also eine gute Idee rund 1100 Leute zu befragen.

Interpretation von Konfidenzintervallen:

- Der Parameter θ ist deterministisch aber unbekannt.
- Wenn wir unser Experiment erneut durchführen würden, bekämen wir neue Daten $x'_1, \ldots, x'_n$ und entsprechend ein neues Konfidenzintervall. Wenn wir unser Experiment 100 mal wiederholen, dann überdeckt das 95% Konfidenzintervall in (durchschnittlich) 95 Fällen den wahren Parameter.



NICHT Interpretation: Nachdem wir Daten $x_1, \ldots, x_n$ beobachtet haben, würden wir sehr gerne sagen, dass der wahre Parameter mit 95% Wahrscheinlichkeit im 95% Konfidenzintervall $(a(x_1, \ldots, x_n), b(x_1, \ldots, x_n))$ liegt. Leider ist diese Interpretation nicht zulässig, der wahre Parameter ist *keine* Zufallsvariable.

Tatsächlich ist diese Art der Interpretation der *Bayes'schen Statistik* vorbehalten, in der man eine sogenannte *a priori* Verteilung auf den möglichen Werten des Parameters θ vorgibt und nach Beobachtung der Daten dann ein *a posteri* Verteilung des Parameters erhält. Zur Unterscheidung spricht man in der Bayes'schen Statistik von *Kredibilitätsintervallen*.

8.6 Lineare Regression

Erinnern Sie sich an das Beispiel 8.2, indem x_i die Monate und y_i die Hauspreise in Oxford bezeichneten. Hier ist unsere Vorstellung, dass

- x_i die erklärende Variable ist, und
- y_i die resultierende Variable.

Beispiel 8.21. Sei x = das BIP pro Kopf und y = die CO_2-Emissionen pro Kopf in verschiedenen Ländern. Wie ändert sich y mit x?

Ein einfaches Modell wäre

$$y = \alpha + \beta x + \text{“Fehler”}.$$

Etwas präziser machen wir die folgende Annahmen.

Definition 8.22. *Ein* lineares Regressionsmodell *ist von der Form*

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

wobei ε_i unabhängige normalverteilte Zufallsvariablen mit Varianz σ^2 sind. In Vektornotationen können wir dieses Modell auch in folgender Form schreiben:

$$Y = \alpha + \beta x + \varepsilon. \tag{8.26}$$

Hier betrachten wir x_i als fest und Y_i als Zufallsvariablen. Deshalb schreiben wir x und Y für die Vektoren $x = (x_1, \ldots, x_n)$ und $Y = (Y_1, \ldots, Y_n)$. Natürlich sind auch ε_i Zufallsvariablen.

Wir betrachten α, β (und vielleicht auch σ^2, obwohl wir es nicht tun) als unbekannte Konstanten, die wir durch Daten schätzen müssen. Unsere Methode der Wahl ist der Maximum-Likelihood Schätzer.

Satz 8.23. *Seien $y = (y_1, \ldots, y_n)$ die beobachteten Werte für $Y = (Y_1, \ldots, Y_n)$. Sei $\bar{x} = (1/n)\sum_{i=1}^n x_i$ und $\bar{y} = (1/n)\sum_{i=1}^n y_i$. Der Maximum-Likelihood Schätzer für α, β erfüllt*

$$(\hat{\alpha}, \hat{\beta}) = \arg\min S(\alpha, \beta)$$

wobei

$$S(\alpha, \beta) = \sum_{i=1}^n (y_i - \beta x_i - \alpha)^2.$$

Explizit erhalten wir

$$\hat{\beta} = \frac{\frac{1}{n}\sum_{i=1}^n x_i y_i - \bar{y}\bar{x}}{\frac{1}{n}\sum_{i=1}^n x_i^2 - \bar{x}^2}$$
$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}.$$

Der Satz besagt, dass $\hat{\alpha}$ und $\hat{\beta}$ durch Minimierung des *quadratischen Fehlers* (quadratic loss) erhalten werden. Aus diesem Grund werden $\hat{\alpha}$ und $\hat{\beta}$ auch **Kleinste-Quadrate-Schätzer** genannt.

Bemerkung 8.24. Wenn wir X als unabhängige, identisch verteilte Zufallsvariablen betrachten würden (was sie genau nicht sind!) dann könnte man aus dem Gesetz der Großen Zahlen schließen, dass

$$\hat{\beta} \approx \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}$$

Beweis des Satzes. Die Zufallsvariablen $Y_i = \alpha + \beta x_i + \varepsilon_i$ sind unabhängig (weil x nicht zufällig ist und ε_i unabhängig sind) und ihre Dichtefunktionen sind durch

$$f_{Y_i}(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(y_i - \alpha - \beta x_i)^2)$$

gegeben. Deshalb ist die Likelihood Funktion gleich

$$L(\alpha, \beta) = \frac{1}{(2\pi\sigma^2)^{n/2}} \prod_{i=1}^n \exp(-\frac{1}{2\sigma^2}(y_i - \alpha - \beta x_i)^2).$$

Die log-Likelihood erfüllt

$$\ell(\alpha, \beta) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2.$$

Das Problem von log-Likelihood Maximierung ist deshalb äquivalent zu dem Problem der Minimierung von

$$S(\alpha, \beta) = \sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$$

wie gewünscht. Außerdem gilt

$$\begin{aligned}\frac{\partial S}{\partial \alpha} &= -2\sum_{i=1}^{n}(y_i - \alpha - \beta x_i) \\ &= -2(n\bar{y} - n\alpha - n\beta\bar{x}) \\ \frac{\partial S}{\partial \beta} &= -2\sum_{i} x_i(y_i - \alpha - \beta x_i).\end{aligned}$$

Wenn wir $\partial S/\partial\alpha = \partial S/\partial\beta = 0$ einsetzen, erhalten wir zwei Gleichungen die wir lösen können:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x},$$

während

$$\sum_i x_i y_i - \sum_i x_i(\bar{y} - \hat{\beta}\bar{x}) - \hat{\beta}\sum_i x_i^2 = 0.$$

Das heißt,

$$\hat{\beta}(\sum_i x_i^2 - \bar{x}\sum_i x_i) = \sum_i x_i y_i - \bar{y}\sum_i x_i.$$

Wir dividieren durch n und erhalten

$$\hat{\beta} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{y}\bar{x}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2},$$

wie gewünscht. □

Die Gerade

$$y = \hat{\alpha} + \hat{\beta}x$$

wird als *Regressionsgerade* bezeichnet. Wir beobachten, dass $(\bar{x}, \bar{y})$ immer auf dieser Linie liegt.

Jetzt untersuchen wir ein Paar Eigenschaften vom MLS $\hat{\alpha}, \hat{\beta}$. Sind $\hat{\alpha}, \hat{\beta}$ biased? Was sind ihre MQA?

Um diese Fragen zu beantworten ist es praktisch das Problem neu zu parametrisieren. Wir setzen

$$w_i = x_i - \bar{x}$$

und schreiben

$$Y = \alpha + \beta w + \varepsilon \tag{8.27}$$

Das ist äquivalent zu (8.26) obwohl die Werte von α und β sich ändern werden. In dieser Form ist (8.27) auch eine Regression und deshalb erhalten wir aus Satz 8.23 dass der MLS

$$\hat{\alpha} = \bar{y}, \hat{\beta} = \frac{\sum_i w_i y_i}{\sum w_i^2} \tag{8.28}$$

erfüllt.

Satz 8.25. *Betrachten wir das Regressionsmodell (8.27) und sei $\hat{\alpha}, \hat{\beta}$ der entsprechende MLS. Dann gilt:*

- $\hat{\alpha} \sim N(\alpha, \sigma^2/n)$
- $\hat{\beta} \sim N(\beta, \frac{\sigma^2}{\sum_i w_i^2})$
- $\hat{\alpha}, \hat{\beta}$ *sind unabhängig.*

Bemerkung 8.26. Bemerken Sie, dass $\hat{\alpha}, \hat{\beta}$ unbiased sind. Wir können auch aus diesem Satz schließen, dass $\hat{\alpha}$ konsistent ist. Falls $\sum_{i=1}^n w_i^2 \to \infty$, ist auch $\hat{\beta}$ konsistent.

Diese zwei Eigenschaften gelten auch für (8.26), Unabhängigkeit jedoch gilt nur in der Form (8.27).

Beweis. Erinnern Sie sich an die explizite Form des MLSes in (8.28). Da $\hat{\alpha} = \bar{Y}$ erhalten wir sofort die Verteilung von $\hat{\alpha}$. $\hat{\beta}$ ist auch eine Linearkombination von unabhängigen normalverteilten Zufallsvariablen, und ist deshalb auch normalverteilt. Es reicht daher, Erwartungswert und Varianz zu berechnen:

$$\begin{aligned}
\mathbb{E}(\hat{\beta}) &= \|w\|_2^{-2} \mathbb{E}(\sum_i w_i Y_i) \\
&= \|w\|_2^{-2} \sum_i w_i \mathbb{E}(Y_i) \\
&= \|w\|_2^{-2} \sum_i w_i (\alpha + \beta w_i) \\
&= \beta.
\end{aligned}$$

Außerdem gilt, da ε_i unabhängig sind,

$$\begin{aligned}
\operatorname{Var}(\hat{\beta}) &= \|w\|_2^{-4} \sum_i w_i^2 \operatorname{Var}(Y_i) \\
&= \|w\|_2^{-4} \sum_i w_i^2 \sigma^2 \\
&= \sigma^2 / \|w\|_2^2,
\end{aligned}$$

wie gewünscht.

Die Unabhängigkeit folgt aus einer wichtigen Eigenschaft von (gemeinsam) normalverteilten Zufallsvariablen, die wir jedoch noch nicht gesehen haben: für solche Zufallsvariablen, impliziert unkorreliert schon unabhängig. Wir setzen diese Eigenschaft voraus und prüfen nur dass $\mathrm{Cov}(\hat{\alpha}, \hat{\beta}) = 0$ ist. Man hat

$$\begin{aligned}
\mathrm{Cov}(\hat{\alpha}, \hat{\beta}) &= \|w\|_2^{-2} \,\mathrm{Cov}(\bar{Y}, \sum_i w_i Y_i) \\
&= \frac{1}{n\|w\|_2^2} \sum_{i,j} \mathrm{Cov}(w_i Y_i, Y_j) \\
&= \frac{1}{n\|w\|_2^2} \sum_{i=1}^{n} w_i \,\mathrm{Var}(Y_i^2) \\
&= \frac{\sigma^2}{n\|w\|_2^2} \sum_i w_i \\
&= 0,
\end{aligned}$$

wie gewünscht. □

8.7 Hypothesentest und Neyman–Pearson lemma

Der Hypothesentest ist einer der wichtigsten Bereiche der Statistik. Eine statistische Hypothese ist eine Behauptung oder Vermutung über die Verteilung einer oder mehrerer Zufallsvariablen. Ein **Test einer statistischen Hypothese** ist eine Regel oder ein Verfahren zur Entscheidung, ob diese Behauptung abgelehnt werden soll.

Ein typisches Beispiel wäre das folgende. Ein*e Wissenschaftler*In hat eine neue Theorie, macht Beobachtungen und zeichnet dann die Daten auf. Gibt es in diesen Daten irgendwelche Beweise, um die Hypothese zu verwerfen, dass die derzeit vorherrschende Theorie richtig ist, und um zu schließen dass die neue Theorie besser geeignet ist?

Wir führen einen theoretischen Rahmen ein, um solche Fragen zu beantworten. Dieser Rahmen wird als **Neyman–Pearson** Rahmen gennant. Angenommen, wir haben Daten $x = (x_1, x_2, \ldots, x_n)$ die sich ergeben nachdem n mal aus einer Verteilung mit Dichtefunktion bzw. Wahrscheinlichkeitsfunktion f gezogen wird. Wir haben zwei Hypothesen über f. Aufgrund der Daten wird die eine akzeptiert, die andere abgelehnt. Die beiden Hypothesen haben jedoch anderen philosophischen Status. Die erste, Nullhypothese genannt, und bezeichnet durch H_0, ist eine konservative Hypothese, die nicht abgelehnt werden darf, ohne dass die Beweise absolut klar sind.

Die zweite wird sonst als die Alternativhypothese bezeichnet und spezifiziert die Art der Abweichung von der Nullhypothese. Sie wird mit H_1 bezeichnet.

Zum Beispiel wäre für die obige Wissenschaftler*in H_0 die vorherrschende Theorie, während H_1 die vorgeschlagene neue Theorie würde. Um die vorherrschende Theorie abzulehnen, müssen die Beweise absolut klar sein: sie müssen eine Art Gegenbeispiel sein, so zu sagen.

Um ein anderes Beispiel zu geben, sollte die Jury in einem Mordprozess als Nullhypothese nehmen, dass die Angeklagt*e unschuldig ist. Bei einem Test of eine (seltene) Krankheit wäre die Nullhypothese der Normalzustand, d.h. dass die getestete Person gesund ist, etc.

Beispiel 8.27. Wir wollen testen, ob eine Münze fair ist (Nullhypothese H_0) d.h. ob sie gleich oft Kopf wie Zahl zeigt. D.h. wir wollen testen, ob eine Münze gemäß $B(1/2)$ verteilt oder nach $B(p)$ für $p \neq 1/2$. Sei $\bar{X}_n = \sum_{i \leq n} X_i/n$ die durchschnittliche Anzahl der Köpfe nach n-maligem Werfen der Münze. Offenbar werden wir H_0 verwerfen, wenn $\bar{X}_n$ deutlich von $1/2$ abweicht. Wie groß soll diese Abweichung sein?

Typischer Weise nimmt man an, dass die Dichte f Element einer Familie $\{f(\cdot, \theta) : \theta \in \Theta\}$ ist. Die beiden Hypothesen werden dann über Teilmengen Θ_0, Θ_1 der Parametermenge Θ beschrieben, wobei H_i dem Fall entspricht, dass $f = f(\cdot, \theta)$ für $\theta \in \Theta_i$:

$$H_0 : \theta \in \Theta_0 \quad \text{gegen} \quad H_1 : \theta \in \Theta_1.$$

Man verlangt hier, dass $\Theta_0 \cap \Theta_1 = \emptyset$. Im Gegensatz dazu kann $\Theta_0 \cup \Theta_1 = \Theta$ erfüllt sein, muß aber nicht.

Wichtige Beispiele von Tests, die nicht von dieser Form sind *Anpassungstests* (engl: *goodness-of-fit*) bei

$$H_0 : f = f_0 \quad \text{gegen} \quad H_1 : f \neq f_0.$$

Häufig wird dies durch asymptotische Betrachtungen der empirischen Verteilungsfunktion realisiert.

Eine weitere Alternative sind Tests der Form

$$H_0 : f = f_0 \quad \text{gegen} \quad H_1 : f = f_1.$$

Eine Hypothese die f komplett festlegt heißt *simpel*, z.B. $\theta = \theta_0$, andernfalls *zusammengesetzt*.

Angenommen wir wollen H_0 gegen H_1 testen. Ein Test wird durch eine *kritische Region* C beschrieben. (Im Münzwurfbeispiel wäre C die Menge aller Folgen von Ausgängen bei denen es deutlich mehr Köpfe gibt.) Das Komplement $\Theta \setminus C$ bezeichnen wir mit $\bar{C}$. Falls

$$\mathbf{x} = (\mathbf{x_1}, \ldots, \mathbf{x_n}) \in \mathbf{C}$$

wir H_0 *verworfen*. Ist andererseits

$$\mathbf{x} = (\mathbf{x_1}, \ldots, \mathbf{x_n}) \in \bar{\mathbf{C}}$$

wird H_0 *akzeptiert* oder, genauer, *nicht zurückgewiesen*.

Beispiel 8.28. Wir betrachten wieder die Münzwurfsituation aus Beispiel **??**. Angenommen die Münze wird 20 mal geworfen. Wir betrachten die Statistik

$$T(X_1, \ldots, X_{20}) := |\bar{X}_{20} - 1/2|.$$

Die Münze ist eher unfair wenn T deutlich größer als 0 ist. Wir wählen daher

$$C = \{(x_1, \ldots, x_{20}) : T(x_1, \ldots, x_{20}) > k\},$$

wobei wir uns noch für einen geeigneten Wert von k entscheiden müssen. (Offenbar passiert es bei recht kleinem k leicht, dass wir eine faire Münze für fair halten. Bei großem k, ist das Problem, dass wir vielleicht eine unfaire Münze nicht als solche erkennen.)

Prinzipiell unterscheiden wir beim Testen von Hypothesen zwei Arten von Fehlern:

1. *Fehler erster Art*: H_0 ist wahr, wird aber verworfen. (Die Münze ist fair, aber wir fühlen uns betrogen. Die Unschuldige kommt lebenslang ins Gefängnis. Der eigentlich Gesunde wird fälschlich als krank identifiziert, vielleicht wird ihm das gesunde Bein abgenommen, oder so.). Man spricht hier auch von *falsch positiv*.

   Da H_0 als 'konservativ' / 'Normalfall' angesehen wird, gilt diese Art Fehler als der schwerwiegende Fehler.

2. *Fehler zweiter Art*: H_0 ist falsch, wird aber akzeptiert. Man spricht hier auch von *falsch negativ*.

Basierend auf der Philosophie, dass Fehler erster Art jedenfalls zu vermeiden sind, fixiert man eine Wahrscheinlichkeit für den Fehler erster Art, typischer Weise 1% oder 5% und definiert dann die kritische Region C sodass der Fehler zweiter Art möglichst minimiert wird.

Falls H_0 simpel ist, $\Theta_0 = \{\theta_0\}$, nennt man die Wahrscheinlichkeit für einen Fehler erster Art *Signifikanzniveau* des Tests. Ist H_0 zusammengesetzt, verwendet man stattdessen

$$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(X \in C)$$

(d.h. man verlangt, dass der Fehler erster Art in jedem Fall möglichst klein bleibt).

Beispiel 8.29. Im Münzbeispiel betrachten wir $\mathbb{P}_{H_0}(|\bar{X}_n - 1/2| > k) = \mathbb{P}_{1/2}(|\bar{X}_n - 1/2| > k)$. Der Fall von 4 oder mehr Würfen Überhang entspricht $k = 3.5/20$. Hier erhalten wir $\mathbb{P}_{1/2}(|X_n - 1/2| > 3.5/20) \approx 0.115$. Das bedeutet, ein so konstruierter Test hat eine Wahrscheinlichkeit für Fehler erster Art von 11.5%. Wenn wir (z.B.) 14 Köpfe sehen, wollen wir also noch nicht auf eine unfaire Münze schließen. Das Ergebnis ist *nicht signifikant*.

Wir betrachten stattdessen einen Überhang von 5 oder mehr. Das entspricht $k = 4.5/20$ womit wir $\mathbb{P}_{1/2}(|X_n - 1/2| > 4.5/20) \approx 0.041 < 0.05$ erhalten.

Zusammenfassend: Um einen Münze auf fairness zu testen, werfen wir sie 20 mal. Wenn wir höchstens 5 Köpfe oder höchstens 5 mal Zahl sehen, werden wir sie als unfair betrachten. Das Signifikanzniveau dieses Tests liegt bei 4.1%.

Eine andere Interpretation dieses Ergebnisses ist dass

$$\mathbb{P}_{1/2}(\bar{X}_{20} - 4.5/20 < 1/2 < \bar{X}_{20} + 4.5/20) \approx 95\%.$$

Das Illustriert den Zusammenhang zwischen Signifikanzniveau und Konfidenzintervall.

Wir wollen auch noch kurz in diesem Beispiel erläutern, was der sogenannte p-Wert ist: Wenn wir (z.B.) 14 Köpfe sehen, können wir uns fragen wie wahrscheinlich eine so große (oder noch größere) Abweichung von den 10 erwarteten Köpfen unter H_0 ist. Die Antwort darauf ist gerade

$$\mathbb{P}_{1/2}(|\bar{X}_n - 1/2| > 3.5/20) \approx 0.115.$$

In diesem Fall ist der p-Wert also 0.115 oder 11.5%. Dieser Wert ist relativ klein und unterstützt die Nullhypothese nicht. Wegen $11.5\% > 5\%$ ist das Ergebnis jedoch *nicht signifikant*, es reicht nicht um die Nullhypothese zu verwerfen.

Die *likelihood* einer einfachen Hypothese H (entsprechend dem Parameter θ^*) ist gegeben durch

$$L_x(H) = f_{\theta^*}(x).$$

Im zusammengesetzten Fall, d.h. einer Hypothese H mit Parameterbereich Θ^* verwendet man

$$L_x(H) = \sup_{\theta \in \Theta^*} f_{\theta^*}(x).$$

Die *likelihood ratio* von zwei Hypothesen H_0, H_1 ist gegeben durch

$$L_x(H_0, H_1) = L_x(H_1)/L_x(H_0).$$

Definition 8.30. *Ein Test bei dem die kritische Region von der Form*

$$C = \{x : L_x(H_1)/L_x(H_0) > k\}$$

für eine Konstante k ist, heißt liklihood ratio Test*.*

Im Fall von simplen Hypothesen sind likelihood ratio tests optimal im Sinne des folgenden Lemmas:

Lemma 8.31 (Neyman-Pearson Lemma)**.** *Seien $H_0 : f = f_0$ und $H_1 : f = f_1$. Hypothesen die gegeneinander getestet werden sollen. Unter allen Tests mit gegebenem Signifikanzniveau α hat der likelihood ratio test (gegeben durch $C = \{x : f_1(x)/f_0(x) > k\}$) den kleinsten Fehler zweiter Art. Hier ist k so gewählt, dass*

$$\alpha = \mathbb{P}_0(X \in C) = \int_C f_0(x)\, dx.$$

Beweis. Gegeben sei ein Test mit Signifikanzniveau α und kritischer Region D sodass

$$\mathbb{P}_0(X \in D) \leq \alpha.$$

Wir betrachten C und k wie in der Formulierung des Lemmas und bemerken dass

$$0 \leq (I_C(x) - I_D(x))(f_1(x) - k f_0(x)).$$

(Wenn der rechte Faktor > 0 ist, dann ist $x \in C$ und $I_C(x) = 1$. Wenn der rechte Faktor < 0 ist, dann ist $x \notin C$ und $I_C(x) = 0$.)

Wir schließen daraus

$$0 \leq \int (I_C(x) - I_D(x))(f_1(x) - k f_0(x))\, dx \tag{8.29}$$

$$= \mathbb{P}_1(X \in C) - \mathbb{P}_1(X \in D) - k[P_0(X \in C) - \mathbb{P}_0(X \in D)] \tag{8.30}$$

$$= \mathbb{P}_1(X \in C) - \mathbb{P}_1(X \in D) - k[\alpha - \mathbb{P}_0(X \in D)] \tag{8.31}$$

$$\leq \mathbb{P}_1(X \in C) - \mathbb{P}_1(X \in D). \tag{8.32}$$

Daher ist $\mathbb{P}_1(X \notin D)$ mindestens so groß wie $\mathbb{P}_1(X \notin C)$. □

Literatur

[1] N. Curien. Yet another proof of the strong law of large numbers. arXiv:2109.04315, 2021.

[2] H.-O. Georgii. Probabilistic aspects of entropy. In A. Greven, G. Keller, and G. Warnecke, editors, *Entropy*, chapter 3, pages 37–54. Princeton University Press, 2003.

[3] D. Kahneman. *Thinking, Fast and Slow.* Farrar, Straus and Giroux, New York, NY, 2011.

[4] V. M. Panaretos. *Statistics for Mathematicians: A Rigorous First Course.* Birkhäuser/Springer, [Cham], 2016.