2 Grundlagen aus der Wahrscheinlichkeitsrechnung

In diesem Kapitel sind die wichtigsten Konzepte der Wahrscheinlichkeitsrechnung zusammengestellt, die für die Zwecke unserer Vorlesung wichtig sind. Sie beschränken sich der Einfachheit halber auf den Fall endlicher und abzählbar unendlicher Wahrscheinlichkeitsräume.

Eine sehr gute Einführung in die Thematik findet sich in den ersten Kapiteln des Buchs „Probability and Computing – Randomized Algorithms and Probabilistic Analysis“ von M. Mitzenmacher und E. Upfal (s. Literaturverzeichnis auf der Webseite).

Ein Wort noch zur Verwendung dieses Kapitels und zu seiner Rolle in der Abschlussprüfung. Eigentlich wird angenommen, dass die folgenden Grundbegriffe der Wahrscheinlichkeitsrechnung aus der Veranstaltung „Stochastik für Informatiker“ bekannt sind. Infolgedessen sind sie nicht in erster Linie Prüfungsstoff.

- Wahrscheinlichkeitsraum, Verteilung
- Zufallsvariable, Erwartungswert, Linearität des E-Wertes, Markov-Ungleichung
- Varianz, Chebychev-Ungleichung, Jensensche Ungleichung
- uniforme Verteilung, Binomialverteilung, geometrische Verteilung
- bedingte Wahrscheinlichkeiten, bedingte Erwartungswerte
- Unabhängigkeit von Ereignissen und Zufallsvariablen, Rechenregeln.

Diese Begriffe werden wiederholend dargestellt, eventuell etwas modifiziert, und mit Beispielen unterlegt. Der Unterschied zur Behandlung in einer Stochastik-Vorlesung ist insbesondere, dass nur endliche und abzählbar unendliche Wahrscheinlichkeitsräume betrachtet werden. Anhand von Beispielen, die nahe an algorithmischen Anwendungen liegen, wird illustriert, wie die abstrakten Konzepte und Rechenregeln in algorithmischen und diskreten Anwendungen verwendet werden, und diese Verwendung wird im konkreten Kontext von Algorithmen und diskreten Strukturen eingeübt.

Wer also die wahrscheinlichkeitstheoretischen Konzepte kennt, kann die Definitionen und bekannten Aussagen in den Abschnitten 2.1–2.5 nur überfliegen, um die hier verwendete Notation kennenzulernen, Unbekanntes zu identifizieren (dies könnte zum Beispiel die Jensensche Ungleichung oder Fakt 2.2.9 sein) und die Beispiele anzusehen und gut zu verstehen. Die Ungleichung vom arithmetischen und geometrischen Mittel (Prop. 2.3.9) sollte man kennen. Verteilungen, die man unbedingt kennen sollte, sind die (diskreten) uniformen Verteilungen, die Binomialverteilungen, die geometrischen Verteilungen. Die Hoeffding-Ungleichung in Abschnitt 2.6 und die Ungleichungen in Abschnitt 2.7 sind sicher neu – sie sind zentrales Werkzeug und auch Thema der Vorlesung. Anhang A.1 beinhaltet eine Sammlung nützlicher Ungleichungen aus der Analysis und der Kombinatorik, die man ohnehin kennen sollte, ohne dass sie direkt Stoff der Vorlesung sind. Anhang A.2 stellt das Konzept der Summierbarkeit von unendlich vielen Zahlen bereit, das für die Modellierung in Kapitel 3 grundlegend ist.

2.1 Grundbegriffe, Beispiele

Definition 2.1.1

Ein **Wahrscheinlichkeitsraum** (**W-Raum**) ist ein Paar (Ω, p), wobei Ω eine endliche oder abzählbar unendliche Menge und $p\colon \Omega \to [0,1]$ eine Funktion ist, mit[1]

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

Wir schreiben oft p_ω statt $p(\omega)$. Eine solche Funktion $p\colon \Omega \to [0,1]$ heißt auch „**Verteilung**“ oder „**Wahrscheinlichkeitsverteilung**“.

Ein Wahrscheinlichkeitsraum ist eine mathematisch exakte Formulierung für das (informale, intuitive) Konzept eines „*Zufallsexperiments*“: Es wird „zufällig“ ein Element aus Ω ausgewählt; dabei ist die Wahrscheinlichkeit, gerade ω zu erhalten, durch $p(\omega)$ gegeben. Die Elemente ω von Ω heißen *Ergebnisse* (gemeint ist „mögliche Ergebnisse des Zufallsexperiments“) oder *Elementarereignisse*. Man teste diese intuitive

[1] Wenn Ω unendlich ist, ist die Schreibweise $\sum_{\omega\in\Omega} p(\omega)$, die Summation ohne Berücksichtigung einer Reihenfolge ausdrückt, in den Mathematikvorlesungen nicht verwendet worden. Die Definition und einige Kommentare hierzu finden sich in Anhang A.2. Man kann die Schreibweise aber auch einfach benutzen und sich vorstellen, dass durch das Bestehen auf absoluter Konvergenz (unabhängig von irgendwelchen Reihenfolgen) nie Probleme beim Umgang mit solchen Summen auftreten können, und die Standardrechenregeln wie bei endlichen Summen gelten.

Auffassung an den folgenden Beispielen.

Beispiele 2.1.2

(a) Zur Modellierung des Zufallsexperiments, einen fairen Würfel einmal zu werfen, benutzt man den Wahrscheinlichkeitsraum (Ω, p) mit $\Omega = \{1, \dots, 6\}$ und $p(\omega) = \frac{1}{6}$ für jedes $\omega \in \{1, \dots, 6\}$.
Um das Werfen einer fairen Münze zu modellieren, wird man (mit „0“ für „Kopf“ und „1“ für „Zahl“) den W-Raum $\Omega = \{0, 1\}$ und $p(\omega) = \frac{1}{2}$ verwenden. Ist die Münze gefälscht, könnte man z. B. $p(0) = 0.55$ und $p(1) = 0.45$ setzen.

(b) Zur Modellierung des Zufallsexperiments, zwei Würfel zu werfen und die Summe der Augenzahlen als Ergebnis zu nehmen, wird man etwa $\Omega = \{2, \dots, 12\}$ und $p(2) = \frac{1}{36}$, $p(3) = \frac{2}{36}$, $p(4) = \frac{3}{36}$, $\dots$, $p(7) = \frac{6}{36}$, $p(8) = \frac{5}{36}$, $\dots$, $p(12) = \frac{1}{36}$ wählen. Man beachte, dass hier die Wahrscheinlichkeiten unterschiedlich sind.

(c) $U \neq \emptyset$ sei eine endliche Menge. Wir modellieren das Zufallsexperiment, ein Element aus U zu wählen, wobei jedes Element die gleichen Chancen haben soll, wie folgt: $\Omega = U$ und $p_\omega = \frac{1}{|U|}$, für alle $\omega \in \Omega$. Diese Wahrscheinlichkeitsverteilung heißt „*uniforme Verteilung*“ oder „(diskrete) Gleichverteilung“ auf U. Gewöhnlich ist implizit diese Verteilung gemeint, wenn über die Wahrscheinlichkeiten der einzelnen Elemente gar nichts gesagt wird oder wenn die Formulierung „wähle zufällig ein Element aus U“ benutzt wird.[2]

(d) Wir wollen wiederholt mit einem Würfel würfeln und warten, bis die erste „6“ erscheint. Das Ergebnis des Experiments soll die Anzahl der benötigten Würfe sein. Um dies zu modellieren, setzen wir $\Omega = \{1, 2, 3, \dots\}$ und $p_i = (\frac{5}{6})^{i-1} \cdot \frac{1}{6}$ als die Wahrscheinlichkeit, dass beim i-ten Versuch zum ersten Mal eine „6“ gewürfelt wird. (Bei den ersten $i - 1$ Versuchen keine „6“, jeweils mit Wahrscheinlichkeit $1 - \frac{1}{6} = \frac{5}{6}$, bei Versuch Nummer i eine „6“, mit Wahrscheinlichkeit $\frac{1}{6}$.) Man sieht, mit der Summenformel für geometrische Reihen:

$$\sum_{i \geq 1} p_i = \sum_{i \geq 1} \left(\frac{5}{6}\right)^{i-1} \cdot \frac{1}{6} = \frac{1}{6} \cdot \sum_{i \geq 1} \left(\frac{5}{6}\right)^{i-1} = \frac{1}{6} \cdot \frac{1}{1 - \frac{5}{6}} = 1.$$

[2] Man nennt die uniforme Verteilung auch *Laplace-Verteilung*, nach Pierre-Simon Laplace (1749–1827), einem französischen Mathematiker, der postulierte, dass man ohne Information, die ein Elementarereignis vor einem anderen bevorzugt, die uniforme Verteilung annehmen sollte (Indifferenzprinzip, `https://de.wikipedia.org/wiki/Indifferenzprinzip`).

Damit haben wir tatsächlich einen Wahrscheinlichkeitsraum definiert. Die hier definierte Verteilung heißt „*geometrische Verteilung* mit Parameter $p = \frac{1}{6}$“. (In Abschnitt 2.5.1 werden geometrische Verteilungen allgemein diskutiert.)

(e) Es sei $U \neq \emptyset$ eine endliche Menge und $n \geq 1$. Der W-Raum (Ω, p) mit

$$\Omega = U^n = \{(a_1, \ldots, a_n) \mid a_1, \ldots, a_n \in U\}$$

und $p_\omega = \frac{1}{|U|^n}$ für $\omega \in \Omega$, das ist also die uniforme Verteilung auf U^n, entspricht dem Zufallsexperiment, bei dem eine Folge von n Elementen aus U zufällig gewählt wird, bzw. n-mal hintereinander ein Element aus U zufällig gewählt wird.

(f) Es sei $U \neq \emptyset$ eine endliche Menge und $1 \leq n \leq |U|$. Wir wollen das Zufallsexperiment „Wähle eine zufällige n-elementige Teilmenge von U“ modellieren. Dazu wählen wir $\Omega = \{S \subseteq U \mid |S| = n\}$ als Grundmenge mit der Verteilung, die durch $p_S = 1/\binom{|U|}{n}$ für alle $S \in \Omega$ gegeben ist.

(g) Für die Durchschnittsanalyse von Sortierverfahren, die n Schlüssel aus dem angeordneten Universum $(U, <)$ sortieren, ist die folgende Verteilung zentral. Für Sortierverfahren, die auf Schlüsseln nur Vergleiche und keine anderen Operationen durchführen, ist der Ablauf des Verfahrens im Wesentlichen durch den „*Ordnungstyp*“ der Eingabe $(a_1, \ldots, a_n) \in U^n$ bestimmt, das ist die Permutation π von $\{1, \ldots, n\}$ mit $a_{\pi(1)} < \cdots < a_{\pi(n)}$. Diese ist eindeutig bestimmt, wenn $a_1, \ldots, a_n$ verschieden sind. Daher betrachten wir

$$\Omega = \{\pi \mid \pi \text{ Permutation von } \{1, \ldots, n\}\},$$

mit der durch $p(\pi) = 1/|\Omega| = 1/n!$ gegebenen Verteilung. Dieser W-Raum entspricht dem Experiment, für n beliebig vorgegebene Elemente von U die Anordnung rein zufällig zu wählen.

(h) Beim Hashing betrachtet man n Schlüssel $x_1, \ldots, x_n$ und n Funktionswerte $h(x_1)$, $\ldots, h(x_n)$ in $[m] := \{0, 1, \ldots, m-1\}$. Es gibt dabei verschiedene Wahrscheinlichkeitsannahmen, die zu verschiedenen Wahrscheinlichkeitsräumen führen. Wenn man etwa die „Uniformitätsannahme“ für eine Hashfunktion macht, meint man damit, dass der Hashwert eines jeden Schlüssels unabhängig von den anderen jeden Wert in $[m]$ mit derselben Wahrscheinlichkeit annimmt. Der zugehörige Wahrscheinlichkeitsraum ist

$$\Omega = [m]^n = \{(v_1, \ldots, v_n) \mid v_1, \ldots, v_n \in [m]\}$$

mit der durch

$$p((v_1, \dots, v_n)) = \frac{1}{m^n}$$

definierten Verteilung. (Das ist derselbe Wahrscheinlichkeitsraum wie der in (e), wenn man $U = [m]$ setzt.)

Definition 2.1.3

Ein **Ereignis** ist eine Menge $A \subseteq \Omega$.
Die **Wahrscheinlichkeit** von A ist $\mathbf{Pr}(\boldsymbol{A}) := \sum_{\omega \in A} p_\omega$.

Bemerkung 2.1.4

Wenn $p\colon \Omega \to [0,1]$ eine Wahrscheinlichkeitsverteilung ist, dann ist $(p_\omega)_{\omega \in A}$ für jedes Ereignis $A \subseteq \Omega$ summierbar (s. Prop. A.2.3(a) im Anhang), also ist $\sum_{\omega \in A} p_\omega$ wohldefiniert.

Notation: Ist φ eine Eigenschaft oder (synonym) eine Aussage, die für ein Ergebnis $\omega \in \Omega$ gelten oder nicht gelten kann, so ist $A = A_\varphi = \{\omega \in \Omega \mid \varphi(\omega)\}$ ein Ereignis. Oft schreibt man hierfür kurz $\{\varphi\}$. Die Wahrscheinlichkeit $\mathbf{Pr}(A) = \mathbf{Pr}(\{\varphi\})$ wird dann als $\mathbf{Pr}(\varphi)$ abgekürzt.

In den folgenden Beispielen sieht man, dass der Name „Ereignis“ und die abkürzende Schreibweise für durch Aussagen gegebene Ereignisse und ihre Wahrscheinlichkeiten recht gut zur Intuition passt. Man beachte, dass in der Notation der W-Raum und auch der Bezug auf einzelne Ergebnisse unterdrückt wird, wann immer es geht.

Beispiel 2.1.5

(a) In Beispiel 2.1.2(b) ist

$$A = \{\omega \in \Omega \mid \omega \geq 6\} = \{\text{Augensumme} \geq 6\}$$

ein Ereignis, das die Situation modelliert, dass die Summe der Augen mindestens 6 beträgt. Man schreibt $\mathbf{Pr}(\text{Augensumme} \geq 6)$ für $\mathbf{Pr}(A)$. Es gilt

$$\mathbf{Pr}(A) = \mathbf{Pr}(\{6,7,8,9,10,11,12\}) = \frac{5}{36} + \frac{6}{36} + \frac{5}{36} + \frac{4}{36} + \frac{3}{36} + \frac{2}{36} + \frac{1}{36} = \frac{13}{18}.$$

(b) In Beispiel 2.1.2(h) ist

$$A = \{(v_1, \dots, v_n) \mid v_1 = v_2 = v_3\}$$

ein Ereignis, das man auch als $\{h(x_1) = h(x_2) = h(x_3)\}$ schreiben kann. Es gilt

$$\mathbf{Pr}(A) = \mathbf{Pr}(h(x_1) = h(x_2) = h(x_3)) = |A|/m^n = m^{n-2}/m^n = 1/m^2.$$

Beachte *allgemein*: Ist (Ω, p) die ***uniforme Verteilung*** auf Ω, d. h. $p_\omega = 1/|\Omega|$ für alle $\omega \in \Omega$, so ist $\mathbf{Pr}(A) = |A|/|\Omega|$.

Fakt 2.1.6

(a) $\mathbf{Pr}(\emptyset) = 0$ („das unmögliche Ereignis“),
$\mathbf{Pr}(\Omega) = 1$ („das sichere Ereignis“),
$\mathbf{Pr}(\overline{A}) = \mathbf{Pr}(\Omega - A) = 1 - \mathbf{Pr}(A)$ („Komplementärereignis“),
$\mathbf{Pr}(\{\omega\}) = p_\omega$, für $\omega \in \Omega$.

(b) Sind $A_1, \ldots, A_n$ *disjunkte* Ereignisse, so ist

$$\mathbf{Pr}(A_1 \cup \cdots \cup A_n) = \sum_{1 \le i \le n} \mathbf{Pr}(A_i) \qquad (\textbf{\textit{Additivität}}).$$

Sind $A_1, A_2, \ldots$ (abzählbar unendlich viele) *disjunkte* Ereignisse, so ist

$$\mathbf{Pr}\Big(\bigcup_{i \ge 1} A_i\Big) = \sum_{i \ge 1} \mathbf{Pr}(A_i) \qquad (\textbf{\textit{σ-Additivität}}).$$

(c) Ist $A_1 \subseteq A_2$, so ist $\mathbf{Pr}(A_1) \le \mathbf{Pr}(A_2)$ (***Monotonie***).

(d) Sind $A_1, \ldots, A_n$ *beliebige* Ereignisse, so ist

$$\mathbf{Pr}(A_1 \cup \cdots \cup A_n) \le \sum_{1 \le i \le n} \mathbf{Pr}(A_i).$$

Sind $A_1, A_2, \ldots$ (abzählbar unendlich viele) Ereignisse, so ist

$$\mathbf{Pr}\Big(\bigcup_{i \ge 1} A_i\Big) \le \sum_{i \ge 1} \mathbf{Pr}(A_i)$$

(***Vereinigungsschranke*** oder englisch ***Union Bound***).

Die Gültigkeit der Aussagen in Fakt 2.1.6 kann man mittels Def. 2.1.3 nachkontrollieren. Man benutzt, dass Assoziativität, Distributivität und Monotonie auch bei

unendlichen Summen gelten, s. Prop. A.2.3 im Anhang.

Formel 2.1.6(d) wird oft folgendermaßen benutzt: Wenn für jedes $\omega \in \Omega$ aus der Aussage $\varphi(\omega)$ die Aussage $\psi(\omega)$ *folgt*, dann gilt $\{\varphi\} \subseteq \{\psi\}$ und damit $\mathbf{Pr}(\varphi) \leq \mathbf{Pr}(\psi)$.

Beispiel 2.1.7
In Beispiel 2.1.2(h) gilt für jedes $v \in [m] = \{0, 1, \ldots, m-1\}$:

$$\mathbf{Pr}(\exists i \in \{1, \ldots, n\} \colon h(x_i) = v) \leq \sum_{1 \leq i \leq n} \mathbf{Pr}(h(x_i) = v) = n \cdot \frac{1}{m}.$$

(*Übung*: Man mache die hier benutzten Ereignisse explizit und benenne die Regeln, die in den einzelnen Rechenschritten angewendet werden.)

2.2 Zufallsvariablen und Erwartungswerte

Definition 2.2.1
Sei (Ω, p) ein W-Raum und R eine beliebige Menge. Eine Funktion $X \colon \Omega \to R$ heißt eine **Zufallsfunktion**. Ist R numerisch (also $R \subseteq \mathbb{R}$), so heißt ein solches X eine **Zufallsvariable** (**ZV**), im Fall $R \subseteq \mathbb{R}^k$ für ein $k \geq 1$ auch ein **Zufallsvektor**.

Die Idee dabei ist natürlich, dass man ein $\omega \in \Omega$ zufällig wählt (gesteuert von der Verteilung $p \colon \Omega \to [0, 1]$), und dass dadurch auch ein zufälliger Wert $X(\omega)$ festgelegt wird.

Zur Schreibweise: Soweit möglich schreibt man X statt $X(\omega)$. Ist beispielsweise $R' \subseteq R$, betrachtet man das Ereignis $\{X \in R'\} = X^{-1}(R') = \{\omega \mid X(\omega) \in R'\}$, und die Wahrscheinlichkeit $\mathbf{Pr}(X \in R')$, usw. Achtung: „X“ sieht aus wie ein Wert, ist aber variabel (mit ω).

Bemerkung 2.2.2
Eine ZV X mit Wertebereich $\{0, 1\}$ bezeichnet man als Indikator(-zufallsvariable). Solche Zufallsvariablen werden mit Hilfe einer Aussage φ wie folgt konstruiert:

$$X(\omega) := \begin{cases} 1, & \text{falls } \varphi(\omega) \text{ wahr ist,} \\ 0, & \text{sonst.} \end{cases}$$

Um Indikatorzufallsvariablen kompakt zu notieren (und nicht jedes Mal die Fallunterscheidung hinschreiben zu müssen) hat sich die **Iverson-Notation** bewährt: Für das X wie oben schreibt man $[\varphi]$. Für die Aussage „Augensumme ≥ 6“ (Beispiel 2.1.5 (a)) könnte man also einen entsprechenden Indikator mit „[Augensumme ≥ 6]“ angeben.

Beispiel 2.2.3
Betrachte Beispiel 2.1.2(h), also $\Omega = [m]^n = \{\omega = (v_1, \ldots, v_n) \mid v_1, \ldots, v_n \in [m]\}$.

(a) Für $1 \leq i \leq n$ ist die Funktion $\omega \mapsto v_i = h(x_i)$ eine Zufallsvariable.

(b) Für $v \in [m]$ ist die Funktion $\omega \mapsto B_v = \{i \mid v_i = v\} = \{i \mid h(x_i) = v\}$ eine Zufallsfunktion (der Wert ist eine „zufällige Menge“ oder „Zufallsmenge“, die den Schlüsseln x_i entspricht, die von h auf den Wert v abgebildet werden); die Funktion $b_v \colon \omega \mapsto |B_v|$ der Anzahl dieser Schlüssel ist eine ZV.

Jede Zufallsfunktion $X \colon \Omega \to R$ induziert einen neuen Wahrscheinlichkeitsraum, wie folgt:

$$\Omega' := X[\Omega] = \{X(\omega) \mid \omega \in \Omega\} \subseteq R;\ p'(\alpha) := \mathbf{Pr}(X = \alpha) \text{ für } \alpha \in \Omega'. \tag{2.2.1}$$

Die Verteilung p' heißt die ***Verteilung von X***. Wenn es bequem ist, kann man auch eine (endliche oder abzählbare) Menge R' mit $X[\Omega] \subseteq R' \subseteq R$ als Grundmenge Ω' benutzen.

Bemerkung 2.2.4
Für jeden Wahrscheinlichkeitsraum (Ω, p) ist die Funktion $p \colon \Omega \to [0, 1]$ Verteilung einer Zufallsvariablen X. Man wählt einfach $X = \mathrm{id}_\Omega$, die Identität, die ω auf ω abbildet, und erhält $\Omega' = \Omega$ und $p' = p$.

Beispiel 2.2.5
(a) Beim Werfen von zwei fairen Würfeln ist folgender Wahrscheinlichkeitsraum mit einer uniformen Verteilung natürlich:

$$\Omega = \{1, \ldots, 6\} \times \{1, \ldots, 6\};\ \ p((i, j)) = \frac{1}{36} \text{ für } (i, j) \in \Omega.$$

Die durch $X((i, j)) := i + j$ definierte Abbildung $X \colon \Omega \to \{2, \ldots, 12\}$ ist eine Zufallsvariable. Die Verteilung von X ist gerade die Verteilung des in Beispiel 2.1.2(b) beschriebenen Wahrscheinlichkeitsraums.

(b) Beim Spiel „Würfeln, bis eine 6 erscheint“ ist folgender Wahrscheinlichkeitsraum natürlich:

$$\Omega = \{(a_1, \ldots, a_i) \mid i \geq 1, a_1 \ldots, a_{i-1} \in \{1, \ldots, 5\}, a_i = 6\};$$
$$p((a_1, \ldots, a_i)) = \left(\frac{1}{6}\right)^i, \text{ für } (a_1 \ldots, a_i) \in \Omega.$$

Ein Elementarereignis ist hier eine Folge von Ergebnissen einzelner Würfe, die abbricht, sobald die erste 6 erschienen ist. Jede solche Folge hat, intuitiv gesehen, die Wahrscheinlichkeit $(1/6)^i$. Die durch $X((a_1, \ldots, a_i)) = i$ gegebene Zufallsvariable zählt die Anzahl dieser Versuche. Ihre Verteilung liefert den Wahrscheinlichkeitsraum aus Beispiel 2.1.2(d).

Beispiel 2.2.6
Wir führen Beispiel 2.2.3 noch etwas weiter. Die Zufallsvariable $b_0 = |B_0|$ induziert eine Verteilung auf $b_0[\Omega] = \{0, 1, \ldots, n\}$. Dabei ist

$$\begin{aligned} p'(i) &= \frac{|\{(v_1, \ldots, v_n) \in \Omega \mid (v_1, \ldots, v_n) \text{ enthält genau } i \text{ Nullen}\}|}{m^n}. \\ &= \binom{n}{i} \cdot \frac{(m-1)^{n-i}}{m^n} = \binom{n}{i} \cdot \left(\frac{1}{m}\right)^i \cdot \left(1 - \frac{1}{m}\right)^{n-i}. \end{aligned}$$

(Dies ist eine *Binomialverteilung.*) Natürlich ergibt sich für jedes $v \in [m]$ anstelle von 0 dieselbe Verteilung.

Der *Erwartungswert* einer Zufallsvariablen ist ein sehr grundlegendes Konzept. Es modelliert den „durchschnittlichen Wert“ der Zufallsvariablen, indem er die Funktionswerte zu Elementarereignissen, mit der entsprechenden Wahrscheinlichkeit gewichtet, aufsummiert. Bei Bedarf schaue man sich das Konzept einer summierbaren Familie von Zahlen (s. Abschnitt A.2) an.

Definition 2.2.7
Für eine Zufallsvariable X, für die $(X(\omega) \cdot p_\omega)_{\omega \in \Omega}$ summierbar ist, definieren wir den **Erwartungswert** von X durch:

$$\mathbf{E}(\boldsymbol{X}) := \sum_{\omega \in \Omega} X(\omega) \cdot p_\omega = \sum_{\alpha \in X[\Omega]} \alpha \cdot \mathbf{Pr}(X = \alpha).$$

Die erste Summe betrachtet die Situation eher vom W-Raum (Ω, p) aus, die zweite eher vom Wertebereich und der Verteilung auf $X[\Omega]$ aus. Für nicht-negative Zufallsvariablen X lässt man mitunter auch den Fall $\mathbf{E}(X) = \infty$ zu (wenn Ω unendlich ist und die Menge $\{\sum_{\omega\in A} X(\omega)\cdot p_\omega \mid A \subseteq \Omega \text{ endlich}\}$ unbeschränkt ist).

Beispiele 2.2.8
(a) Es sei $\Omega = \mathbb{N}^+ = \{1, 2, 3, \dots\}$ und $p_i = 2^{-i}$ für $i \in \mathbb{N}^+$. Dann hat die Zufallsvariable X mit $X(i) = (-1)^i \cdot i^2$ für $i \in \mathbb{N}^+$ einen Erwartungswert (weil $\sum_{i=1}^\infty i^2/2^i$ konvergent ist), die Zufallsvariable Y mit $Y(i) = (-2)^i$, $i \in \mathbb{N}^+$, dagegen nicht (weil $\sum_{i=1}^\infty 2^i/2^i = \sum_{i\geq 1} 1$ divergent ist, also beliebig große endliche Teilsummen besitzt).
(b) Man erinnere sich an die bekannte Formel $\sum_{i=1}^\infty \frac{1}{i(i+1)} = 1$. (Das liegt daran, dass $\frac{1}{i(i+1)} = \frac{1}{i} - \frac{1}{i+1}$ gilt.) Mit $\Omega = \mathbb{N}^+$ und $p_i = \frac{1}{i(i+1)}$ für $i \in \mathbb{N}^+$ erhalten wir also einen W-Raum. In diesem W-Raum hat die Zufallsvariable X mit $X(i) = (-1)^{i-1}$ für $i \in \mathbb{N}^+$ einen Erwartungswert, nämlich $\mathbf{E}(X) = \sum_{i\geq 1} \frac{(-1)^{i-1}}{i(i+1)}$ $(= 2\ln 2 - 1)$. Dagegen hat die Zufallsvariable Y mit $Y(i) = (-1)^{i-1}\cdot i$ keinen Erwartungswert, da die Reihe $\sum_{i=1}^\infty \frac{(-1)^{i-1}\cdot i}{i(i+1)} = \sum_{i=1}^\infty \frac{(-1)^{i-1}}{i+1} = \frac{1}{2} - \frac{1}{3} + \frac{1}{4} - \frac{1}{5} \pm \dots$, zwar bedingt konvergent ist (mit Grenzwert $1 - \ln 2$), aber nicht summierbar ist.

Die zweite Darstellung des Erwartungswertes in Definition 2.2.7 lässt sich durch Umstellen und Gruppieren von Summen beweisen. Ein solches Umstellen und Umklammern ist hier kein Problem, weil alle auftretenden Reihen summierbar sind, siehe Anhang A.2. Man kann diese zweite Darstellung auch so auffassen: Man betrachtet die Verteilung von X, die jeder Zahl α im Wertebereich $X[\Omega]$ die Wahrscheinlichkeit $p'(\alpha) = \mathbf{Pr}(X = \alpha)$ zuordnet, und bildet den Mittelwert dieser Zahlen, gewichtet mit diesen Wahrscheinlichkeiten.

Die folgende Formel für nicht-negative, ganzzahlige Zufallsvariablen X ist äußerst nützlich, wenn sich $\mathbf{Pr}(X \geq i)$ bequem berechnen oder abschätzen lässt. Wir werden sie häufig benutzen.

Fakt 2.2.9
Sei $X\colon \Omega \to \mathbb{N}$ eine Zufallsvariable, deren Erwartungswert existiert. Dann gilt:

$$\mathbf{E}(X) = \sum_{i\geq 1} \mathbf{Pr}(X \geq i) \qquad \big(= \sum_{i\geq 0} \mathbf{Pr}(X \geq i+1)\big).$$

Beweis. Setze $p_j = \mathbf{Pr}(X = j)$, $q_i = \mathbf{Pr}(X \geq i)$. Dann gilt: $q_i = \sum_{j \geq i} p_j$, also

$$\mathbf{E}(X) = \sum_{j \geq 0} j \cdot p_j = \sum_{j \geq 1} j \cdot p_j = \sum_{j \geq 1} \sum_{1 \leq i \leq j} p_j = \sum_{i \geq 1} \sum_{j \geq i} p_j = \sum_{i \geq 1} q_i.$$

□

Beispiel: In Beispiel 2.1.2(d) (Würfeln, bis die erste „6“ erscheint) definieren wir die Zufallsvariable $X :=$ Anzahl der Würfe bis zur ersten 6 (einschließlich). Technisch ist das in diesem Wahrscheinlichkeitsraum einfach $X(i) := i$, für $i \geq 1$. Intuitiv sieht man, dass $\mathbf{Pr}(X \geq i) = (\frac{5}{6})^{i-1}$ sein muss (Misserfolg in den ersten $i-1$ Würfen). Zur Sicherheit rechnen wir das nach:

$$\begin{aligned}\mathbf{Pr}(X \geq i) &= \sum_{j \geq i} \mathbf{Pr}(X = j) = \sum_{j \geq i} \left(\frac{5}{6}\right)^{j-1} \cdot \frac{1}{6} = \frac{1}{6} \cdot \left(\frac{5}{6}\right)^{i-1} \cdot \sum_{j \geq i} \left(\frac{5}{6}\right)^{j-i} \\ &= \frac{1}{6} \cdot \left(\frac{5}{6}\right)^{i-1} \cdot \frac{1}{1 - \frac{5}{6}} = \left(\frac{5}{6}\right)^{i-1}.\end{aligned}$$

Mit Fakt 2.2.9 ergibt sich:

$$\mathbf{E}(X) = \sum_{i \geq 1} \mathbf{Pr}(X \geq i) = \sum_{i \geq 1} \left(\frac{5}{6}\right)^{i-1} = \frac{1}{1 - \frac{5}{6}} = 6.$$

Die Verallgemeinerung dieser Situation auf beliebige Erfolgswahrscheinlichkeiten p anstelle von $\frac{1}{6}$ führt zu der geometrischen Verteilung zu Parameter p, siehe Abschnitt 2.5.1.

Fakt 2.2.10

Für beliebige Zufallsvariablen $X, Y, X_1, \ldots, X_n$ gilt (unter der Voraussetzung, dass alle Erwartungswerte definiert sind):

(a) $X \leq Y$ (d. h. $\forall \omega \in \Omega$: $X(\omega) \leq Y(\omega)$) $\Rightarrow \mathbf{E}(X) \leq \mathbf{E}(Y)$ (***Monotonie***).

(b) $\mathbf{E}(\alpha X + \beta Y) = \alpha \mathbf{E}(X) + \beta \mathbf{E}(Y)$, für beliebige $\alpha, \beta \in \mathbb{R}$ (***Linearität des Erwartungswertes I***).

(c) $\mathbf{E}(X_1 + \cdots + X_n) = \mathbf{E}(X_1) + \cdots + \mathbf{E}(X_n)$ (***Linearität des Erwartungswertes II***).

(d) $X \in \{0, 1\}$ (d. h. $\forall \omega \in \Omega$: $X(\omega) \in \{0, 1\}$) $\Rightarrow \mathbf{E}(X) = \mathbf{Pr}(X = 1)$.

Die *Beweise* von (a), (b), (c) sind (einfache) Anwendungen der Rechenregeln für Summen. Für (d) beobachtet man $\mathbf{E}(X) = \mathbf{Pr}(X = 0) \cdot 0 + \mathbf{Pr}(X = 1) \cdot 1$.

Bemerkung 2.2.11
Für eine Indikatorvariable $[\varphi]$ gilt nach Fakt 2.2.10 $\mathbf{E}([\varphi]) = \mathbf{Pr}(\varphi)$.

Beispiel 2.2.12
Betrachte Bsp. 2.2.3(b). Wir berechnen $\mathbf{E}(|B_v|)$ mit Hilfe der Indikatorvariablen $[h(x_i) = v]$, für $i \in \{1, 2, \ldots, n\}$. Klar: $|B_v| = [h(x_1) = v] + \cdots + [h(x_n) = v]$. Also gilt

$$\mathbf{E}(|B_v|) = \sum_{1 \le i \le n} \mathbf{E}([h(x_i) = v]) = \sum_{1 \le i \le n} \mathbf{Pr}(h(x_i) = v) = \sum_{1 \le i \le n} \frac{1}{m} = \frac{n}{m}.$$

Ein anderes, sehr typisches Beispiel für die Anwendung der Technik der Zerlegung einer Zufallsvariablen in Summanden, die Indikatorvariablen sind, ist die Analyse von Quicksort in Abschnitt 1.4. (Hier ist der richtige Zeitpunkt, diese Analyse nochmals mit vollem Verständnis durchzusehen.)

2.3 Varianz und Ungleichungen von Markov, Chebychev und Jensen

Der Zweck der folgenden fundamentalen Ungleichung ist, bei gegebenem Erwartungswert einer nichtnegativen Zufallsvariablen Z die Wahrscheinlichkeit dafür zu begrenzen, dass Z sehr große Werte hat.

Fakt 2.3.1 (*Markoff/Markov-Ungleichung*)
Es sei $Z \ge 0$ eine beliebige Zufallsvariable, und $t > 0$ sei beliebig. Dann gilt:

$$\mathbf{Pr}(Z \ge t) \le \frac{\mathbf{E}(Z)}{t}.$$

Beweis. Offenbar gilt $Z \ge t \cdot [Z \ge t]$ (eine Ungleichung für jedes $\omega \in \Omega$), also auch $\mathbf{E}(Z) \ge t \cdot \mathbf{E}([Z \ge t]) = t \cdot \mathbf{Pr}(Z \ge t)$ (mit Monotonie und Linearität des Erwartungswertes und Bem. 2.2.11). Dividieren durch t liefert die Behauptung. □

Definition 2.3.2
Für eine beliebige Zufallsvariable X, für die $\mathbf{E}(X^2)$ existiert, definieren wir[3] die **Varianz** von X als

$$\mathbf{Var}(X) := \mathbf{E}((X - \mathbf{E}(X))^2).$$

Bemerkung 2.3.3
Für jedes $a \in \mathbb{R}$ gilt $\mathbf{Var}(X - a) = \mathbf{Var}(X)$. Insbesondere haben wir für $X' := X - \mathbf{E}(X)$ die Beziehungen $\mathbf{E}(X') = 0$ und $\mathbf{Var}(X') = \mathbf{Var}(X)$. Weiter gilt $\mathbf{Var}(a \cdot X) = a^2 \cdot \mathbf{Var}(X)$, für jedes $a \in \mathbb{R}$.

Man sieht sofort, dass gilt:

$$\mathbf{Var}(X) = \mathbf{E}(X^2 - 2X\mathbf{E}(X) + \mathbf{E}(X)^2) = \mathbf{E}(X^2) - 2\mathbf{E}(X)^2 + \mathbf{E}(X)^2 = \mathbf{E}(X^2) - \mathbf{E}(X)^2. \tag{2.3.2}$$

Als Erwartungswert von $(X - \mathbf{E}(X))^2 \geq 0$ ist $\mathbf{Var}(X) \geq 0$. Mit (2.3.2) folgt

$$\mathbf{E}(X)^2 \leq \mathbf{E}(X^2) \tag{2.3.3}$$

für jede Zufallsvariable X, deren Varianz existiert.

Die Varianz misst in einem gewissen Sinn die Tendenz einer Zufallsvariablen, Werte anzunehmen, die weit von ihrem Erwartungswert entfernt sind. (Schön ist immer, wenn sich die Werte einer Zufallsvariablen ganz in der Nähe ihres Erwartungswertes befinden. Dieser Situation entspricht eine kleine Varianz.) Durch das Quadrieren in der Definition der Varianz werden große Entfernungen stark betont. Eine der zentralen Anwendungen der Varianz ist die folgende Ungleichung, die bei gegebener Varianz $\mathbf{Var}(X)$ die Wahrscheinlichkeit dafür begrenzt, dass X weit von seinem Erwartungswert entfernt liegt. Für den Beweis wendet man einfach auf die Zufallsvariable $Z = (X - \mathbf{E}(X))^2 \geq 0$ die Markov-Ungleichung an.

Fakt 2.3.4 (*Chebychev/Tschebyscheff-Ungleichung*)
Es sei X eine Zufallsvariable, deren Varianz existiert. Dann gilt für jedes $t > 0$:

$$\mathbf{Pr}(|X - \mathbf{E}(X)| \geq t) \leq \frac{\mathbf{Var}(X)}{t^2}.$$

[3] Wenn $\mathbf{E}(X^2) = \sum_{\omega \in \Omega} p_\omega X(\omega)^2$ definiert ist, dann ist auch $\mathbf{E}(X)$ definiert. Wieso?

Beweis. Setze $Z := (X - \mathbf{E}(X))^2$. Dann gilt nach der Markov-Ungleichung:

$$\mathbf{Pr}(|X - \mathbf{E}(X)| \geq t) = \mathbf{Pr}(Z \geq t^2) \leq \frac{\mathbf{E}(Z)}{t^2} = \frac{\mathbf{Var}(X)}{t^2}.$$

□

Wir können die Markov-Ungleichung verallgemeinern:

Proposition 2.3.5

X sei eine beliebige Zufallsvariable, $D \subseteq \mathbb{R}$, $f\colon D \to \mathbb{R}^+$ sei monoton wachsend mit $D = \mathrm{Def}(f) \supseteq X[\Omega]$, so dass $\mathbf{E}(f(X))$ existiert. Dann gilt für jedes $t \in D$:

$$\mathbf{Pr}(X \geq t) \leq \frac{\mathbf{E}(f(X))}{f(t)}.$$

Beweis: Man wendet die Markov-Ungleichung auf die Zufallsvariable $f(X)$ an, und verwendet, dass wegen der Monotonie von f die Aussagen $X \geq t$ und $f(X) \geq f(t)$ äquivalent sind. □

Beispiele 2.3.6
Sei X eine Zufallsvariable.

(a) Sei $\alpha > 0$ beliebig, so dass $\mathbf{E}(|X|^\alpha)$ existiert. Dann gilt für $t > 0$:

$$\mathbf{Pr}(X \geq t) \leq \frac{\mathbf{E}(|X|^\alpha)}{t^\alpha}.$$

(b) Sei $k \geq 2$ eine gerade natürliche Zahl, so dass $\mathbf{E}(X^k)$ existiert. Dann gilt für $t > 0$:

$$\mathbf{Pr}(|X - \mathbf{E}(X)| \geq t) \leq \frac{\mathbf{E}((X - \mathbf{E}(X))^k)}{t^k}.$$

(Hier wird Prop. 2.3.5 auf die Zufallsvariable $Z = |X - \mathbf{E}(X)|$ und $f(x) = x^k$ angewendet.)

(c) Seien $t, c > 0$ beliebig, und sei X eine Zufallsvariable, deren Varianz existiert. Dann gilt

$$\mathbf{Pr}((X + c)^2 \geq (t + c)^2) \leq \frac{\mathbf{E}((X + c)^2)}{(t + c)^2}.$$

Diese Ungleichung kann man für den Beweis der Chebychev-Cantelli-Ungleichung (Prop. 2.7.2) benutzen (siehe Übung).

(d) Sei X reellwertig, sei $a > 0$, und sei $\mathbf{E}(e^{aX}) < \infty$. Dann gilt für jedes $t \in \mathbb{R}$:

$$\mathbf{Pr}(X \geq t) \leq \frac{\mathbf{E}(e^{aX})}{e^{at}}.$$

(Dies ist die ursprüngliche **„Chernoff-Schranke“** von 1952. Wir werden sie weiter unten benutzen, um eine spezialisierte Folgerung, die *Hoeffding-Schranke*, zu beweisen.)

Ungleichung (2.3.3) besagt, dass stets $\mathbf{E}(X)^2 \leq \mathbf{E}(X^2)$ gilt. Diese Ungleichung wollen wir verallgemeinern, indem wir anstelle der Funktion $x \mapsto x^2$ irgendeine *konvexe* Funktion benutzen. Wir erinnern an die Definition von konvexen Funktionen.

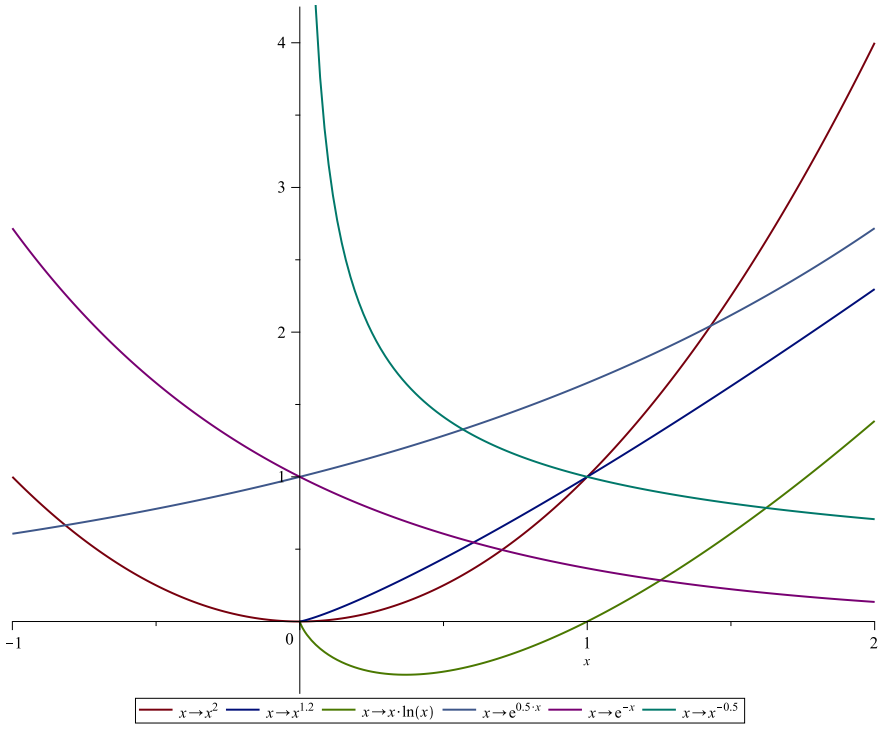

Abbildung 2.3.1: Einige konvexe Funktionen: $\mathbb{R} \ni x \mapsto x^2$, $[0,\infty) \ni x \mapsto x^{1.2}$, $[0,\infty) \ni x \mapsto x \ln x$, $\mathbb{R} \ni x \mapsto e^{x/2}$, $\mathbb{R} \ni x \mapsto e^{-x}$, $(0,\infty) \ni x \mapsto x^{-1/2}$.

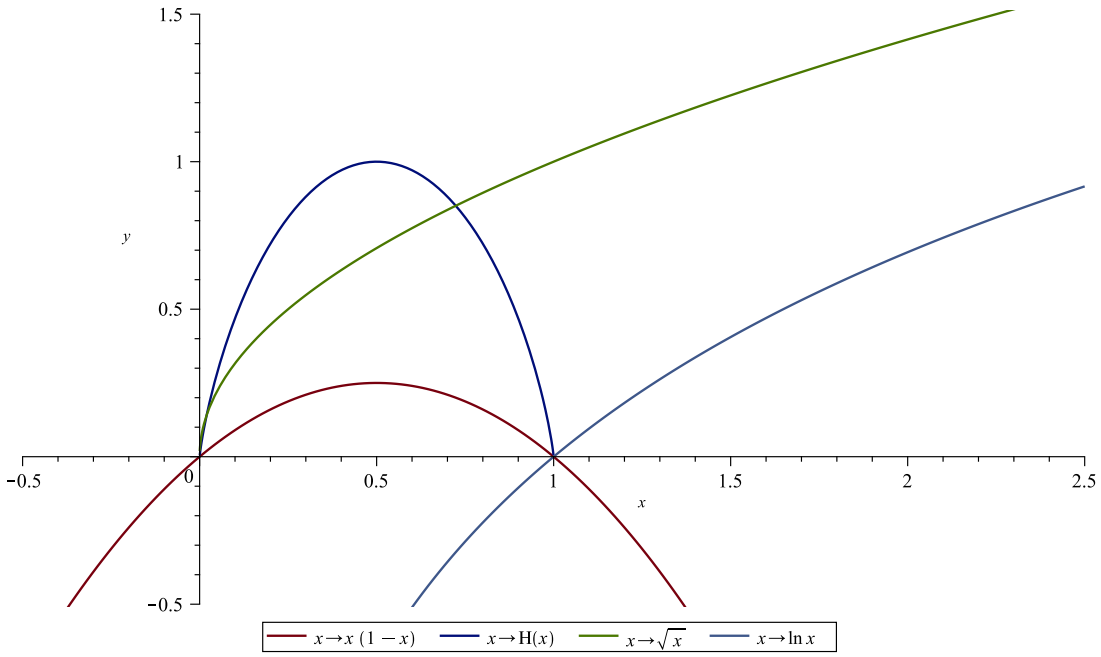

Abbildung 2.3.2: Einige konkave Funktionen: $\mathbb{R} \ni x \mapsto x(1-x)$, $[0,1] \ni x \mapsto \mathrm{H}(x) = -x\log_2 x - (1-x)\log_2(1-x)$ (binäre Entropie), $[0,\infty) \ni x \mapsto \sqrt{x}$, $(0,\infty) \ni x \mapsto \ln x$.

Definition 2.3.7

$D \subseteq \mathbb{R}$ sei ein Intervall. Eine Funktion $f\colon D \to \mathbb{R}$ heißt **konvex**, wenn gilt:

$$f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y), \text{ für alle } x, y \in D \text{ und } \lambda \in [0,1].$$

Eine Funktion f heißt **konkav**, wenn $-f$ konvex ist.

Anschaulich, geometrisch gesehen ist eine Funktion konvex, wenn in jedem Teilintervall $[x,y]$ des Definitionsbereichs der Graph der Funktion unter der Verbindungsstrecke der Punkte $(x,f(x))$ und $(y,f(y))$ verläuft. – Aus der Schule oder aus der Analysis weiß man, dass für die Konvexität hinreichend ist, dass $f''(x)$ in D (bzw. im Inneren von D) existiert und nicht negativ ist. (Aber Achtung: Die Existenz der zweiten Ableitung ist nicht notwendig. Zum Beispiel ist die Funktion $x \mapsto |x|$ auf $\mathbb{R}$ konvex, aber in $x = 0$ hat sie keine Ableitung.)

Beispiele:

(i) Die Funktion $f\colon x \mapsto x^2$ ist konvex in $\mathbb{R}$. Allgemeiner gilt dies für $x \mapsto x^{2d}$, für jede natürliche Zahl $d > 0$.

(ii) Wenn $\alpha \in \mathbb{R}$, $\alpha \geq 1$, dann ist die Funktion $f_\alpha\colon x \mapsto x^\alpha$ konvex in $[0, \infty)$.

(iii) Wenn $\alpha \in \mathbb{R}$, $0 < \alpha \leq 1$, dann ist die Funktion $f_\alpha\colon x \mapsto x^\alpha$ konkav in $[0, \infty)$.

(iv) Wenn $\alpha \in \mathbb{R}$, $\alpha > 0$, dann ist die Funktion $g_\alpha\colon x \mapsto x^{-\alpha}$ konvex in $(0, \infty)$. (Differenziere zweimal: $g'_\alpha(x) = -\alpha/x^{\alpha+1}$, und dann: $g''_\alpha(x) = \alpha(\alpha+1)/x^{\alpha+2}$. Dies ist immer positiv.)

(v) Die Funktion $h\colon x \mapsto x \ln x$ ist konvex in $[0, \infty)$. (Differenziere zweimal: $h'(x) = \ln x + 1$, und $h''(x) = x^{-1} > 0$, für $x > 0$.)

(vi) Für $t \in \mathbb{R}$ ist die Funktion $k\colon x \mapsto e^{tx}$ konvex in $\mathbb{R}$.

(vii) Die Funktion $\mathrm{H}\colon x \mapsto -x \log_2 x - (1-x)\log_2(1-x)$ (*binäre Entropie* von x) ist konkav in $[0, 1]$.

Siehe Abb. 2.3.1 und Abb. 2.3.2 für Beispiele konvexer und konkaver Funktionen.

Proposition 2.3.8 (*Jensensche Ungleichung, allgemeine Form*)

Es sei X eine reellwertige Zufallsvariable und f eine Funktion mit $X[\Omega] \subseteq D := \mathrm{Def}(f)$. Wenn $\mathbf{E}(X)$ und $\mathbf{E}(f(X))$ definiert sind, dann gilt:

(a) Wenn f konvex ist: $f(\mathbf{E}(X)) \leq \mathbf{E}(f(X))$.

(b) Wenn f konkav ist: $f(\mathbf{E}(X)) \geq \mathbf{E}(f(X))$.

Beispiele: Unter der Voraussetzung, dass jeweils die Erwartungswerte definiert sind, gilt:

(i) $\mathbf{E}(X)^{2d} \leq \mathbf{E}(X^{2d})$.

(ii) Für $\alpha \geq 1$ und $X \geq 0$ gilt: $\mathbf{E}(X)^\alpha \leq \mathbf{E}(X^\alpha)$.

(iii) Für $0 < \alpha \leq 1$ und $X \geq 0$ gilt: $\mathbf{E}(X)^\alpha \geq \mathbf{E}(X^\alpha)$.

(iv) Für $\alpha > 0$ und $X > 0$ gilt $\mathbf{E}(X)^{-\alpha} \leq \mathbf{E}(X^{-\alpha})$.

(v) Für $X \geq 0$ gilt $\mathbf{E}(X)\ln(\mathbf{E}(X)) \leq \mathbf{E}(X \ln X)$.

(vi) Für $t \in \mathbb{R}$ gilt $e^{t\mathbf{E}(X)} \leq \mathbf{E}(e^{tX})$.

Beweis. (Jensensche Ungleichung.) Wir beweisen nur (a). ((b) folgt durch Multiplikation der Ungleichung mit -1.) Setze $x_0 := \mathbf{E}(X)$. Dann ist $x_0 \in D$, denn D ist ein Intervall. Nach einer Grundeigenschaft von konvexen Funktionen, die man in der Analysis beweist, hat der Graph von f im Punkt $(x_0, f(x_0))$ eine „untere Stützgerade“, das ist eine Gerade, die durch den Punkt verläuft und stets unterhalb des Funktionsgraphen bleibt. Das heißt: Es gibt ein $\alpha \in \mathbb{R}$ (die Steigung der Stützgeraden) derart dass

$$f(x_0) + \alpha(x - x_0) \leq f(x) \text{ , für alle } x \in \mathrm{Def}(f) \text{ .}$$

(Wenn f differenzierbar ist, wählt man $\alpha = f'(x_0)$.) Daraus folgt, mit der Linearität und der Monotonie des Erwartungswertes:

$$f(x_0) + \alpha(\mathbf{E}(X) - x_0) = \mathbf{E}(f(x_0) + \alpha(X - x_0)) \leq \mathbf{E}(f(X)).$$

Da $x_0 = \mathbf{E}(X)$ gewählt wurde, folgt die behauptete Ungleichung. □

Die Jensensche Ungleichung ist eine recht allgemeine Konvexitätsaussage. Um ihre Kraft zu demonstrieren, beweisen wir mit ihrer Hilfe die bekannte Ungleichung zwischen dem arithmetischen und dem geometrischen Mittel:

Proposition 2.3.9 (*Arithmetisches versus geometrisches Mittel*)
Für $a_1, \ldots, a_n \geq 0$ gilt:

$$\frac{a_1 + \cdots + a_n}{n} \geq (a_1 \cdots a_n)^{1/n}.$$

Allgemeiner: Wenn zudem $p_1, \ldots, p_n \geq 0$ sind mit $p_1 + \cdots + p_n = 1$, dann gilt:

$$p_1 a_1 + \cdots + p_n a_n \geq a_1^{p_1} \cdots a_n^{p_n}.$$

Beweis. Wir können o. B. d. A. annehmen, dass alle a_i strikt positiv sind. Dann betrachten wir eine Zufallsvariable X, die die Werte $a_1, \ldots, a_n$ mit Wahrscheinlichkeiten

$p_1, \ldots, p_n$ annimmt, sowie die konkave Funktion $f(t) = \ln t$ (mit $\mathrm{Def}(f) = (0, \infty)$). Nach Prop. 2.3.8(b) gilt $f(\mathbf{E}(X)) \geq \mathbf{E}(f(X))$. Wenn man dies ausschreibt und die Logarithmus-Rechenregeln anwendet, ergibt sich

$$\ln(p_1 a_1 + \cdots + p_n a_n) \geq p_1 \ln(a_1) + \cdots + p_n \ln(a_n) = \ln(a_1^{p_1} \cdots a_n^{p_n}).$$

Die Monotonie der Logarithmusfunktion liefert die Behauptung. □

2.4 Bedingte Wahrscheinlichkeiten und bedingte Erwartungswerte

Definition 2.4.1
Sei $B \subseteq \Omega$ ein Ereignis mit $\mathbf{Pr}(B) > 0$. Für beliebige Ereignisse A setzen wir

$$\mathbf{Pr}(A \mid B) := \frac{\mathbf{Pr}(A \cap B)}{\mathbf{Pr}(B)},$$

und nennen dies die **bedingte Wahrscheinlichkeit von** A (unter der Bedingung B), für beliebige Ereignisse $A \subseteq \Omega$.

Eine Routinerechnung zeigt, dass Ω mit der durch $\mathbf{Pr}(\cdot \mid B)$ definierten Verteilung ebenfalls ein Wahrscheinlichkeitsraum ist. (Elementarwahrscheinlichkeiten: $p_\omega^B = p_\omega / \mathbf{Pr}(B)$ für $\omega \in B$ und $p_\omega^B = 0$ für $\omega \notin B$.) Auch in diesem Wahrscheinlichkeitsraum lassen sich Erwartungswerte von Zufallsvariablen X bilden (geschrieben $E(X \mid B)$). Man sieht leicht:

$$\mathbf{Pr}(B \mid B) = \mathbf{Pr}(\Omega \mid B) = 1; \quad \mathbf{E}(X \mid B) = \frac{1}{\mathbf{Pr}(B)} \cdot \sum_{\omega \in B} p_\omega X(\omega).$$

Fakt 2.4.2 (*Basisformel für bedingte Wahrscheinlichkeiten*)

$$\mathbf{Pr}(A \cap B) = \mathbf{Pr}(A \mid B) \cdot \mathbf{Pr}(B), \text{ für } A, B \subseteq \Omega \text{ mit } \mathbf{Pr}(B) > 0.$$

Im Fall $\mathbf{Pr}(B) = 0$ ist $\mathbf{Pr}(A \mid B)$ nicht definiert. Solange man bedingte Wahrscheinlichkeiten nur über die Basisformel aus Fakt 2.4.2 benutzt, kann man aber bei $\mathbf{Pr}(B) = 0$ so tun, als ob $\mathbf{Pr}(A \mid B)$ einen beliebigen Wert hätte. Die Formel kann man auf den Durchschnitt mehrerer Ereignisse verallgemeinern:

$$\mathbf{Pr}(A_1 \cap \cdots \cap A_n) = \mathbf{Pr}(A_1)\mathbf{Pr}(A_2 \mid A_1)\mathbf{Pr}(A_3 \mid A_1 \cap A_2) \cdots \mathbf{Pr}(A_n \mid A_1 \cap \cdots \cap A_{n-1}).$$

Diese Formel wurde in der Analyse des MinCut-Algorithmus in Abschnitt 1.1 benutzt.

Im Kontext von bedingten Wahrscheinlichkeiten gibt es recht einfache, aber grundlegende Formeln, die auch in der Analyse von Algorithmen benutzt werden. Dabei geht es darum, ein Ereignis oder eine Zufallsvariable in disjunkten Teilen des Wahrscheinlichkeitsraumes zu analysieren und dann die Ergebnisse zusammenzufassen. Das erste Ergebnis heißt etwas hochtrabend *Satz von der totalen Wahrscheinlichkeit*, das zweite hat keinen Namen, tut aber in etwa dasselbe für Erwartungswerte.

Fakt 2.4.3 (*Satz von der totalen Wahrscheinlichkeit*)

Für eine (endliche oder abzählbar unendliche) Indexmenge I seien B_i, $i \in I$, Ereignisse, die eine disjunkte Zerlegung des W-Raumes bilden, d. h. $B_i \cap B_j = \emptyset$ für $i, j \in I$ mit $i \neq j$ und $\bigcup_{i \in I} B_i = \Omega$. Dann gilt für jedes Ereignis A und jede Zufallsvariable X, deren Erwartungswert existiert:

$$\mathbf{Pr}(A) = \sum_{i \in I} \mathbf{Pr}(A \mid B_i)\mathbf{Pr}(B_i) \text{ und}$$
$$\mathbf{E}(X) = \sum_{i \in I} \mathbf{E}(X \mid B_i)\mathbf{Pr}(B_i).$$

Die Formel $\mathbf{E}(X) = \sum_{\alpha \in X[\Omega]} \alpha \cdot \mathbf{Pr}(X = \alpha)$ aus Definition 2.2.7 ist ein Spezialfall der zweiten Gleichung: $B_\alpha = \{X = \alpha\}$ und $\mathbf{E}(X \mid X = \alpha) = \alpha$. – In der Analyse von Algorithmen findet man oft das folgende Argumentationsmuster: Man legt eine Bedingung für das Ereignis A bzw. die Zufallsvariable X fest; damit wird eines der Ereignisse B_i ausgewählt. Unter dieser Bedingung beweist man $\mathbf{Pr}(A \mid B_i) \leq a$ bzw. $\mathbf{E}(X \mid B_i) \leq b$, für eine „obere Schranke“ a bzw. b. Fakt 2.4.3 liefert dann sofort $\mathbf{Pr}(A) \leq a$ bzw. $\mathbf{E}(X) \leq b$ (weil $\sum_{i \in I} \mathbf{Pr}(B_i) = 1$ gilt).

2.5 Unabhängigkeit von Ereignissen und von Zufallsvariablen

Definition 2.5.1

(a) Ereignisse A und B heißen **unabhängig**, falls $\mathbf{Pr}(A \cap B) = \mathbf{Pr}(A)\mathbf{Pr}(B)$.

(b) Ereignisse $A_1, \ldots, A_n$ heißen **unabhängig**, falls

$$\mathbf{Pr}\left(\bigcap_{i\in I} A_i \cap \bigcap_{i\in J} (\Omega - A_i)\right) = \prod_{i\in I} \mathbf{Pr}(A_i) \cdot \prod_{i\in J} (1 - \mathbf{Pr}(A_i)),$$

für beliebige *disjunkte* Mengen $I, J \subseteq \{1, \ldots, n\}$.

(c) Eine Familie $(A_i)_{i\in K}$ von Ereignissen heißt **unabhängig**, wenn für jede endliche Teilmenge $J \subseteq K$ die Familie $(A_i)_{i\in J}$ im Sinn von (b) unabhängig ist.

Bemerkung: Wenn $\mathbf{Pr}(A) = 0$ oder $\mathbf{Pr}(A) = 1$, dann sind A und B auf jeden Fall unabhängig. – Zwei Ereignisse A und B mit $\mathbf{Pr}(B) > 0$ sind unabhängig genau dann wenn $\mathbf{Pr}(A \mid B) = \mathbf{Pr}(A)$ gilt. (Denn nach Definition 2.5.1(a) ist $\mathbf{Pr}(A \mid B) \cdot \mathbf{Pr}(B) = \mathbf{Pr}(A \cap B)$.) Das ergibt die übliche intuitive Erklärung der Unabhängigkeit, nämlich dass sich die Wahrscheinlichkeit von A durch das „Wissen“, dass B eingetreten ist, nicht ändert.

In vielen Büchern findet man auch eine (auf den ersten Blick) andere Form von Definition 2.5.1(b): Man nennt $A_1, \ldots, A_n$ unabhängig, falls

$$\mathbf{Pr}\left(\bigcap_{i\in I} A_i\right) = \prod_{i\in I} \mathbf{Pr}\left(A_i\right), \tag{2.5.4}$$

für beliebige Teilmengen I von $\{1, 2, \ldots, n\}$. Diese Definition und Definition 2.5.1(b) sind jedoch äquivalent. Unsere Definition hat den Vorteil, dass man sofort Aussagen machen kann, bei denen Komplementereignisse $\overline{A_i} = \Omega - A_i$ vorkommen.

Beispiel 2.5.2
(a) In Bsp. 2.1.2(h) (Hashing) sind die Ereignisse $\{v_1 = v_1^0\}, \ldots, \{v_n = v_n^0\}$ unabhängig, für beliebige feste Werte $v_1^0, \ldots, v_n^0 \in \{0, \ldots, m-1\}$.

(b) In Bsp. 2.1.2(h) (Hashing) sind die Ereignisse $\{v_1 \neq 0\}, \ldots, \{v_n \neq 0\}$ unabhängig.

Definition 2.5.3

(a) Zufallsfunktionen $X_i\colon \Omega \to R_i$, $1 \le i \le n$, heißen **unabhängig**, wenn für beliebige $R_i' \subseteq R_i$ die Ereignisse $\{X_1 \in R_1'\}, \dots, \{X_n \in R_n'\}$ unabhängig sind. Dies gilt genau dann, wenn

$$\mathbf{Pr}(X_i \in R_i' \text{ für } 1 \le i \le n) = \prod_{1 \le i \le n} \mathbf{Pr}(X_i \in R_i'),$$

für beliebige $R_i' \subseteq R_i$.

(b) Eine Familie $(X_i)_{i \in K}$ von Zufallsfunktionen ist **unabhängig**, wenn für jede endliche Teilmenge $J \subseteq K$ die Familie $(X_i)_{i \in J}$ im Sinn von (a) unabhängig ist.

Die folgende Charakterisierung der Unabhängigkeit für diskrete Wahrscheinlichkeitsräume ist durch die naheliegenden Summationen zu beweisen.

Bemerkung 2.5.4
Zufallsfunktionen $X_i\colon \Omega \to R_i$, $1 \le i \le n$, sind unabhängig genau dann wenn für beliebige $r_i \in R_i$ gilt

$$\mathbf{Pr}(X_i = r_i \text{ für } 1 \le i \le n) = \prod_{1 \le i \le n} \mathbf{Pr}(X_i = r_i).$$

Fakt 2.5.5
Sind $X_1, \dots, X_n$ unabhängig und sind $g_i\colon R_i \to S_i$ beliebig, $1 \le i \le n$, dann sind die Zufallsfunktionen $g_1 \circ X_1, \dots, g_n \circ X_n$ unabhängig.

Beispiel 2.5.6
Sind (Ω_i, p^i), $1 \le i \le n$, W-Räume, so wird durch (Ω, p) mit $\Omega := \Omega_1 \times \cdots \times \Omega_n$, $p := p^1 \times \cdots \times p^n$, wo $p(\omega) = p^1(\omega_1) \cdot \ldots \cdot p^n(\omega_n)$, für $\omega = (\omega_1, \dots, \omega_n) \in \Omega$, ein neuer W-Raum (der „*Produktraum*“) definiert. Auf (Ω, p) sind die n Projektionsfunktionen $X_i\colon (\omega_1, \dots, \omega_n) \mapsto \omega_i \in \Omega_i$ unabhängig; nach Fakt 2.5.5 ist also jede Folge $Y_1, \dots, Y_n$ von Zufallsfunktionen, wo $Y_i = g_i \circ X_i$ (d. h. Y_i hängt *nur* von der i-ten Komponente ω_i ab), unabhängig. (*Beispiel*: Der Wahrscheinlichkeitsraum in Beispiel 2.1.2(h) ist ein Produktraum.)

Fakt 2.5.7

„Bei Unabhängigkeit multiplizieren sich Erwartungswerte, Varianzen addieren sich.“[4]

(a) Sind $X_1, \ldots, X_n$ *unabhängige* Zufallsvariable, so gilt

$$\mathbf{E}(X_1 \cdot \ldots \cdot X_n) = \prod_{1 \leq i \leq n} \mathbf{E}(X_i).$$

(b) Sind $X_1, \ldots, X_n$ *unabhängige* Zufallsvariable, so gilt

$$\mathbf{Var}(X_1 + \cdots + X_n) = \sum_{1 \leq i \leq n} \mathbf{Var}(X_i).$$

Dies gilt sogar, wenn nur X_i und X_j unabhängig sind für $i \neq j$ (*paarweise Unabhängigkeit*).

Beweis. (a) Wir beweisen die Aussage für zwei Zufallsvariable X und Y. Die Verallgemeinerung auf n Zufallsvariable ergibt sich durch vollständige Induktion.

$$\begin{aligned}
\mathbf{E}(X)\mathbf{E}(Y) &= \left(\sum_{\alpha \in X[\Omega]} \alpha \cdot \mathbf{Pr}(X = \alpha) \right) \left(\sum_{\beta \in Y[\Omega]} \beta \cdot \mathbf{Pr}(Y = \beta) \right) \\
&= \sum_{\substack{\alpha \in X[\Omega] \\ \beta \in Y[\Omega]}} \alpha\beta \cdot \mathbf{Pr}(X = \alpha)\mathbf{Pr}(Y = \beta) \\
&= \sum_{\delta \in XY[\Omega]} \left(\sum_{\substack{\alpha \in X[\Omega] \\ \beta \in Y[\Omega] \\ \alpha\beta = \delta}} \alpha\beta \cdot \mathbf{Pr}(X = \alpha)\mathbf{Pr}(Y = \beta) \right) \\
&\overset{(*)}{=} \sum_{\delta \in XY[\Omega]} \delta \cdot \left(\sum_{\substack{\alpha \in X[\Omega] \\ \beta \in Y[\Omega] \\ \alpha\beta = \delta}} \mathbf{Pr}(X = \alpha \wedge Y = \beta) \right)
\end{aligned}$$

[4] Additivität von Erwartungswerten gilt immer, siehe Fakt 2.2.10(c).

$$\overset{(**)}{=} \sum_{\delta \in XY[\Omega]} \delta \cdot \mathbf{Pr}(XY = \delta) = \mathbf{E}(XY).$$

Für $(*)$ wird die Unabhängigkeit benutzt, für $(**)$ die disjunkte Zerlegung

$$\{XY = \delta\} = \bigcup_{\substack{\alpha \in X[\Omega] \\ \beta \in Y[\Omega] \\ \alpha\beta = \delta}} \{X = \alpha \wedge Y = \beta\}.$$

(b) Definiere $X_i' := X_i - \mathbf{E}(X_i)$, für $1 \leq i \leq n$, und $X' = X_1' + \cdots + X_n' = X - \mathbf{E}(X)$. Dann gilt $\mathbf{E}(X_i') = 0$ und $\mathbf{Var}(X_i') = \mathbf{Var}(X_i)$, für $1 \leq i \leq n$, sowie $\mathbf{E}(X') = 0$ und $\mathbf{Var}(X') = \mathbf{Var}(X)$. Das heißt, dass wir o. B. d. A. annehmen können, dass $\mathbf{E}(X_i) = 0$ und $\mathbf{Var}(X_i) = \mathbf{E}(X_i^2)$ und $\mathbf{Var}(X) = \mathbf{E}(X^2)$ gelten. Wir haben dann:

$$\begin{aligned}
\mathbf{Var}(X) &= \mathbf{E}\Big(\big(\sum_{1 \leq i \leq n} X_i \big)^2 \Big) = \mathbf{E}\Big(\sum_{1 \leq i,j \leq n} X_i X_j \Big) = \sum_{1 \leq i,j \leq n} \mathbf{E}(X_i X_j) \\
&= \sum_{1 \leq i \leq n} \mathbf{E}(X_i^2) + \sum_{1 \leq i \neq j \leq n} \mathbf{E}\big(X_i X_j\big) \\
&\overset{(\dagger)}{=} \sum_{1 \leq i \leq n} \mathbf{E}(X_i^2) + \sum_{1 \leq i \neq j \leq n} \underbrace{\mathbf{E}(X_i)}_{=0} \underbrace{\mathbf{E}(X_j)}_{=0} = \sum_{1 \leq i \leq n} \mathbf{Var}(X_i).
\end{aligned}$$

Für $(\dagger)$ wird die Unabhängigkeit von X_i und X_j benutzt (aber auch nicht mehr). □

Beachte noch: Wenn X_i 0-1-wertig ist, ist $X_i^2 = X_i$, also $\mathbf{E}(X_i^2) = \mathbf{E}(X_i)$. Damit erhält man für $X = X_1 + \cdots + X_n$ die folgende nützliche Ungleichung, wenn die X_i paarweise unabhängig sind:

$$\mathbf{Var}(X) = \sum_{1 \leq i \leq n} \mathbf{Var}(X_i) = \sum_{1 \leq i \leq n} (\mathbf{E}(X_i^2) - \mathbf{E}(X_i)^2) \leq \sum_{1 \leq i \leq n} \mathbf{E}(X_i) = \mathbf{E}(X).$$

Gleichheit gilt nur, wenn alle X_i gleich 0 sind.

2.5.1 Zwei Verteilungen

Die Binomialverteilung(en). Die anschauliche Situation, die zu einer Binomialverteilung führt, ist folgende: Wir führen n Experimente durch, wobei bei jedem mit Wahrscheinlichkeit p „Erfolg“ und mit Wahrscheinlichkeit $q = 1 - p$ „Misserfolg“ auftritt. Die Experimente sind unabhängig. (Veranschaulicht wird dies mit dem

n-maligen Werfen einer Münze, die mit Wahrscheinlichkeit p „Kopf“ und mit Wahrscheinlichkeit $q = 1 - p$ „Zahl“ zeigt.) Uns interessiert die Anzahl der „Erfolge“ – eine Zufallsvariable mit Werten in $\{0, \ldots, n\}$. Die Verteilung dieser Zufallsvariablen heißt die *Binomialverteilung zu n und p*, kurz $B(n, p)$. (Es gibt also unendlich viele Binomialverteilungen.)

Technisch betrachtet man den Wahrscheinlichkeitsraum $(\Omega, p^{n,p})$ mit $\Omega = \{0, 1\}^n$, wobei die Idee der Unabhängigkeit durch die Produktverteilung realisiert wird:

$$p^{n,p}((a_1, \ldots, a_n)) = \prod_{1 \leq i \leq n} p^{a_i}(1-p)^{1-a_i}, \text{ für } (a_1, \ldots, a_n) \in \{0, 1\}^n.$$

Die 1-Positionen werden mit Wahrscheinlichkeit p realisiert, die 0-Positionen mit Wahrscheinlichkeit $q = 1 - p$. Die Projektionen $X_i((a_1, \ldots, a_n)) = a_i$ sind dann unabhängige 0-1-wertige Zufallsvariablen mit $\mathbf{Pr}(X_i = 1) = p$. Die Anzahl der „Erfolge“ wird durch die Summe $X = X_1 + \cdots + X_n$ modelliert. Die Verteilung von X auf $\mathbb{N}$ ist dann $B(n, p)$.

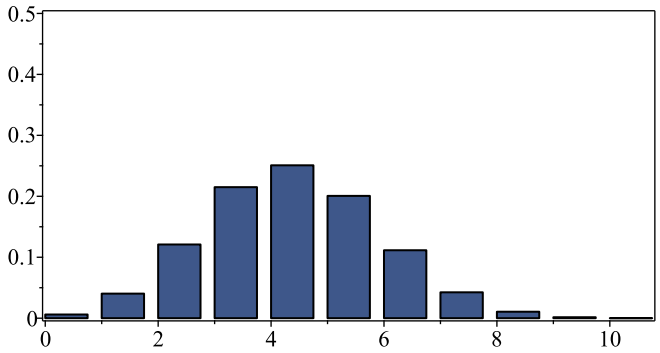

Abbildung 2.5.3: Binomialverteilung $B(10, \frac{2}{5})$. Höhe des Balkens rechts von k ist die Wahrscheinlichkeit $\binom{n}{k}(\frac{2}{5})^k(\frac{3}{5})^{10-k}$ für $k = 0, \ldots, 10$.

Was ist $\mathbf{Pr}(X = k)$, für ein $k \in \mathbb{N}$? Werte $k > n$ kann X nicht annehmen, also ist $\mathbf{Pr}(X = k) = 0$ für $k > n$. Für $k \in \{0, \ldots, n\}$ gilt: $X((a_1, \ldots, a_n)) = k$ genau dann wenn es genau k Positionen i mit $a_i = 1$ gibt. Es gibt genau $\binom{n}{k}$ viele Folgen $(a_1, \ldots, a_n)$ mit genau k Einsen; jede dieser Folgen hat Wahrscheinlichkeit $p^k(1-p)^{n-k}$. Daher gilt

$$\mathbf{Pr}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \text{ für } 0 \leq k \leq n.$$

In Abb. 2.5.3 ist die Wahrscheinlichkeitsfunktion für die Binomialverteilung $B(10, 0.4)$ bildlich dargestellt. – Erwartungswert und Varianz einer binomialverteilten Zufallsvariablen sind nicht schwer zu berechnen.

Fakt 2.5.8
Sei X eine Zufallsvariable, die $B(n,p)$-verteilt ist, und sei $q = 1 - p$. Dann gilt:

$$\mathbf{E}(X) = np \quad \text{und} \quad \mathbf{Var}(X) = np(1-p) = npq.$$

Beweis. Nach der Konstruktion der Binomialverteilung können wir $X = X_1 + \cdots + X_n$ schreiben, für unabhängige 0-1-wertige Zufallsvariablen $X_1, \ldots, X_n$ mit Erwartungswert p. Es folgt $\mathbf{E}(X) = \sum_{1 \le i \le n} \mathbf{E}(X_i) = \sum_{1 \le i \le n} p = np$. Für die Varianz berechnen wir zunächst (beachte $X_i^2 = X_i$, weil X_i 0-1-wertig ist):

$$\mathbf{Var}(X_i) = \mathbf{E}(X_i^2) - \mathbf{E}(X_i)^2 = \mathbf{E}(X_i) - \mathbf{E}(X_i)^2 = p - p^2 = pq.$$

Da die X_i unabhängig sind, addieren sich nach Fakt 2.5.7(b) die Varianzen und wir erhalten

$$\mathbf{Var}(X) = \sum_{1 \le i \le n} \mathbf{Var}(X_i) = \sum_{1 \le i \le n} pq = npq.$$

□

Bemerkung 2.5.9 Schwaches Gesetz der großen Zahlen (Bernoulli)
Wir werfen wiederholt und unabhängig eine Münze, bei der mit Wahrscheinlichkeit p „Kopf“ (entspricht „1“) und mit Wahrscheinlichkeit $q = 1 - p$ „Zahl“ (entspricht „0“) auftritt. Wie wir gerade gesehen haben, ist die Anzahl der „Kopf“-Ergebnisse eine $B(n,p)$-verteilte Zufallsvariable, wenn n die Anzahl der Würfe ist. Wir stellen uns hier auf den Standpunkt, dass wir p nicht kennen und *schätzen* wollen. Es ist naheliegend, die relative Häufigkeit des Ergebnisses „Kopf“ als Schätzwert für p zu nehmen. Das schwache Gesetz der großen Zahlen besagt, dass dieser Wert nur mit kleiner Wahrscheinlichkeit weit von p entfernt ist – mit umso kleinerer Wahrscheinlichkeit, je größer die Anzahl der Münzwürfe ist. Technisch sieht das so aus: Sei $n \in \mathbb{N}$ beliebig. Für $1 \le i \le n$ sei X_i eine 0-1-wertige Zufallsvariable mit $\mathbf{Pr}(X_i = 1) = p$, und diese Zufallsvariablen seien unabhängig.[5] Der Schätzwert für p ist die Zufallsvariable

$$Y_n := \frac{X_1 + \cdots + X_n}{n}.$$

[5] Unsere abzählbar unendlichen Wahrscheinlichkeitsräume lassen es nicht zu, eine unendliche unabhängige Folge $X_1, X_2, \ldots$ zu betrachten. Daher nehmen wir einen W-Raum für jedes n.

Das *Schwache Gesetz der großen Zahlen* (Jakob I Bernoulli, 1655–1705) besagt dann, dass für jedes $\varepsilon > 0$ gilt: $\lim_{n\to\infty} \mathbf{Pr}(|Y_n - p| \geq \varepsilon) = 0$. Genauer gilt:

$$\mathbf{Pr}(|Y_n - p| \geq \varepsilon) \leq \frac{1}{4n\varepsilon^2}. \tag{2.5.5}$$

Der Beweis beruht auf der Chebychev-Ungleichung. Wir haben $\mathbf{E}(X_i) = p$ und $\mathbf{Var}(X_i) = pq = p(1-p) \leq \frac{1}{4}$. Nach Fakt 2.5.7(b) folgt[6] $\mathbf{Var}(Y_n) = (1/n^2)\mathbf{Var}(X_1 + \cdots + X_n) = (1/n^2) \cdot npq = pq/n \leq 1/(4n)$. Nun können wir die Chebychev-Ungleichung anwenden und erhalten:

$$\mathbf{Pr}(|Y_n - p| \geq \varepsilon) \leq \frac{\mathbf{Var}(Y_n)}{\varepsilon^2} \leq \frac{1}{4n\varepsilon^2},$$

wie behauptet.

Wir bemerken, dass das schwache Gesetz der großen Zahlen schon bei nur paarweiser Unabhängigkeit gilt (und in vielen anderen Situationen mit schwächeren Voraussetzungen).

Anwendung: Wir nehmen $p = \frac{1}{2}$ an, also einen fairen Münzwurf. Wieviele Münzwürfe genügen, damit der Schätzwert Y_n mit genügend großer Wahrscheinlichkeit im Intervall [0.45..0.55] liegt? In diesem Fall ist $\varepsilon = \frac{1}{20}$ zu wählen. Dann ist $1/(4\varepsilon^2) = 100$. Wir erhalten $\mathbf{Pr}(\left|Y_n - \frac{1}{2}\right| \geq \frac{1}{20}) \leq \frac{1}{4n\varepsilon^2} = \frac{100}{n}$. Mit 2000 Münzwürfen ist die Wahrscheinlichkeit, dass die Anzahl der Ergebnisse „Kopf“ nicht zwischen 900 und 1100 liegt, höchstens 0.05, mit 10000 Münzwürfen ist die Wahrscheinlichkeit, dass man nicht zwischen 4500 und 5500 mal „Kopf“ erhält, höchstens 0.01.

Die geometrische(n) Verteilung(en). Wieder führen wir wiederholt und unabhängig identische Bernoulli-Experimente durch, also Experimente, bei denen mit Wahrscheinlichkeit p „Erfolg“ und mit Wahrscheinlichkeit $q = 1 - p$ „Misserfolg“ auftritt. Nur ist hier die Anzahl der Experimente unbeschränkt. Wir halten an, sobald „Erfolg“ eintritt. Die Zufallsvariable, die uns interessiert, ist die Anzahl der Versuche. In Beispielen 2.1.2(d) und 2.2.5(b) hatten wir die Situation „Würfeln, bis die erste 6 erscheint“ betrachtet. Wir können hier einen ähnlichen W-Raum wie in Beispiel 2.2.5(b) benutzen:[7] $\Omega = \{(a_1, \ldots, a_i) \mid i \geq 1, a_1 = \cdots = a_{i-1} = 0 \text{ und } a_i = 1\}$

[6] Wir benutzen hier eine weitere einfache, aber nützliche Ungleichung: Für $0 \leq t \leq 1$ gilt $t(1-t) \leq \frac{1}{4}$. Das liegt daran, dass die quadratische Funktion $t \mapsto t(1-t)$ bei $t_0 = \frac{1}{2}$ ihr Maximum annimmt.

[7] Natürlich entspricht dieser Raum genau der Menge $\mathbb{N}^+$ mit der Verteilung analog zu Beispiel 2.1.2(d).

mit der Wahrscheinlichkeitsfunktion $p((a_1,\dots,a_i)) = (1-p)^{i-1}p$. Dass die Versuche unabhängig sind, wird durch die Multiplikation der Einzelwahrscheinlichkeiten modelliert. Die Zufallsvariable X bildet $(a_1,\dots,a_i)$ auf i ab. Die Verteilung von X ist eine W-Verteilung auf $\mathbb{N}$, sie heißt *geometrische Verteilung zu Parameter* p und ist gegeben durch $\mathbf{Pr}(X=0)=0$, $\mathbf{Pr}(X=k)=q^{k-1}p$ für $k\geq 1$. Eine Illustration einer geometrischen Verteilung ist in Abb. 2.5.4 angegeben. Wir berechnen den

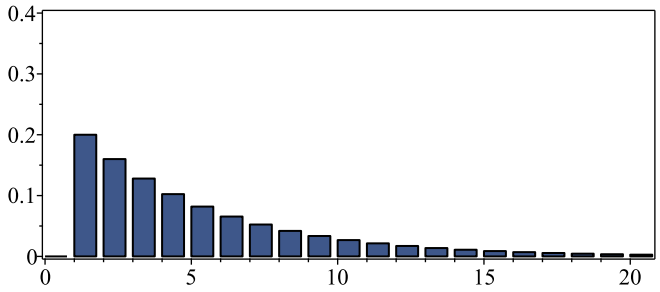

Abbildung 2.5.4: Geometrische Verteilung zum Parameter $p=0.2$, Wahrscheinlichkeiten für $k=0,\dots,20$. Höhe des Balkens rechts von k ist $\mathbf{Pr}(X=k)=0.8^{k-1}\cdot 0.2$.

Erwartungswert, mit Hilfe von Fakt 2.2.9. Zunächst rechnen wir (mit $p=1-q$):

$$\mathbf{Pr}(X\geq k)=\sum_{i\geq k} q^{i-1}p = pq^{k-1}\sum_{i\geq k} q^{i-k} = pq^{k-1}\cdot\frac{1}{1-q}=q^{k-1}.$$

(Dies entspricht der Tatsache, dass die Anzahl der Versuche genau dann $\geq k$ ist, wenn die ersten $k-1$ Versuche alle mit „Misserfolg“ enden, und der Intuition, dass die Wahrscheinlichkeit hierfür q^{k-1} ist.) Anwendung von Fakt 2.2.9 liefert:

$$\mathbf{E}(X)=\sum_{k\geq 1}\mathbf{Pr}(X\geq k)=\sum_{k\geq 1} q^{k-1}=\frac{1}{1-q}=\frac{1}{p}. \tag{2.5.6}$$

Wir wollen noch die Varianz von X bestimmen. Wir stellen dazu zunächst fest, dass $\mathbf{E}(X^2)=\sum_{k\geq 1}k^2q^{k-1}p<\infty$ gilt, sobald $q<1$, d. h. sobald die Erfolgswahrscheinlichkeit p positiv ist. Nach (2.3.2) gilt $\mathbf{Var}(X)=\mathbf{E}(X^2)-\mathbf{E}(X)^2$. Da wir den Erwartungswert kennen, können wir uns auf $\mathbf{E}(X^2)$ konzentrieren.

Wir benötigen eine kleine Vorüberlegung zur Differentiation von Potenzreihen. Betrachte $f(x):=\frac{1}{1-x}=\sum_{k\geq 0}x^k$, absolut konvergent für $|x|<1$. Aus der Analysis

weiß man, dass man gliedweise differenzieren darf. Das ergibt $f'(x) = \sum_{k\geq 1} kx^{k-1}$ und $f''(x) = \sum_{k\geq 2} k(k-1)x^{k-2}$, für x im Konvergenzbereich. Damit:

$$\mathbf{E}(X^2) = \sum_{k\geq 1} k^2 q^{k-1} p = p \sum_{k\geq 2} k(k-1)q^{k-1} + p \sum_{k\geq 1} kq^{k-1} = pq \cdot f''(q) + p \cdot f'(q).$$

Mit den geschlossenen Formeln

$$f'(x) = \frac{1}{(1-x)^2} \quad \text{und} \quad f''(x) = \frac{2}{(1-x)^3}, \quad \text{für } |x| < 1,$$

für die Ableitungen erhalten wir

$$\mathbf{E}(X^2) = pq \cdot \frac{2}{(1-q)^3} + p \cdot \frac{1}{(1-q)^2} = \frac{2pq}{p^3} + \frac{p}{p^2} = \frac{2(1-p)+p}{p^2} = \frac{2-p}{p^2}.$$

Damit ergibt sich die Varianz für die geometrische Verteilung mit Parameter p:

$$\mathbf{Var}(X) = \mathbf{E}(X^2) - \mathbf{E}(X)^2 = \frac{2-p}{p^2} - \left(\frac{1}{p}\right)^2 = \frac{1-p}{p^2}. \tag{2.5.7}$$

2.6 Die Hoeffding-Ungleichung

Die Hoeffding-Ungleichung (oder Chernoff-Hoeffding-Ungleichung) gibt, wie die Chebychev-Ungleichung, eine Schranke für die Wahrscheinlichkeit an, dass eine Zufallsvariable weit von ihrem Erwartungswert entfernt liegt. Sie liefert jedoch ungleich schärfere Abschätzungen als die Chebychev-Ungleichung, wenn X eine Summe von unabhängigen Zufallsvariablen ist, die alle jeweils nur einen kleinen Wertebereich überstreichen. Standardbeispiel für diese Situation ist die Binomialverteilung als Verteilung einer Summe von n unabhängigen 0-1-wertigen Zufallsvariablen.

Satz 2.6.1 (*Hoeffding-Ungleichung*)

$X_1, \ldots, X_n$ seien unabhängige Zufallsvariable mit Werten im Intervall $[0, 1]$. Definiere

$$\begin{aligned} X &:= X_1 + \cdots + X_n\,; \\ m &:= \mathbf{E}(X)\,. \end{aligned}$$

Dann gilt:[8]

$$\mathbf{Pr}(X \geq r) \leq \left(\frac{m}{r}\right)^r \left(\frac{n-m}{n-r}\right)^{n-r}, \text{ für } m \leq r \leq n; \tag{2.6.8}$$

$$\mathbf{Pr}(X \leq s) \leq \left(\frac{m}{s}\right)^s \left(\frac{n-m}{n-s}\right)^{n-s}, \text{ für } 0 \leq s \leq m. \tag{2.6.9}$$

Man sieht sofort, dass das am Anfang von Abschnitt 2.5.1 diskutierte wiederholte Bernoulli-Experiment mit n Münzwürfen, wo jeder mit Wahrscheinlichkeit p das Ergebnis „Kopf“ liefert, direkt zur Situation des Satzes führt. Die Zufallsvariable X_i ist das Ergebnis des i-ten Münzwurfs, und X zählt, wie oft „Kopf“ aufgetreten ist. In diesem Fall ist $m = \mathbf{E}(X) = np$.

Die Hoeffding-Ungleichung gehört zu der Familie der „tail inequalities“, das sind Ungleichungen, die obere Schranken für die Wahrscheinlichkeit liefern, dass Zufallsvariable Werte weit weg von ihrem Erwartungwert annehmen. Wir werden weiter unten sehen, dass die Hoeffding-Schranke relativ kräftig ist, wenn $m = \mathbf{E}(X)$ nicht zu klein ist. Grob gesprochen: Summen von *vielen* (auf $[0, 1]$) *beschränkten unabhängigen* Zufallsvariablen sind eng um ihren Erwartungswert konzentriert. Der Anwendungsbereich ist viel weiter als nur wiederholte Bernoulli-Experimente: Über die einzelnen X_i wird nichts weiter angenommen, als dass sie in $[0, 1]$ eingeschlossen sind. Insbesondere können sie auch ganz unterschiedliche Verteilungen haben und irgendwelche reellen Werte annehmen.

Beweis. (Von Satz 2.6.1.) Wir setzen $p := m/n$ und beweisen zunächst (2.6.8). Der Fall $r = m$ ist trivial, weil auf der linken Seite eine Wahrscheinlichkeit steht, auf der rechten Seite 1. Es sei jetzt also $m < r \leq n$ beliebig, aber fest. – Da der Beweis etwas länger ist, wird er in Schritte untergliedert.

[8] Lesehilfe: Wir benutzen die Konvention, dass $\alpha^0 = 1$ für alle $\alpha \geq 0$ gilt; Faktoren $(\alpha/\beta)^\beta$ haben also für $\alpha > \beta = 0$ den Wert 1.

1. Schritt: Chernoff-Schranke. – Betrachte eine beliebige reelle Zahl $u > 1$. Wir wenden Prop. 2.3.5, mit der monoton wachsenden Funktion $t \mapsto u^t$ an und erhalten

$$\mathbf{Pr}(X \geq r) \leq \frac{\mathbf{E}\big(u^X\big)}{u^r} = u^{-r} \cdot \mathbf{E}\big(u^{X_1+\cdots+X_n}\big) = u^{-r} \cdot \mathbf{E}\Big(\prod_{1\leq i\leq n} u^{X_i}\Big). \tag{2.6.10}$$

2. Schritt: Multipliziere Erwartungswerte. – Weil $X_1, \ldots, X_n$ unabhängig sind, sind auch $u^{X_1}, \ldots, u^{X_n}$ unabhängig (Fakt 2.5.5). Daher (Fakt 2.5.7(a)) multiplizieren sich in (2.6.10) die Erwartungswerte, und wir erhalten:

$$\mathbf{Pr}(X \geq r) \leq u^{-r} \cdot \prod_{1\leq i\leq n} \mathbf{E}(u^{X_i}). \tag{2.6.11}$$

3. Schritt: Abschätzung der einzelnen Erwartungswerte $\mathbf{E}(u^{X_i})$*, per Konvexität.* –

Lemma 2.6.2
Sei $u > 1$ beliebig. Dann gilt:

(i) $u^x \leq 1 + (u-1)x$, für $0 \leq x \leq 1$.

(ii) Für jede Zufallsvariable Y mit $0 \leq Y \leq 1$ gilt $\mathbf{E}(u^Y) \leq 1 + (u-1)\mathbf{E}(Y)$.

Beweis. (i) Da $u > 1$, ist die Funktion $x \mapsto u^x$ konvex. Daher verläuft der Graph dieser Funktion in $[0,1]$ unterhalb der Strecke durch $(0, u^0) = (0,1)$ und $(1, u^1) = (1, u)$, die durch $x \mapsto 1 + (u-1)x$ gegeben ist. Das liefert genau (i).
(ii) Wegen (i) gilt $u^Y \leq 1 + (u-1)Y$ als Ungleichung zwischen Zufallsvariablen. Die Behauptung folgt wegen der Monotonie und der Linearität des Erwartungswertes. □

Mit Lemma 2.6.2(ii) erhalten wir aus (2.6.11):

$$\mathbf{Pr}(X \geq r) \leq u^{-r} \cdot \prod_{1\leq i\leq n} \left(1 + (u-1)\mathbf{E}(X_i)\right). \tag{2.6.12}$$

4. Schritt: Arithmetisches und geometrisches Mittel. – Auf das Produkt in (2.6.12) wenden wir die Ungleichung zwischen dem arithmetischen und dem geometrischen Mittel (Prop. 2.3.9) an, mit $a_i = 1 + (u-1)\mathbf{E}(X_i)$. Dies liefert:

$$\mathbf{Pr}(X \geq r) \leq u^{-r} \cdot \left(\frac{1}{n} \sum_{1\leq i\leq n} \left(1 + (u-1)\mathbf{E}(X_i)\right)\right)^n.$$

Nun erinnern wir uns, dass $X = X_1 + \cdots + X_n$ und $m = \mathbf{E}(X) = np$ ist, und erhalten

$$\mathbf{Pr}(X \geq r) \leq u^{-r} \cdot \left(1 + (u-1) \cdot \frac{m}{n}\right)^n = u^{-r} \cdot (1 - p + pu)^n. \tag{2.6.13}$$

5. Schritt: Minimiere Schranke durch Variation von u. – Die rechte Seite $g(u) = u^{-r}(1-p+pu)^n$ in (2.6.13) hängt von dem beliebigen Wert $u > 1$ ab. Um diese Schranke möglichst gut auszunutzen, variieren wir u und suchen die Stelle, an der $g(u)$ minimal wird. Dazu verwenden wir die üblichen Techniken aus der Analysis. Ein bisschen Bruchrechnen erledigt den Rest.

Im Grenzfall $r = n$ gilt $g(u) = (u^{-1}(1-p)+p)^n \to p^n$ für $u \to \infty$, und damit $\mathbf{Pr}(X \geq r) \leq p^n = (m/r)^r$. Damit ist (2.6.8) für $r = n$ bewiesen, und wir können ab hier $m < r < n$ annehmen.

Wir orientieren uns grob: $\lim_{u\to 1} g(u) = 1$ und $\lim_{u\to\infty} g(u) = \infty$. Irgendwo dazwischen vermuten wir ein globales (also auch lokales) Minimum. Wir differenzieren mit der Produktregel:

$$g'(u) = -ru^{-(r+1)}(1-p+pu)^n + u^{-r}n(1-p+pu)^{n-1}p.$$

Nullsetzen von $g'(u)$ und Multiplizieren der entstehenden Gleichung mit $u^{r+1}(1-p+pu)^{-n+1}$ $(\neq 0)$ führt zu der Bedingung $r(1-p+pu) = npu$ für die Nullstelle u_0 von $g'(u)$ (an der wir das Minimum vermuten), und damit zu der Lösung

$$u_0 = \frac{r(1-p)}{(n-r)p} = \frac{r(1-\frac{m}{n})}{(n-r)\frac{m}{n}} = \frac{r(n-m)}{m(n-r)} > 1.$$

Einsetzen von u_0 in $g(u)$ liefert:

$$g(u_0) = \left(\frac{m(n-r)}{r(n-m)}\right)^r \cdot \left(1 - \frac{m}{n} + \frac{r(n-m)}{n(n-r)}\right)^n = \left(\frac{m(n-r)}{r(n-m)}\right)^r \cdot \left(\frac{n-m}{n-r}\right)^n.$$

Mit (2.6.13) ergibt sich

$$\mathbf{Pr}(X \geq r) \leq g(u_0) = \left(\frac{m}{r}\right)^r \cdot \left(\frac{n-m}{n-r}\right)^{n-r},$$

und das ist (2.6.8).

Um (2.6.9) zu beweisen, benutzen wir (2.6.8) auf geschickte Weise. Wir definieren „Komplementär-Zufallsvariablen“

$$\overline{X_i} := 1 - X_i, \quad 1 \leq i \leq n, \quad \text{und} \quad \overline{X} := \overline{X_1} + \cdots + \overline{X_n} = n - X.$$

Weiter setzen wir $\overline{m} := \mathbf{E}(\overline{X}) = \mathbf{E}(n - X) = n - \mathbf{E}(X) = n - m$ und $\overline{r} := n - s$. Dann gilt $\overline{m} \leq \overline{r} \leq n$. Die Zufallsvariablen $\overline{X_i}$, $1 \leq i \leq n$, sind wieder unabhängig, mit Werten in $[0, 1]$. Wir können also (2.6.8) anwenden und erhalten:

$$\mathbf{Pr}(\overline{X} \geq \overline{r}) \leq \left(\frac{\overline{m}}{\overline{r}}\right)^{\overline{r}} \cdot \left(\frac{n - \overline{m}}{n - \overline{r}}\right)^{n-\overline{r}}.$$

Wenn man diese Ungleichung wieder in die „X_i-Notation“ überführt, ergibt sich:

$$\mathbf{Pr}(X \leq s) \leq \left(\frac{n - m}{n - s}\right)^{n-s} \cdot \left(\frac{m}{s}\right)^{s},$$

und das ist gerade (2.6.9). □

Im Folgenden geben wir nützliche Varianten der Hoeffding-Ungleichung an.

Korollar 2.6.3
In der Situation von Satz 2.6.1 gilt:

$$\mathbf{Pr}(X \geq (1+\varepsilon)m) \leq \left(\frac{e^{\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}}\right)^{m}, \text{ für } 0 \leq \varepsilon \leq \tfrac{n}{m} - 1; \tag{2.6.14}$$

$$\mathbf{Pr}(X \leq (1-\varepsilon)m) \leq \left(\frac{e^{-\varepsilon}}{(1-\varepsilon)^{1-\varepsilon}}\right)^{m}, \text{ für } 0 \leq \varepsilon \leq 1. \tag{2.6.15}$$

Beweis. Für (2.6.14) setzen wir $r := (1+\varepsilon)m$ und wenden (2.6.8) an, um zu erhalten:

$$\mathbf{Pr}(X \geq (1+\varepsilon)m) \leq \left(\frac{1}{1+\varepsilon}\right)^{(1+\varepsilon)m} \left(\frac{n-m}{n-r}\right)^{n-r}.$$

Wir müssen nur noch zeigen, dass der zweite Faktor nicht größer als $e^{\varepsilon m}$ ist. Für $r = n$ ist dies trivial, weil der zweite Faktor 1 ist; sei ab hier $r < n$. Die Ungleichung $(1 + \frac{x}{y})^y \leq e^x$, die nach Prop. A.1.2(b) für $y > 0$ und $x \geq -y$ gültig ist, liefert:

$$\left(\frac{n-m}{n-r}\right)^{n-r} = \left(1 + \frac{r-m}{n-r}\right)^{n-r} \leq e^{r-m} = e^{\varepsilon m}.$$

Für (2.6.15) setzen wir $s := (1-\varepsilon)m$, und erhalten mit (2.6.9):

$$\mathbf{Pr}(X \leq (1-\varepsilon)m) \leq \left(\frac{1}{1-\varepsilon}\right)^{(1-\varepsilon)m} \left(\frac{n-m}{n-s}\right)^{n-s}.$$

Die eben genannte Ungleichung aus Prop. A.1.2(b) liefert für den zweiten Faktor:

$$\left(\frac{n-m}{n-s}\right)^{n-s} = \left(1-\frac{m-s}{n-s}\right)^{n-s} \leq e^{-(m-s)} = e^{-\varepsilon m}.$$

□

Korollar 2.6.3 besagt Folgendes: Wenn man eine tolerierbare prozentuale Abweichung (z. B. $\varepsilon = 0.01$, was 1 Prozent entspricht) vorgibt, dann ist die Wahrscheinlichkeit, dass X weiter als diese Toleranz von seinem Erwartungswert $m = \mathbf{E}(X)$ abweicht, durch eine in m exponentiell fallende Funktion beschränkt. Je kleiner ε wird, desto näher an 1 liegt die Basis dieser Exponentialfunktion. Der Verlauf der Funktion $\varepsilon \mapsto \frac{e^{\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}}$ ist in Abb. 2.6.5 angegeben.

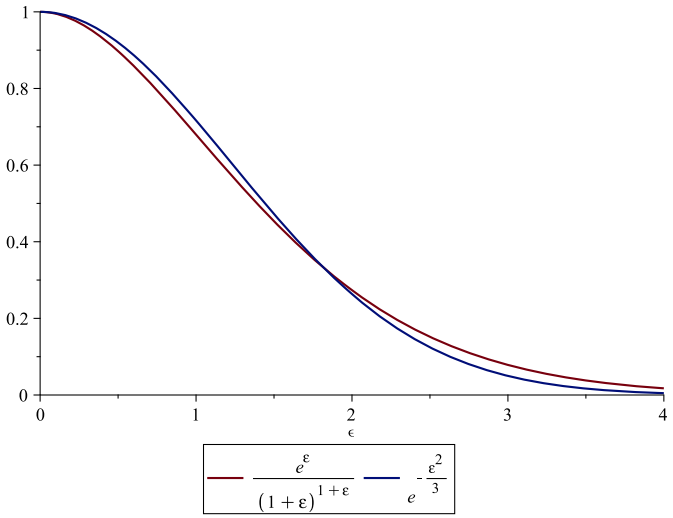

Abbildung 2.6.5: Funktionen $\varepsilon \mapsto \frac{e^{\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}}$ und $\varepsilon \mapsto e^{-\varepsilon^2/3}$. Die erste Funktion ist für $0 \leq \varepsilon \leq 1.8$ durch die zweite beschränkt.

Wir notieren noch weitere nützliche und häufig benutzte Formen der Hoeffding-Ungleichungen.

Korollar 2.6.4
In der Situation von Korollar 2.6.3 gilt:

$$\mathbf{Pr}(X \geq (1+\varepsilon)m) \leq e^{-\varepsilon^2 m/3}, \text{ für } 0 \leq \varepsilon \leq 1.8; \tag{2.6.16}$$

$$\mathbf{Pr}(X \geq (1+\varepsilon)m) \leq e^{-\varepsilon^2 m/4}, \text{ für } 0 \leq \varepsilon \leq 4.1; \tag{2.6.17}$$

$$\mathbf{Pr}(X \leq (1-\varepsilon)m) \leq e^{-\varepsilon^2 m/2}, \text{ für } 0 \leq \varepsilon \leq 1; \tag{2.6.18}$$

$$\mathbf{Pr}(|X-m| \geq \varepsilon m) \leq 2e^{-\varepsilon^2 m/3}, \text{ für } 0 \leq \varepsilon \leq 1. \tag{2.6.19}$$

$$r \geq 5m \Rightarrow \mathbf{Pr}(X \geq r) \leq 2^{-r}. \tag{2.6.20}$$

Für eine Illustration der beiden Funktionen $e^{\varepsilon}/(1+\varepsilon)^{1+\varepsilon}$ und $e^{-\varepsilon^2/3}$ siehe Abb. 2.6.5. Aus dem Bild erkennt man auch, dass sich für $\varepsilon > 1.85$ die Beziehung zwischen den Funktionen umdreht. Der *Beweis* von (2.6.16) und (2.6.17) besteht in einer Diskussion des Verlaufs der Funktionen

$$\varepsilon \mapsto \ln\left(\frac{e^{-\varepsilon^2/K}}{e^{\varepsilon}/(1+\varepsilon)^{1+\varepsilon}}\right) = -\varepsilon^2/K - \varepsilon + (1+\varepsilon)\ln(1+\varepsilon),$$

für $K = 3$ und $K = 4$, aus der hervorgeht, dass diese Funktion im Intervall $[1, 1.8]$ (für $K = 3$) bzw. $[1, 4.1]$ (für $K = 4$) nicht negativ ist. Wir führen dies nicht im Detail durch, sondern begnügen uns mit einem Blick auf ein Bild für $K = 3$ (s. Abb. 2.6.6). Damit folgt die Behauptung direkt aus (2.6.14). Die dritte Ungleichung (2.6.17) folgt

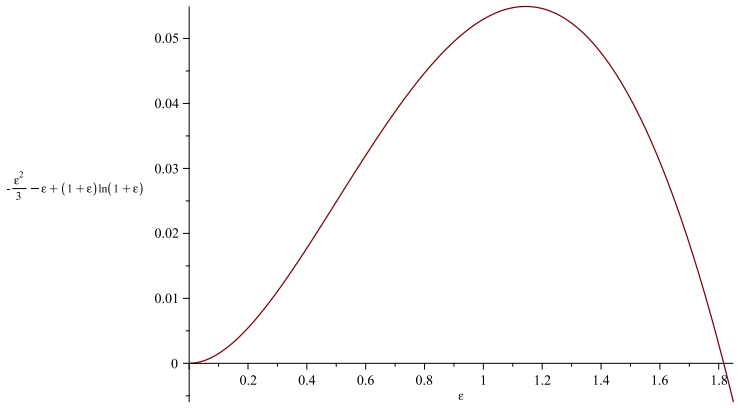

Abbildung 2.6.6: Funktion $\varepsilon \mapsto -\varepsilon^2/3 - \varepsilon + (1+\varepsilon)\ln(1+\varepsilon)$ ist in $[1, 1.8]$ nicht negativ.

ähnlich aus der Beobachtung, dass die Funktion $\varepsilon \mapsto \ln\bigl(e^{-\varepsilon^2/2}/(e^{-\varepsilon}/(1-\varepsilon)^{1-\varepsilon})\bigr) = -\varepsilon^2/2+\varepsilon+(1-\varepsilon)\ln(1-\varepsilon)$ im Intervall $[0,1]$ nicht negativ ist. Wir führen dies nicht durch. Die vierte Ungleichung (2.6.19) folgt mit der Vereinigungsschranke aus (2.6.16) und (2.6.18). Für die fünfte und letzte Ungleichung kehren wir zu $r = (1+\varepsilon)m$ und zu Schranke (2.6.14) zurück, um zu erhalten:

$$\mathbf{Pr}(X \geq r) \leq \left(\frac{e^{\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}}\right)^m = \left(\left(\frac{e^{r/m-1}}{(r/m)^{r/m}}\right)^{m/r}\right)^r = \left(\frac{e^{1-m/r}}{r/m}\right)^r.$$

Die Funktion $g\colon t \mapsto e^{1-1/t}/t$ fällt strikt monoton mit $t \geq 1$. (Man betrachtet $\frac{\mathrm{d}}{\mathrm{d}t}\ln(g(t)) = \frac{\mathrm{d}}{\mathrm{d}t}(1-1/t-\ln t) = 1/t^2 - 1/t < 0$, für $t > 1$.) Wir betrachten $t_0 = 2e-1 = 4.43656... < 4.5$. Es gilt, mit Prop. A.1.3 im Anhang:

$$g(t_0) = \frac{e^{1-\frac{1}{2e-1}}}{2e-1} \leq \frac{\frac{e}{1+\frac{1}{2e-1}}}{2e-1} = \frac{1}{2}.$$

Also gilt $g(t) < \frac{1}{2}$ für alle $t \geq 2e-1$, und das liefert $\mathbf{Pr}(X \geq r) \leq (\frac{1}{2})^r$ für $r \geq (2e-1)m$.

Bemerkung 2.6.5
Um die Stärke der Hoeffding-Ungleichung(en) zu demonstrieren, betrachten wir nochmals das n-fache Werfen einer fairen Münze. Die Zufallsvariable X zählt, wie oft „Kopf“ auftaucht. In Bem. 2.5.9 (Schwaches Gesetz der großen Zahlen) wird gezeigt, wie mit der Chebychev-Ungleichung die Wahrscheinlichkeit für eine große (relative) Abweichung vom Erwartungswert beschränkt werden kann. Wir vergleichen dies mit Schranken, die sich aus der Hoeffding-Ungleichung ergeben.

Seien also $X_1, \ldots, X_n$ unabhängige 0-1-wertige Zufallsvariablen mit $\mathbf{E}(X_i) = \frac{1}{2}$ für alle i. Sei $X = X_1 + \cdots + X_n$. Sei $m = \mathbf{E}(X) = n/2$. Für eine beliebige Konstante $c > 0$ betrachten wir die (zu m) relative Abweichung $\varepsilon := \varepsilon(n) := 2\sqrt{c(\ln n)/n}$. Der Rest der Diskussion trifft nur für die n zu, für die $\varepsilon(n) \leq 1.8$ gilt, was keine ernsthafte Einschränkung ist. (Dieser relativen Abweichung entspricht eine absolute Abweichung von $\varepsilon m = 2\sqrt{c(\ln n)/n} \cdot \frac{n}{2} = \sqrt{cn\ln n}$. Wir sehen gleich in der Rechnung, weshalb es interessant ist, eine so merkwürdige Formel zu wählen.)

Wir wenden das schwache Gesetz der großen Zahlen mit $p = q = \frac{1}{2}$ an, wobei man achtgeben muss, dass in der Formulierung dort die Abweichung relativ zu n formuliert ist. Wir erhalten aus (2.5.5):

$$\mathbf{Pr}(|X - \mathbf{E}(X)| \geq \varepsilon m) = \mathbf{Pr}\left(\left|\tfrac{1}{n}X - \mathbf{E}(\tfrac{1}{n}X)\right| \geq \varepsilon/2\right) \leq \frac{1}{4(\varepsilon/2)^2 n} = \frac{1}{4c\ln n}.$$

Für jedes feste c geht diese Schranke mit $n \to \infty$ gegen 0, aber relativ gemächlich. Die Hoeffding-Ungleichung in der Form (2.6.19) liefert:

$$\mathbf{Pr}(|X - \mathbf{E}(X)| \geq \varepsilon m) \leq 2e^{-\left(2\sqrt{c(\ln n)/n}\right)^2 (n/2)/3} = 2e^{-4c(\ln n)/6} = \frac{2}{n^{2c/3}}.$$

Mit der Wahl $c = \frac{3}{2}$ ist die Schranke $2/n$, mit der Wahl $c = 3$ wird sie $2/n^2$. Dies wird mit wachsendem n rasch sehr klein.

Ein Zahlenbeispiel soll den Unterschied verdeutlichen. Wähle $c = 3/2$. Mit $n = 1\,000\,000$ gilt $\ln n = 13.8155...$, also $\varepsilon m = \sqrt{cn \ln n} \leq 4553$. Mit dem schwachen Gesetz der großen Zahlen erhalten wir für die Wahrscheinlichkeit, dass die beobachtete Anzahl von „Kopf"-Ergebnissen um mehr als 4553 (das ist weniger als 1%) vom Erwartungswert $m = 500\,000$ abweicht:

$$\mathbf{Pr}(|X - \mathbf{E}(X)| \geq 4553) \leq \frac{1}{4c \ln n} \approx 0.0121.$$

Mit der Hoeffding-Ungleichung ergibt sich

$$\mathbf{Pr}(|X - \mathbf{E}(X)| \geq 4553) \leq \frac{2}{n} = 0.000002.$$

Mit $c = 3$ vergrößert man die tolerierte Abweichung auf 6438, immer noch weniger als 1.3%. Die vom schwachen Gesetz der großen Zahlen gelieferte Wahrscheinlichkeitsschranke verringert sich auf ≈ 0.0061. Die von der Hoeffding-Ungleichung gelieferte Schranke fällt jedoch auf den winzigen Wert $2 \cdot 10^{-12}$.

Noch zwei Bemerkungen zum Schluss: (i) Auch die Hoeffding-Ungleichung verhindert nicht, dass Abweichungen von X von m um $\sqrt{n}$ oder mehr mit konstanter Wahrscheinlichkeit vorkommen. Dies ist einfach eine Eigenschaft der Binomialverteilung $B(n, \frac{1}{2})$ und hängt damit zusammen, dass die *Standardabweichung* $\sqrt{\mathbf{Var}(X)}$ gleich $\frac{1}{2}\sqrt{n}$ ist. (ii) Dass die Hoeffding-Ungleichung schärfere Schranken liefert, hat auch damit zu tun, dass das schwache Gesetz der großen Zahlen unter viel schwächeren Bedingungen als der vollständigen Unabhängigkeit gilt. (Paarweise Unabhängigkeit genügt!)

2.7 Weitere Ungleichungen

Proposition 2.7.1 (*Cauchy-Schwarz-Ungleichung*)
Für Zufallsvariablen X und Y, deren Erwartungswert und Varianz definiert ist, gilt:

$$|\mathbf{E}(XY)| \leq \sqrt{\mathbf{E}(X^2)\mathbf{E}(Y^2)}.$$

Beweis: Wir zeigen: $\mathbf{E}(XY)^2 \leq \mathbf{E}(X^2)\mathbf{E}(Y^2)$. – Für $\lambda \in \mathbb{R}$ betrachte

$$f(\lambda) := \mathbf{E}((\lambda X + Y)^2) = \lambda^2\mathbf{E}(X^2) + 2\lambda\mathbf{E}(XY) + \mathbf{E}(Y^2) = \alpha\lambda^2 + 2\beta\lambda + \mathbf{E}(Y^2),$$

mit $\alpha := \mathbf{E}(X^2) \geq 0$ und $\beta := \mathbf{E}(XY)$. Weil $(\lambda X + Y)^2 \geq 0$, gilt $f(\lambda) \geq 0$ für alle $\lambda \in \mathbb{R}$. Wenn $\alpha = 0$ ist, haben wir $2\beta\lambda + \mathbf{E}(Y^2) \geq 0$ für alle $\lambda \in \mathbb{R}$, was nur für $\mathbf{E}(XY) = \beta = 0$ möglich ist. In diesem Fall gilt also die Ungleichung. Nun nehmen wir $\alpha > 0$ an. Für $\lambda_0 := -\beta/\alpha$ (Minimalstelle der Funktion $f(\lambda)$) gilt:

$$0 \leq f(\lambda_0) = \alpha\left(-\frac{\beta}{\alpha}\right)^2 + 2\beta\left(-\frac{\beta}{\alpha}\right) + \mathbf{E}(Y^2) = -\frac{\beta^2}{\alpha} + \mathbf{E}(Y^2),$$

also $\mathbf{E}(XY)^2 = \beta^2 \leq \alpha\mathbf{E}(Y^2) = \mathbf{E}(X^2)\mathbf{E}(Y^2)$. □

Nicht ganz ideal an der Chebychev-Ungleichung (Fakt 2.3.4) ist, dass sie nur für $t > \sqrt{\mathbf{Var}(X)}$ nützliche Information liefert. Für kleinere t ist die Schranke $\mathbf{Var}(X)/t^2$ größer oder gleich 1, also trivial. Oft hilft die folgende Variante.

Proposition 2.7.2 (*Chebychev-Cantelli-Ungleichung*)
Es sei X eine Zufallsvariable mit $\mathbf{E}(X^2) < \infty$. Dann gilt für alle $t > 0$:

$$\mathbf{Pr}(X \geq \mathbf{E}(X) + t) \leq \frac{\mathbf{Var}(X)}{\mathbf{Var}(X) + t^2} \text{ und } \mathbf{Pr}(X \leq \mathbf{E}(X) - t) \leq \frac{\mathbf{Var}(X)}{\mathbf{Var}(X) + t^2}.$$

Beweis:[9] Die zweite Ungleichung folgt, indem man die erste auf die Zufallsvariable $X' = \mathbf{E}(X) - X$ anwendet, die dieselbe Varianz hat wie X. Wir zeigen die erste Ungleichung. Wir können o. B. d. A. annehmen, dass $\mathbf{E}(X) = 0$ ist (sonst betrachte

[9] Wir geben hier einen Beweis mit der Cauchy-Schwarz-Ungleichung an. In der Übung wird die Ungleichung auf direktem Weg bewiesen.

$X' = X - \mathbf{E}(X)$); dann ist $\mathbf{Var}(X) = \mathbf{E}(X^2)$. Man erinnere sich an die Iverson-Notation: $[X < t]$ ist die charakteristische Funktion des Ereignisses $\{X < t\}$, usw. Für alle $t \in \mathbb{R}$ gilt offenbar: $t - X \leq (t - X) \cdot [X < t]$, also

$$t = \mathbf{E}(t - X) \leq \mathbf{E}((t - X) \cdot [X < t]).$$

Für $t > 0$ können wir dann mit der Cauchy-Schwarz-Ungleichung wie folgt weiterrechnen:

$$\begin{aligned} t^2 &\leq \mathbf{E}((t-X)^2)\mathbf{E}([X<t]^2) \\ &= \mathbf{E}((t-X)^2)\mathbf{Pr}(X<t) \\ &= (\mathbf{Var}(X) + t^2)\mathbf{Pr}(X<t). \end{aligned}$$

Umstellen ergibt:

$$\mathbf{Pr}(X \geq t) = 1 - \mathbf{Pr}(X < t) \leq 1 - \frac{t^2}{\mathbf{Var}(X) + t^2} = \frac{\mathbf{Var}(X)}{\mathbf{Var}(X) + t^2},$$

wie gewünscht. □

Bemerkung: Wir vergleichen Prop. 2.7.2 mit der Chebychev-Ungleichung (Fakt 2.3.4). Für die Wahrscheinlichkeit einer beidseitigen Abweichung liefert die Chebychev-Ungleichung engere Schranken; sie wirkt aber nur für $t > \sqrt{\mathbf{Var}(X)}$. Die Chebychev-Cantelli-Ungleichung ist geeignet, wenn man die Wahrscheinlichkeit der Abweichung nur nach einer Seite begrenzen will; sie wirkt für alle $t > 0$.

Wir leiten noch eine obere und eine untere Schranke für $\mathbf{Pr}(X > 0)$ her, falls $X \not\equiv 0$ eine Zufallsvariable mit Werten in $\mathbb{N}$ ist.

Proposition 2.7.3

Für eine Zufallsvariable X mit Werten in $\mathbb{N}$, die nicht konstant 0 ist und deren Erwartungswert und Varianz definiert ist, gilt:

$$\frac{\mathbf{E}(X)^2}{\mathbf{E}(X^2)} \leq \mathbf{Pr}(X > 0) \leq \mathbf{E}(X).$$

Beweis: Die zweite Ungleichung folgt direkt aus der Markov-Ungleichung, da wegen der Ganzzahligkeit von X die Gleichung $\mathbf{Pr}(X > 0) = \mathbf{Pr}(X \geq 1)$ gilt. Weiter

beobachten wir, mit Prop. 2.7.2 (Chebychev-Cantelli-Ungleichung):

$$1 - \mathbf{Pr}(X > 0) = \mathbf{Pr}(X = 0) = \mathbf{Pr}(X \leq \mathbf{E}(X) - \mathbf{E}(X)) \leq \frac{\mathbf{Var}(X)}{\mathbf{Var}(X) + \mathbf{E}(X)^2}$$
$$= \frac{\mathbf{E}(X^2) - \mathbf{E}(X)^2}{\mathbf{E}(X^2)} = 1 - \frac{\mathbf{E}(X)^2}{\mathbf{E}(X^2)}.$$

Umstellen liefert die erste Ungleichung. □

Wenn X Summe von 0-1-wertigen Zufallsvariablen ist, kann man alternativ mit folgender Ungleichung die Wahrscheinlichkeit für $\mathbf{Pr}(X > 0)$ nach unten abschätzen.

Proposition 2.7.4 (*Conditional Expectation Inequality, CEI*)
Für beliebige Zufallsvariablen $X_1, X_2, \ldots, X_n$ mit Werten in $\{0, 1\}$ gilt:

$$\mathbf{Pr}(X_1 + \cdots + X_n > 0) \geq \sum_{1 \leq i \leq n} \frac{\mathbf{Pr}(X_i = 1)}{\mathbf{E}(X \mid X_i = 1)}.$$

Beweis: Sei $X = X_1 + \cdots + X_n$. Wir wählen die Zufallsvariable Y so, dass $X \cdot Y = [X > 0]$; sei dazu $Y(\omega) = 1/X(\omega)$, falls $X(\omega) > 0$ und $Y(\omega) = 0$, falls $X(\omega) = 0$. Dann gilt:

$$\begin{aligned}
\mathbf{Pr}(X > 0) &= \mathbf{E}(X \cdot Y) \qquad \text{(Wahl von } Y\text{)} \\
&= \sum_{1 \leq i \leq n} \mathbf{E}(X_i \cdot Y) \\
&\overset{(1)}{=} \sum_{1 \leq i \leq n} \mathbf{Pr}(X_i = 1) \cdot \mathbf{E}(\tfrac{1}{X} \mid X_i = 1) \\
&\overset{(2)}{\geq} \sum_{1 \leq i \leq n} \frac{\mathbf{Pr}(X_i = 1)}{\mathbf{E}(X \mid X_i = 1)}.
\end{aligned}$$

Für (1) benutzt man, dass $\mathbf{E}(X_i \cdot Y \mid X_i = 1) = \mathbf{E}(Y \mid X_i = 1)$ und $\mathbf{E}(X_i \cdot Y \mid X_i = 0) = 0$ gilt. Für (2) wendet man die Jensensche Ungleichung (Prop. 2.3.8(a)) auf die für $x > 0$ konvexe Funktion $x \mapsto \frac{1}{x}$ und die Zufallsvariable X mit dem auf $\{X_i = 1\}$ bedingten Wahrscheinlichkeitsraum an. Dies liefert $\mathbf{E}(\frac{1}{X} \mid X_i = 1) \geq 1/\mathbf{E}(X \mid X_i = 1)$. □

A Anhang

A.1 Ungleichungen aus der Analysis und der Kombinatorik

Proposition A.1.1

Für alle $x \in \mathbb{R}$ gilt $1 + x \leq e^x$, mit Gleichheit genau für $x = 0$.

Beweis: Die Funktion $f(x) = e^x - (1+x)$ besitzt die Ableitung $f'(x) = e^x - 1$ und die zweite Ableitung $f''(x) = e^x > 0$. Die Ableitung ist also strikt monoton wachsend; sie hat bei $x = 0$ ihre einzige Nullstelle. Daraus folgt, dass f an der Stelle $x = 0$ ein globales Minimum hat, d. h., es gilt $e^x - (1+x) \geq f(0) = 0$ für alle x, mit Gleichheit genau für $x = 0$. □

Diese Behauptung wird oft für den Fall (kleiner) negativer Werte angewendet; sie liest sich dann $1 - x \leq e^{-x}$, für $x > 0$ beliebig.

Proposition A.1.2

(a) Für alle $y > 0$ und alle $x \geq -1$ gilt: $(1 + x)^y \leq e^{xy}$.

(b) Für alle $y > 0$ und alle $x \geq -y$ gilt: $\left(1 + \frac{x}{y}\right)^y \leq e^x$.

Beweis. (a) Es gilt $0 \leq 1 + x \leq e^x$ (nach Prop. A.1.1). Weil $y > 0$ ist, ist die Funktion $u \mapsto u^y$ in $[0, \infty)$ monoton wachsend. Damit: $(1 + x)^y \leq (e^x)^y = e^{xy}$, also (a). Aussage (b) folgt aus (a), indem man $x' = x/y \geq -1$ betrachtet. □

Proposition A.1.3

Für alle $x \in \mathbb{R}$ mit $|x| < 1$ gilt $e^x \leq \frac{1}{1-x}$, mit Gleichheit genau für $x = 0$.

Beweis: Sei $g(x) := \ln(e^x(1 - x)) = x + \ln(1 - x)$. Wir zeigen: $g(x) \leq 0$ für $x \in \mathbb{R}, |x| < 1$, mit Gleichheit genau für $x = 0$. Differenzieren liefert $g'(x) = 1 - \frac{1}{1-x}$, eine Funktion, die in $(-1, 1)$ strikt monoton fallend ist und in $x = 0$ eine Nullstelle hat. Daraus folgt, dass $g(x)$ für $x < 0$ strikt monoton wächst und für $x > 0$ strikt monoton fällt. Weil zudem $g(0) = 0$ gilt, folgt die Behauptung. □

Proposition A.1.4

Für alle $x > 0$ gilt $\ln x \leq x - 1$, mit Gleichheit genau für $x = 1$.

Beweis: Setze $y := x - 1$. Dann lautet die Behauptung: $\ln(1+y) \leq y$ für alle $y > -1$, mit Gleichheit genau für $y = 0$. Wir wenden die Exponentialfunktion an, und erhalten, dass Folgendes äquivalent zur Behauptung ist: $1 + y \leq e^y$ für alle $y > -1$, mit Gleichheit genau für $y = 0$. Diese Aussage folgt aber aus Prop. A.1.1. □

Proposition A.1.5

Für alle $n, k \in \mathbb{N}$, $0 \leq k \leq n$, gilt:

$$\binom{n}{k} \leq \frac{n^n}{k^k (n-k)^{n-k}} = \frac{1}{(\alpha^\alpha (1-\alpha)^{1-\alpha})^n},$$

wobei $\alpha = \frac{k}{n}$. Weiterhin:

$$\binom{n}{k} \leq \left(\frac{en}{k}\right)^k.$$

Beweis: Für $k = 0$ und $k = n$ ist nichts zu zeigen – linke und rechte Seite sind beide gleich 1. Sonst gilt nach der binomischen Formel:

$$n^n = (k + (n-k))^n = \sum_{0 \leq i \leq n} \binom{n}{i} k^i (n-k)^{n-i} \geq \binom{n}{k} k^k (n-k)^{n-k},$$

und daher $\binom{n}{k} \leq \frac{n^n}{k^k (n-k)^{n-k}}$. Die zweite Ungleichung folgt, weil

$$\left(\frac{n}{n-k}\right)^{n-k} = \left(1 + \frac{k}{n-k}\right)^{n-k} < e^k,$$

mit Prop. A.1.2(b). □

A.2 Summierbarkeit

Bei der Diskussion von abzählbar unendlichen Wahrscheinlichkeitsräumen ist es bequem, „Summen“ von unendlich vielen Summanden zu betrachten, ohne dass eine

bestimmte Reihenfolge festgelegt ist. Wir notieren grundlegende Definitionen, Fakten und Schreibweisen.

Erinnerung: In der Analysis diskutiert man das Konzept einer absolut konvergenten Reihe. Dabei heißt eine Reihe $\sum_{i=1}^{\infty} a_i$ mit reellen Summanden $a_1, a_2, \ldots$ *absolut konvergent*, wenn $\sum_{i=1}^{\infty} |a_i| < \infty$, das heißt, wenn die Zahlenmenge $\{\sum_{1\le i\le n} |a_i| \mid n \ge 1\}$ eine obere Schranke in $\mathbb{R}$ hat. Zunächst ist dann $a := \sum_{i=1}^{\infty} a_i$ definiert, aber zusätzlich gilt noch, dass eine Umordnung der Summationsreihenfolge den Wert der Reihe nicht ändert, das heißt: Wenn $\pi\colon \mathbb{N} \to \mathbb{N}$ eine Bijektion ist, dann gilt auch $\sum_{i=1}^{\infty} a_{\pi(i)} = a$.

Beispiel: Die Reihe $\sum_{i=0}^{\infty} x^i$ ist absolut konvergent für jedes x mit $|x| < 1$. (Das sieht man am besten mit dem Quotientenkriterium ein.) Hingegen ist die Reihe $\sum_{i=1}^{\infty}(-1)^i/i$ zwar konvergent, aber nicht absolut konvergent (denn $\sum_{i=1}^{\infty} 1/i$ ist divergent).

In der Wahrscheinlichkeitsrechnung (mit diskreten W-Räumen) sind Diskussionen über Summationsreihenfolgen sehr unbequem. Man vermeidet dies mit einem geeigneten Konzept.[10] Im Folgenden betrachten wir Indexmengen I ohne bestimmte Reihenfolge. Wenn jedem $i \in I$ eine (reelle) Zahl a_i zugeordnet ist, möchten wir nach der „Summe“ $\sum_{i\in I} a_i$ der „Familie“ $(a_i)_{i\in I}$ fragen. Was soll das sein? Wenn I endlich ist, gibt es kein Problem: Beim Addieren der endlich vielen Zahlen in beliebiger Reihenfolge kommt immer dasselbe Ergebnis heraus, wegen der Kommutativität der Addition, und mit $\sum_{i\in I} a_i$ bezeichnen wir die Summe. Aber was ist, wenn I unendlich ist?

Definition A.2.1
Die Familie $(a_i)_{i\in I}$ heißt *summierbar*, wenn $\left\{\sum_{i\in J} |a_i| \;\middle|\; J \subseteq I \text{ endlich}\right\}$ nach oben beschränkt ist.

Lemma A.2.2
$(a_i)_{i\in I}$ ist summierbar $\Leftrightarrow$ es gibt eine eindeutig bestimmte Zahl a mit folgender Eigenschaft:

$$\forall \varepsilon > 0\ \exists J \subseteq I \text{ endlich}\ \ \forall K\colon J \subseteq K \subseteq I, K \text{ endlich}\ \Rightarrow \left|\sum_{i\in K} a_i - a\right| \le \varepsilon. \tag{A.2.21}$$

[10] In allgemeinen, nicht diskreten Wahrscheinlichkeitsräumen verschwinden die Summen, und ihre Rolle wird von Integralen übernommen.

Im Fall der Summierbarkeit heißt die Zahl a aus dem Lemma die *Summe* von $(a_i)_{i\in I}$, geschrieben $\sum_{i\in I} a_i$.

Beweis. „⇒“: Die Eindeutigkeit ist nicht schwer einzusehen. Wir zeigen nur die Existenz. Setze $I^+ := \{i \in I \mid a_i > 0\}$ und $I^- := \{i \in I \mid a_i < 0\}$. Weil $(a_i)_{i\in I}$ summierbar ist, sind $\left\{\sum_{i\in J} a_i \;\middle|\; J \subseteq I^+ \text{ endlich}\right\}$ und $\left\{\sum_{i\in J}(-a_i) \;\middle|\; J \subseteq I^- \text{ endlich}\right\}$ beide nach oben beschränkt. (Man lässt $J = \emptyset$ zu, daher enthalten die beiden Mengen mindestens die Zahl 0.) Es sei s^+ das Supremum der ersten Menge (zu I^+) und s^- das Supremum der zweiten Menge. Wir definieren $a := s^+ - s^-$ und zeigen, dass dieses a die Aussage des Lemmas erfüllt. Sei $\varepsilon > 0$ gegeben. Wähle eine endliche Menge $J^+ \subseteq I^+$ mit $\sum_{i\in J^+} a_i \geq s^+ - \frac{1}{2}\varepsilon$, und eine endliche Menge $J^- \subseteq I^-$ mit $\sum_{i\in J^-}(-a_i) \geq s^- - \frac{1}{2}\varepsilon$. Definiere $J := J^+ \cup J^-$. Wenn K eine endliche Menge mit $J \subseteq K \subseteq I$ ist, dann kann sich $\sum_{i\in K} a_i = \sum_{i\in K\cap I^+} a_i - \sum_{i\in K\cap I^-}(-a_i)$ um nicht mehr als $2 \cdot \frac{1}{2}\varepsilon = \varepsilon$ von a unterscheiden, weil die erste Summe zwischen $s^+ - \frac{1}{2}\varepsilon$ und s^+ und die zweite Summe zwischen $s^- - \frac{1}{2}\varepsilon$ und s^- liegt.

„⇐“: Wenn die Menge $\{\sum_{i\in J} |a_i| \mid J \subseteq I \text{ endlich}\}$ nicht nach oben beschränkt ist, dann ist, mit I^+ und I^- wie im Teil „⇒“, die Menge $\{\sum_{i\in K} a_i \mid K \subseteq I^+ \text{ endlich}\}$ oder $\{\sum_{i\in K}(-a_i) \mid K \subseteq I^- \text{ endlich}\}$ nicht nach oben beschränkt. Wir nehmen z. B. den ersten Fall an. Sei nun $J \subseteq I$ endlich. Weil $\{\sum_{i\in K} a_i \mid K \subseteq I^+ \text{ endlich}\}$ nicht nach oben beschränkt ist, gibt es endliche Teilmengen $K \subseteq I^+$ mit $\sum_{i\in K} a_i$ beliebig groß, also ist $\{\sum_{i\in J\cup K} a_i \mid K \subseteq I^+ \text{ endlich}\}$ nicht nach oben beschränkt. Daher kann (A.2.21) für kein a erfüllt sein. □

Wir werden die Theorie der summierbaren Familien hier nicht weiter ausarbeiten, sondern belassen es dabei, einige Rechenregeln zu notieren, ohne Beweis. Diese besagen, dass sich solche Summen durch die „eingebaute“ absolute Konvergenz gegenüber Operationen sehr gutmütig verhalten, und man mit ihnen im Wesentlichen wie mit endlichen Summen rechnen kann.

Proposition A.2.3

(a) **(Majorisierung und Auswahl)** Wenn $(a_i)_{i\in I}$ summierbar ist und $(b_i)_{i\in I}$ erfüllt $|b_i| \leq |a_i|$ für jedes i, dann ist auch $(b_i)_{i\in I}$ summierbar. Insbesondere: Wenn $(a_i)_{i\in I}$ summierbar und $J \subseteq I$ beliebig ist, dann ist auch die Teilfamilie $(a_i)_{i\in J}$ summierbar.

(b) **(Monotonie)** Wenn $(a_i)_{i\in I}$ und $(b_i)_{i\in I}$ summierbar sind und es gilt $a_i \leq b_i$ für jedes i, dann gilt

$$\sum_{i\in I} a_i \leq \sum_{i\in I} b_i.$$

(c) **(Linearität)** Wenn $(a_i)_{i\in I}$ und $(b_i)_{i\in I}$ summierbar sind, und $\alpha, \beta \in \mathbb{R}$, dann ist auch $(\alpha a_i + \beta b_i)_{i\in I}$ summierbar, und es gilt

$$\sum_{i\in I}(\alpha a_i + \beta b_i) = \alpha\Big(\sum_{i\in I} a_i\Big) + \beta\Big(\sum_{i\in I} b_i\Big).$$

(d) **(Distributivität)** Wenn $(a_i)_{i\in I}$ und $(b_j)_{j\in J}$ summierbar sind, dann ist auch $(a_i \cdot b_j)_{(i,j)\in I\times J}$ summierbar, und es gilt

$$\sum_{(i,j)\in I\times J} (a_i \cdot b_j) = \Big(\sum_{i\in I} a_i\Big) \cdot \Big(\sum_{j\in J} b_j\Big).$$

(e) **(Assoziativität)** Wenn $(a_i)_{i\in I}$ summierbar ist und $(I_k)_{k\in K}$ eine beliebige disjunkte Zerlegung von I ist (d. h. $I_k \cap I_{k'} = \emptyset$ für $k \neq k'$ und $\bigcup_{k\in K} I_k = I$), dann gilt

$$\sum_{i\in I} a_i = \sum_{k\in K}\Big(\sum_{i\in I_k} a_i\Big).$$

(Insbesondere sind alle vorkommenden Summen definiert.)

In unserer Diskussion haben wir nie gesagt, dass die Indexmenge I abzählbar unendlich sein muss. Es könnte also zum Beispiel auch $I = \mathbb{R}$ sein. Die folgende Tatsache zeigt aber, dass diese Verallgemeinerung gegenüber absolut konvergenten Reihen mit Indexmenge $\mathbb{N}$ nur scheinbar ist.

Lemma A.2.4
Wenn $(a_i)_{i\in I}$ summierbar ist, dann ist die Menge $J_{\neq 0} = \{i \in I \mid a_i \neq 0\}$ endlich oder abzählbar unendlich.

Beweis. Für $n \geq 1$ setze $I_n := \{i \in I \mid |a_i| \geq \frac{1}{n}\}$. Weil $(a_i)_{i\in I}$ summierbar ist, ist jede der Mengen I_n endlich, und daher ist $J_{\neq 0}$ als Teilmenge von $\bigcup_{n\geq 1} I_n$ entweder selbst endlich oder abzählbar unendlich. □