Technische Universität Ilmenau
Fakultät für Informatik und Automatisierung
Univ.-Prof. Dr. Martin Dietzfelbinger
09.04.2018

2 Grundlagen aus der Wahrscheinlichkeitsrechnung

In diesem Abschnitt 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 im Buch „Probability and Computing – Randomized Algorithms and Probabilistic Analysis“ von M. Mitzenmacher und E. Upfal.

2.1 Grundbegriffe, Beispiele

Definition 2.1.1 *Ein* ***Wahrscheinlichkeitsraum*** *(****W-Raum****) ist ein Paar (Ω, p), wo Ω eine endliche oder abzählbar unendliche Menge und $p\colon \Omega \to [0,1]$ eine Funktion ist, mit $\sum_{\omega\in\Omega} p(\omega) = 1$. Wir schreiben oft p_ω statt $p(\omega)$. Eine solche Funktion $p\colon \Omega \to [0,1]$ heißt auch „****Verteilung****“ oder „****Wahrscheinlichkeitsverteilung****“.*

Bemerkung 2.1.2 Man weiß, dass in der Situation der Definition für jedes $A \subseteq \Omega$ die Reihe $\sum_{\omega\in A} p_\omega$ absolut (d. h. ohne Rücksicht auf die Summationsreihenfolge) konvergiert, also einen wohldefinierten Wert hat.

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_ω gegeben. Man teste diese intuitive Auffassung an den folgenden Beispielen.

Beispiele 2.1.3 (a) Zur Modellierung des Zufallsexperiments, einen fairen Würfel einmal zu werfen, benutzt man den Wahrscheinlichkeitsraum (Ω, p) mit $\Omega =$

$\{1,\ldots,6\}$ und $p(\omega)=\frac{1}{6}$ für jedes $\omega\in\{1,\ldots,6\}$.
Bei einer fairen Münze 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 Resultat zu nehmen, wird man etwa $\Omega=\{2,\ldots,12\}$ und $p(2)=\frac{1}{36}$, $p(3)=\frac{2}{36}$, …, $p(7)=\frac{6}{36}$, …, $p(12)=\frac{1}{36}$ 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 $\omega\in\Omega$. Man spricht von der „*uniformen Verteilung*“ 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.

(d) Wir wollen wiederholt mit einem Würfel würfeln und warten, bis die erste „6“ erscheint. Dazu setzen wir $\Omega=\{1,2,3,\ldots\}$ 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. Man sieht, mit der Summenformel für geometrische Reihen:

$$\textstyle\sum_{i\geq1}p_i=\frac{1}{6}\cdot\sum_{i\geq1}(\frac{5}{6})^{i-1}=\frac{1}{6}\cdot\frac{1}{1-\frac{5}{6}}=1.$$

Damit haben wir tatsächlich einen Wahrscheinlichkeitsraum definiert.

(e) Es sei $U\neq\emptyset$ eine endliche Menge und $n\geq1$. Der W-Raum (Ω,p) mit

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

und $p_\omega=\frac{1}{|U|^n}$, $\omega\in\Omega$, 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 $U\neq\emptyset$ endlich, $1\leq n\leq|U|$. Die Menge $\Omega=\{A\subseteq U\mid|A|=n\}$ mit der Verteilung, die durch $p_\omega=\binom{|U|}{n}^{-1}$ für alle $\omega\in\Omega$ gegeben ist, definiert einen W-Raum, der das Zufallsexperiment „Wähle eine zufällige n-elementige Teilmenge von U“ modelliert.

(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 Raum entspricht dem Experiment, für n beliebig vorgegebene Elemente von U die Anordnung zufällig zu wählen.

(h) Beim Hashing betrachtet man n Schlüssel $x_1, \ldots, x_n$ und n Funktionswerte $h(x_1), \ldots, h(x_n) \in [m] := \{0, 1, \ldots, m-1\}$. Man macht 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 $\{0, 1, \ldots, m-1\}$ mit derselben Wahrscheinlichkeit annimmt. Der zugehörige Wahrscheinlichkeitsraum ist

$$\Omega = [m]^n = \{0, 1, \ldots, m-1\}^n = \{(v_1, \ldots, v_n) \mid v_1, \ldots, v_n \in \{0, 1, \ldots, m-1\}\}$$

mit der durch

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

definierten Verteilung.
(Dies ist derselbe Wahrscheinlichkeitsraum wie der in (e), wenn man $U = \{0, 1, \ldots, m-1\}$ setzt.)

Definition 2.1.4 *Ein **Ereignis** ist eine Menge $A \subseteq \Omega$.*
*Die **Wahrscheinlichkeit** (**W.**) von A ist* $\mathbf{Pr}(A) := \sum_{\omega \in A} p_\omega$.

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

In den folgenden Beispielen sieht man, dass der intuitive Name „Ereignis“ und die abkürzende Schreibweise für durch Aussagen gegebene Ereignisse und ihre Wahrscheinlichkeiten recht gut passt. Man beachte, dass in der Notation der W-Raum immer unterdrückt wird.

Beispiel 2.1.5 (a) In Beispiel 2.1.3 (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 $\mathbf{Pr}(A) = \frac{26}{36}$.

(b) In Beispiel 2.1.3 (h) ist

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

ein Ereignis, das man auch $\{h(x_1) = h(x_2) = h(x_3)\}$ 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 ***Gleichverteilung*** (oder ***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) $\mathbf{Pr}(\emptyset) = 0$, $\mathbf{Pr}(\Omega) = 1$, $\mathbf{Pr}(\{\omega\}) = p_\omega$, $\mathbf{Pr}(\Omega - A) = 1 - \mathbf{Pr}(A)$.

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

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

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

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

(***Vereinigungs-Schranke*** oder englisch ***union bound***.)

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

Formel 2.1.6(d) wird oft folgendermaßen benutzt: Wenn aus der Aussage $\varphi(\omega)$ die Aussage $\psi(\omega)$ *folgt*, dann gilt $\mathbf{Pr}(\varphi) \le \mathbf{Pr}(\psi)$.

Die Aussagen von Fakt 2.1.6 sind leicht mittels Def. 2.1.4 nachzukontrollieren.

Beispiel 2.1.7 In Beispiel 2.1.3 (h) ist für jedes $v \in [m] = \{0, 1, \ldots, m-1\}$

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

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

2.2 Zufallsvariablen und Erwartungswerte

Definition 2.2.1 *Ist R eine Menge, so heißt eine Funktion $X\colon \Omega \to R$ 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 $X(\omega)$. Beispiel: Ist $R' \subseteq R$, betrachtet man das Ereignis $\{X \in R'\} = \{\omega \mid X(\omega) \in R'\}$, und die Wahrscheinlichkeit $\mathbf{Pr}(X \in R')$, usw.

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 obige Fallunterscheidung angeben zu müssen) hat sich die **Iverson-Notation** bewährt: Statt X 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 wieder Beispiel 2.1.3 (h). ($\omega = (v_1, \ldots, v_n)$.)

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

(b) Für $0 \leq v < m$ 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 Zufallsvariable X induziert einen neuen Wahrscheinlichkeitsraum, wie folgt:

$$\Omega' := X[\Omega] = \{X(\omega) \mid \omega \in \Omega\};\ p'(\alpha) := \mathbf{Pr}(X = \alpha) \text{ für } \alpha \in \Omega'. \qquad (1)$$

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

Bemerkung 2.2.4 Für jeden Wahrscheinlichkeitsraum (Ω, p) ist p Verteilung einer passenden Zufallsvariablen. 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 Würfeln ist folgender Wahrscheinlichkeitsraum sehr natürlich:

$$\Omega = \{1, \ldots, 6\}^2;\ p((i,j)) = \frac{1}{36} \text{ für } (i,j) \in \Omega.$$

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 2.1.3(b) beschriebenen Wahrscheinlichkeitsraums.

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

$$\begin{aligned} \Omega &= \{(a_1, \ldots, a_i) \mid i \geq 1, a_1 \ldots, a_{i-1} \in \{1, \ldots, 5\}, a_i = 6\}; \\ p((a_1, \ldots, a_i)) &= \frac{1}{6^i}. \end{aligned}$$

Ein Elementarereignis ist hier eine Folge von Würfen mit ihren Ergebnissen, 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 2.1.3(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

$$\begin{aligned} p'(i) &= \frac{|\{(v_1, \ldots, v_n) \in \Omega \mid (v_1, \ldots, v_n) \text{ enthält genau } i \text{ Nullen}\}|}{m^n}. && (2) \\ &= \binom{n}{i} \cdot \frac{(m-1)^{n-i}}{m^n} = \binom{n}{i} \cdot \left(\frac{1}{m}\right)^i \cdot \left(1 - \frac{1}{m}\right)^{n-i}. && (3) \end{aligned}$$

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

Definition 2.2.7 *Der* ***Erwartungswert*** *einer ZV $X \geq 0$ ist*

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

Wenn X auch negative Werte annehmen kann, betrachten wir den Erwartungswert $\mathbf{E}(X)$ von X nur dann, wenn $\sum_{\omega\in\Omega} |X(\omega)| \cdot p_\omega < \infty$. In diesem Fall ist der Wert der Summe $\sum_{\omega\in\Omega} X(\omega) \cdot p_\omega$ von der Summationsreihenfolge unabhängig.

Die zweite Darstellung des Erwartungswertes in Definition 2.2.7 lässt sich leicht durch Umstellen von Summen bzw. Reihen beweisen, was hier kein Problem ist, weil alle Reihen absolut konvergieren. Man kann die zweite Darstellung auch so auffassen: Man betrachtet die Verteilung von X, die jeder Zahl $\alpha \in X[\Omega]$ eine Wahrscheinlichkeit $p'(\alpha)$ zuordnet, und bildet den Mittelwert dieser Zahlen, gewichtet mit diesen Wahrscheinlichkeiten.

Fakt 2.2.8 Ist $X\colon \Omega \to \mathbb{N}$ eine Zufallsvariable, so gilt:

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

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 2.1.3(d) (Würfeln, bis die erste „6“ erscheint) definieren wir die Zufallsvariable $X :=$ Anzahl der Würfe bis zur ersten 6. Technisch ist das in diesem Wahrscheinlichkeitsraum einfach $X(i) := i$, für $i \geq 1$. Nach dem Fakt von eben wissen wir:

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

Wir müssen also nur $\mathbf{Pr}(X \geq i)$ bestimmen. Intuitiv sieht man, dass dies $(\frac{5}{6})^{i-1}$ sein muss (Misserfolg in den ersten $i-1$ Würfen). Man kann dies auch aus der Definition der Werte $p_j = (\frac{5}{6})^{j-1} \cdot \frac{1}{6}$ ausrechnen. Damit ist dann

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

Fakt 2.2.9 Für beliebige Zufallsvariable $X, Y, X_1, \ldots, X_n$ 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) $\mathbf{E}(\alpha X + \beta Y) = \alpha \mathbf{E}(X) + \beta \mathbf{E}(Y)$.

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

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

Die *Beweise* dieser Rechenregeln sind einfache Übungsaufgaben.

Bemerkung 2.2.10 Für eine Indikatorvariable $[\varphi]$ gilt

$$\mathbf{E}([\varphi]) = 0 \cdot \mathbf{Pr}([\varphi] = 0) + 1 \cdot \mathbf{Pr}([\varphi] = 1) = \mathbf{Pr}(\varphi) \quad (= \mathbf{Pr}(\varphi \text{ wahr})).$$

Beispiel 2.2.11 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: $|B_v| = [h_1 = v] + \cdots + [h_n = v]$. Also gilt

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

2.3 Varianz und Ungleichungen von Markov, Chebychev und Jensen

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

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

Beweis. Offenbar gilt $Z \geq t \cdot [Z \geq t]$, also auch $\mathbf{E}(Z) \geq t \cdot \mathbf{E}([Z \geq t]) = t\mathbf{Pr}(Z \geq t)$. Dividieren durch t liefert die Behauptung. ∎

Definition 2.3.2 *Für eine beliebige Zufallsvariable X mit $\mathbf{E}(X^2) < \infty$ definieren wir die **Varianz** von X als*

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

Bemerkung: 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 $\mathbf{Var}(a \cdot X) = a^2\mathbf{Var}(X)$, für jede Konstante a.

Man sieht sofort, dass gilt:

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

Folgerung: Da $\mathbf{Var}(X)$ Erwartungswert von $(X - \mathbf{E}(X))^2 \geq 0$ ist, ist $\mathbf{Var}(X) \geq 0$.

Daraus folgt

$$\boxed{\mathbf{E}(X)^2 \leq \mathbf{E}(X^2)} \tag{4}$$

für jede Zufallsvariable X, deren Varianz existiert.

Wenn wir auf die Zufallsvariable $Z = (X - \mathbf{E}(X))^2 \geq 0$ die Markov-Ungleichung anwenden, erhalten wir:

Fakt 2.3.3 (Chebychev/Tschebyscheff-Ungleichung) Es sei X eine Zufallsvariable mit $\mathbf{E}(X^2) < \infty$. Dann gilt für jedes $t > 0$:

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

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 2.3.4 *X sei eine beliebige Zufallsvariable, $D \subseteq \mathbb{R}$, $f\colon D \to \mathbb{R}^+$ sei monoton mit $D = Def(f) \supseteq X(\Omega)$, 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: Sei X eine Zufallsvariable, für die $\mathbf{Var}(X)$ existiert.

- Sei $\alpha > 0$ beliebig. Dann gilt

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

- Sei $k \geq 2$ eine gerade ganze Zahl und $t \geq 0$. Dann gilt:

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

  (Hier wird 2.3.4 auf die Zufallsvariable $Z = |X - \mathbf{E}(X)|$ und $f(x) = x^k$ angewendet.)

- Seien $t, c > 0$ beliebig. Dann gilt

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

  Diese Ungleichung kann man für den Beweis einer Variante der Chebychev-Ungleichung benutzen (siehe Übung).

- Sei X reellwertig, sei $a > 0$, und sei $\mathbf{E}(e^{aX})$ definiert. Dann gilt

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

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

Wir haben oben gesehen, dass stets $\mathbf{E}(X)^2 \leq \mathbf{E}(X^2)$ gilt. Anstelle der Funktion $x \mapsto x^2$ kann man jede beliebige *konvexe* Funktion benutzen.

Definition 2.3.5 $D \subseteq \mathbb{R}$ sei ein Intervall. Eine Funktion $f\colon D \to \mathbb{R}$ heißt **konvex**, wenn gilt*:*

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

Sie heißt **konkav**, wenn $-f$ konvex ist.

Grob gesprochen ist eine Funktion konvex, wenn an jeder Stelle der Graph der Funktion unter jeder Sekante dieses Funktionsgraphen 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 positiv ist.

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

(iii) Wenn $\alpha \in \mathbb{R}$, $0 < \alpha < 1$, 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: $g''_\alpha(x) = -\alpha(-(\alpha+1))/x^{\alpha+2}$. 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 $h''(x) = x^{-1} > 0$.)

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

Proposition 2.3.6 (Jensensche Ungleichung, allgemeine Form) *Es sei X eine reellwertige Zufallsvariable und f eine Funktion mit $D = Def(f) \supseteq X(\Omega)$. 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 $\alpha > 1$ und $X \geq 0$ gilt: $\mathbf{E}(X)^{\alpha} \leq \mathbf{E}(X^{\alpha})$.

(iii) Für $0 < \alpha < 1$ 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 der Jensenschen Ungleichung: Wir beweisen nur (a). ((b) folgt durch Multiplikation der Ungleichung mit -1.) Setze $x_0 := \mathbf{E}(X)$. Dann ist $x_0 \in \mathrm{Def}(f)$. 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:

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

Nach der Wahl von x_0 folgt die behauptete Ungleichung. □

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

Proposition 2.3.7 (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. 2.3.6(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, bedingte Erwartungswerte

Definition 2.4.1 *Ist $A \subseteq \Omega$ ein Ereignis mit $\mathbf{Pr}(A) > 0$, setzen wir*

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

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

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

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

Fakt 2.4.2 Basisformel für bedingte Wahrscheinlichkeiten:

$$\mathbf{Pr}(A \cap B) = \mathbf{Pr}(A)\mathbf{Pr}(B \mid A).$$

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

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

2.5 Unabhängigkeit bei Ereignissen und 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 $I, J \subseteq \{1, \ldots, n\}$, $I \cap J = \emptyset$.

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

In vielen Büchern findet man auch eine (auf den ersten Blick) andere Form von Definition 2.5.1(b): Man spricht von Unabhängigkeit, falls

$$\mathbf{Pr}\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} \mathbf{Pr}\left(A_i\right) \tag{5}$$

für beliebige Teilmengen I von $\{1, 2, \ldots, n\}$ gilt. Diese Definition und Definition 2.5.1(b) sind jedoch äquivalent. Unsere Definition 2.5.1(b) bietet sogar einen Vorteil, denn man kann sofort Aussagen machen, bei denen „Gegenereignisse“ $\bar{A}_i$ vorkommen.

Beispiel 2.5.2 (a) In Bsp. 2.1.3 (h) sind die Ereignisse $\{v_1 = v_1^0\}, \ldots, \{v_n = v_n^0\}$ unabhängig, für beliebige $v_1^0, \ldots, v_n^0 \in \{0, \ldots, m-1\}$ fest.

(b) In Bsp. 2.1.3 (h) sind die Ereignisse $\{v_1 \neq 0\}, \ldots, \{v_n \neq 0\}$ unabhängig.

Definition 2.5.3 ***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* $\{X_1 \in R_1'\}, \ldots, \{X_n \in R_n'\}$ *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 $R_i' \subseteq R_i$*.)*

Fakt 2.5.4 Sind $X_1, \ldots, X_n$ unabhängig und sind $g_i\colon R_i \to S_i$ beliebig, $1 \le i \le n$, dann sind die Zufallsfunktionen $g_1 \circ X_1, \ldots, g_n \circ X_n$ unabhängig.

Beispiel 2.5.5 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 $p(\omega_1, \ldots, \omega_n) = p^1(\omega_1) \cdot \ldots \cdot p^n(\omega_n)$, für $\omega = (\omega_1, \ldots, \omega_n) \in \Omega$, ein neuer W-Raum (der „Produktraum“) definiert. In Ω sind die n Projektionsfunktionen $X_i\colon \omega = (\omega_1, \ldots, \omega_n) \mapsto \omega_i \in \Omega_i$ unabhängig; nach Fakt 2.5.4 ist also jede Folge $Y_1, \ldots, Y_n$ 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 2.1.3(h) ist ein Produktraum.)

Fakt 2.5.6 „Bei Unabhängigkeit multiplizieren sich Erwartungswerte, Varianzen addieren sich.“ [a]

(a) Sind $X_1, \ldots, X_n$ *unabhängige* Zufallsvariable, so gilt $\mathbf{E}(X_1 \cdot \ldots \cdot X_n) = \prod_{1 \le i \le n} \mathbf{E}(X_i)$.

(b) Sind $X_1, \ldots, X_n$ *unabhängige* Zufallsvariable, so gilt $\mathbf{Var}(X_1 + \cdots + X_n) = \sum_{1 \le i \le n} \mathbf{Var}(X_i)$. Dies gilt sogar, wenn nur X_i und X_j unabhängig sind für $i \ne j$ (*paarweise Unabhängigkeit*).

[a] Additivität von Erwartungswerten gilt immer, siehe Fakt 2.2.9(c).

Beweis. (a) Wir beweisen die Aussage für zwei Zufallsvariable X und Y. Die

Verallgemeinerung auf n Zufallsvariable ergibt sich durch vollständige Induktion.

$$\begin{aligned}
\mathbf{E}(X \cdot Y) &= \sum_{\omega \in \Omega} p_\omega X(\omega) Y(\omega) \\
&= \sum_{\alpha \in X[\Omega]} \sum_{\beta \in Y[\Omega]} \alpha\beta \cdot \mathbf{Pr}(X = \alpha \wedge Y = \beta) \\
&= \sum_{\alpha \in X[\Omega]} \sum_{\beta \in Y[\Omega]} \alpha\beta \cdot \mathbf{Pr}(X = \alpha) \cdot \mathbf{Pr}(Y = \beta) \\
&= \Big(\sum_{\alpha \in X[\Omega]} \alpha \mathbf{Pr}(X = \alpha) \Big) \Big(\sum_{\beta \in Y[\Omega]} \beta \mathbf{Pr}(Y = \beta) \Big) \\
&= \mathbf{E}(X)\mathbf{E}(Y).
\end{aligned}$$

(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:

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

■

Bemerkung: Für den Beweis von Teil (b) haben wir nicht die volle Unabhängigkeit eingesetzt, sondern nur die ***paarweise Unabhängigkeit*** von $X_1, \ldots, X_n$. Sie lieferte die Gleichheit $\mathbf{E}(X_i X_j) = \mathbf{E}(X_i)\mathbf{E}(X_j)$.

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.)

Anwendung: 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“ (entpricht „0“) auftritt. Wir nehmen die relative Häufigkeit von „Kopf“ als Schätzwert für p. Das schwache Gesetz der Großen Zahl besagt, dass dieser Wert nur mit kleiner Wahrscheinlichkeit weit von p entfernt ist – umso kleinerer Wahrscheinlichkeit, je größer die Anzahl der Würfe ist. Mathematisch sieht das so aus: X_i ist eine 0-1-wertige Zufallsvariable mit $\mathbf{Pr}(X_i = 1) = p$, für $i = 1, \ldots, n$, und diese Zufallsvariablen sind unabhängig. Der Schätzwert ist die Zufallsvariable

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

Dann gilt für jedes $\varepsilon > 0$: $\lim_{n\to\infty} \mathbf{Pr}(|Y_n - p| \geq \varepsilon) = 0$. Genauer gilt:

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

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 2.5.6(b) folgt $\mathbf{Var}(Y_n) = (1/n^2)\mathbf{Var}(X_1 + \cdots + X_n) = (1/n^2) \cdot npq = pq/n \leq 1/(4n)$. 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 dieses schwache Gesetz der großen Zahlen schon bei nur paarweiser Unabhängigkeit gilt (und in anderen Situationen mit schwächeren Voraussetzungen).

2.6 Die Hoeffding-Ungleichung

Die Hoeffding-Ungleichung oder Chernoff-Hoeffding-Ungleichung lässt sich ebenfalls auf Bernoulli-Experimente anwenden. Sie liefert jedoch ungleich schärfere Abschätzungen für die Wahrscheinlichkeit, dass $(X_1 + \cdots + X_n)/n$ weit von seinem Erwartungswert abweicht.

Satz 2.6.1 (Hoeffding) *$X_1, \ldots, X_n$ 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:

$$\mathbf{Pr}(X \geq m+a) \leq \left(\frac{m}{m+a}\right)^{m+a} \left(\frac{n-m}{n-(m+a)}\right)^{n-(m+a)}, \text{ für } 0{\leq}a{\leq}n{-}m; \quad (6)$$

$$\mathbf{Pr}(X \leq m-b) \leq \left(\frac{m}{m-b}\right)^{m-b} \left(\frac{n-m}{n-(m-b)}\right)^{n-(m-b)}, \text{ für } 0 \leq b \leq m. \quad (7)$$

Bevor wir diese Ungleichungen beweisen, wollen wir sie ein wenig diskutieren.[1] Man beachte zuerst, dass das soeben diskutierte Bernoulli-Experiment mit n Münzwürfen direkt zur Situation des Satzes führt. Die Zufallsvariable X zählt, wie oft „Kopf“ aufgetreten ist.

Die Hoeffding-Ungleichung gehört zu der Familie der „tail inequalities“, das sind Ungleichungen, die Schranken dafür liefern, dass Zufallsvariable Werte weit weg von ihrem Erwartungwert annehmen. Wir werden weiter unten sehen, dass die Hoeffding-Schranke relativ kräftig ist, wenn m nicht zu klein ist: Summen von *vielen* (auf $[0,1]$) *beschränkten unabhängigen* Zufallsvariablen sind eng um ihren Erwartungwert konzentriert. Man beachte, dass über die einzelnen X_i nichts weiter angenommen wird, als dass sie in $[0,1]$ eingeschlossen sind. Insbesondere können sie auch ganz unterschiedliche Verteilungen haben.

Korollar 2.6.2 *In der Situation von Satz 1 gilt:*

$$\mathbf{Pr}(X \geq m+a) \leq \left(\frac{m}{m+a}\right)^{m+a} e^a, \text{ für } 0 \leq a \leq n-m; \quad (8)$$

$$\mathbf{Pr}(X \leq m-b) \leq \left(\frac{m}{m-b}\right)^{m-b} e^{-b}, \text{ für } 0 \leq b \leq m. \quad (9)$$

Der Beweis von Korollar 2.6.2 ist sehr einfach, wenn man sich die folgende auch

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

sonst nützliche Ungleichung in Erinnerung ruft (siehe Prop. A.0.2(b) im Anhang):

$$\left(1+\frac{x}{y}\right)^y < e^x \text{ für } y>0 \text{ und } x \geq -y. \tag{10}$$

Wenn man (10) für den zweiten Faktor

$$\left(\frac{n-m}{n-(m+a)}\right)^{n-(m+a)} = \left(1+\frac{a}{n-(m+a)}\right)^{n-(m+a)}$$

in (6) einsetzt, ergibt sich (8) für $0 \leq a < n-m$. Der Fall $a = n-m$ ist noch einfacher. (Ungleichung (10) liefert eine Schranke von $(m/n)^n$, weniger als $(m/n)^n e^{n-m}$.)

An dieser Stelle beobachten wir, dass (wieder mit (10)) der Faktor $(\frac{m}{m+a})^{m+a} = (1-\frac{a}{m+a})^{m+a}$ in (8) immer kleiner als e^{-a} ist. Daher ist die rechte Seite in (8) kleiner als 1 und stellt damit eine echte Schranke für eine Wahrscheinlichkeit dar.

Ganz analog ergibt sich (9) aus (7) mit Hilfe von (10). □

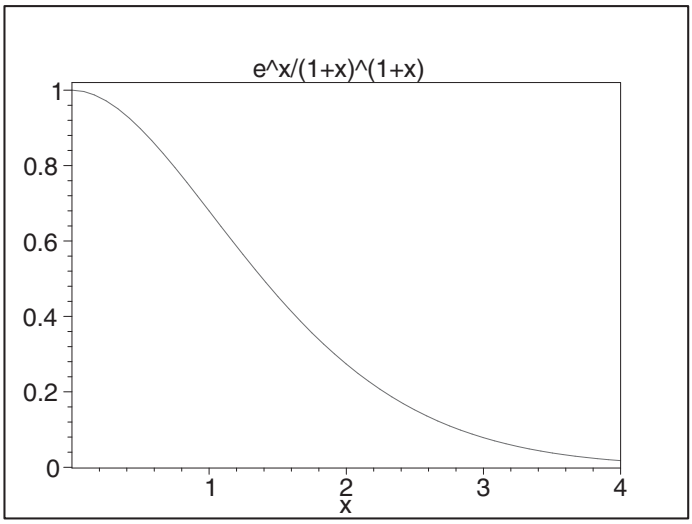
Die folgende Version ergibt sich aus Korollar 2.6.2 einfach dadurch, dass man $\varepsilon = \frac{a}{m}$ bzw. $\varepsilon = \frac{b}{m}$ setzt.

Korollar 2.6.3 *In der Situation des Satzes gilt:*

$$\mathbf{Pr}(X \geq (1+\varepsilon)m) \leq \left(\frac{e^{\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}}\right)^m, \text{ für } 0 \leq \varepsilon \leq \tfrac{n}{m}-1; \tag{11}$$

$$\mathbf{Pr}(X \leq (1-\varepsilon)m) \leq \left(\frac{e^{-\varepsilon}}{(1-\varepsilon)^{1-\varepsilon}}\right)^m, \text{ für } 0 \leq \varepsilon \leq 1. \tag{12}$$

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. Um einen Eindruck von dem Verlauf der Funktion $\varepsilon \mapsto \frac{e^{\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}}$ zu bekommen, betrachte man die folgende Skizze:



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

Korollar 2.6.4 *In der Situation von Korollar 2.6.3 gilt*:

$$\mathbf{Pr}(X \geq (1+\varepsilon)m) \leq e^{-\varepsilon^2 m/3}, \text{ für } 0 \leq \varepsilon \leq 1.8; \tag{13}$$

$$\mathbf{Pr}(X \geq (1+\varepsilon)m) \leq e^{-\varepsilon^2 m/4}, \text{ für } 0 \leq \varepsilon \leq 4.1; \tag{14}$$

$$\mathbf{Pr}(X \leq (1-\varepsilon)m) \leq e^{-\varepsilon^2 m/2}, \text{ für } 0 \leq \varepsilon \leq 1. \tag{15}$$

Der *Beweis* von (13) und (14) 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 (s. Abb. 1 und 2).

Damit folgt die Behauptung direkt aus (11). Die dritte Ungleichung (14) folgt ähnlich aus der Beobachtung, dass die Funktion $\varepsilon \mapsto \ln\big(e^{-\varepsilon^2/2}/(e^{-\varepsilon}/(1-\varepsilon)^{(1-\varepsilon)})\big) = -\varepsilon^2/2 + \varepsilon + (1-\varepsilon)\ln(1-\varepsilon))$ im Intervall $[0, 1]$ nicht negativ ist, und mit (12) (s. Abb. 3).

Nun kommen wir endlich zum ***Beweis*** von Formel (6) aus Satz 2.6.1.

Der Fall $a = 0$ ist trivial, weil auf der linken Seite von (6) eine Wahrscheinlichkeit

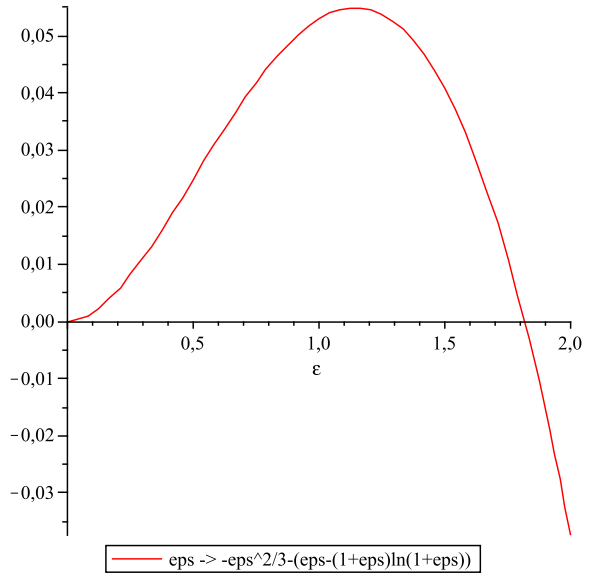


Abbildung 1: Funktion $\varepsilon \mapsto -\varepsilon^2/3-(\varepsilon-(1+\varepsilon)\ln(1+\varepsilon))$: in $[1, 1.8]$ nicht negativ.

steht, auf der rechten Seite 1. Es sei also $0 < a \leq n - m$ beliebig, aber fest. Für jedes beliebige $t > 0$ erhält man durch Anwenden der Markov-Ungleichung auf die Zufallsvariable e^{tX} Folgendes:[2]

$$\begin{aligned} \mathbf{Pr}(X \geq m + a) &= \mathbf{Pr}(e^{tX} \geq e^{t(m+a)}) \\ &\leq \frac{\mathbf{E}(e^{tX})}{e^{t(m+a)}} \\ &= e^{-t(m+a)} \cdot \mathbf{E}\Big(\prod_{1\leq i\leq n} e^{tX_i}\Big). \end{aligned}$$

Weil $X_1, \ldots, X_n$ unabhängig sind, sind auch $e^{tX_1}, \ldots, e^{tX_n}$ unabhängig (Fakt 2.5.4)). Daher (Fakt 2.5.6(a)) multiplizieren sich die Erwartungswerte, und wir erhalten:

$$\begin{aligned} \mathbf{Pr}(X \geq m + a) &\leq e^{-t(m+a)} \cdot \mathbf{E}\Big(\prod_{1\leq i\leq n} e^{tX_i}\Big) \\ &= e^{-t(m+a)} \cdot \prod_{1\leq i\leq n} \mathbf{E}(e^{tX_i}). \end{aligned} \tag{16}$$

Was können wir über die Zahlen $\mathbf{E}(e^{tX_i})$ sagen?

[2]Die erste Ungleichung ist die „Chernoff-Schranke“, ein Spezialfall der verallgemeinerten Markov-Ungleichung (Prop. 2.3.4). Weil sie zentral für die Hoeffding-Schranke ist, heißen die Ungleichungen in diesem Kapitel oft auch Chernoff-Schranken oder Chernoff-Hoeffding-Schranken.

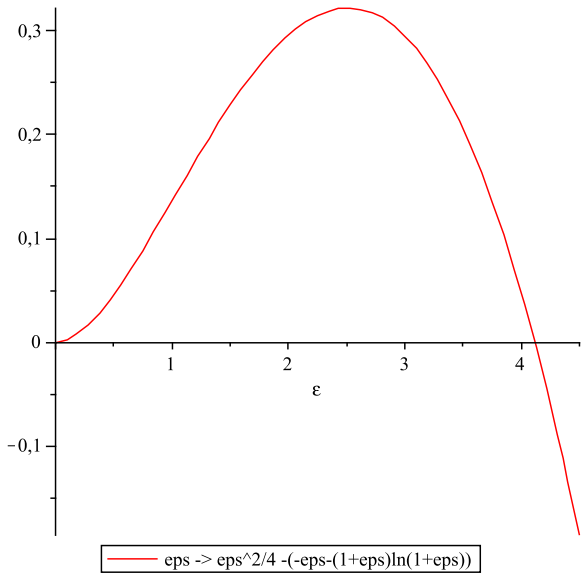


Abbildung 2: Funktion $\varepsilon \mapsto -\varepsilon^2/4 - \varepsilon + (1+\varepsilon)\ln(1+\varepsilon)$: in $[1, 4.1]$ nicht negativ.

Lemma 2.6.5 *Sei $t > 0$ beliebig. Dann gilt*:

(i) $e^{tx} \leq 1 + x(e^t - 1)$, *für* $0 \leq x \leq 1$.

(ii) *Ist Y eine Zufallsvariable mit $0 \leq Y \leq 1$, dann gilt* $\mathbf{E}(e^{tY}) \leq 1 + \mathbf{E}(Y)(e^t - 1)$.

Beweis: (i) Die Funktion $g\colon x \mapsto (e^t)^x$ ist konvex. Das heißt, dass der Graph der Funktion unterhalb der Sekante durch $(0,1)$ und $(1, e^t)$ verläuft, also für $0 \leq x \leq 1$ gilt:

$$e^{tx} = (e^t)^x \leq g(0) + x \cdot (g(1) - g(0)) = 1 + x(e^t - 1).$$

(ii) Wegen (i) gilt $e^{tY} \leq 1 + Y(e^t - 1)$, als Ungleichung zwischen den Zufallsvariablen e^{tY} und $1 + Y(e^t - 1)$. Die Behauptung folgt nun wegen der Monotonie und der Linearität der Erwartungswerte. □

Mit Lemma 2.6.5 erhalten wir aus (16):

$$\mathbf{Pr}(X \geq m + a) \leq e^{-t(m+a)} \cdot \prod_{1 \leq i \leq n} (1 + \mathbf{E}(X_i)(e^t - 1)). \tag{17}$$

Um dies zu vereinfachen, benutzen wir die Ungleichung zwischen dem arithmetischen und dem geometrischen Mittel, Proposition 2.3.7. Wenn wir diese Ungleichung in (17) auf die nichtnegativen Zahlen $a_i = 1 + \mathbf{E}(X_i)(e^t - 1)$, $1 \leq i \leq n$,

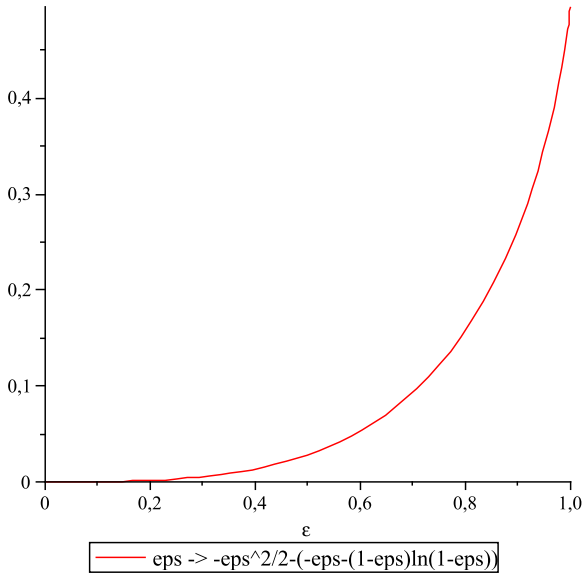


Abbildung 3: Funktion $\varepsilon \mapsto -\varepsilon^2/2 + \varepsilon + (1-\varepsilon)\ln(1-\varepsilon)$: in $[0,1]$ nicht negativ.

anwenden, und uns erinnern, dass $X = X_1 + \cdots + X_n$ und $m = \mathbf{E}(X)$ ist, ergibt sich

$$\begin{aligned}
\mathbf{Pr}(X \geq m+a) &\leq e^{-t(m+a)} \cdot \left(\frac{1}{n} \sum_{1 \leq i \leq n} (1 + \mathbf{E}(X_i)(e^t - 1)) \right)^n \\
&= e^{-t(m+a)} \cdot \left(1 + \frac{m(e^t-1)}{n} \right)^n \qquad (18) \\
&= \left(e^{-t(m+a)/n} \cdot \left(1 + \frac{m(e^t-1)}{n} \right) \right)^n \\
&= \left(z^{-(m+a)/n} \cdot \left(1 + \frac{m(z-1)}{n} \right) \right)^n, \qquad (19)
\end{aligned}$$

mit $z = e^t$.

Um aus (19) optimalen Nutzen zu ziehen, werden wir den bisher noch freien Parameter t bzw. $z = e^t$ so festlegen, dass der „Kern“

$$z^{-(m+a)/n} \cdot \left(1 + \frac{m}{n}(z-1) \right) \qquad (20)$$

möglichst klein wird. Wir betrachten aber zuerst den Spezialfall $a = n - m$. Dann ist (20) gleich $z^{-1} \cdot (1 + m(z-1)/n)$, was für $z \to \infty$ gegen m/n konvergiert, und $\mathbf{Pr}(X \geq n) = \mathbf{Pr}(X \geq m + (n-m)) \leq (m/n)^n = (m/(m+a))^{m+a}$. Damit ist

(6) für $a = n - m$ bewiesen, und wir können ab hier $0 < a < n - m$ annehmen. Wir wählen

$$z = \frac{m+a}{m} \cdot \frac{n-m}{n-(m+a)}. \tag{21}$$

(Dass dieser Wert tatsächlich (19) minimiert, sieht man durch Differenzieren.) Man sieht, dass (wegen $a > 0$) beide Faktoren in (21) größer als 1 sind, so dass sich $z > 1$ und $t > 0$ ergibt. Wir setzen (21) in die obere Schranke (18) ein und erhalten

$$\mathbf{Pr}(X \geq m+a) \leq \left(\frac{m(n-(m+a))}{(m+a)(n-m)}\right)^{m+a} \cdot \left(1 - \frac{m}{n} + \frac{(m+a)(n-m)}{n(n-(m+a))}\right)^{n}. \tag{22}$$

Eine leichte (Bruch-)Rechnung ergibt, dass der zweite Faktor in (22) einfach $\left(\frac{n-m}{n-(m+a)}\right)^n$ ist. Passendes Zusammenfassen ergibt dann

$$\mathbf{Pr}(X \geq m+a) \quad \leq \quad \left(\frac{m}{m+a}\right)^{m+a} \cdot \left(\frac{n-m}{n-(m+a)}\right)^{n-(m+a)}, \tag{23}$$

und das ist (6).

Um schließlich (7) zu beweisen, könnte man „analog“ vorgehen. Stattdessen führen wir diesen Fall aber auf (6) zurück, wie folgt. Wir definieren Zufallsvariable

$$\overline{X_i} = 1 - X_i, \;\; 1 \leq i \leq n,$$

und

$$\overline{X} = \overline{X_1} + \cdots + \overline{X_n},$$

und $\overline{m} = \mathbf{E}(\overline{X}) = \mathbf{E}(n - X) = n - \mathbf{E}(X) = n - m$. Weiter setzen wir $\overline{a} = b$ (dann ist $0 \leq \overline{a} < m = n - \overline{m}$). Nun wenden wir (6) an und erhalten:

$$\mathbf{Pr}(\overline{X} \geq \overline{m} + \overline{a}) \quad \leq \quad \left(\frac{\overline{m}}{\overline{m}+\overline{a}}\right)^{\overline{m}+\overline{a}} \cdot \left(\frac{n-\overline{m}}{n-(\overline{m}+\overline{a})}\right)^{n-(\overline{m}+\overline{a})}. \tag{24}$$

Wenn man diese Ungleichung wieder in die „X_i-Notation“ überführt, ergibt sich wegen $\overline{m} + \overline{a} = n - m + b = n - (m - b)$ und $n - \overline{m} = m$:

$$\mathbf{Pr}(X \leq m - b) \quad \leq \quad \left(\frac{n-m}{n-(m-b)}\right)^{n-(m-b)} \cdot \left(\frac{m}{m-b}\right)^{m-b}, \tag{25}$$

und das ist gerade (7).

2.7 Weitere Ungleichungen

Die folgende Behauptung ist eine Verallgemeinerung von Ungleichung (4) auf zwei Zufallsvariable:

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

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

Wenn $\mathbf{E}(Y^2) = 0$ ist, dann ist $\mathbf{Pr}(Y \neq 0) = 0$, also $\mathbf{E}(XY) = 0$, und die Ungleichung gilt trivialerweise. Sonst sucht man die Minimalstelle der quadratischen Funktion f (durch Differenzieren und Null-Setzen) und findet sie bei $\lambda_0 = -\mathbf{E}(XY)/\mathbf{E}(Y^2)$. Der Wert $f(\lambda_0)$ ist als Erwartungswert einer nichtnegativen Zufallsvariablen selbst nicht negativ, also gilt

$$0 \leq f(\lambda_0) = \mathbf{E}(X^2) - \frac{2\mathbf{E}(XY)^2}{\mathbf{E}(Y^2)} + \frac{\mathbf{E}(XY)^2}{\mathbf{E}(Y^2)} = \mathbf{E}(X^2) - \frac{\mathbf{E}(XY)^2}{\mathbf{E}(Y^2)}\ ;$$

daraus folgt $\mathbf{E}(XY)^2 \leq \mathbf{E}(X^2)\mathbf{E}(Y^2)$. □

Nicht ganz ideal an der Chebychev-Ungleichung (Fakt 2.3.3) 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 $t \geq 0$:*

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

Beweis:[3] 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 $X' = X - \mathbf{E}(X)$); dann ist $\mathbf{Var}(X) = \mathbf{E}(X^2)$. Man erinnere sich an

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

die Iverson-Notation: $[X \leq t]$ ist die charakteristische Funktion des Ereignisses $\{X \leq t\}$, 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 $t \geq 0$ 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 Proposition 2.7.2 mit der Chebychev-Ungleichung (Fakt 2.3.3). 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 wollen noch Ungleichung (4) benutzen, um eine Schranke für $\mathbf{Pr}(X \neq 0)$ herzuleiten, falls X eine Zufallsvariable ist, die Werte in den natürlichen Zahlen annimmt (und nicht konstant 0 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:*

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

Beweis: Die zweite Ungleichung folgt aus der Markov-Ungleichung, da $\mathbf{Pr}(X \neq 0) = \mathbf{Pr}(X \geq 1)$. Für die erste Ungleichung wenden wir Ungleichung (4) mit der auf $\{X \neq 0\}$ bedingten Wahrscheinlichkeit an:

$$\mathbf{E}(X \mid X \neq 0)^2 \leq \mathbf{E}(X^2 \mid X \neq 0).$$

Weiter gilt

$$\mathbf{E}(X \mid X \neq 0)^2 = \Big(\frac{\mathbf{E}(X)}{\mathbf{Pr}(X \neq 0)}\Big)^2 \quad \text{und} \quad \mathbf{E}(X^2 \mid X \neq 0) = \frac{\mathbf{E}(X^2)}{\mathbf{Pr}(X \neq 0)}.$$

Kombinieren dieser (Un-)Gleichungen, Kürzen und Umstellen liefert die Behauptung. □

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) *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. 2.3.6(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 Ungleichungen aus der Analysis und der Kombinatorik

Proposition A.0.1
Für alle $x \in \mathbb{R}$: $1 + x \leq e^x$, 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 hat bei $x = 0$ ihre einzige Nullstelle und ist strikt monoton wachsend. 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 $x = 0$. □

Proposition A.0.2

(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) Wenn $x = -1$, ist die linke Seite 0, die rechte ist $e^{-y} > 0$. Sei nun $x > -1$, also $1 + x > 0$. Dann folgt aus Prop. A.0.1 mit der Monotonie der Funktion $u \mapsto u^z$ die Ungleichung $(1+x)^y \leq (e^x)^y = e^{xy}$. (b) folgt aus (a), indem man $x' = x/y \geq -1$ betrachtet. □

Proposition A.0.3

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

Beweis: $e^x = \sum_{i \geq 0} \frac{x^i}{i!} \leq \sum_{i \geq 0} x^i = \frac{1}{1-x}$. □

Proposition A.0.4

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

Beweis: Betrachte $f(x) = (x - 1) - \ln x$, für $x > 0$. Die Ableitungen sind $f'(x) = 1 - 1/x$ und $f''(x) = x^{-2} > 0$. Daher hat f ein globales Minimum an der Stelle $x = 1$. Es folgt $(x - 1) - \ln x \geq f(1) = 0$, mit Gleichheit genau für $x = 1$. □

Proposition A.0.5 *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 – die rechte Seite ist 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

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

mit Prop. A.0.2(b). □