1 Diskrete Wahrscheinlichkeitsräume

Die Stochastik unterscheidet sich in mancher Hinsicht von anderen Zweigen der Mathematik. Viele ihrer Definitionen, Konzepte und Resultate sind ohne ihren Bezug auf Probleme des "täglichen Lebens", oder andere Naturwissenschaften, etwa die theoretische Physik, nur schwer zu verstehen. Andererseits ist die Wahrscheinlichkeitstheorie (im weiteren kurz:W.-Theorie) eine eigene, rigorose mathematische Theorie mit vielen Bezügen zu anderen mathematischen Disziplinen.

So ist es nicht weiter verwunderlich, daß man mit dem Begriff der *Wahrscheinlichkeit* in vielen Bereichen des täglichen und wissenschaftlichen Lebens konfrontiert ist:

1. In Wetterberichten heißt es, daß die Wahrscheinlichkeit für Regen bei 20% liegt.
2. Eine erste Hochrechnung nach einer Wahl besagt, daß wahrscheinlich ca. 9% der Bevölkerung grün gewählt haben.
3. Die Leukämiewahrscheinlichkeit beträgt ca. 0.5 Promille.
4. Die Wahrscheinlichkeit beim Würfeln eine 6 zu werfen wird als $\frac{1}{6}$ angenommen.
5. In der Quantenmechanik ist die Wahrscheinlichkeit gewisser Ereignisse proportional zum Integral des Quadrats ihrer Wellenfunktion.

Hierbei fällt auf, daß wir es auf den ersten Blick mit unterschiedlichen Arten zu tun haben, den Begriff Wahrscheinlichkeit zu gebrauchen (z.B. wird er sowohl auf zukünftige als auch auf vergangene Ereignisse angewandt). Um diese verschiedenen Arten unter einen Hut zu bekommen, wollen wir unter der Wahrscheinlichkeit eines Ereignisses, dessen Ausgang uns unbekannt ist, zunächst einmal ein Maß für die Gewißheit seines Eintretens (bzw. dafür, daß es eingetreten ist) verstehen. Diese Definition impliziert natürlich, daß die Wahrscheinlichkeit eines Ereignisses von meinem (unseren) subjektiven Kenntnisstand abhängt. Das ist auch durchaus sinnvoll. Beispielsweise kann in Beispiel (2) der Wahlleiter schon die Information über den Ausgang der Wahl besitzen – das Ereignis "8-10% aller Wähler haben grün gewählt" hat für ihn also Wahrscheinlichkeit 1 oder 0 – während man als Fernsehzuschauer noch auf Hochrechnungen, d.h. auf unsichere Informationen, angewiesen ist. Trotzdem ist es selbstverständlich aber so, daß jeder Beobachter glaubt mit der von seinem Kenntnisstand aus bestmöglichen Approximation der "wahren" Wahrscheinlichkeit zu arbeiten. Es soll hier noch bemerkt werden, daß die W.-Theorie selber, sofern wir uns erst einmal eine solche verschafft haben, unsensibel gegenüber dieser Subjektivität der Wahl einer Wahrscheinlichkeit ist. Die W.-Theorie findet quasi zu "einem späteren Zeitpunkt" statt: Zunächst bildet man ein Modell des Vorganges, den man analysieren möchte (und hier findet die Festlegung der Wahrscheinlichkeit eines Ereignisses statt), dann tritt die W.-Theorie auf den Plan und beschreibt, welches Verhalten das gewählte Modell aufweisen sollte.

Dieser erste Versuch Wahrscheinlichkeit zu definieren erfüllt offenbar nicht die Kriterien, die man an eine mathematisch saubere Definition stellen würde. Beispielsweise ist nicht klar inwieweit eine Wahrscheinlichkeit von 50% ein kleineres Maß an Sicherheit bedeutet als eine Wahrscheinlichkeit von 75%.

Ein Versuch in diese Richtung würde die Wahrscheinlichkeit eines Ereignisses E als den *Erwartungswert der relativen Häufigkeit* des Eintretens von E definieren, also als den Quotienten aus der Zahl der Fälle in denen E eingetreten ist und der Gesamtlänge der "Versuchsreihe", den man bei einer sehr langen Reihe gleichartiger Situationen erwarten würde. Diese Definition – obgleich sie auf einem "wahren" Sachverhalt beruht (gemeint ist das Gesetz der großen Zahlen, das wir in einem späteren Kapitel kennenlernen werden) – krankt aber an verschiedenen Defiziten. Sieht man einmal von dem praktischen Einwand ab, daß es eventuell unmöglich oder nur schwer möglich ist, eine große Zahl indentischer und unabhängiger Situationen herzustellen, so bleibt doch das schwerwiegende Hindernis, daß man für eine vernünftige Definition eines Erwartungswertes zunächst eine Definition der Wahrscheinlichkeit benötigt (wir werden dies im Lauf des dritten Kapitels kennenlernen). Man hat versucht, dieses Problem durch Wahl von sogenannten zufälligen Folgen von Ereignissen zu umgehen, doch stellte sich heraus, daß schon die Definition des Begriffs einer zufälligen Folge von Ereignissen beinahe das komplette Problem einer mathematischen Grundlegung der W.-Theorie beinhaltet.

Es waren grob gesprochen diese Gründe, die dazu führten, daß die mathematische Fundierung der Wahrscheinlichkeitstheorie lange Zeit ein offenes Problem war (das sogar als sechstes Problem Eingang in die berühmten Hilbertschen Probleme fand – genauer formulierte *Hilbert* (1862-1943) in seiner berümt gewordenen Rede auf dem Weltkongress 1900 das sechste Problem als dasjenige die theoretische Physik und die Wahrscheinlichkeitstheorie zu axiomatisieren), obschon die ersten wahrscheinlichkeitstheoretischen Resultate schon sehr viel älter sind. Das Problem der Axiomatisierung wurde schließlich 1933 von *A.N. Kolmogoroff* (1903-1987) gelöst.

Grundlage seiner Axiomatisierung bilden ein paar einfache Beobachtungen über relative Häufigkeiten. Um diese zu formulieren, führen wir zunächst die Menge Ω aller möglichen Ausgänge eines zufälligen Experiments ein (unter einem zufälligen Experiment oder Zufallsexperiment wollen wir gerade einen Vorgang verstehen, dessen Ausgang uns unbekannt ist). Um größere Schwierigkeiten, die bei beliebiger Wahl von Ω auftreten können, aus dem Wege zu gehen, sei bis auf weiteres Ω eine abzählbare Menge. Teilmengen von Ω heißen dann in der Wahrscheinlichkeitstheorie *Ereignisse (events)*. Die üblichen Mengenoperationen haben in der Wahrscheinlichkeitstheorie folgende Bedeutung:

Sprache der Ereignisse	*Mengenschreib- bzw. Sprechweise*
A, B, C sind Ereignisse	A, B, C sind Teilmengen von Ω
A und B	$A \cap B$
A oder B	$A \cup B$
nicht A	$A^c = \Omega \setminus A$
A und B sind unvereinbar	$A \cap B = \emptyset$
A impliziert B	$A \subset B$.

Grundlegend ist nun folgende

(1.1) Beobachtung. Es sei Ω eine abzählbare Menge und auf Ω führe man ein Zufallsexperiment n mal durch. Für $A \subset \Omega$ sei die relative Häufigkeit $r(A)$ definiert als die Anzahl

der Fälle, in denen A eingetreten ist, geteilt durch n. Dann gilt für jedes n und jede abzählbare Indexmenge I, so daß die Familie der Ereignisse $(A_i)_{i\in I}$ paarweise unvereinbar ist

1. $r(\Omega) = 1$
2. $r(\bigcup_{i\in I} A_i) = \sum_{i\in I} r(A_i)$.

Eine Wahrscheinlichkeit ist nun eine Mengenfunktion, die sich wie relative Häufigkeiten verhält, genauer:

(1.2) Definition. Es sei Ω eine abzählbare Menge. Eine Wahrscheinlichkeit (*probability*) auf Ω ist eine Mengenfunktion $P : \mathcal{P}(\Omega) \to [0,1]$ von der Potenzmenge $\mathcal{P}(\Omega)$ von Ω in das Einheitsintervall mit

1. $P(\Omega) = 1$
2. $P(\bigcup_{i\in I} A_i) = \sum_{i\in I} P(A_i)$ für jede abzählbare Indexmenge I und jede paarweise unvereinbare Familie von Ereignissen $(A_i)_{i\in I}$.

Das Paar (Ω, P) heißt *Wahrscheinlichkeitsraum (probability space)*.

(1.3) Beispiele.

1. Beim Würfeln mit 2 Würfeln besteht die Menge Ω offenbar aus allen möglichen Kombinationen von Augenzahlen. Ω besteht in diesem Fall aus 36 Elementen: $\Omega = \{(1,1),(1,2),\ldots,(6,6)\} = \{1,2,3,4,5,6\}^2$. Wir setzen $P(\{(i,j)\}) = 1/36$ für jedes sogenannte Elementarereignis (i,j). Für jedes Ereignis A ist daher $P(A) = |A|/36$, wobei $|A|$ die Anzahl der Elemente in A ist. Sei z. B. $A = \{(1,1),(2,2),\ldots,(6,6)\}$ das Ereignis, daß die Augenzahlen gleich sind. Dann ist $P(A) = 6/36 = 1/6$.

2. In einem Kartenspiel mit einer geraden Anzahl ($= 2n$) von Karten befinden sich 2 Joker. Nach guter Mischung werden die Karten in zwei gleich große Haufen aufgeteilt. Wie groß ist die Wahrscheinlichkeit, daß beide Joker im gleichen Haufen sind? Wir wählen $\Omega = \{\,(i,j) \in \{1,2,\ldots,2n\}^2 : i \neq j\}$. Hierbei ist $\{(i,j)\} \subset \Omega$ das Ereignis, daß sich der erste Joker am Platz i und der zweite am Platz j befindet. Nach guter Mischung hat jedes dieser Ereignisse die Wahrscheinlichkeit $P(\{(i,j)\}) = 1/|\Omega| = 1/2n(2n-1)$. Das uns interessierende Ereignis ist

$$A = \{\,(i,j) \in \{1,2,\ldots,n\}^2 : i \neq j\} \cup \{(i,j) \in \{n+1,\ldots,2n\}^2 : i \neq j\}.$$

   Dieses enthält $2 \cdot n(n-1)$ "Elementarereignisse" (i,j). Somit ist

$$P(A) = \frac{2n(n-1)}{2n(2n-1)} = \frac{n-1}{2n-1}.$$

3. Eine Münze wird n-mal geworfen. Ω sei die Menge der n-Tupel, bestehend aus „Zahl“ und „Kopf“. Somit ist $|\Omega| = 2^n$. Haben alle n-Tupel gleiche Wahrscheinlichkeiten, so hat jedes Element von Ω Wahrscheinlichkeit 2^{-n}. Es sei A_k das Ereignis, daß k-mal „Zahl“ fällt. Es gilt also $P(A_k) = |A_k|2^{-n}$. Die Anzahl $|A_k|$ wird weiter unten bestimmt.

4. *Auto-Ziege Problem:* Ein Spielleiter konfrontiert einen Spieler mit drei verschlossenen Türen; hinter einer steht ein Auto, hinter den anderen je eine Ziege. Der Spieler muß sich für eine Tür entscheiden und dies dem Leiter verkünden. Dieser öffnet daraufhin eine der beiden anderen Türen und zeigt eine Ziege. Dann fragt er den Spieler, ob er sich für die ungeöffnete Tür umentscheiden möchte, die der Spieler nicht gewählt hatte. Ist es von Vorteil zu tauschen (angenommen, der Spieler hat Interesse an dem Auto)? Auch der geübte Spieler neigt zu der falschen Antwort, daß ein Tausch irrelevant ist. Wir werden dies im folgenden analysieren. Angenommen, der Spieler entscheidet sich, in jedem Fall zu tauschen. Die Tür mit dem Auto dahinter sei mit 1 gekennzeichnet, die beiden anderen mit 2 und 3. Eine Möglichkeit, ein Spiel zu beschreiben, ist die Angabe eines 4-Tupels (u, v, w, x), wobei u die gewählte Tür des Spielers, v die des Spielleiters und w die Tür, zu der der Spieler auf jeden Fall wechselt, beschreibt. x beschreibe dann den Ausgang des Spiels, also den Gewinn (G) oder Verlust (V) des Autos. Der Stichprobenraum hat dann die folgende Gestalt:

$$S = \{(1, 2, 3, V), (1, 3, 2, V), (2, 3, 1, G), (3, 2, 1, G)\}.$$

Natürlich nehmen wir an, daß alle drei Türen mit gleicher Wahrscheinlichkeit 1/3 gewählt werden können. Bei Wahl der Tür 1 mit Wahrscheinlichkeit 1/3 führt ein Wechsel der Entscheidung natürlich zum Verlust des Spiels. Bei Wahl der Tür 2 oder 3 ergibt der Wechsel einen Gewinn. Also ist die Wahrscheinlichkeit, das Auto zu gewinnen, $1/3 + 1/3 = 2/3$. Unter der Annahme, der Spieler tausche generell nicht, hat der Stichprobenraum die Gestalt:

$$S = \{(1, 2, 1, G), (1, 3, 1, G), (2, 3, 2, V), (3, 2, 3, V)\}.$$

Hier ergibt sich eine Wahrscheinlichkeit von 2/3 zu verlieren.

Es sei an dieser Stelle erwähnt, daß es aufgrund der Eigenschaft (2) in der Definition von P und der Abzählbarkeit von Ω natürlich genügt, die Wahrscheinlichkeit auf den einzelnen Elementen von Ω, den sogenannten *Elementarereignissen* (*elementary events, sample points*) festzulegen (dies ist auch in einigen der Beispiele unter (1.3) so geschehen). Genauer gilt:

(1.4) Lemma. Es sei Ω eine abzählbare Menge $(p(\omega))_{\omega\in\Omega}$ eine Folge positiver Zahlen mit

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

Dann ist durch $P(\{\omega\}) := p(\omega)$ eine Wahrscheinlichkeit auf Ω eindeutig definiert.

Ist umgekehrt P eine Wahrscheinlichkeit auf Ω, so induziert diese durch $p(\omega) := P(\{\omega\})$ eine Folge mit obigen Eigenschaften.

Beweis. Man setze einfach für $A \subset \Omega$

$$P(A) := \sum_{\omega \in A} p(\omega)$$

und sehe, daß dies eine Wahrscheinlichkeit auf Ω definiert. □

Bemerkung. Da alle $p(\omega) \geq 0$ sind, spielt selbst im Fall, wo Ω unendlich ist, die Reihenfolge der Summation in $\sum_{\omega \in \Omega} p(\omega)$ keine Rolle. Genau genommen handelt es sich dann um einen Grenzwert. Man wählt zunächst eine Abzählung $\omega_1, \omega_2, \ldots$ der Elemente von Ω. Dann ist $\sum_{\omega \in \Omega} p(\omega) = \lim_{n \to \infty} \sum_{i=1}^{n} p(\omega_i)$, wobei der Grenzwert nicht von der gewählten Abzählung abhängt, da die $p(\omega) \geq 0$ sind, die Summe also absolut konvergiert.

Wir haben also gesehen, daß eine Folge $p := (p(\omega))_{\omega \in \Omega}$ mit $\sum_{\omega \in \Omega} p(\omega) = 1$ eineindeutig einer Wahrscheinlichkeit P auf Ω entspricht. Wir werden daher oft auch (Ω, p) synonym für (Ω, P) verwenden, wenn P durch p induziert ist.

In konkreten Situationen wählt man Ω oft so, daß die einzelnen Elementarereignisse $\omega \in \Omega$ gleich wahrscheinlich sind, was natürlich nur möglich ist, wenn Ω endlich ist. Man spricht dann von einem *Laplace-Experiment*. Einige Beispiele dazu:

(1.5) Beispiele.

1. Beim Würfeln mit einem Würfel wählt man $\Omega = \{1, 2, 3, 4, 5, 6\}$. Dabei ist $i \in \Omega$ das Elementarereignis, daß die Zahl i geworfen wird. Ist der Würfel nicht gezinkt, so wird man $p(i) = 1/6$ für alle $i \in \Omega$ setzen.

2. Als Elementarereignisse beim Würfeln mit 2 Würfeln fassen wir alle möglichen Kombinationen von Augenzahlen auf (siehe auch Beispiel 1.3 (1)). Ω besteht in diesem Fall aus 36 Elementarereignissen: $\Omega = \{(1,1), (1,2), \ldots, (6,6)\} = \{1, 2, 3, 4, 5, 6\}^2$. Wir setzen $p((i,j)) = 1/36$ für jedes Elementarereignis.

3. Ein Stapel mit n Karten wird gut gemischt. Wir denken uns die Karten von 1 bis n durchnumeriert. Die Elementarereignisse sind die möglichen Reihenfolgen dieser n Karten, etwa bei $n = 3$:

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

   Bei guter Mischung wird man jede Reihenfolge als gleich wahrscheinlich betrachten können. Jedes Elementarereignis hat dann Wahrscheinlichkeit $\frac{1}{n!}$.

4. *Urnenmodell:* In einer Schachtel (Urne) befinden sich r rote und s schwarze Kugeln. Eine Kugel wird zufällig herausgenommen. Mit welcher Wahrscheinlichkeit ist sie rot?
   Wir denken uns die Kugeln von 1 bis $r + s$ durchnumeriert. Die Kugeln mit den Nummern 1 bis r sind rot; die anderen schwarz. Für Ω nehmen wir die Menge $\{1, 2, \ldots, r + s\}$. Dann ist $i \in \Omega$ das Elementarereignis, daß die Kugel i gezogen wird. Diese Elementarereignisse sind nach guter Mischung gleich wahrscheinlich, haben also Wahrscheinlichkeit $\frac{1}{r+s}$. Unser Ereignis enthält r Elementarereignisse. Seine Wahrscheinlichkeit ist also $r/(r + s)$.

Wahrscheinlichkeiten genügen einigen einfachen Regeln, die untenstehend aufgelistet sind

(1.6) Satz. Es sei (Ω, p) ein W-Raum.

1. Für jedes Ereignis A gilt $0 \leq P(A) \leq 1$.
2. $P(\emptyset) = 0$, $P(\Omega) = 1$.
3. Sind Ereignisse A_i für $i \in \mathbb{N}$ paarweise disjunkt (d.h. $A_i \cap A_j = \emptyset$ für $i \neq j$), so gilt $P(\bigcup_{i\in\mathbb{N}} A_i) = \sum_{i=1}^{\infty} P(A_i)$ (*abzählbar additiv, countable additive*).
4. In (3) ohne die Voraussetzung, daß die A_i paarweise disjunkt sind, gilt noch $P(\bigcup_{i\in\mathbb{N}} A_i) \leq \sum_{i=1}^{\infty} P(A_i)$ (*abzählbar subadditiv, countable subadditive*).
5. $A \subset B \Rightarrow P(B) = P(A) + P(B \setminus A)$.
6. $A \subset B \Rightarrow P(A) \leq P(B)$ (*monoton*).
7. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Bemerkung. Gilt $A_{n+1} = A_{n+2} = \cdots = \emptyset$ für ein $n \geq 1$, so besagen (3) und (4)

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) \quad \text{bzw.} \quad P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i)$$

(*endlich additiv bzw. subadditiv*).

Beweis. (1), (2) und (3) folgen sofort aus der Definition.
(4): Jedes $\omega \in \bigcup_{i=1}^{\infty} A_i$ gehört zu mindestens einem der A_i. Demzufolge gilt

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{\omega\in\bigcup_{i=1}^{\infty} A_i} p(\omega) \leq \sum_{i=1}^{\infty} \sum_{\omega\in A_i} p(\omega) = \sum_{i=1}^{\infty} P(A_i).$$

(5) Es gelten $B = A \cup (B \setminus A)$ und $A \cap (B \setminus A) = \emptyset$. Somit ist nach (3) $P(B) = P(A) + P(B \setminus A)$.
(6) folgt aus (5) und $P(B \setminus A) \geq 0$.
(7) Wir haben folgende Zerlegungen in disjunkte Teilmengen:

$$A \cup B = (A \setminus B) \cup B$$

und

$$A = (A \setminus B) \cup (A \cap B).$$

Nach (3) gilt:

$$\begin{aligned} P(A \cup B) &= P(A \setminus B) + P(B), \\ P(A) &= P(A \setminus B) + P(A \cap B). \end{aligned}$$

Subtrahiert man die zweite Gleichung von der ersten, so folgt (7). □

Die Festlegung der Wahrscheinlichkeiten der Elementarereignisse ist ein außermathematisches Problem. In den bisherigen Beispielen hatten die Elementarereignisse jeweils alle die gleichen Wahrscheinlichkeiten. Dies ist vernünftig, wenn alle Elementarereignisse als ,,gleich möglich" erscheinen, oder wenn kein Grund für eine Ungleichbehandlung der Elementarereignisse vorliegt. Tatsächlich wählt man die Zerlegung in Elementarereignisse oft unter diesem Gesichtspunkt.

Ein Beispiel dazu: Jemand wirft zwei Würfel. Interessiert er sich nur für die Augensumme, so kann er als Elementarereignisse die möglichen Ergebnisse dafür nehmen: $\Omega = \{2, 3, 4, \ldots, 12\}$. Es ist offensichtlich, daß diese Elementarereignisse nicht gleichwertig sind. Deshalb nimmt man besser die Elementarereignisse aus (1.5 (2)).

In vielen Fällen wäre die Festlegung, daß alle Elementarereignisse gleich wahrscheinlich sind, aber ganz unsinnig, beispielsweise wenn man die Wahrscheinlichkeit modelliert, daß ein produziertes Werkstück defekt ist.

Nun ein Beispiel mit einem unendlichen W∴Raum:

(1.7) Beispiel. Eine Münze wird so lange geworfen, bis zum erstenmal ,,Kopf" fällt. Wir wählen als Ω die natürlichen Zahlen $\mathbb{N}$. Das Elementarereignis $i \in \mathbb{N}$ bedeutet, daß zum erstenmal beim i-ten Wurf ,,Kopf" fällt. Wie groß ist $p(i)$? Daß i eintritt, ist auch ein Elementarereignis in unserem Beispiel (1.3 (3)), nämlich, daß zunächst $(i-1)$-mal ,,Zahl" fällt und dann ,,Kopf". Somit ist $p(i) = 2^{-i}$. Die $p(i)$ erfüllen die Bedingung in Lemma (1.4): $\sum_{i\in\mathbb{N}} p(i) = 1$. Also ist (Ω, p) ein W∴Raum.

In unserem Modell ist das Ereignis, daß ,,Kopf" nie fällt, das unmögliche Ereignis. (*Vorsicht:* Es gibt in der Literatur andere Modelle – mit überabzählbarem W∴Raum – wo dieses Ereignis zwar Wahrscheinlichkeit 0 hat, aber nicht unmöglich ist.)

Die Bestimmung der Wahrscheinlichkeit von Durchschnitten ist in der Regel einfacher als die von Vereinigungen. Eine Verallgemeinerung von (1.6 (7)) sieht wie folgt aus: $A_1, \ldots, A_n$ seien n Ereignisse. $A_1 \cup \cdots \cup A_n$ ist das Ereignis, daß mindestens eines der A_i eintritt.

(1.8) Satz (*Ein- und Ausschlußprinzip, inclusion-exclusion principle).*
Für $A_1, \ldots, A_n \subset \Omega$ gilt

$$P(A_1 \cup \ldots \cup A_n) = \sum_{i=1}^{n} P(A_i) - \sum_{i_1<i_2} P(A_{i_1} \cap A_{i_2}) + \sum_{i_1<i_2<i_3} P(A_{i_1} \cap A_{i_2} \cap A_{i_3}) - \ldots + (-1)^{n-1} P(A_1 \cap A_2 \cap \ldots \cap A_n).$$

Beweis. Induktion nach n: Für $n = 2$ ist dies (1.6 (7)).

Induktionsschluß:

$$P(A_1 \cup \ldots \cup A_{n+1}) = P(A_1 \cup \ldots \cup A_n) + P(A_{n+1}) - P((A_1 \cup \ldots \cup A_n) \cap A_{n+1})$$

nach (1.6 (7))

$$= \sum_{i=1}^{n+1} P(A_i) - \sum_{1\le i_1<i_2\le n} P(A_{i_1}\cap A_{i_2}) + \sum_{1\le i_1<i_2<i_3\le n} P(A_{i_1}\cap A_{i_2}\cap A_{i_3}) - \ldots - P((A_1\cap A_{n+1})\cup(A_2\cap A_{n+1})\cup\cdots\cup(A_n\cap A_{n+1}))$$

nach Induktionsvoraussetzung und dem Distributivgesetz für Mengenoperationen. Wendet man auf den letzten Summanden nochmals die Induktionsvoraussetzung an, so folgt die Behauptung. □

Exkurs zu Abzählmethoden

Zur Berechnung der Wahrscheinlichkeiten in Laplace-Experimenten sind die folgenden kombinatorischen Ergebnisse von Nutzen. In einer Urne seinen n Kugeln mit $1, 2, \ldots, n$ numeriert. Es werden k Kugeln zufällig gezogen. Können Kugeln mehrfach gezogen werden (man legt also die gezogene Kugel jeweils zurück), spricht man von einer *Stichprobe mit Zurücklegen*; kann jede Kugel nur einmal auftreten von einer *Stichprobe ohne Zurücklegen*. Eine Ziehung kann durch ein k-Tupel $(\omega_1, \ldots, \omega_k)$ angegeben werden, wobei ω_i die Nummer der bei der i'ten Ziehung gezogenen Kugel ist. Es kommt hier auf die Reihenfolge an, und man spricht von einer *Stichprobe in Reihenfolge*. Kommt es hingegen nur auf die Anzahl der einzelnen Kugeln an, spricht man von einer *Stichprobe ohne Reihenfolge* und notiert in gewöhnlichen Mengenklammern $\{\omega_1, \ldots, \omega_k\}$.

Man kann nun 4 Stichprobenräume unterscheiden, deren Elemente gezählt werden sollen. Sei $A = \{1, \ldots, n\}$.

1. (*Stichprobe in Reihenfolge mit Zurücklegen*) Man wählt hier den Stichprobenraum

   $$\Omega_1 = \{\omega = (\omega_1, \ldots, \omega_k) : \omega_i \in A, i = 1, \ldots, k\} = A^k.$$

   Offensichtlich gilt $|\Omega_1| = n^k$.

2. (*Stichprobe in Reihenfolge ohne Zurücklegen*) Hier ist der Stichprobenraum

   $$\Omega_2 = \{\omega = (\omega_1, \ldots, \omega_k) : \omega_i \in A,\ \omega_i \neq \omega_j \text{ für } i \neq j,\ 1 \le i, j \le k\}.$$

   Es dient uns nun ein vermutlich wohlbekanntes *Abzählprinzip*: Sei Ω die Menge von k-Tupeln $\omega = (\omega_1, \ldots, \omega_k)$, aufzufassen als Ergebnisse eines aus k Teilexperimenten bestehenden zufälligen Experiments. Gibt es für das i'te Teilexperiment r_i mögliche Ausgänge, und ist für jedes i die Zahl r_i unabhängig von den Ausgängen der früheren Teilexperimente, dann ist

   $$|\Omega| = r_1\, r_2 \cdots r_k.$$

   Dies sieht man einfach via einer Induktion. Es folgt nun unmittelbar: $|\Omega_2| = n(n-1)(n-2)\cdots(n-k+1)$. Speziell für $n = k$ besteht Ω_2 aus der Menge der *Permutationen* von $\{1, \ldots, n\}$ und es gilt $|\Omega_2| = n! := n(n-1)(n-2)\cdots 2 \cdot 1$.

3. (*Stichprobe ohne Reihenfolge ohne Zurücklegen*) Hier hat der Stichprobenraum die Form

$$\Omega_3 = \{\{\omega_1, \ldots, \omega_k\} : \omega_i \in A, \omega_i \neq \omega_j, (i \neq j)\}.$$

Dieser Raum läßt sich nun einfach beschreiben, indem man in Ω_2 die folgende Äquivalenzrelation einführt: $(\omega_1, \ldots, \omega_k) \sim (\omega'_1, \ldots, \omega'_k)$ genau dann, wenn es eine Permutation π von $\{1, \ldots, k\}$ gibt mit $\omega'_i = \omega_{\pi i}$ für $i = 1, \ldots, k$. Die Elemente von Ω_3 sind nun die Äquivalenzklassen. Da jede Äquivalenzklasse $k!$ Elemente hat, folgt $|\Omega_2| = k!|\Omega_3|$. Man schreibt

$$|\Omega_3| = \binom{n}{k} := \frac{n!}{k!(n-k)!}$$

(*Binomialkoeffizient*) für $1 \leq k \leq n$. $\binom{n}{k}$ ist die Anzahl der Teilmengen der Mächtigkeit k von einer Menge der Mächtigkeit n. Da jede Menge genau eine Teilmenge der Mächtigkeit 0 hat (die leere Menge), setzt man $\binom{n}{0} = 1$. Setzt man nun noch $0! = 1$, gilt die obige Definitionsgleichung des Binomialkoeffizienten auch für $k = 0$. Es sei bemerkt, daß man jede obige Äquivalenzklasse zum Beispiel durch den Repräsentanten $(\omega_1, \ldots, \omega_k)$ mit $\omega_1 < \omega_2 < \ldots < \omega_k$ beschreiben kann.

In Beispiel (1.5)(3) ist also $|A_k| = \binom{n}{k}$.

4. (*Stichprobe ohne Reihenfolge mit Zurücklegen*) Hier wählt man die Menge der Äquivalenzklassen unter der oben eingeführten Relation im Stichprobenraum Ω_1 als Stichprobenraum. Man wählt als Repräsentanten einer jeden Klasse ein Tupel mit $\omega_1 \leq \omega_2 \leq \ldots \leq \omega_k$, so daß man die Darstellung

$$\Omega_4 = \{\omega = (\omega_1, \ldots, \omega_k) \in A^k : \omega_1 \leq \omega_2 \leq \ldots \leq \omega_k\}$$

erhält. Ordnet man jedem Element $(\omega_1, \ldots, \omega_k)$ der Menge Ω_4 die Folge $(\omega'_1, \ldots, \omega'_k)$ mit $\omega'_i = \omega_i + i - 1$ zu, so wird der Stichprobenraum bijektiv auf die Menge $\{(\omega'_1, \ldots, \omega'_k) \in B^k : \omega'_1 < \omega'_2 < \ldots \omega'_k\}$ mit $B = \{1, 2, \ldots, n + k - 1\}$ abgebildet, und nach Fall (3) folgt:

$$|\Omega_4| = \binom{n+k-1}{k}.$$

Eine erste Anwendung haben diese Abzählverfahren bei der Berechnungen gewisser Wahrscheinlichkeiten in wesentlichen physikalischen Verteilungen.

Die Maxwell–Boltzmannsche und die Bose–Einsteinsche Statistik Diese sogenannten Statistiken beschreiben Verteilungen in der statistischen Physik, genauer die Verteilungen von n Teilchen in einem abstrakten Raum, dem sogenannten Phasenraum. Zerteilt man diesen Raum in N Zellen, so ist die entsprechende Verteilung dadurch festgelegt, daß man bestimmt, was die Wahrscheinlichkeit ist, in einer bestimmten Zelle k Teilchen zu finden. Nimmt man an, daß die Teilchen unterscheidbar sind, so ergibt sich die Maxwell–Boltzmann–Statistik, die jeder Verteilung die Wahrscheinlichkeit $\frac{1}{N^n}$ zurordnet. Für das Ereignis in einer bestimmten Zelle genau k Teilchen vorzufinden haben wir dann $\binom{n}{k} \times (N-1)^{n-k}$ Möglichkeiten, da es $\binom{n}{k}$ Möglichkeiten gibt, die k Teilchen für die Zelle auszuwählen und sich die restlichen $n-k$ Teilchen auf $(N-1)^{n-k}$ verschiedene Möglichkeiten

auf die $N-1$ restlichen Zellen verteilen lassen. Dies ergibt eine Wahrscheinlichkeit P_k des Ereignisses genau k Teilchen in einer bestimmten Zelle vorzufinden von

$$P_k = \binom{n}{k} \big(\frac{1}{N}\big)^k \big(1 - \frac{1}{N}\big)^{n-k}.$$

Diese Verteilung wird uns im Laufe der Vorlesung noch einige Male begegnen.

Die Maxwell–Boltzmann–Statistik hat sich beispielsweise für Gasmoleküle als der richtige Ansatz erwiesen. Für einige Elementarteilchen, z.B. Photonen oder Elektronen hingegen hat es bewährt, die Teilchen als ununterscheidbar anzunehmen. Wir können daher auch nur noch zwei Verteilungen unterscheiden, wenn sie sich in der Besetzungszahl mindestens einer (und damit mindestens zwei) Zelle(n) unterscheiden. Dies ist der Ansatz der *Einstein–Bose–Statistik*. Man überlegt sich schnell, daß dies dem Ziehen mit Zurücklegen ohne Beachtung der Reihenfolge entspricht, man also $\binom{N+n-1}{n}$ Elementarereignisse hat. Die entsprechende Laplace–Wahrscheinlichkeit eines Elementarereignisses ist damit gegeben durch $\frac{1}{\binom{N+n-1}{n}}$. Nun berechnen wir nach der Einstein–Bose–Statistik die Wahrscheinlichkeit dafür, daß in einer vorgegebenen Zelle k Teilchen liegen. Dafür genügt es, die Anzahl der Möglichkeiten zu bestimmen, in denen dieses Ereignis eintritt. Diese Anzahl ist gleich der Anzahl der Möglichkeiten, daß in den übrigen $N-1$ Zellen $n-k$ Teilchen liegen, also gleich $\binom{N+n-k-2}{n-k}$; die gesuchte Wahrscheinlichkeit ist daher gegeben durch

$$\frac{\binom{N+n-k-2}{n-k}}{\binom{N+n-1}{n}}.$$

Es sei hier noch abschließend erwähnt, daß auch die Bose–Einstein–Statistik nicht allgemein gültig ist. Für einige Elementarteilchen wendet man daher noch das Pauli–Prinzip an, um zur sogenannten *Fermi–Dirac–Statistik* zu gelangen.

2 Bedingte Wahrscheinlichkeiten, Unabhängigkeit

Ein wichtiges Werkzeug in der Wahrscheinlichkeitstheorie ist die sogenannte ,,bedingte Wahrscheinlichkeit". Dazu ein Beispiel:

Sei Ω die Menge der Einwohner Bielefelds. Ein Reporter des WDR befragt einen rein zufällig herausgegriffenen Bielefelder nach seiner Meinung zur Einführung von Studiengebühren. Wir nehmen an, daß jeder Einwohner die gleiche Chance hat, befragt zu werden. Ist N die Anzahl der Einwohner, so ist die Wahrscheinlichkeit dafür, daß ein bestimmter Einwohner befragt wird, $1/N$. Natürlich ist es sehr wahrscheinlich, daß Studierende der Einführung von Studiengebühren skeptischer gegenüberstehen als Nichtstudierende. Es sei B die Menge der Bielefelder Studierenden. Es gilt daher $P(B) = |B|/N$. Sei A die Menge der Bielefelder, die die Einführung befürwortet. Es gilt dann $P(A) = |A|/N$, während der relative Anteil der Studierenden, die die Studiengebühren befürworten, sich berechnet als $|A \cap B|/|B| = P(A \cap B)/P(B)$. Man bezeichnet dies als bedingte Wahrscheinlichkeit von A gegeben B. Sie unterscheidet sich für gewöhnlich von der "unbedingten" Wahrscheinlichkeit $P(A)$.

Allgemein definieren wir:

(2.1) Definition. Sei $B \subset \Omega$ ein Ereignis mit $P(B) > 0$. Für jedes Ereignis $A \subset \Omega$ heißt $P(A|B) := P(A \cap B)/P(B)$ die *bedingte Wahrscheinlichkeit (conditional probability)* für A gegeben B .

Der nachfolgende Satz enthält einige einfache Eigenschaften, die zeigen, daß es sich bei der bedingten Wahrscheinlichkeit in der Tat um eine Wahrscheinlichkeit handelt.

(2.2) Satz. Es seien $A, B \subset \Omega$ und $P(B) > 0$. Dann gilt:

1. $A \supset B \Rightarrow P(A|B) = 1$.
2. $B \cap A = \emptyset \Rightarrow P(A|B) = 0$.
3. Sind die Ereignisse A_i, $i \in \mathbb{N}$, paarweise disjunkt, so gilt
$$P\left(\bigcup_{i=1}^{\infty} A_i \middle| B\right) = \sum_{i=1}^{\infty} P(A_i|B).$$
4. $P(A^c|B) = 1 - P(A|B)$.

Beweis. (1), (2) folgen sofort aus der Definition.

(3)

$$\begin{aligned} P\left(\bigcup_{i=1}^{\infty} A_i \middle| B\right) &= \frac{1}{P(B)} P\left(\left(\bigcup_{i=1}^{\infty} A_i\right) \cap B\right) = \frac{1}{P(B)} P\left(\bigcup_{i=1}^{\infty} (A_i \cap B)\right) \\ &= \sum_{i=1}^{\infty} \frac{P(A_i \cap B)}{P(B)} = \sum_{i=1}^{\infty} P(A_i|B). \end{aligned}$$

(4) Wegen $A \cap A^c = \emptyset$ gilt nach (3)

$$P(A|B) + P(A^c|B) = P(A \cup A^c|B) = P(\Omega|B) = 1.$$

□

(2.3) Bemerkung. Sei (Ω, p) ein endlicher Wahrscheinlichkeitsraum, und alle Elementarereignisse seien gleich wahrscheinlich (Laplace-Experiment). Dann gilt für $A, B \subset \Omega$ und $B \neq \emptyset$

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

d.h; die bedingten Wahrscheinlichkeiten lassen sich über die Mächtigkeiten der Ereignisse bestimmen.

(2.4) Beispiele.

1. Wie groß ist die Wahrscheinlichkeit, daß beim Werfen mit zwei Würfeln einer der beiden eine 2 zeigt, gegeben die Augensumme ist 6? Sei B das Ereignis ,,Die Augensumme ist 6.“, also

   $$B = \{(1,5), (2,4), (3,3), (4,2), (5,1)\},$$

   und A das Ereignis ,,Mindestens einer der Würfel zeigt 2.“:

   $$A = \{(2,1), (2,2), (2,3), (2,4), (2,5), (2,6), (1,2), (3,2), (4,2), (5,2), (6,2)\}.$$

   Dann gilt $A \cap B = \{(2,4), (4,2)\}$ und $P(A|B) = 2/5$. Zum Vergleich: Die unbedingte Wahrscheinlichkeit ist $P(A) = 11/36 < P(A|B)$.

2. Es seien drei Kästen mit je zwei Schubladen gegeben, in denen je eine Gold (G)- bzw. eine Silbermünze (S) in der folgenden Aufteilung liege: $\Omega = \{[G,G], [G,S], [S,S]\}$. Zufällig wird ein Kasten gewählt, und dann zufällig eine Schublade geöffnet. In dieser liege eine Goldmünze. Wie groß ist die Wahrscheinlichkeit dafür, daß in der anderen Schublade dieses Kastens eine Silbermünze liegt? Die zufällige Wahl sei jeweils ein Laplace-Experiment. Wir numerieren die Kästen und Schubladen und wählen als Stichprobenraum $\Omega = \{1,2,3\} \times \{1,2\}$ und setzen $P(\{(i,j)\}) = 1/3 \cdot 1/2 = 1/6$. Dann ist die gesuchte Wahrscheinlichkeit $P(A|B)$ mit $B = \{(1,1), (1,2), (2,1)\}$ (Züge, so daß in der Schublade eine Goldmünze liegt) und $A = \{(2,1), (3,1), (3,2)\}$ (Züge, so daß in der anderen Schublade eine Silbermünze liegt). Es gilt $P(A|B) = (1/6)/(1/2) = 1/3$. Welchen Wert für die Wahrscheinlichkeit hätte man vor der Rechnung erwartet?

In der bisherigen Diskussion haben wir die bedingten Wahrscheinlichkeiten auf die unbedingten zurückgeführt. Es ist jedoch oft wichtiger, umgekehrt Wahrscheinlichkeiten aus gewissen bedingten Wahrscheinlichkeiten zu berechnen. Die grundlegende Idee dabei ist es den zugrunde liegenden W.-Raum mit Hilfe einer Bedingung in disjunkte Teilräume

zu zerlegen, auf diesen die bedingten Wahrscheinlichkeiten zu berechnen und diese dann wieder mit geeigneten Gewichten zusammenzufügen. Ein Beispiel dazu:

(2.5) Beispiel. Ein Gesundheitslexikon sagt, daß es sich beim Auftreten eines Symptoms S um 2 Krankheiten K oder K^c handeln kann. Diese sind insgesamt unterschiedlich häufig: Sie treten im Verhältnis 7:93 auf. Andererseits zeigt sich das Symptom S, wenn K vorliegt in 92% aller Fälle, bei Vorliegen von K^c nur in 8.5% aller Fälle. Mit welcher Wahrscheinlichkeit ist nun eine Person, bei der S festgestellt wird, an K erkrankt ?

Zunächst einmal ist es plausibel, daß wir die Wahrscheinlichkeit für das Auftreten von S berechnen können als

$$P(S) = P(K)P(S|K) + P(K^c)P(S|K^c).$$

Dem liegt der folgende allgemeine Satz zugrunde:

(2.6) Satz *(Formel von der totalen Wahrscheinlichkeit).* Es seien $B_1, \ldots, B_n$ paarweise disjunkte Ereignisse. Dann gilt für alle $A \subset \bigcup_{j=1}^n B_j$

$$P(A) = \sum_{j=1}^{n} P(A|B_j)P(B_j).$$

(Sollte $P(B_j) = 0$ sein, so wird der entsprechende Summand $P(A|B_j)P(B_j)$ als Null definiert.)

Beweis. Wegen $A = \bigcup_{j=1}^n (A \cap B_j)$ und der Disjunktheit der $A \cap B_j$ gilt:

$$P(A) = P\left(\bigcup_{j=1}^{n}(A \cap B_j)\right) = \sum_{j=1}^{n} P(A \cap B_j) = \sum_{j=1}^{n} P(A|B_j)P(B_j).$$

□

Nun können wir auch das ursprüngliche Problem lösen. Gesucht ist $P(K|S)$ bei gegebenem $P(K) = 0.07; P(K^c) = 0.93; P(S|K) = 0.92; P(S|K^c) = 0.085$. Nun ist nach obigen Satz

$$\begin{aligned} P(K|S) &= \frac{P(K \cap S)}{P(S)} = \frac{P(K)P(S|K)}{P(K)P(S|K) + P(K^c)P(S|K^c)} \\ &= \frac{0.92 \times 0.07}{0.92 \times 0.07 + 0.085 \times 0.93} = 0.4489. \end{aligned}$$

Dies ist ein Spezialfall der sogenannten *Bayes-Formel:*

(2.7) Satz. Unter den Voraussetzungen von (2.6) und $P(A) > 0$ gilt

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^n P(A|B_j)P(B_j)}.$$

Die von *Thomas Bayes* (1702-1761) hergeleitete Formel wurde 1763 veröffentlicht.
Beweis.

$$\begin{aligned} P(B_i|A) &= \frac{P(B_i \cap A)}{P(A)} \\ &= \frac{P(A|B_i)P(B_i)}{P(A)} \\ &= \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{n} P(A|B_j)P(B_j)} \end{aligned}$$

nach Satz (2.6). □

Wird die Wahrscheinlichkeit für ein Ereignis A durch ein anderes Ereignis B mit $P(B) > 0$ nicht beeinflußt, im Sinne, daß $P(A|B) = P(A)$ gilt, so heißen A und B unabhängig. Es ist bequemer, dies symmetrisch in A und B zu definieren und auf die Voraussetzung $P(B) > 0$ zu verzichten:

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

Diese Definition spiegelt genau unsere intuitive Vorstellung von Unabhängigkeit wider. Es gilt offensichtlich $P(A|B) = P(A)$ dann und nur dann, wenn A und B unabhängig sind (vorausgesetzt, daß $P(B) > 0$ ist).

Unabhängigkeit von endlichen vielen Ereignissen wird wie folgt definiert:

(2.9) Definition. Die Ereignisse $A_1, \ldots, A_n$ heißen *unabhängig*, wenn für jede Auswahl von Indizes $\{i_1, \ldots, i_k\} \subset \{1, \ldots, n\}$ gilt:

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

(2.10) Bemerkungen.

1. Sind $A_1, \ldots, A_n$ unabhängige Ereignisse und ist $\{i_1, \ldots, i_m\}$ eine Teilmenge von $\{1, \ldots, n\}$, so sind offensichtlich $A_{i_1}, A_{i_2}, \ldots, A_{i_m}$ unabhängig.

2. Die Forderung $P(A_1 \cap \cdots \cap A_n) = P(A_1) \cdots P(A_n)$ allein ist keine befriedigende Definition der Unabhängigkeit (für $n \geq 3$), denn damit wäre die Eigenschaft in Teil (1) nicht erfüllt. Dazu ein Beispiel: Es seien $\Omega = \{1, 2\}$ und $p(1) = p(2) = 1/2$ sowie $A_1 = \{1\}$, $A_2 = \{2\}$ und $A_3 = \emptyset$. Dann gilt $P(A_1 \cap A_2 \cap A_3) = P(\emptyset) = 0 = P(A_1)P(A_2)P(A_3)$, aber natürlich ist $P(A_1 \cap A_2) \neq P(A_1)P(A_2)$.

3. Paarweise Unabhängigkeit, d.h. $P(A_i \cap A_j) = P(A_i)P(A_j)$ für $i \neq j$, impliziert nicht Unabhängigkeit. Wieder ein künstliches Beispiel dazu: Es seien $\Omega = \{1, 2, 3, 4\}$ und $p(i) = 1/4$ für jedes $i \in \Omega$ sowie $A_1 = \{1, 2\}$, $A_2 = \{2, 3\}$ und $A_3 = \{3, 1\}$. Dann ist $P(A_1 \cap A_2 \cap A_3) = 0 \neq P(A_1)P(A_2)P(A_3)$; jedoch sind A_1, A_2, A_3 paarweise unabhängig.

4. Die Ausdrucksweise „Die Ereignisse $A_1, \ldots, A_n$ sind unabhängig“, die auch hier verwendet wird, ist nicht ganz genau und führt in gewissen Situation zu Mißverständnissen. Unabhängigkeit ist keine Eigenschaft von Mengen von Ereignissen, sondern eine Eigenschaft von n-Tupeln von Ereignissen, die allerdings nicht von der Reihenfolge dieser Ereignisse im Tupel abhängt. Für ein Ereignis A ist das 1-Tupel (A) nach unserer Definition stets unabhängig, das Paar (A, A) jedoch nicht. (A, A) ist genau dann unabhängig, wenn $P(A) = P(A \cap A) = P(A)P(A)$, d.h. $P(A) \in \{0, 1\}$ gilt.

Zur bequemen Formulierung des nachfolgenden Ergebnisses führen wir die Bezeichnung $A^1 := A$ für $A \subset \Omega$ ein, A^c ist wie üblich das Komplement.

(2.11) Lemma. *Die Ereignisse $A_1, \ldots, A_n$ sind genau dann unabhängig, wenn für alle $(k_1, \ldots, k_n) \in \{1, c\}^n$*

$$P\Big(\bigcap_{j=1}^{n} A_j^{k_j}\Big) = \prod_{j=1}^{n} P(A_j^{k_j})$$

gilt. Hierbei ist $\{1, c\}^n$ die Menge der n-Tupel mit den Komponenten 1 und c.

Beweis (I). Unter der Voraussetzung der Unabhängigkeit zeigen wir die obige Gleichung mit Induktion nach n:

$n = 1$: Offensichtlich gilt $P(A^1) = P(A^1)$ und $P(A^c) = P(A^c)$.

Induktionsschluß $n \to n+1$: Die Ereignisse $A_1, \ldots, A_{n+1}$ seien unabhängig. Wir beweisen die obige Gleichung (für $n+1$) mit Induktion nach der Anzahl m der Komplementzeichen in $(k_1, \ldots, k_{n+1})$. Für $m = 0$ folgt sie aus der Unabhängigkeit. Induktionsschluß $m \to m+1$ für $0 \le m < n+1$: Es seien $m + 1 \ge 1$ Komplementzeichen in $(k_1, \ldots, k_{n+1})$. Durch Permutation der Ereignisse können wir annehmen, daß $k_{n+1} = c$ ist.

$$P\Big(\bigcap_{j=1}^{n+1} A_j^{k_j}\Big) = P\Big(\bigcap_{j=1}^{n} A_j^{k_j} \cap A_{n+1}^c\Big) = P\Big(\bigcap_{j=1}^{n} A_j^{k_j}\Big) - P\Big(\bigcap_{j=1}^{n} A_j^{k_j} \cap A_{n+1}\Big).$$

Der erste Summand ist nach der Induktionsvoraussetzung an n gleich $\prod_{j=1}^{n} P(A_j^{k_j})$, der zweite nach der Induktionsvoraussetzung an m gleich $\big(\prod_{j=1}^{n} P(A_j^{k_j})\big)P(A_{n+1})$. Damit folgt, wie gewünscht,

$$P\Big(\bigcap_{j=1}^{n+1} A_j^{k_j}\Big) = \prod_{j=1}^{n+1} P(A_j^{k_j}).$$

(II) Wir zeigen die Umkehrung: Die obige Gleichung in (2.11) gelte für alle Tupel $(k_1, \ldots, k_n) \in \{1, c\}^n$. Wir zeigen die Unabhängigkeit von $A_1, \ldots, A_n$.

Sei $\{i_1, \ldots, i_k\} \subset \{1, \ldots, n\}$ und $\{j_1, \ldots, j_m\}$ sei das Komplement dieser Menge in $\{1, \ldots, n\}$. Dann läßt sich $A_{i_1} \cap \cdots \cap A_{i_k}$ als Vereinigung paarweise disjunkter Mengen wie folgt schreiben:

$$\bigcup_{(k_1,\ldots,k_m)\in\{1,c\}^m} A_{i_1} \cap \cdots \cap A_{i_k} \cap A_{j_1}^{k_1} \cap \cdots \cap A_{j_m}^{k_m}.$$

Die Wahrscheinlichkeit davon ist nach unserer Voraussetzung gleich

$$\sum_{(k_1,\ldots,k_m)\in\{1,c\}^m} P(A_{i_1})\cdots P(A_{i_k})P(A_{j_1}^{k_1})\cdots P(A_{j_m}^{k_m}) = P(A_{i_1})\cdots P(A_{i_k}).$$

□

Als Beispiel betrachten wir das übliche Modell für das n-malige Werfen einer Münze.

(2.12) Satz. Wir bezeichnen mit B_k das Ereignis, daß der k-te Wurf ,,Kopf" ist. Die Ereignisse $B_1,\ldots,B_n$ sind unabhängig.

Beweis. Es gilt $P(B_j) = P(B_j^c) = 1/2$ für alle $j \in \{1,\ldots,n\}$. Für jedes n-Tupel $(k_1,\ldots,k_n) \in \{1,c\}^n$ gilt $P(B_1^{k_1}\cap\cdots\cap B_n^{k_n}) = 2^{-n} = \prod_{j=1}^n P(B_j^{k_j})$. Nach (2.11) sind $B_1,\ldots,B_n$ unabhängig. □

Offenbar ist der n–fache Münzwurf äquivalent zu einem Zufallsexperiment, welches mit gleicher Wahrscheinlichkeit in Erfolg (abgekürzt durch E) oder Mißerfolg (abgekürzt durch M) endet und das wir n Mal unabhängig durchführen. Dieses Modell ist allerdings – wie schon in Kapitel 1 diskutiert – nicht immer realistisch. Die naheliegende Verallgemeinerung ist die, anzunehmen, daß E und M nicht notwendig gleich wahrscheinlich sind; das Ereignis E tritt mit Wahrscheinlichkeit $0 \le p \le 1$ auf. Der entsprechende W.-Raum ist $\Omega = \{E,M\}^n$, d. h. die Menge der E-M-Folgen der Länge n. Die Wahrscheinlichkeiten der Elementarereignisse $\omega = (\omega_1,\ldots,\omega_n) \in \Omega$ sind gegeben durch $p(\omega) = p^k(1-p)^{n-k}$, wobei k die Anzahl der E's in der Folge $\omega_1,\ldots,\omega_n$ bezeichnet (wir werden diesen Ansatz im nächsten Satz rechtfertigen).

(2.13) Definition. Das durch diesen W-Raum beschriebene Zufallsexperiment heißt *Bernoulli-Experiment* der Länge n mit ,,Erfolgswahrscheinlichkeit" p.

Wir wollen die Wahrscheinlichkeit von einigen besonders wichtigen Ereignissen im Bernoulli-Experiment berechnen. Für $k \in \{0,1,\ldots,n\}$ sei A_k das Ereignis, daß insgesamt k Erfolge eintreten. In unserer Beschreibung des Bernoulli-Experiments enthält A_k diejenigen Elementarereignisse, in denen k mal E vorkommt. Davon gibt es so viele, wie es Möglichkeiten gibt, die k erfolgreich ausgegangenen Experimente auszuwählen, also $\binom{n}{k}$. Jedes hat Wahrscheinlichkeit $p^k(1-p)^{n-k}$. Somit ist $P(A_k) = \binom{n}{k}p^k(1-p)^{n-k}$.

Diese Wahrscheinlichkeit kürzt man meist mit $b(k;n,p)$ ab. Die $b(k;n,p)$ sind erwartungsgemäß am größten, wenn k in der Nähe von np liegt. Für großes n sind sie jedoch klein (höchstens von der Größenordnung $1/\sqrt{n}$). Eine ausführliche Analyse der Größen $b(k;n,p)$ wird später gegeben werden.

Beispiel: Ein Würfel wird n-mal geworfen. Die Wahrscheinlichkeit dafür, daß k-mal die Sechs erscheint, ist $b(k;n,1/6)$.

Eine andere, äquivalente Möglichkeit die Binomialverteilung zu erhalten ist ein sogenanntes Urnenmodell.

(2.14) Beispiel *Ziehen mit Zurücklegen (sampling with replacement).*
Eine Schachtel (Urne) enthält r rote und s schwarze Kugeln. Es werden n Kugeln nacheinander zufällig entnommen. Dabei wird jede sofort wieder zurückgelegt und die Schachtel neu gemischt. Die Elementarereignisse seien die Rot-Schwarz-Folgen der Länge n. Es ist klar, daß unter idealen Bedingungen die einzelnen Ziehungen unabhängig sind, daß dies also ein Bernoulli-Experiment der Länge n mit Erfolgswahrscheinlichkeit $p = \frac{r}{r+s}$ ist. Die Wahrscheinlichkeit des Ereignisses A_k, genau k-mal Rot zu ziehen, ist somit

$$P(A_k) = \binom{n}{k} \left(\frac{r}{r+s}\right)^k \left(\frac{s}{r+s}\right)^{n-k}.$$

Eine eingehendere Betrachtung der obigen Beispiele legt die Vermutung nahe, daß Unabhängigkeit eng mit den sogenannten *Produkträumen* zusammenhängt. Wir werden dies gleich beweisen. Zunächst aber müssen wir sagen, was wir überhaupt unter einem Produktraum verstehen wollen. Dazu seien $(\Omega_1, p_1), \ldots, (\Omega_n, p_n)$ diskrete W.-Räume. Wir konstruieren daraus einen neuen W.-Raum (Ω, p) mit $\Omega = \Omega_1 \times \cdots \times \Omega_n$. Für jedes $\omega = (\omega_1, \ldots, \omega_n) \in \Omega$ definieren wir $p(\omega) = p_1(\omega_1)p_2(\omega_2)\cdots p_n(\omega_n)$. Offensichtlich gilt $\sum_{\omega\in\Omega} p(\omega) = 1$.

(2.15) Definition. (Ω, p) heißt der *Produktraum (product space)* der W.-Räume (Ω_i, p_i), $1 \le i \le n$.

Zu $A \subset \Omega_i$ definieren wir das Ereignis $A^{(i)} = \{(\omega_1, \ldots, \omega_n) \in \Omega : \omega_i \in A\} \subset \Omega$.

(2.16) Satz. Sind $A_i \subset \Omega_i$ für $1 \le i \le n$, so sind die Ereignisse $A_1^{(1)}, \ldots, A_n^{(n)}$ im W.-Raum (Ω, p) unabhängig.

Beweis. Es gilt $A_i^{(i)c} = \{\omega \in \Omega : \omega_i \in A_i^c\} = A_i^{c(i)}$. Die 2^n Gleichungen in Lemma (2.11) sind also nachgewiesen, wenn

$$P\big(A_1^{(1)} \cap \cdots \cap A_n^{(n)}\big) = P(A_1^{(1)}) \cdots P(A_n^{(n)})$$

für alle möglichen $A_i \subset \Omega_i$, $1 \le i \le n$, gilt. Die linke Seite dieser Gleichung ist gleich

$$\begin{aligned}\sum_{\omega\in A_1^{(1)}\cap\cdots\cap A_n^{(n)}} p(\omega) &= \sum_{\omega_1\in A_1} \cdots \sum_{\omega_n\in A_n} p_1(\omega_1)\cdots p_n(\omega_n) \\ &= \prod_{j=1}^n \sum_{\omega_j\in A_j} p_j(\omega_j) = \prod_{j=1}^n \sum_{\omega\in A_j^{(j)}} p(\omega) = \prod_{j=1}^n P(A_j^{(j)}).\end{aligned}$$

□

Der Produktraum liefert somit ein Modell für eine unabhängige Hintereinanderreihung von n einzelnen Zufallsexperimenten, insbesondere ist offenbar die oben eingeführte Binomialverteilung das Resultat eines n-fachen (nicht notwendig fairen) Münzwurfes. Sie spielt eine zentrale Rolle in der diskreten W.-Theorie.

Natürlich mag man einwenden, daß das Ziehen *mit* Zurücklegen für einige Anwendungen nicht besonders interessant ist. Beispielsweise wird man sich bei einer Meinungsumfrage tunlichst hüten, dieselbe Person mehrfach zu befragen. Das mathematische Modell hierfür liefert ein weiteres Urnenmodell:

(2.17) Beispiel *Ziehen ohne Zurücklegen (sampling without replacement).*
Wir betrachten dieselbe Situation wie in Beispiel (2.14) mit dem Unterschied, daß die gezogenen Kugeln nicht wieder zurückgelegt werden. Es muß nun natürlich $n \leq r+s$ sein. Die einzelnen Ziehungen sind nicht mehr unabhängig, da ihr Ausgang die Zusammensetzung der Schachtel und damit die nachfolgenden Ziehungen beeinflußt.

Sei A_k wieder das Ereignis, daß k rote Kugeln gezogen werden. Wir setzen voraus, daß $0 \leq k \leq r$ und $0 \leq n-k \leq s$ gilt, sonst ist A_k das unmögliche Ereignis. Um $P(A_k)$ zu bestimmen, muß ein geeigneter Wahrscheinlichkeitsraum festgelegt werden. Als Elementarereignis betrachten wir die Menge der n-elementigen Teilmengen der $r+s$ Kugeln. Wie viele darunter gehören zu A_k? Es gibt $\binom{r}{k}$ Möglichkeiten, die k Kugeln aus den roten auszuwählen, und $\binom{s}{n-k}$ Möglichkeiten für die schwarzen Kugeln, also enthält A_k genau $\binom{r}{k}\binom{s}{n-k}$ Elementarereignisse. Es gilt also

$$P(A_k) = \binom{r}{k}\binom{s}{n-k} \Big/ \binom{r+s}{n},$$

offensichtlich ein anderer Wert als im Modell mit Zurücklegen. Man nennt dies auch die *hypergeometrische Wahrscheinlichkeitsverteilung (hypergeometric probability distribution).*

Obschon die Binomialverteilung und die hypergeometrische Verteilung unterschiedliche Wahrscheinlichkeiten für das Ereignis k Erfolge zu haben liefern, kann man mutmaßen, daß der Unterschied klein ist, sofern r und s groß sind. Dies ist plausibel, denn in diesem Fall ist die Wahrscheinlichkeit, eine Kugel doppelt zu ziehen klein (und dies ist ja die Ursache für die Abhängigkeit der einzelnen Ziehungen bei der hypergeometrischen Verteilung). $P(A_k)$ (in der hypergeometrischen Verteilung) kann in der Tat durch die Wahrscheinlichkeit $b(k;n,p)$ mit $p = r/(r+s)$ angenähert werden, sofern $n = r+s$ groß ist. Genauer:

(2.18) Satz.

$$\lim_{\substack{r,s\to\infty \\ r/(r+s)\to p}} \binom{r}{k}\binom{s}{n-k} \Big/ \binom{r+s}{n} = \binom{n}{k} p^k (1-p)^{n-k}.$$

Beweis. Die Größen auf der linken Seite sind gleich

$$\frac{n!}{k!(n-k)!} \, \frac{r(r-1)\cdots(r-k+1)s(s-1)\cdots(s-n+k+1)}{(r+s)(r+s-1)\cdots(r+s-n+1)}$$
$$\to \binom{n}{k} p^k (1-p)^{n-k} \quad \text{für } r,s \to \infty, \ \frac{r}{r+s} \to p.$$

□

Wir schließen dieses Kapitel mit einer Anwendung der bedingten Wahrscheinlichkeiten bei genetischen Modellen:

Hardy-Weinberg Theorem : Gene sogenannter ,,diploide“ Organismen treten paarweise auf und sind die Träger der vererblichen Eigenschaften. In einem einfachen Fall nehmen die Gene zwei Formen an, die man die Allele A und a nennt. Als Kombinationen sind dann die *Genotypen* AA, Aa und aa möglich. Zu einem bestimmten Zeitpunkt sei nun in einer Bevölkerung der Genotyp AA mit relativer Häufigkeit $u > 0$ vorhanden, der Genotyp Aa mit der relativen Häufugkeit $2v > 0$, und aa mit relativer Häufigkeit $w > 0$. Dann ist $u + 2v + w = 1$. Wir nehmen an, daß das Gen nicht geschlechtsgebunden ist. Bei jeder Fortpflanzung überträgt jedes Elternteil ein Gen seines Genpaares, und zwar mit Wahrscheinlichkeit $1/2$ auf den Nachkommen und für beide Elternteile unabhängig voneinander (zufällige Zeugung). Bei unabhängiger Auswahl von Mutter und Vater beträgt die Wahrscheinlichkeit, daß beide Genotyp AA haben, dann u^2. Die folgende Tabelle gibt die möglichen Kombinationen der Genotypen sowie die Wahrscheinlichkeit P_{AA} an, daß diese Kombination von Genotypen zu einem Nachkommen vom Genotyp AA führt:

Vater	Mutter	relative Häufigkeit	P_{AA}
AA	AA	u^2	1
AA	Aa	$2uv$	$1/2$
Aa	AA	$2uv$	$1/2$
Aa	Aa	$4v^2$	$1/4$

Mit der Formel von der totalen Wahrscheinlichkeit ergibt sich somit in der ersten Nachkommengeneration der Genotyp AA mit Wahrscheinlichkeit $P_1(AA) = (u+v)^2$. Analog ergibt sich $P_1(aa) = (w+v)^2$ und somit $P_1(Aa) = 1-(u+v)^2-(w+v)^2 = 2(u+v)(v+w)$. Wir fassen diese Wahrscheinlichkeiten als die relativen Häufigkeiten der nächsten Generation auf: $u_1 = (u+v)^2$, $2v_1 = 2(u+v)(v+w)$, $w_1 = (v+w)^2$. Dann folgt für die darauffolgende Generation $u_2 = (u_1+v_1)^2$, $2v_2 = 2(u_1+v_1)(v_1+w_1)$, $w_2 = (v_1+w_1)^2$. Durch Einsetzten sieht man $u_2 = ((u+v)^2 + (u+v)(v+w))^2 = (u+v)^2 = u_1$ und aus Symmetriegründen $w_2 = w_1$, und damit auch $v_2 = v_1$. Durch Induktion folgt für die k-te Generation:

$$u_k = (u+v)^2,\ 2v_k = 2(u+v)(v+w),\ w_k = (v+w)^2.$$

Die Häufigkeitsverteilung der Genotypen ist also in allen Nachkommengenerationen gleich. Diese Aussage stammt von dem Mathematiker *Godfrey Harold Hardy* (1877-1947) und dem Physiker *Wilhelm Weinberg* (1862-1937) aus dem Jahre 1908.

3 Zufallsgrößen, Gesetz der großen Zahlen

Zu Beginn dieses Kapitels sei noch einmal daran erinnert, wie wir im vergangenen Abschnitt vom Bernoulli-Experiment zur Binomialverteilung gekommen sind. Während das Bernoulli-Experiment auf dem Wahrscheinlichkeitsraum $(\Omega, \tilde{p})$ lebte, wobei $\Omega = \{0,1\}^N$ und $\tilde{p}(\omega) = p^k(1-p)^{n-k}$ für ein ω mit k Einsen war, war die Binomialverteilung $b(k;n,p)$ eine Wahrscheinlichkeit auf der Menge $\{0,\ldots,n\}$. Der Zusammenhang zwischen beiden ist der, daß man für $b(k;n,p)$ die Wahrscheinlichkeiten im Bernoulli-Experiment für sämtliche ω mit k Einsen quasi aufsammelt. Formal entspricht das einer Abbildung $X : \Omega \to \mathbb{N}$, wobei wir zusätzlich jedem $n \in \mathbb{N}$ die Summe der Wahrscheinlichkeiten seiner Urbilder zuordnen. Dies ist das Konzept der Zufallsvariablen.

(3.1) Definition Sei (Ω, p) ein (diskreter) W.-Raum. Dann heißt eine Abbildung $X : \Omega \to \mathbb{R}$ eine *(diskrete) Zufallsvariable oder Zufallsgröße ((discrete) random variable).*

Wir beobachten, daß für die formale Definition einer Zufallsvariablen p zunächst völlig belanglos ist. Eine Zufallsgröße ist einfach eine Abbildung und keine ,,zufällige" Abbildung. Natürlich werden wir jedoch nun die Eigenschaften von X im Zusammenhang mit p untersuchen. Die zentrale Idee hierbei wird immer sein, daß eine Zufallsvariable "wesentliche Eigenschaften" eines W.-Raumes herausfiltert.

Es bezeichne $X(\Omega)$ das Bild von Ω unter X, d. h. die höchstens abzählbare Menge reeller Zahlen $\{\, X(\omega) : \omega \in \Omega \,\}$. Für $A \subset \mathbb{R}$ ist $X^{-1}(A) = \{\, \omega \in \Omega : X(\omega) \in A \,\}$ eine Teilmenge von Ω, d. h. ein Ereignis. Wir nennen dies das Ereignis, ,,daß X einen Wert in A annimmt". Wir benutzen die folgenden Kurzschreibweisen:

$$\begin{aligned} \{X \in A\} &:= \{\, \omega \in \Omega : X(\omega) \in A \,\} = X^{-1}(A), \\ \{X = z\} &:= \{\, \omega \in \Omega : X(\omega) = z \,\} = X^{-1}(\{z\}), \\ \{X \le z\} &:= \{\, \omega \in \Omega : X(\omega) \le z \,\} = X^{-1}((-\infty, z]), \qquad \text{etc.} \end{aligned}$$

Statt $P(\{X \in A\})$, $P(\{X = z\})$ schreiben wir einfach $P(X \in A)$, $P(X = z)$, etc. Wir schreiben meistens ein Komma anstelle von ,,und" bzw. des mengentheoretischen Durchschnitts innerhalb der Klammer in $P(\quad)$. Sind etwa X, Y Zufallsgrößen und $A, B \subset \mathbb{R}$, so schreiben wir $P(X \in A,\, Y \in B)$ für $P(\{X \in A\} \cap \{Y \in B\})$ oder noch ausführlicher $P(\{\, \omega : X(\omega) \in A \text{ und } Y(\omega) \in B \,\})$.

(3.2) Beispiele.

1. Es sei X die Augensumme beim zweimaligen Werfen eines Würfels. Zur formalen Beschreibung dieses Versuchs betrachten wir den W.-Raum (Ω, p) mit $\Omega = \{1,2,3,4,5,6\}^2$ und der Gleichverteilung p, also $p((i,j)) = 1/36$ für alle $(i,j) \in \Omega$. Die Zufallsgröße $X : \Omega \to \mathbb{R}$ mit $X((i,j)) = i + j$ für alle $(i,j) \in \Omega$ beschreibt dann die Augensumme, und es gilt z. B.

$$P(X = 3) = P(\{(1,2),(2,1)\}) = 1/18$$

und

$$P(X \le 4) = P(\{(1,1),(1,2),(2,1),(1,3),(2,2),(3,1)\}) = 1/6.$$

2. Es bezeichne X die Anzahl der Erfolge in einem Bernoulli-Experiment der Länge n. Setzen wir $X_i = 1$, falls der i-te Versuch ein Erfolg ist, und $X_i = 0$ sonst $(1 \leq i \leq n)$, so folgt $X = \sum_{i=1}^{n} X_i$.

3. Für eine beliebige Teilmenge $A \subset \Omega$ definieren wir die *Indikatorfunktion* 1_A von A durch
$$1_A(\omega) = \begin{cases} 1 & \text{falls } \omega \in A, \\ 0 & \text{falls } \omega \notin A. \end{cases}$$

Sei $X : \Omega \to \mathbb{R}$ eine Zufallsgröße. Für $z \in X(\Omega)$ sei $f(z) := P(X = z)$. Da die Ereignisse $\{X = z\}$ für verschiedene $z \in X(\Omega)$ sich gegenseitig ausschließen und

$$\Omega = \bigcup_{z \in X(\Omega)} \{X = z\}$$

gilt, folgt

$$\sum_{z \in X(\Omega)} f(z) = 1.$$

$(X(\Omega), f)$ ist somit ein W.-Raum (dies entspricht der eingangs gemachten Beobachtung für die Binomialverteilung).

(3.3) Definition. f heißt die *Verteilung (distribution)* der Zufallsgröße X.

Aus der Verteilung einer Zufallsgröße läßt sich $P(X \in A)$ für jede Teilmenge A von $\mathbb{R}$ berechnen:

$$P(X \in A) = \sum_{z \in A \cap X(\Omega)} f(z).$$

Verteilungen sind jedoch oft kompliziert und in vielen praktisch wichtigen Beispielen nicht explizit berechenbar. Zunächst einige Beispiele, bei denen die Verteilung einfach angegeben werden kann:

Beispiel (3.2 (1)) (Augensumme beim zweimaligen Würfeln): $X(\Omega) = \{2, 3, 4, \ldots, 12\}$,

$$\begin{aligned} f(2) = f(12) &= \tfrac{1}{36}, & f(5) = f(9) &= \tfrac{1}{9}, \\ f(3) = f(11) &= \tfrac{1}{18}, & f(6) = f(8) &= \tfrac{5}{36}, \\ f(4) = f(10) &= \tfrac{1}{12}, & f(7) &= \tfrac{1}{6}. \end{aligned}$$

Binomialverteilte Zufallsgrößen:

Sei X die Anzahl der Erfolge in einem Bernoulli-Experiment der Länge n und Erfolgswahrscheinlichkeit p. Dann ist, wie wir schon in Kapitel 2 berechnet haben:

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

(3.4) Definition. Eine Zufallsgröße mit obiger Verteilung heißt *binomialverteilt* mit Parametern p und n.

Diese Zufallsvariablen werden im weiteren Verlauf der Vorlesung noch eine zentrale Rolle einnehmen.

Offensichtlich kann man auf einem W.-Raum (Ω, p) sehr viele verschiedene Zufallsvariablen definieren. Um diese zu unterscheiden, muß man ihre Verteilungen unterscheiden. Da sich die exakte Verteilung in vielen Beispielen nur schwer oder gar nicht explizit berechnen läßt, ist es wichtig, daß es gewisse Kenngrößen von Zufallsgrößen gibt, die oft einfacher zu berechnen oder abzuschätzen sind, und die wichtige Informationen über die Zufallsgröße enthalten. Die wichtigste dieser Größen ist der Erwartungswert, der angibt, wo die Zufallsgröße "im Mittel" liegt.

(3.5) Definition. Sei X eine Zufallsgröße. Man sagt, daß der *Erwartungswert (expected value, expectation) von X existiert*, falls $\sum_{z\in X(\Omega)} |z|P(X=z) < \infty$ ist. Der *Erwartungswert von X* ist dann definiert durch

$$E(X) = \sum_{z\in X(\Omega)} zP(X=z).$$

Wir definieren also $E(X)$ nur, wenn die Reihe absolut konvergiert. Der Wert der Reihe $\sum_{z\in X(\Omega)} zP(X=z)$ hängt dann nicht von der Reihenfolge der Summation ab. Es muß hervorgehoben werden, daß der Erwartungswert einer Zufallsgröße nur von deren Verteilung abhängt. Zwei verschiedene Zufallsgrößen mit derselben Verteilung haben also denselben Erwartungswert – unabhängig von ihrem Startraum (Ω, p). Wir lassen die Klammern oft weg und schreiben EX statt $E(X)$.

Man kann statt über $X(\Omega)$ auch über Ω summieren:

(3.6) Lemma. *Der Erwartungswert von X existiert genau dann, wenn die Reihe $\sum_{\omega\in\Omega} p(\omega)X(\omega)$ absolut konvergiert. In diesem Falle gilt $E(X) = \sum_{\omega\in\Omega} p(\omega)X(\omega)$.*

Beweis.

$$\begin{aligned}\sum_{z\in X(\Omega)} |z|P(X=z) &= \sum_{z\in X(\Omega)} |z| \sum_{\omega: X(\omega)=z} p(\omega) \\ &= \sum_{(z,\omega): X(\omega)=z} |z|p(\omega) = \sum_{\omega\in\Omega} |X(\omega)|p(\omega).\end{aligned}$$

Somit folgt der erste Teil der Behauptung; der zweite ergibt sich mit einer Wiederholung der obigen Rechnung ohne Absolutzeichen. □

(3.7) Satz.

1. Ist $c \in \mathbb{R}$ und X die konstante Abbildung nach c (d. h. $X(\omega) = c$ für alle $\omega \in \Omega$), so gilt $EX = c$.
2. $X_1, \ldots, X_n$ seien (auf einem gemeinsamen W.-Raum definierte) Zufallsgrößen, deren Erwartungswerte existieren, und $a_1, \ldots, a_n$ seien reelle Zahlen. Ferner sei $a_1X_1 + a_2X_2 + \cdots + a_nX_n$ die Zufallsgröße, deren Wert an der Stelle $\omega \in \Omega$ gleich $a_1X_1(\omega) + a_2X_2(\omega) + \cdots + a_nX_n(\omega)$ ist. Dann existiert $E(a_1X_1 + \cdots + a_nX_n)$ und ist gleich $a_1EX_1 + \cdots + a_nEX_n$. *("Der Erwartungswert ist linear".)*

3. X, Y seien Zufallsgrößen. Gilt $X \leq Y$ und existiert der Erwartungswert von Y, so gilt $EX \leq EY$. *("Der Erwartungswert ist monoton".)*

Beweis.
(1) und (3) sind nach der Definition des Erwartungswertes evident.
(2) Wir benutzen (3.6):

$$\sum_\omega p(\omega)|a_1X_1(\omega) + \ldots + a_nX_n(\omega)|$$
$$\leq |a_1| \sum_\omega p(\omega)|X_1(\omega)| + \ldots + |a_n| \sum_\omega p(\omega)|X_n(\omega)| < \infty.$$

Somit existiert der Erwartungswert und es gilt

$$\begin{aligned} E(a_1X_1 + \cdots + a_nX_n) &= \sum_\omega p(\omega)(a_1X_1(\omega) + \cdots + a_nX_n(\omega)) \\ &= a_1 \sum_\omega p(\omega)X_1(\omega) + \ldots + a_n \sum_\omega p(\omega)X_n(\omega) \\ &= a_1EX_1 + \ldots + a_nEX_n. \end{aligned}$$

□

Bemerkung. Die Menge aller Zufallsgrößen, die auf Ω definiert sind, ist einfach $\mathbb{R}^\Omega$ und in natürlicher Weise ein $\mathbb{R}$-Vektorraum. Die Menge der Zufallsgrößen, deren Erwartungswert existiert, ist nach (3.7 (2)) ein Unterraum von $\mathbb{R}^\Omega$. Man bezeichnet ihn oft als $L_1(\Omega, p)$. Der Erwartungswert ist eine lineare Abbildung von $L_1(\Omega, p)$ nach $\mathbb{R}$, also ein Element des Dualraumes von $L_1(\Omega, p)$.

(3.8) Beispiele.

1. Der Erwartungswert der Indikatorfunktion 1_A von $A \subset \Omega$ ist $E(1_A) = P(A)$, denn $A = \{\omega : 1_A(\omega) = 1\}$ und also $E(1_A) = 0 \cdot P(A^c) + 1 \cdot P(A)$.

2. X *binomialverteilt* mit Parametern p, n:
   Wir schreiben X als $X_1 + \cdots + X_n$, wobei $X_i = 1$ ist, wenn der i-te Versuch von Erfolg gekrönt war, und andernfalls $X_i = 0$. Es gilt $E(X_i) = P(X_i = 1) = p$ und somit $E(X) = np$.

Die alleinige Kenntnis von Erwartungswerten ist im allgemeinen wenig nützlich, wenn nicht gleichzeitig bekannt ist, daß die Zufallsgröße mit hoher Wahrscheinlichkeit „nahe“ beim Erwartungswert liegt.

Dazu ein Beispiel: Ist $P(X = 0) = P(X = 1) = 1/2$, so ist $EX = 1/2$, aber dies gibt im Grunde wenig Information über X. Anderseits: Sei X die mittlere Anzahl der Kopfwürfe bei einem Münzwurf-Experiment der Länge 1000, d. h. die Anzahl der Kopfwürfe / 1000. Aus Beispiel (3.8 (2)) wissen wir, daß ebenfalls $EX = 1/2$ gilt. Jedermann „ist bekannt“,

daß X mit großer Wahrscheinlichkeit nahe bei $1/2$ liegt. Dies ist der Inhalt des Gesetzes der großen Zahlen, das wir weiter unten gleich diskutieren und beweisen werden. Die Verteilung von X ist hier ziemlich scharf um EX konzentriert. Ohne solche ,,Maßkonzentrationsphänomene“ wäre jede statistische Umfrage beispielsweise sinnlos.

Ein Maß für die Abweichung, die eine Zufallsgröße von ihrem Erwartungswert hat, ist die sogenannte Varianz:

(3.9) Definition. Es sei X eine Zufallsgröße mit existierendem Erwartungswert EX. Dann heißt

$$V(X) := \sum_{z \in X(\Omega)} (z - EX)^2 P(X = z)$$

die *Varianz (variance)* von X und $S(X) := +\sqrt{V(X)}$ die *Standardabweichung (standard deviation)* von X, falls die auftretende (möglicherweise unendliche) Reihe konvergiert. Die Varianz ist stets nicht negativ, da die Glieder in der obigen Reihe alle größer oder gleich Null sind. Man sagt oft auch, die Varianz sei unendlich, wenn die Reihe divergiert.

Für die Diskussion der Varianz und auch in anderen Zusammenhängen ist die nachstehende Folgerung aus (3.6) nützlich:

(3.10) Lemma. *$X_1, \ldots, X_k$ seien (auf einem gemeinsamen W.-Raum definierte) Zufallsgrößen, und g sei eine Abbildung von $X_1(\Omega) \times \cdots \times X_k(\Omega)$ nach $\mathbb{R}$. Dann ist $X := g(X_1, \ldots, X_k) = g \circ (X_1, \ldots, X_k)$ eine Zufallsgröße, deren Erwartungswert genau dann existiert, wenn*

$$\sum_{x_1 \in X_1(\Omega)} \cdots \sum_{x_k \in X_k(\Omega)} |g(x_1, \ldots, x_k)| P(X_1 = x_1, \ldots, X_k = x_k) < \infty$$

gilt. In diesem Fall gilt

$$E(X) = \sum_{x_1 \in X_1(\Omega)} \cdots \sum_{x_k \in X_k(\Omega)} g(x_1, \ldots, x_k) P(X_1 = x_1, \ldots, X_k = x_k).$$

Beweis. Wir betrachten den neuen W.-Raum (Ω', p') mit $\Omega' = X_1(\Omega) \times \cdots \times X_k(\Omega)$ und $p'(x_1, \ldots, x_k) = P(X_1 = x_1, \ldots, X_k = x_k)$. Auf diesem W.-Raum definieren wir die Zufallsgröße $g : \Omega' \to \mathbb{R}$. Für $z \in g(\Omega') = X(\Omega)$ gilt

$$P'(g = z) = \sum_{\substack{(x_1, \ldots, x_k) \in \Omega' \\ g(x_1, \ldots, x_k) = z}} p'(x_1, \ldots, x_k) = \sum_{\substack{\omega \in \Omega \\ X(\omega) = z}} p(\omega) = P(X = z).$$

g und X haben also dieselbe Verteilung. Unser Lemma folgt nun sofort aus (3.6). □

(3.11) Lemma.

1. *$V(X)$ ist der Erwartungswert der Zufallsgröße $\omega \mapsto (X(\omega) - EX)^2$.*

2. *$V(X)$ existiert genau dann, wenn $E(X^2)$ existiert.*

3. *Existiert $V(X)$, so gibt $V(X) = E(X^2) - (EX)^2$.*

4. *Für $a, b \in \mathbb{R}$ gilt $V(a + bX) = b^2 V(X)$.*

5. *Sind X und Y Zufallsgrößen, deren Varianzen existieren, so existiert die Varianz von $X + Y$.*

Beweis.

1. folgt aus (3.10) mit $k = 1$ und $g(x) = (x - EX)^2$.

2. Falls $V(X)$ existiert, so existiert EX (per Definition).
   Wegen $z^2 \leq 2(EX)^2 + 2(z - EX)^2$ für $z \in \mathbb{R}$ folgt

$$\sum_{z \in X(\Omega)} z^2 P(X = z) \leq 2(EX)^2 + 2 \sum_{z \in X(\Omega)} (z - EX)^2 P(X = z) < \infty.$$

   Nach (3.10) existiert dann $E(X^2)$.
   Falls $E(X^2)$ existiert, so folgt

$$\begin{aligned} \sum_{z \in X(\Omega)} |z| P(X = z) &= \sum_{\substack{z \in X(\Omega) \\ |z| \leq 1}} |z| P(X = z) + \sum_{\substack{z \in X(\Omega) \\ |z| > 1}} |z| P(X = z) \\ &\leq 1 + \sum_{z \in X(\Omega)} z^2 P(X = z) < \infty. \end{aligned}$$

   Somit existiert EX. Wegen $(z - EX)^2 \leq 2(EX)^2 + 2z^2$ folgt die Existenz von $V(X)$ wie oben.

3. $V(X) = E((X - EX)^2) = E(X^2 - 2(EX)X + (EX)^2) =$
   $E(X^2) - 2EX \times EX + (EX)^2 = E(X^2) - (EX)^2$.

4. folgt sofort aus (1) und der Linearität des Erwartungswertes.

5. Es gilt $(X(\omega) + Y(\omega))^2 \leq 2X(\omega)^2 + 2Y(\omega)^2$ für alle $\omega \in \Omega$. Nach (2) folgt dann die Existenz von $V(X + Y)$.

□

Im allgemeinen gilt $V(X + Y) \neq V(X) + V(Y)$ (die Varianz ist also nicht linear). Eine einfache Rechnung ergibt nämlich

$$\begin{aligned} V(X + Y) &= E\big(((X + Y) - E(X + Y))^2\big) \\ &= E\big((X - EX)^2\big) + E\big((Y - EY)^2\big) + 2E\big((X - EX)(Y - EY)\big) \\ &= V(X) + V(Y) + 2E\big((X - EX)(Y - EY)\big), \end{aligned} \tag{3.1}$$

und der letzte Summand ist in vielen Fällen ungleich Null, z. B. für $X = Y$, $V(X) \neq 0$. Dennoch ist der Fall, wo für zwei Zufallsgrößen X und Y die Gleichung $V(X+Y) = V(X) + V(Y)$ gilt, von besonderem Interesse. Wir werden dies weiter unten diskutieren.

(3.12) Definition. Sind X und Y zwei Zufallsgrößen, so wird die *Kovarianz (covariance)* zwischen X und Y definiert durch $\operatorname{cov}(X,Y) = E((X - EX)(Y - EY))$, falls alle in diesem Ausdruck vorkommenden Erwartungswerte existieren.

(3.13) Bemerkung. Eine analoge Überlegung wie im Beweis von (3.11 (2)) zeigt, daß $\operatorname{cov}(X,Y)$ genau dann existiert, wenn $E(X)$, $E(Y)$ und $E(XY)$ existieren. In diesem Fall gilt

$$\operatorname{cov}(X,Y) = E(XY) - E(X)E(Y).$$

(3.14) Lemma. *Seien X und Y Zufallsgrößen, für die $\operatorname{cov}(X,Y)$ existiert. Dann gelten $\operatorname{cov}(X,Y) = \operatorname{cov}(Y,X)$ und $\operatorname{cov}(\lambda X, \mu Y) = \lambda\mu \operatorname{cov}(X,Y)$ für alle $\lambda, \mu \in \mathbb{R}$.*

Beweis. Definition und Linearität des Erwartungswerts. □

Die Gleichung (3.1) kann wie folgt verallgemeinert werden:

(3.15) Satz. Seien $X_1, \ldots, X_n$ Zufallsgrößen mit existierenden Varianzen und Kovarianzen. Dann gilt

$$V\Big(\sum_{i=1}^{n} X_i\Big) = \sum_{i=1}^{n} V(X_i) + \sum_{\substack{i,j=1 \\ i\neq j}}^{n} \operatorname{cov}(X_i, X_j).$$

Beweis.

$$\begin{aligned} V\Big(\sum_{i=1}^{n} X_i\Big) &= E\bigg(\Big(\sum_{i=1}^{n} X_i - E\Big(\sum_{i=1}^{n} X_i\Big)\Big)^2\bigg) = E\bigg(\Big(\sum_{i=1}^{n}(X_i - EX_i)\Big)^2\bigg) \\ &= \sum_{i,j=1}^{n} E\big((X_i - EX_i)(X_j - EX_j)\big) = \sum_{i=1}^{n} V(X_i) + \sum_{\substack{i,j=1 \\ i\neq j}}^{n} \operatorname{cov}(X_i, X_j). \end{aligned}$$

□

(3.16) Satz. Existieren $V(X)$ und $V(Y)$, so existiert $\operatorname{cov}(X,Y)$ und es gilt

$$|\operatorname{cov}(X,Y)| \leq S(X)S(Y) \qquad (S(X) := +\sqrt{V(X)}).$$

Beweis. Für alle $\omega \in \Omega$ gilt $2|X(\omega)Y(\omega)| \leq X^2(\omega) + Y^2(\omega)$. Daraus und aus (3.11 (2)) folgt die Existenz von $E(XY)$ und nach der Bemerkung (3.13) auch die von $\operatorname{cov}(X,Y)$. Für $\lambda, \mu \in \mathbb{R}$ folgt aus (3.14) und (3.15):

$$0 \leq V(\lambda X + \mu Y) = \lambda^2 V(X) + 2\lambda\mu \operatorname{cov}(X,Y) + \mu^2 V(Y).$$

Als Funktion von $(\lambda, \mu) \in \mathbb{R}^2$ definiert dies also eine positiv semidefinite quadratische Form. Demzufolge ist

$$\det \begin{pmatrix} V(X) & \operatorname{cov}(X,Y) \\ \operatorname{cov}(X,Y) & V(Y) \end{pmatrix} \geq 0.$$

Dies impliziert die Aussage. □

(3.17) Bemerkung. Der Vollständigkeit halber sei auf den folgenden Sachverhalt hingewiesen. Die Existenz von $\operatorname{cov}(X,Y)$ setzt nach (3.13) die Existenz von EX, EY und $E(XY)$ voraus und folgt nach dem obigen Satz aus der Existenz von $V(X)$ und $V(Y)$. Letzteres ist jedoch dafür nicht notwendig: Es gibt Zufallsgrößen mit existierender Kovarianz, deren Varianzen nicht existieren.

(3.18) Definition. Die Zufallsgrößen X und Y heißen *unkorreliert (uncorrelated)*, wenn $\operatorname{cov}(X,Y)$ existiert und gleich null ist. Sind die Zufallsgrößen $X_1, \ldots, X_n$ paarweise unkorreliert und existeren die Varianzen, so gilt nach (3.15)

$$V\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} V(X_i)$$

(*Gleichheit nach Irénée Jules Bienaymé* (1796-1878)). Die für uns zunächst wichtigste Klasse von unkorrelierten Zufallsgrößen sind unabhängige:

(3.19) Definition. n diskrete Zufallsgrößen $X_1, \ldots, X_n$ heißen *unabhängig*, wenn

$$P(X_1 = z_1, \ldots, X_n = z_n) = P(X_1 = z_1) \cdots P(X_n = z_n)$$

für alle $z_i \in X_i(\Omega)$, $i \in \{1, \ldots, n\}$ gilt.

Der folgende Satz stellt einen Zusammenhang zwischen der Unabhängigkeit von Zufallsvariablen und der Unabhängigkeit von Ereignissen her.

(3.20) Satz. Die folgenden vier Aussagen über die diskreten Zufallsgrößen $X_1, X_2, \ldots, X_n$ sind äquivalent

(a) $X_1, \ldots, X_n$ sind unabhängig.

(b) Für alle $A_1, \ldots, A_n \subset \mathbb{R}$ gilt

$$P(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n) = P(X_1 \in A_1) \times \cdots \times P(X_n \in A_n).$$

(c) Für alle $A_1, \ldots, A_n \subset \mathbb{R}$ sind die Ereignisse $\{X_1 \in A_1\}, \ldots, \{X_n \in A_n\}$ unabhängig.

(d) Für $z_1 \in X_1(\Omega), \ldots, z_n \in X_n(\Omega)$ sind die Ereignisse $\{X_1 = z_1\}, \ldots, \{X_n = z_n\}$ unabhängig.

Beweis. (a)⇒(b): Summation der Gleichung in (3.20) über $(z_1, \ldots, z_n) \in A_1 \times A_2 \times \cdots \times A_n$.
(b)⇒(c): Nach (2.11) ist zu zeigen, daß für $(i_1, \ldots, i_n) \in \{1, c\}^n$ die Gleichung

$$P\left(\bigcap_{j=1}^{n} \{X_j \in A_j\}^{i_j}\right) = \prod_{j=1}^{n} P(\{X_j \in A_j\}^{i_j})$$

gilt, wobei $\{X_j \in A_j\}^1 := \{X_j \in A_j\}$ ist. Nun ist jedoch $\{X_j \in A_j\}^c = \{X_j \in A_j^c\}$. Wir können deshalb einfach (b) mit A_j oder A_j^c anstelle von A_j anwenden.
(c)⇒(d) ist trivial und (d)⇒(a) ergibt sich aus der Definition. □

(3.21) Satz. Sind die Zufallsgrößen $X_1, \ldots, X_n$ unabhängig, und sind $f_i : \mathbb{R} \to \mathbb{R}$ für $i = 1, \ldots, n$ beliebige Funktionen, so sind die Zufallsgrößen $Y_i = f_i \circ X_i$, $i = 1, \ldots, n$, unabhängig.

Beweis. Für beliebige $y_1, \ldots, y_n \in \mathbb{R}$ sei $A_i = \{x_i \in \mathbb{R} : f_i(x_i) = y_i\}$. Dann ist $\{Y_i = y_i\} = \{X_i \in A_i\}$. Die Aussage folgt somit aus Satz (3.20). □

(3.22) Satz. Zwei unabhängige Zufallsgrößen, deren Erwartungswerte existieren, sind unkorreliert.

Beweis. Sind X und Y unabhängig, so folgt

$$\begin{aligned}\sum_{x \in X(\Omega)} \sum_{y \in Y(\Omega)} |xy| P(X = x, Y = y) &= \sum_x \sum_y |x|\,|y| P(X = x) P(Y = y) \\ &= \left(\sum_x |x| P(X = x)\right)\left(\sum_y |y| P(Y = y)\right) < \infty.\end{aligned}$$

Nach (3.10) mit $k = 2$ und $g(x, y) = xy$ folgt die Existenz von $E(XY)$. Eine Wiederholung der obigen Rechnung ohne Absolutzeichen ergibt $E(XY) = E(X)E(Y)$. Nach (3.13) folgt daraus die Unkorreliertheit von X und Y. □

(3.23) Bemerkung. Derselbe Beweis ergibt für n Zufallsgrößen $X_1, \ldots, X_n$, die unabhängig sind und deren Erwartungswerte existieren, daß der Erwartungswert von $\prod_{i=1}^n X_i$ existiert und gleich $\prod_{i=1}^n EX_i$ ist.

(3.24) Beispiele.

1. Wir betrachten ein Bernoulli-Experiment mit Parametern n, p und setzen $X_i = 1$, falls der i-te Versuch ein Erfolg ist, und $X_i = 0$ sonst ($1 \le i \le n$). Dann gilt $V(X_i) = E(X_i^2) - (EX_i)^2 = p - p^2 = p(1 - p)$. Die Unabhängigkeit von $X_1, \ldots, X_n$ folgt aus der Definition. Nach (3.22) sind die X_i paarweise unkorreliert. Nach (3.15) folgt für die Anzahl $X = \sum_{i=1}^n X_i$ der Erfolge

$$V(X) = \sum_{i=1}^{n} V(X_i) = np(1 - p)$$

und somit $S(X) = \sqrt{np(1-p)}$.

2. Um an einem Beispiel zu zeigen, daß die Umkehrung von (3.22) nicht gilt, wählen wir $\Omega = \{-1, 0, 1\}$ mit der Gleichverteilung und definieren die Zufallsgröße X durch $X(\omega) = \omega$ für alle $\omega \in \Omega$. Dann gelten $E(X) = 0$, $E(|X|) = 2/3$ und $E(X|X|) = 0$, also sind X und $|X|$ nach (3.13) unkorreliert. Offensichtlich sind X und $|X|$ aber abhängig, denn zum Beispiel ist $\{X = 1, |X| = 0\}$ das unmögliche Ereignis, aber $P(X = 1)P(|X| = 0)$ ist gleich $1/9$.

3. Ein Stapel mit n numerierten Karten wird zufällig in eine Reihe gelegt. Alle $n!$ Möglichkeiten mögen gleich wahrscheinlich sein. S_n bezeichne nun die Anzahl der Karten, die in Bezug auf die natürliche Anordnung an „ihrem“ Platz liegen. S_n nimmt also Werte in $\{0, 1, \ldots, n\}$ an. In einer Übungsaufgabe wird die Verteilung von S_n bestimmt. Von ihr kann man Erwartungwert und Varianz ableiten. Wir berechnen diese Werte hier direkt: Dazu sei X_k die Zufallsgröße mit Werten 1 oder 0 je nachdem, ob die Karte mit der Nummer k am k-ten Platz liegt oder nicht. Dann ist $S_n = X_1 + X_2 + \cdots + X_n$. Jede Karte ist mit Wahrscheinlichkeit $1/n$ am k-ten Platz, also ist $P(X_k = 1) = 1/n$ und $P(X_k = 0) = (n-1)/n$ und somit $E(X_k) = 1/n$. Damit folgt $E(S_n) = 1$. Im Durchschnitt liegt also eine Karte an ihrem Platz. Weiter ist $V(X_k) = 1/n - (1/n)^2 = (n-1)/n^2$. Das Produkt X_jX_k nimmt die Werte 0 und 1 an. Der Wert 1 entspricht dem Ereignis, daß die Karten mit Nummer j und k an ihrem Platz liegen, was mit Wahrscheinlichkeit $1/n(n-1)$ geschieht. Daher ist $E(X_jX_k) = 1/(n(n-1))$. Nach Bemerkung (3.13) ist $\operatorname{cov}(X_j, X_k) = 1/(n(n-1)) - 1/n^2 = 1/(n^2(n-1))$. Nach Satz (3.15) folgt damit

$$V(S_n) = n\frac{n-1}{n^2} + 2\binom{n}{2}\frac{1}{n^2(n-1)} = 1.$$

Die Standardabweichung ist ein Maß dafür, wie weit X von $E(X)$ mit nicht zu kleiner Wahrscheinlichkeit abweichen kann. Diese sehr vage Aussage wird durch die sogenannte *Tschebyscheff-Ungleichung* präzisiert. *Pafnuty Lwowitsch Tschebyscheff* (1821-1894) bewies diese Ungleichung 1867. Wir beweisen zunächst eine etas allgemeinere Version dieser Ungleichung, die später noch nützlich sein wird:

(3.25) Satz. *(Markoff-Ungleichung, Markov-inequality)* Es sei ϕ eine auf $[0, \infty)$ definierte, nichtnegative monoton wachsende Funktion. Es sei X eine Zufallsgröße, für die der Erwartungswert $E(\phi(|X|))$ existiert. Dann gilt für jedes $a > 0$ mit
$\phi(a) > 0$

$$P(|X| \geq a) \leq \frac{E(\phi(|X|))}{\phi(a)}.$$

Beweis.

$$\begin{aligned} P(|X| \geq a) &= \sum_{\substack{x \in X(\Omega) \\ |x| \geq a}} P(X = x) \leq \sum_{\substack{x \in X(\Omega) \\ \phi(|x|) \geq \phi(a)}} \frac{\phi(|x|)}{\phi(a)} P(X = x) \\ &\leq \sum_{x \in X(\Omega)} \frac{\phi(|x|)}{\phi(a)} P(X = x) = \frac{E(\phi(|X|))}{\phi(a)}. \end{aligned}$$

□

(3.26) Satz. *(Tschebyscheff-Ungleichung, Chebyshev-inequality)* Sei X eine Zufallsgröße, deren Erwartungswert EX und Varianz $V(X)$ existieren. Dann gilt für jedes $a > 0$

$$P(|X - EX| \geq a) \leq \frac{V(X)}{a^2}.$$

Beweis. Mit $\phi(x) = x^2$ folgt aus Satz (3.25)

$$P(|X - EX| \geq a) = P((X - EX)^2 \geq a^2) \leq \frac{1}{a^2} E((X - EX)^2) = \frac{V(X)}{a^2}.$$

□

Beispiel: Sei $a > 0$ und X eine Zufallsgröße, die als Werte $-a$, $+a$ und 0 annimmt und deren Verteilung gegeben ist durch $P(X = -a) = P(X = +a) = 1/(2a^2)$ und $P(X = 0) = 1 - 1/a^2$. Wir erhalten $E(X) = 0$ und $V(X) = 1$ und damit

$$P(|X - E(X)| \geq a) = P(|X| \geq a) = P(X = -a) + P(X = +a) = \frac{1}{a^2}.$$

Dieses Beispiel zeigt, daß die Tschebyscheff-Ungleichung im allgemeinen nicht verbessert werden kann. Dennoch ist sie in vielen Fällen keine sehr gute Abschätzung. Für viele Zufallsgrößen können Abweichungen vom Erwartungswert sehr viel besser als mit der Tschebyscheff-Ungleichung abgeschätzt werden. Wir werden dies in einem der nächsten Kapitel intensiver diskutieren.

Die Tschebyscheff-Ungleichung ist gut genug, um das nachfolgende Gesetz der großen Zahlen zu beweisen. Es wurde vermutlich bereits im Jahre 1689 von *Jakob Bernoulli* (1654-1705) für den Fall des n-maligen Münzwurfes bewiesen. Dieses Theorem steht in der *Ars conjectandi*, welche erst acht Jahre nach Bernoullis Tod, mit einem Vorwort seines Neffen Nikolaus versehen, 1713 in Basel erschien:

(3.27) Satz. *(Schwaches Gesetz der großen Zahlen, weak law of large numbers)* Es seien für jedes $n \in \mathbb{N}$ auf einem diskreten Wahrscheinlichkeitsraum paarweise unkorrelierte Zufallsgrößen $X_1, X_2, \ldots, X_n$ gegeben, die von n abhängen dürfen, die aber alle den gleichen Erwartungswert E und die gleiche Varianz V besitzen. Sei $S_n := X_1 + \cdots + X_n$, und $\bar{S}_n = \frac{S_n}{n}$ sei die Folge der Mittelwerte. Dann gilt für jedes $\varepsilon > 0$

$$\lim_{n \to \infty} P(|\bar{S}_n - E| \geq \varepsilon) = 0.$$

Beweis. Aus (3.26), (3.11 (4)) und (3.15) folgt

$$P(|\bar{S}_n - E| \geq \varepsilon) \leq \frac{1}{\varepsilon^2} V(\bar{S}_n) = \frac{1}{n^2 \varepsilon^2} V(S_n) = \frac{1}{n^2 \varepsilon^2} nV \to 0 \quad \text{für } n \to \infty.$$

□

Interpretation. Falls wir beliebig oft ein Experiment wiederholen und annehmen, daß die Ergebnisse (Zufallsgrößen) paarweise voneinander unabhängig oder mindestens unkorreliert sind, so ist die Wahrscheinlichkeit für ein Abweichen der Mittelwerte der ersten n Experimente vom Erwartungswert schließlich (d. h. für hinreichend große n) beliebig klein.

(3.28) Bemerkung. Die Voraussetzungen des Satzes muten etwas umständlich an. Wieso setzen wir nicht einfach voraus, daß $(X_i)_{i\in\mathbb{N}}$ eine Folge von unkorrelierten Zufallsgrößen ist? Die Antwort ist einfach, daß wir (im Moment) keine Möglichkeiten haben, eine derartige unendliche Folge auf einem abzählbaren Wahrscheinlichkeitsraum zu definieren (außer im ganz trivialen Fall, wo die X_i alle konstant sind). Im Satz (3.27) setzen wir jedoch nur voraus, daß für jedes n ein W.-Raum $\Omega^{(n)}$ existiert, auf dem die $X_1, \ldots, X_n$ existieren. Wenn wir ganz pedantisch wären, sollten wir deshalb $X_1^{(n)}, \ldots, X_n^{(n)}$ schreiben. Es macht keine Schwierigkeiten, eine solche Folge von W.-Räumen und die dazugehörenden Zufallsgrößen als mathematisch präzis definierte Objekte zu konstruieren:

Es seien $f_1, \ldots, f_n$ beliebige W.-Verteilungen auf abzählbaren Teilmengen A_i von $\mathbb{R}$ (d. h. $f_i : A_i \to [0,1]$ mit $\sum_{x\in A_i} f_i(x) = 1$). Wir konstruieren einen W.-Raum (Ω, p) und unabhängige Zufallsgrößen X_i mit $X_i(\Omega) = A_i$ und Verteilungen f_i wie folgt:

Sei $\Omega = A_1 \times \cdots \times A_n$. Für $\omega = (\omega_1, \ldots, \omega_n) \in \Omega$ setzen wir $X_i(\omega) = \omega_i$ für alle i in $\{1, \ldots, n\}$ und $p(\omega) = f_1(\omega_1) f_2(\omega_2) \cdots f_n(\omega_n)$. Per Konstruktion sind $X_1, \ldots, X_n$ unabhängig, also auch unkorreliert. Haben die f_i alle denselben Erwartungswert und dieselbe Varianz (z. B. wenn sie alle gleich sind), so haben die X_i alle denselben Erwartungswert und dieselbe Varianz. Diese Konstruktion können wir für jedes n durchführen.

Der Satz (3.27) läßt sich natürlich auf binomialverteilte Zufallsgrößen anwenden, denn diese lassen sich ja in der Form $X_1 + \cdots + X_n$ schreiben, wobei die $X_1, \ldots, X_n$ unabhängig, also auch unkorreliert sind. Es ist instruktiv, sich die Aussage für diesen Fall zu veranschaulichen: Seien also die X_i unabhängig mit $P(X_i = 1) = p$, $P(X_i = 0) = 1 - p$, und sei $S_n = X_1 + \cdots + X_n$ also binomialverteilt mit Parametern n, p. Dann ist $E(X_i) = p$ und $V(X_i) = p(1-p)$. Aus (3.27) folgt also, daß für jedes $\varepsilon > 0$

$$\begin{aligned} P\Big(\Big|\frac{S_n}{n} - p\Big| \geq \varepsilon\Big) &= P(|S_n - np| \geq n\varepsilon) \\ &= \sum_{k:|k-np|\geq n\varepsilon} P(S_n = k) = \sum_{k:|k-np|\geq n\varepsilon} \binom{n}{k} p^k (1-p)^{n-k} \end{aligned}$$

mit $n \to \infty$ gegen 0 konvergiert.

Man muß sich jedoch darüber im klaren sein, daß keineswegs etwa $P(S_n \neq np)$ gegen null konvergiert. In der Tat konvergiert $P(|S_n - np| \geq r)$ gegen 1 für jede Zahl $r > 0$. Nicht S_n liegt mit großer Wahrscheinlichkeit (für große n) in der Nähe von np, sondern S_n/n in der Nähe von p. Wir werden diese Sachverhalte in einem späteren Kapitel präzisieren.

Der Satz (3.27) heißt schwaches Gesetz der großen Zahlen, um es vom sogenannten *starken Gesetz der großen Zahlen (strong law of large numbers)* zu unterscheiden. Dieses besagt

$$P\Big(\lim_{n\to\infty} \frac{S_n}{n} \text{ existiert und ist } = E\Big) = 1. \tag{3.2}$$

Die Gleichung (3.2) macht jedoch nur Sinn, wenn *alle* X_i, $i \in \mathbb{N}$, auf *einem* W.-Raum definiert sind. Die Konstruktion eines solchen W.-Raumes macht aber, vielleicht unerwartet, erhebliche Probleme.

Eine Anwendung des schwachen Gesetzes der großen Zahlen führt zu der folgenden von *Sergej Natanowitsch Bernstein* (1880-1968) gegebenen Beweisvariante des Approximationssatzes von *Karl Weierstrass* (1815-1897). Dieser Satz besagt ja, daß man jede stetige reelle Funktion f auf dem Einheitsintervall $[0,1]$ durch Polynome, definiert auf $[0,1]$, gleichmäßig approximieren kann. Wir betrachten nun das sogenannte Bernstein-Polynom zu f:

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

Wenn S_n eine binomialverteilte Zufallsgröße mit Parametern x und n bezeichnet, so folgt mit $\bar{S}_n := S_n/n$ unmittelbar $E(f(\bar{S}_n)) = B_n^f(x)$. Da jedes obige f auf $[0,1]$ gleichmäßig stetig ist, gibt es zu jedem $\varepsilon > 0$ ein $\delta(\varepsilon) > 0$ derart, daß für alle $x, y \in [0,1]$ gilt: $|x-y| < \delta(\varepsilon) \Rightarrow |f(x) - f(y)| < \varepsilon$. Nach der Tschebyscheff-Ungleichung folgt

$$P(|\bar{S}_n - x| \geq \delta) \leq \frac{x(1-x)}{n\delta^2} \leq \frac{1}{4n\delta^2},$$

denn $4x(1-x) = 1 - (2x-1)^2 \leq 1$. Es folgt somit die Abschätzung

$$\begin{aligned} |B_n^f(x) - f(x)| &= |E(f(\bar{S}_n) - f(x))| \leq E(|f(\bar{S}_n) - f(x)|) \\ &\leq 2\sup_u |f(u)| P(|\bar{S}_n - x| > \delta) + \sup_{|u-v|\leq\delta} |f(u) - f(v)| P(|\bar{S}_n - x| \leq \delta). \end{aligned}$$

Der erste Term ist durch $\frac{1}{2n\delta^2}\sup_u |f(u)|$ beschränkt, der zweite Term durch ε für $\delta \leq \delta(\varepsilon)$, da f gleichmäßig stetig ist. Indem man also zunächst $\delta = \delta(\varepsilon)$ und dann $n = n(\delta, \varepsilon)$ wählt, erhält man $\sup_x |B_n^f(x) - f(x)| \leq \varepsilon$. Somit ist gezeigt, daß für jede stetige reelle Funktion auf $[0,1]$ die Folge $(B_n^f)_{n\in\mathbb{N}}$ der zugehörigen Bernstein-Polynome gleichmäßig auf $[0,1]$ gegen f konvergiert. Die Bedeutung dieses probabilistischen Ansatzes für einen Beweis des Approximationssatzes von Weierstrass liegt im kanonischen Auffinden der explizit angebbaren Polynomfolge $(B_n^f)_{n\in\mathbb{N}}$.

Wir betrachten zwei weitere Anwendung des schwachen Gesetzes der großen Zahlen:

vorteilhaftes Spiel, bei dem man auf Dauer verliert:
Ein Spiel heißt *fair*, wenn in jeder Runde der erwartete Verlust gleich dem erwarteten Gewinn ist. Ist der erwartete Gewinn jeweils größer, heißt das Spiel *vorteilhaft*. Überraschend mag nun sein, daß es vorteilhafte Spiele gibt, bei denen man auf Dauer verliert. Ein erstes Beispiel wurde bereits 1945 von *William Feller* (1906-1970) gegeben. Wir betrachten hier ein von *Ulrich Krengel* ausgearbeitetes Beispiel. Sei $X_0 = 1$ das Startkapital. Man wirft in jeder Runde eine Münze. Da Kapital X_n nach der n-ten Runde sei $X_{n-1}/2$, wenn Kopf im n-ten Wurf fällt, sonst $5X_{n-1}/3$. Somit ist das Spiel vorteilhaft. Mit $Y_n = 1/2$ bei Kopf im n-ten Wurf und $Y_n = 5/3$ sonst folgt die Darstellung $X_n = Y_1 \cdot Y_2 \cdots Y_n$. Aus der Unabhängigkeit der Y_i folgt mit $E(Y_i) = (1/2)(1/2) + (1/2)(5/3) = 13/12$:

$E(X_n) = (13/12)^n \to \infty$. Wenn nun μ den Erwartungwert von $\log Y_i$ bezeichnet, besagt das Gesetz der großen Zahlen

$$P\left(\left|\frac{1}{n}(\log Y_1 + \cdots + \log Y_n) - \mu\right| \leq \varepsilon\right) \to 1.$$

Dies gilt insbesondere für $\varepsilon = -\mu/2$, denn $\mu = (\log(1/2) + \log(5/3))/2 < 0$, also $P(1/n \log X_n - \mu \leq -\mu/2) \to 1$. Also ist mit großer Wahrscheinlichleit $X_n \leq \exp(\mu n/2)$, was wegen $\mu < 0$ gegen Null strebt. Der Kapitalstand strebt also auf lange Sicht ziemlich schnell gegen Null.

Normale Zahlen: Wir betrachten das Intervall $[0, 1]$ und stellen jede Zahl $x \in [0, 1]$ mit Hilfe ihrer Dezimalentwicklung $x = 0.a_1a_2a_3\ldots$, mit $a_i \in \{0, \ldots, 9\}$ dar. Hält man die ersten n Ziffern $a_1, \ldots, a_n$ fest, so ergeben die Zahlen, die mit $0.a_1a_2\ldots a_n$ beginnen, ein Intervall der Länge 10^{-n}. Diese Intervalle sind für unterschiedliche Wahlen von $a_1, a_2, \ldots, a_n$ disjunkt. Wir betrachten nun ein Zufallsexperiment, das solche Intervalle mit Wahrscheinlichkeit 10^{-n} konstruiert. Zu diesem Zweck ziehen wir für jedes i die Ziffer a_i mit Wahrscheinlichkeit $1/10$. Dann hat in der Tat jedes der obigen Intervalle Wahrscheinlichkeit 10^{-n}. Wir bezeichnen für $j \in \{0, \ldots, 9\}$ mit $\nu_n^{(j)}(x)$ die absolute Häufigkeit des Auftretens der Ziffer j in den ersten n Stellen der Ziffer x. Für ein festes $\delta > 0$ besagt nun das schwache Gesetz der großen Zahlen, daß für jedes $\delta > 0$ die Menge der obigen Intervalle für die gilt $|\nu_n^{(j)}(x) - 1/10| \leq \delta$ für alle $0 \leq j \leq 9$ eine Länge hat die asymptotisch für $n \to \infty$ gegen 1 konvergiert. Die relative Häufigkeit der Ziffern in der Dezimalentwicklung solcher Zahlen ist also annähernd gleich. Solche Zahlen heißen normale Zahlen. Das Gesetz der großen Zahlen besagt also, daß die normalen Zahlen in einer Vereinigung von Intervallen liegen, die asymptotisch Länge 1 haben.

4 Normalapproximation der Binomialverteilung

Es sei daran erinnert, daß eine Zufallsgröße X mit der Verteilung

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

binomialverteilt heißt.

Die exakten Werte für $b(k; n, p)$ lassen sich bei festem p allerdings nur für moderat große n und k ($n = 100$ und $k = 50$ ist z.B. schon nicht mehr so leicht) berechnen. Im Falle großer n hilft uns aber eine Version des *Zentralen Grenzwertsatzes*, einer Art Naturgesetz, das die asymptotische Verteilung einer großen Klasse von Variablen angibt.

Die Basis für diese Approximation ist die *Stirlingsche Formel*, die von *James Stirling* (1692-1770) bewiesen wurde:

(4.1) Satz.

$$\lim_{n\to\infty} n!/(\sqrt{2\pi} n^{n+1/2} e^{-n}) = 1.$$

Für einen Beweis: Siehe etwa O. Forster: Analysis 1 §20 Satz 6.

Man bemerke, daß die Stirlingsche Formel nicht bedeutet, daß $|n! - \sqrt{2\pi} n^{n+1/2} e^{-n}|$ gegen 0 konvergiert, im Gegenteil. Es gilt

$$\lim_{n\to\infty} |n! - \sqrt{2\pi} n^{n+1/2} e^{-n}| = \infty.$$

Die erste Frage, die man sich stellen sollte ist die, in welchem Sinne man eigentlich einen Limes von $b(k; n, p)$ sinnvoll definieren kann. Dazu bemerken wir zunächst, daß

$$\frac{b(k+1; n, p)}{b(k; n, p)} = \frac{(n-k)p}{(k+1)(1-p)}$$

ist und daher

$$\frac{b(k+1; n, p)}{b(k; n, p)} < 1 \Leftrightarrow k + 1 > (n+1)p.$$

Die Funktion $k \mapsto b(k; n, p)$ nimmt also ihr Maximum genau bei $k = [n+1]p$ an. Nun ist aber mit Hilfe der Stirlingschen Formal sofort klar, daß

$$\begin{aligned} b([n+1(p)]; n, p) \simeq b(np; n, p) &\simeq \frac{(\frac{n}{e})^n \sqrt{2\pi n} p^{np} (1-p)^{n-np}}{(\frac{np}{e})^{np} \sqrt{2\pi np} (\frac{n-np}{e})^{n-np} \sqrt{2\pi(n-np)}} \\ &= \sqrt{\frac{1}{2\pi np(1-p)}}, \end{aligned}$$

wobei wir für zwei Folgen a_n und b_n schreiben $a_n \simeq b_n$, falls $\lim_{n\to\infty} \frac{a_n}{b_n} = 1$. Also ist für jedes k und p $\lim_{n\to\infty} b(k; n, p) = 0$. Diese Aussage ist eben so wahr wie unnütz. Im wesentlichen bedeutet sie, daß man, um einen "vernünftigen Grenzwert" zu erhalten, nicht einzelne Wahrscheinlichkeiten $b(k; n, p)$ anschauen sollte, sondern die Wahrscheinlichkeit

für ganze Bereiche, also: $\sum_{\alpha_n a+c_n \le k \le \alpha_n b+c_n} b(k;n,p)$, wobei $a, b \in \mathbb{R}$ reelle Zahlen sind und α_n, c_n Funktionen von n sind. Wie aber soll man α_n, c_n wählen? Zunächst ist klar, daß die Binomialverteilung $b(\cdot;n,p)$ den Erwartungswert np hat, es also ratsam ist, $c_n = np$ zu wählen, damit die obige Summe für reelle a, b von einer relevanten Größenordnung ist. Andererseits zeigt die obige Rechnung, daß $\max_k b(k;n,p) \simeq \sqrt{\frac{1}{2\pi np(1-p)}}$ ist. Nimmt man an, daß die Terme $b(k;n,p)$ für k nahe bei np von derselben Ordnung sind, so liegt es nahe $\alpha_n = \sqrt{n}$ oder besser $\alpha_n = p(1-p)\sqrt{n}$ zu wählen (um ein Resultat zu erhalten, das von p nicht abhängt). Letzteres wird in der Tat unsere Wahl sein.

Der erste Schritt zur Herleitung eines Grenzwertsatzes für die Binomialverteilung wird sein, daß wir zunächst die $b(k;n,p)$ einzeln genauer unter die Lupe nehmen. Wir werden sehen, daß diese für relevante k tatsächlich von der Ordnung $1/\sqrt{n}$ sind und darüber hinaus zeigt sich, daß dann $b(k;n,p)$ durch eine schöne Funktion approximiert werden kann. Dazu setzen wir

$$x_k := x_k(n,p) := \frac{k-np}{\sqrt{np(1-p)}}.$$

x_k hängt natürlich von n und p ab, was wir in der Notation jedoch nicht gesondert betonen. Wir kürzen $1-p$ meist durch q ab.

(4.2) Satz. *(lokaler Grenzwertsatz, local limit theorem)* Es seien $0 < p < 1$, $q = 1-p$ und $(a_n)_{n\in\mathbb{N}} > 0$ eine Folge reeller Zahlen mit $\lim_{n\to\infty} a_n^3/\sqrt{n} = 0$. Dann gilt

$$\lim_{n\to\infty} \sup_{k;|x_k|\le a_n} \left| \frac{\sqrt{2\pi npq}\, b(k;n,p)}{e^{-x_k^2/2}} - 1 \right| = 0.$$

(4.3) Bemerkungen.

1. Ist $a_n = A$ eine beliebige, aber feste positive Konstante, so folgt aus dem obigen Satz unmittelbar
$$\lim_{n\to\infty} \sup_{k:|x_k|\le A} \left| \frac{\sqrt{2\pi npq}\, b(k;n,p)}{e^{-x_k^2/2}} - 1 \right| = 0.$$

2. Wir schreiben nachfolgend stets $b(k;n,p) \sim \frac{1}{\sqrt{2\pi npq}} e^{-x_k^2/2}$ für die obige gleichmäßige Konvergenz. Allgemeiner: Sind $\alpha(k,n)$, $\beta(k,n) > 0$ für $n \in \mathbb{N}_0$, $0 \le k \le n$, so bedeutet (während des untenstehenden Beweises) $\alpha(k,n) \sim \beta(k,n)$, daß für die obige Folge $(a_n)_{n\in\mathbb{N}} > 0$
$$\lim_{n\to\infty} \sup_{k:|x_k|\le a_n} \left| \frac{\alpha(k,n)}{\beta(k,n)} - 1 \right| = 0$$
gilt.

3. Wir überzeugen uns vom folgenden Sachverhalt, der im Beweis von (4.2) mehrfach verwendet wird:
$$\alpha(k,n) \sim \beta(k,n),\ \ \alpha'(k,n) \sim \beta'(k,n) \ \ \Rightarrow \ \ \alpha(k,n)\alpha'(k,n) \sim \beta(k,n)\beta'(k,n).$$

Beweis.

$$\begin{aligned}\left|\frac{\alpha(k,n)\alpha'(k,n)}{\beta(k,n)\beta'(k,n)}-1\right| &\le \frac{\alpha'(k,n)}{\beta'(k,n)}\left|\frac{\alpha(k,n)}{\beta(k,n)}-1\right|+\left|\frac{\alpha'(k,n)}{\beta'(k,n)}-1\right| \\ &\le \left|\frac{\alpha'(k,n)}{\beta'(k,n)}-1\right|\left|\frac{\alpha(k,n)}{\beta(k,n)}-1\right|+\left|\frac{\alpha'(k,n)}{\beta'(k,n)}-1\right|+\left|\frac{\alpha(k,n)}{\beta(k,n)}-1\right|.\end{aligned}$$

Daraus folgt die Aussage sofort. □

Beweis von Satz (4.2). Es gilt

$$k = np + \sqrt{npq}\,x_k,\ \ n-k = nq - \sqrt{npq}\,x_k,$$

also

$$k \sim np,\ \ n-k \sim nq.$$

Mit Hilfe der Stirlingschen Formel folgt:

$$\begin{aligned}b(k;n,p) &\sim \frac{(\frac{n}{e})^n\sqrt{2\pi n}p^kq^{n-k}}{(\frac{k}{e})^k\sqrt{2\pi k}(\frac{n-k}{e})^{n-k}\sqrt{2\pi(n-k)}} \\ &= \sqrt{\frac{n}{2\pi k(n-k)}}\varphi(n,k) \sim \frac{1}{\sqrt{2\pi npq}}\varphi(n,k),\end{aligned}$$

wobei wir $\varphi(n,k)$ für $(\frac{np}{k})^k(\frac{nq}{n-k})^{n-k}$ schreiben. Es ist nun

$$-\log\varphi(n,k) = nH(k/n|\,p),$$

wobei

$$H(x|\,p) = x\log\Big(\frac{x}{p}\Big) + (1-x)\log\Big(\frac{1-x}{1-p}\Big)$$

(diese Funktion heißt *relative Entropie* von x bezüglich p; sie wird im Rahmen des Studiums der großen Abweichungen (Kapitel 6) eine zentrale Rolle spielen). Wir wollen diese Funktion nun um den Wert p Taylor entwickeln. Es ist $H'(p|\,p) = 0$ und $H''(p|\,p) = 1/p + 1/q = 1/(pq)$. Damit folgt

$$H(x|\,p) = \frac{(x-p)^2}{2pq} + \psi(x-p),$$

wobei ψ das Restglied in der Taylorentwicklung bezeichnet. Insbesondere gilt in jedem endlichen Intervall, das p enthält eine Abschätzung

$$|\psi(x-p)| \le c|x-p|^3$$

mit einer geeigneten Konstanten c. Wir erhalten somit

$$\left|-\log\varphi(n,k) - \frac{n\big(\frac{k}{n}-p\big)^2}{2pq}\right| \le cn\Big|\frac{k}{n}-p\Big|^3.$$

Aus der Definition der x_k erhält man für eine geeignete Konstante $0 < c' < \infty$ folgt

$$\frac{|k-np|^3}{n^2} = c'\frac{|x_k|^3}{\sqrt{n}}.$$

Wählen wir nun ein k mit $|x_k| \le a_n$, so konvergiert aufgrund der Bedingung an die Folge $(a_n)_{n\in\mathbb{N}}$ die rechte Seite der Ungleichung gegen 0. Da nun aber

$$\frac{n\left(\frac{k}{n}-p\right)^2}{2pq} = \frac{x_k^2}{2},$$

erhalten wir

$$\lim_{n\to\infty} \sup_{k:|x_k|\le a_n} \left| \frac{\varphi(n,k)}{e^{-x_k^2/2}} - 1 \right| = 0.$$

Damit ist der Satz gezeigt. □

Ein Rechenbeispiel dazu:
Jemand wirft 1200-mal einen Würfel. Mit welcher Wahrscheinlichkeit hat er genau 200-mal eine 6? Mit welcher Wahrscheinlichkeit 250-mal?

Wir berechnen x_k für $k = 200, 250$, $n = 1200$, $p = 1/6$.

$$x_{200} = 0, \quad x_{250} = \frac{5\sqrt{6}}{\sqrt{10}} = 3.873$$
$$b(200; 1200, 1/6) \cong 0.0309019$$
$$b(250; 1200, 1/6) \cong 0.0000170913.$$

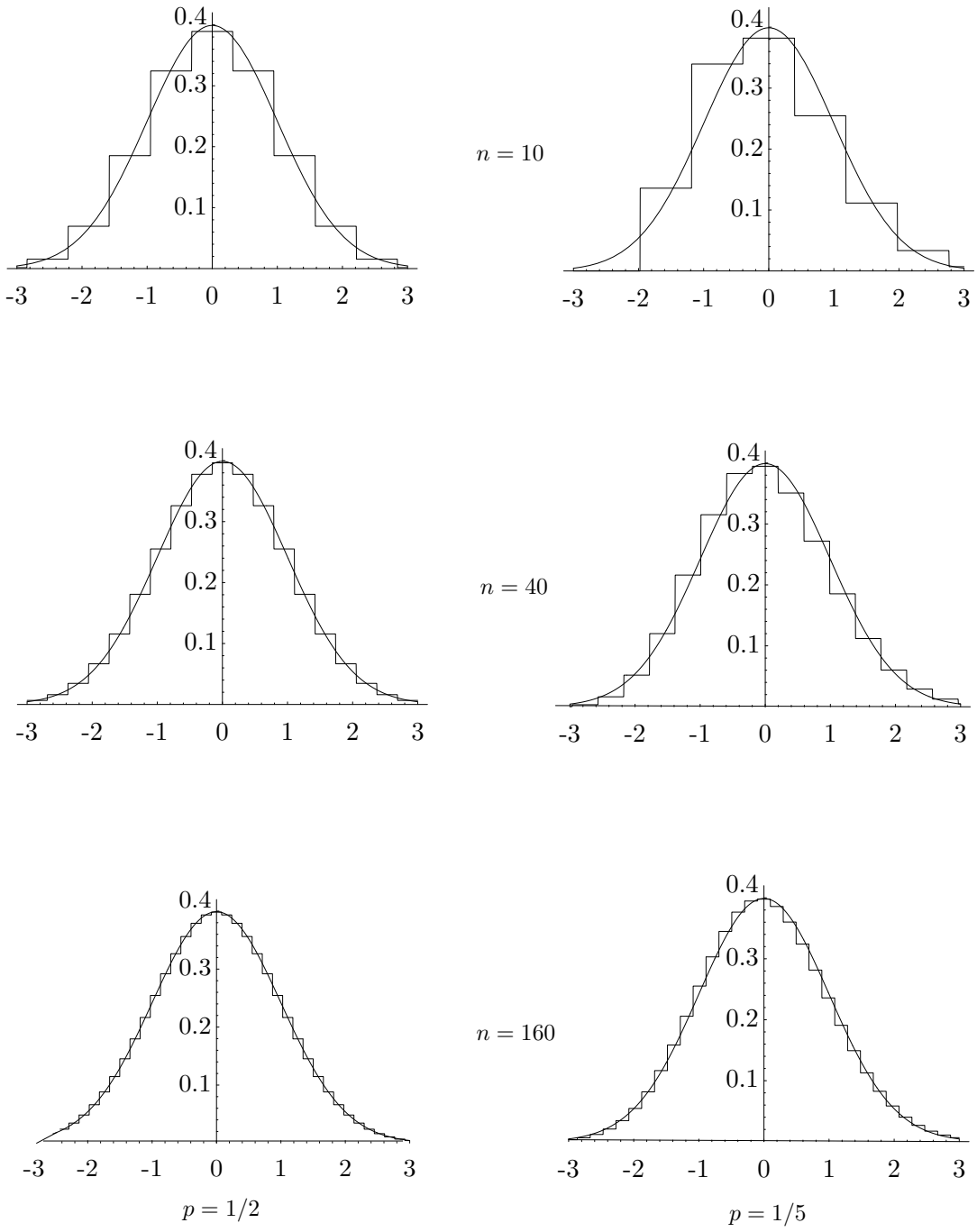
Wie üblich muß hier bemerkt werden, daß ein reines Limesresultat für die Güte einer Approximation wie in obigem Rechenbeispiel zunächst natürlich gar nichts aussagt. Gefragt sind konkrete Abschätzungen des Fehlers. Dies ist ein technisch aufwendiges Feld, in das wir in dieser Vorlesung nicht eintreten werden.

Nachfolgend ist eine numerische Illustration von (4.19) angegeben:

Die sechs Bilder illustrieren die Konvergenz der Binomialverteilung gegen die Funktion $\varphi(x) = (2\pi)^{-1/2}\exp(-x^2/2)$. Hier ist jeweils die Funktion $\varphi(x)$ zusammen mit dem skalierten Histogramm

$$f_{n,p}(x) = \begin{cases} \sqrt{np(1-p)}b(k;n,p), & \text{falls } k \in \{0,1,\dots,n\} \text{mit } |x-x_k| < \frac{1}{2\sqrt{np(1-p)}}, \\ 0 & \text{andernfalls,} \end{cases}$$

der Binomialverteilung $b(\cdot\,; n, p)$ gezeichnet; in der linken Spalte der symmetrische Fall mit $p = 1/2$, in der rechten Spalte der asymmetrische Fall $p = 1/5$.



Nun kommen wir dazu, die schon eingangs diskutierten "Bereichswahrscheinlichkeiten" zu approximieren.

(4.4) Satz. *(von de Moivre-Laplace)* Für beliebige reelle Zahlen a und b mit $a < b$ gilt:

$$\lim_{n\to\infty} P\left(a \le \frac{S_n - np}{\sqrt{npq}} \le b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2} dx. \tag{4.1}$$

Beweis. Die zentrale Idee des Beweises ist es für die einzelnen Summanden der linken Seite von (4.1) die Approximation aus dem lokalen Grenzwertsatz einzusetzen und zu sehen, daß dies eine Riemannsumme für das Integral auf der rechten Seite von (4.1) liefert.

Sei also $k \in \{0, \ldots, n\}$. Dann ist $\{S_n = k\} = \{(S_n - np)/\sqrt{npq} = x_k\}$. Also ist die links stehende Wahrscheinlichkeit gleich

$$\sum_{k:a \le x_k \le b} P(S_n = k) = \sum_{k:a \le x_k \le b} b(k; n, p).$$

Wir setzen nun für jeden Summanden auf der rechten Seite seinen in Satz (4.2) angegebenen asymptotischen Wert ein und berücksichtigen, daß $x_{k+1} - x_k = \frac{1}{\sqrt{npq}}$ ist. Die Summe dieser Größen nennen wir R_n:

$$R_n = \frac{1}{\sqrt{2\pi}} \sum_{k:a \le x_k \le b} e^{-x_k^2/2}(x_{k+1} - x_k).$$

Unter Verwendung der Gleichmäßigkeit der Konvergenz in Satz (4.2) sieht man sofort, daß der Quotient von $P(a \le \frac{S_n - np}{\sqrt{npq}} \le b)$ und dem obenstehenden Ausdruck gegen 1 konvergiert, das heißt, es existiert eine Nullfolge $(\varepsilon_n)_{n \in \mathbb{N}}$, $\varepsilon_n > 0$ mit

$$R_n(1 - \varepsilon_n) \le P\left(a \le \frac{S_n - np}{\sqrt{npq}} \le b\right) \le R_n(1 + \varepsilon_n). \tag{4.2}$$

k und x_k entsprechen einander bikjektiv, und wenn k von 0 bis n läuft, dann variiert x_k im Intervall $[-\sqrt{np/q}, \sqrt{nq/p}]$ mit der Schrittweite $x_{k+1} - x_k = 1/\sqrt{npq}$. Für hinreichend große n umfaßt dieses Intervall das gegebene Intervall $[a, b]$, und die in $[a, b]$ fallenden Punkte x_k teilen dieses in Teilintervalle derselben Länge $1/\sqrt{npq}$. Wenn nun der kleinste und der größte Wert von k mit $a \le x_k \le b$ gleich j bzw. l ist, dann ist

$$x_{j-1} < a \le x_j < x_{j+1} < \ldots < x_{l-1} < x_l \le b < x_{l+1}$$

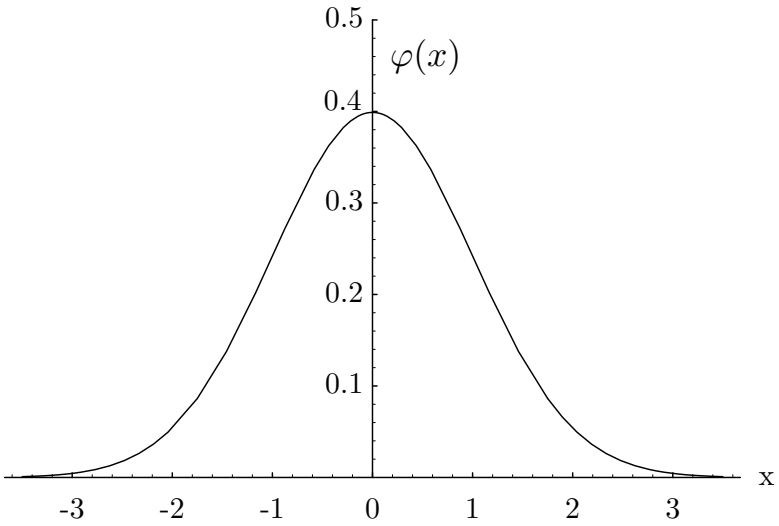
und die obige Summe läßt sich schreiben als

$$\sum_{k=j}^{l} \varphi(x_k)(x_{k+1} - x_k),$$

wobei $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ ist. Das ist eine Riemannsche Summe für das bestimmte Integral $\int_a^b \varphi(x)dx$. Somit konvergiert R_n mit $n \to \infty$ gegen das Integral in der Behauptung des Satzes. Dieser folgt nun sofort mit (4.2). □

Abraham de Moivre (1667–1754) veröffentlichte dieses Ergebnis in seiner „*Doctrine of Chances*“ 1714. *Pierre Simon Marquis de Laplace* (1749-1827) erweiterte das Ergebnis und wies dessen Bedeutung in seiner „*Théorie analytique des probabilités*“ 1812 nach. Es handelt sich um den zuerst bekanntgewordenen Spezialfall des sogenannten *Zentralen Grenzwertsatzes (central limit theorem).*

Die Funktion $x \to \varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ heißt auch Gaußsche Glockenkurve, wegen des glockenförmigen Verlaufs ihres Graphen.



Die Verteilung, die durch das Integral unter der Glockenkurve gegeben ist, heißt auch Standard–Normalverteilung und wird oft mir $\mathcal{N}(0,1)$ abgekürzt.

Die Integrale $\int_a^b \varphi(x)dx$ sind leider nicht in geschlossener Form mit Hilfe von Polynomen, rationalen Funktionen, Wurzelausdrücken oder elementaren transzendenten Funktionen (wie sin, cos, exp, etc.) darstellbar.

Es gilt offenbar für $a < b$

$$\int_a^b \varphi(x)dx = \int_{-\infty}^b \varphi(x)dx - \int_{-\infty}^a \varphi(x)dx = \Phi(b) - \Phi(a),$$

wobei wir $\Phi(y) := \int_{-\infty}^y \varphi(x)dx$ gesetzt haben. Wie nicht anders zu erwarten ist, gilt

$$\int_{-\infty}^{\infty} \varphi(x)dx = 1. \tag{4.3}$$

Der Beweis, den man üblicherweise in der Analysis für diese Tatsache gibt, benutzt Polarkoordinaten. Wir geben hier einen Beweis, der sich darauf stützt, daß wir den Satz von de-Moivre-Laplace schon kennen: Wir verwenden (4.4) und setzen $S_n^* := \frac{S_n - np}{\sqrt{npq}}$. (Für das Argument hier spielt p keine Rolle; wir können z.B. $p = 1/2$ nehmen.) Sei $a > 0$. Dann ist

$$1 = P(-a \le S_n^* \le a) + P(|S_n^*| > a).$$

Nach der Tschebyscheff-Ungleichung gilt:

$$P(|S_n^*| > a) \le \frac{1}{a^2} \operatorname{Var}(S_n^*) = \frac{1}{a^2}.$$

Nach (4.4) gilt

$$\lim_{n\to\infty} P(-a \le S_n^* \le a) = \int_{-a}^a \varphi(x)dx.$$

Demzufolge ist

$$1 - \frac{1}{a^2} \le \int_{-a}^a \varphi(x)dx \le 1$$

für jedes $a > 0$, womit (4.6) bewiesen ist.

(4.7) Bemerkung. (a) Wegen $\lim_{n\to\infty} \sup_k P(S_n = k) = 0$ ist es natürlich gleichgültig, ob in der Aussage von (4.21) $\leq$ oder $<$ steht.

(b) Es gilt für $a \in \mathbb{R}$:

$$\begin{aligned} \lim_{n\to\infty} P\left(\frac{S_n - np}{\sqrt{npq}} \leq a\right) &= \Phi(a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2} dx, \\ \lim_{n\to\infty} P\left(\frac{S_n - np}{\sqrt{npq}} \geq a\right) &= 1 - \Phi(a). \end{aligned}$$

Beweis von (b). Wir beweisen die erste Gleichung; die zweite folgt analog. Wegen der Symmetrie von φ und (4.6) gilt:

$$\Phi(x) = \int_{-\infty}^{x} \varphi(u)du = 1 - \int_{x}^{\infty} \varphi(u)du = 1 - \int_{-\infty}^{-x} \varphi(u)du = 1 - \Phi(-x).$$

Wir setzen wieder $S_n^* = \frac{S_n - np}{\sqrt{npq}}$ und wählen $b > 0$ so groß, daß $-b < a$ gilt. Dann ist nach (4.4)

$$\begin{aligned} \limsup_{n\to\infty} P(S_n^* \leq a) &= \limsup_{n\to\infty} (P(-b \leq S_n^* \leq a) + P(S_n^* < -b)) \\ &= \limsup_{n\to\infty} (P(-b \leq S_n^* \leq a) + (1 - P(S_n^* \geq -b))) \\ &\leq \limsup_{n\to\infty} (P(-b \leq S_n^* \leq a) + (1 - P(-b \leq S_n^* \leq b))) \\ &= \Phi(a) - \Phi(-b) + (1 - \Phi(b) + \Phi(-b)) \\ &= \Phi(a) + \Phi(-b) \\ \liminf_{n\to\infty} P(S_n^* \leq a) &\geq \liminf_{n\to\infty} P(-b \leq S_n^* \leq a) \\ &= \Phi(a) - \Phi(-b). \end{aligned}$$

Wegen $\Phi(-b) \to 0$ für $b \to \infty$ folgt die gewünschte Aussage. □

Der Satz (4.4) ist eine Präzisierung des Gesetzes der großen Zahlen, welches besagt, daß für jedes $\varepsilon > 0$ $\lim_{n\to\infty} P(\left|\frac{S_n}{n} - p\right| \geq \varepsilon) = 0$ ist. Letzteres können wir sofort auch aus (4.4) herleiten:

$$\begin{aligned} P\left(\left|\frac{S_n}{n} - p\right| \leq \varepsilon\right) &= P\left(-\varepsilon \leq \frac{S_n}{n} - p \leq \varepsilon\right) \\ &= P\left(-\frac{\sqrt{n}\varepsilon}{\sqrt{pq}} \leq \frac{S_n - np}{\sqrt{npq}} \leq \frac{\sqrt{n}\varepsilon}{\sqrt{pq}}\right) \geq P\left(a \leq \frac{S_n - np}{\sqrt{npq}} \leq b\right), \end{aligned}$$

sofern n so groß ist, daß $\sqrt{n}\varepsilon/\sqrt{pq} \geq b$ und $-\sqrt{n}\varepsilon/\sqrt{pq} \leq a$ sind. Für beliebige Zahlen $a, b \in \mathbb{R}$ ist dies aber für genügend große n der Fall. Somit ist $\lim_{n\to\infty} P(|\frac{S_n}{n} - p| \leq \varepsilon) = 1$ für jedes $\varepsilon > 0$.

Dieser Beweis ist natürlich insgesamt wesentlich aufwendiger als der in Kapitel 3 angegebene. (4.4) ist jedoch sehr viel informativer als das Gesetz der großen Zahlen.

Tabelle der Verteilungsfunktion $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du$ für $x \geq 0$. Wir hatten bereits gesehen, daß für $x \leq 0$ gilt: $\Phi(x) = 1 - \Phi(-x)$.

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7356	0.7389	0.7421	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7793	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8364	0.8389
1.0	0.8413	0.8437	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8687	0.8708	0.8728	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8906	0.8925	0.8943	0.8962	0.8979	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9146	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9278	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9624	0.9633
1.8	0.9641	0.9648	0.9656	0.9664	0.9671	0.9678	0.9685	0.9692	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9761	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9874	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9895	0.9898	0.9901	0.9903	0.9906	0.9908	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9926	0.9928	0.9930	0.9932	0.9934	0.9936
2.5	0.9938	0.9939	0.9941	0.9943	0.9944	0.9946	0.9947	0.9949	0.9950	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9958	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9973
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986
3.0	0.9986	0.9987	0.9987	0.9988	0.9988	0.9988	0.9989	0.9990	0.9989	0.9990

Wir wollen nun sehen, dass das Grenzwertverhalten des Satzes von de Moivre/Laplace ein Spezialfall eines viel allgemeineren Phänomens ist, eines Satzes, der neben dem Gesetz der großen Zahlen ein zweites "Naturgesetz" der Stochastik darstellt. Wie wir dies schon im Satz von de Moivre/Laplace kennengelernt haben, befasst sich dieser Satz mit der Konvergenz von Verteilungen $P_n(\bullet) = P[X_n \in \bullet]$ für geeignete Zufallsvariablen X_n. Es läge sicherlich nahe davon zu sprechen, dass eine Folge von Verteilungen P_n gegen eine

Grenzverteilung P_0 konvergiert, falls

$$P_n(\{x\}) \underset{n\to\infty}{\longrightarrow} P_0(\{x\}) \quad \forall\, x \in \mathbb{R}$$

bzw.

$$F_n(x) := P_n((-\infty, x]) = \sum_{y\le x} P_n(y) \to F_0(x)$$

für eine geeignete (Verteilungs-) Funktion F_0 gilt.

Das folgende Beispiel zeigt, dass diese Begriffsbildung nicht das Gewünschte liefert.

(4.8) Beispiel. Seien X_n Zufallsvariablen die im Punkt $\frac{1}{n}$ konzentriert sind, d. h. für alle $n \in \mathbb{N}$ gelte

$$P(X_n = \frac{1}{n}) = 1.$$

Die P_n sind entsprechend Deltafunktionen in $\frac{1}{n}$:

$$P_n(\{x\}) = \delta_{x-\frac{1}{n}}.$$

Es ist anschaulich klar, dass die P_n gegen die Dirac-Verteilung in der 0 konvergieren. Dies würde der obige Konvergenzbegriff aber nicht leisten, denn

$$\lim_{n\to\infty} P_n(\{0\}) = 0 \neq 1 = P_0(\{0\}),$$

wenn P_0 gerade die Dirac-Verteilung in der 0 ist. Entsprechend gilt auch

$$\lim_{n\to\infty} F_n(0) = 0 \neq 1 = F_0(0).$$

Die Schwierigkeit ist hierbei offenbar, dass, der Limes F_0 gerade im Punkt 0 unstetig ist. Um diese Schwierigkeit zu umgehen, verlangt man für den neuen Konvergenzbegriff nur das Folgende:

(4.9) Definition. Eine Folge von Verteilungsfunktionen F_n von Wahrscheilichkeiten P_n auf $\mathbb{R}$ heißt verteilungskonvergent gegen F_0, falls F_0 eine Verteilungsfunktion ist, d. h. falls gilt

a) F_0 ist monoton wachsend;

b) F_0 ist rechtsseitig stetig;

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

und falls

$$F_n(x) \to F_0(x)$$

für alle x, in denen F_0 stetig ist, gilt. Ist F_0 die Verteilungsfunktion einer Wahrscheinlichkeit P_0 auf $\mathbb{R}$, so schreiben wir

$$P_n \xrightarrow{\mathcal{D}} P_0.$$

(4.10) Beispiel. Für die Funktion F_n, F_0 aus dem Eingangsbeispiel gilt

$$F_n(x) = \begin{cases} 1 & x \geq 1/n \\ 0 & x < 1/n \end{cases} \quad \text{also} \quad \lim_{n\to\infty} F_n(x) = \begin{cases} 1 & x > 0 \\ 0 & x \leq 0 \end{cases}$$

Dies impliziert die Verteilungskonvergenz von F_n gegen F_0.

Es ist interessant, diesen neuen Begriff zu vergleichen mit der Konvergenz von Zufallsvariaben X_n gegen eine Zufallsvariable X_0 in Wahrscheinlichkeit. Letzteres bedeutet, dass analog zum Gesetz der großen Zahlen gilt

$$P(|X_n - X_0| \geq \varepsilon) \to 0 \quad \forall\, \varepsilon > 0.$$

Wir werden sehen, dass der Begriff der Verteilungskonvergenz schwächer ist als der Begriff der Konvergenz in Wahrscheinlichkeit:

(4.11) Satz. Es seien $(X_n)_n$ Zufallsvariablen mit

$$X_n \to X_0 \quad \text{in Wahrscheinlichkeit.}$$

Dann konvergiert P^{X_n}, die Verteilung von X_n, in Verteilung gegen P^{X_0}.

Beweis: Wir schreiben

$$F_n := P^{X_n} \quad \text{bzw.} \quad F_0 := P^{X_0}.$$

Es sei x ein Stetigkeitspunkt von F_0 und $\varepsilon > 0$. Dann gibt es ein $\delta > 0$ mit

$$F_0(x) - \varepsilon \leq F_0(x - \delta) = P(X_0 \leq x - \delta)$$

und

$$F_0(x + \delta) \leq F_0(x) + \varepsilon.$$

Nun gilt aber für alle $n \in \mathbb{N}$:

$$\{X_0 \leq x - \delta\} \subseteq \{X_n < x\} \cup \{|X_n - x| \geq \delta\},$$

da $X_n \geq x$ und $|X_n - X_0| < \delta$ folgt

$$X_0 = (X_0 - X_n) + X_n > x - \delta.$$

Hieraus folgt

$$F_0(x) \leq F_0(x - \delta) + \varepsilon \leq F_n(x) + P(|X_n - X_0| \geq \delta) + \varepsilon.$$

Analog gilt

$$\{X_n \leq x\} \subset \{X_0 < x + \delta\} \cup \{|X_n - X_0| \geq \delta\},$$

also auch

$$F_n(x) \leq F_0(x) + P(|X_n - X_0| \geq \delta) + \varepsilon.$$

Insgesamt erhält man:

$$|F_n(x) - F_0(x)| \leq \varepsilon + P(|X_n - X_0| \geq \delta).$$

Da der letzte Term für $n \to \infty$ verschwindet, folgt die Behauptung. □

(4.12) Bemerkung. Die Umkehrung des vorhergehenden Satzes gilt in der Regel nicht, wie dieses Beispiel zeigt. X sei eine Zufallsvariable mit

$$P(X = 1) = P(X = -1) = \frac{1}{2}.$$

(X_n) sei eine Folge von Zufallsvariablen mit

$$X_{2n} = X \quad \text{und} \quad X_{2n+1} = -X \quad \forall\, n \in \mathbb{N}.$$

Da $P^{X_n} = P^X$ für alle $n \in \mathbb{N}$ gilt, ist X_n natürlich verteilungskovnergent gegen X. Andererseits gilt

$$P(|X_{2n+1} - X| \geq 1) = 1 \quad \forall\, n \in \mathbb{N}.$$

Wir werden diesen Begriff in der Wahrscheinlichkeitstheorie noch genauer betrachten. Für den Moment begnügen wir uns mit einer hinreichenden Bedingung für die Verteilungskonvergenz.

(4.13) Satz. Es seien P_n diskrete Wahrscheinlichkeitsverteilungen über $\mathbb{R}$ und $F_0 : \mathbb{R} \to \mathbb{R}$ differenzierbar, monoton wachsend mit

$$\lim_{x \to -\infty} F_0(x) = 0 \quad \text{und} \quad \lim_{x \to \infty} F_0(x) = 1.$$

Gilt dann

$$\lim_{n \to \infty} \sum_{x \in \mathbb{R}} f(x) P_n(\{x\}) = \int_{-\infty}^{\infty} f(x) F_0'(x) dx,$$

so für alle stetigen Funktionen $f : \mathbb{R} \to \mathbb{R}$ mit existenten Limiten $\lim_{x \to \pm\infty} f(x)$, so ist P_n verteilungskonvergent und es gilt

$$F_n \to F_0 \quad \text{in Verteilung.}$$

Beweis. Sei $g : \mathbb{R} \to \mathbb{R}$ definiert durch

$$g(x) = 1_{(-\infty,0]}(x) + (1 - x) 1_{(0,1)}(x).$$

g ist stetig und es gilt $\lim_{x \to \infty} g(x) = 0,\ \lim_{x \to -\infty} g(x) = 1$. Selbiges gilt für die Funktionen

$$f_k(x) := g(kx).$$

Für die zu P_n gehörigen Verteilungsfunktion F_n gilt dann zum einen für alle $x \in \mathbb{R}$ und $k \in \mathbb{N}$

$$\begin{aligned}
\limsup_{n \to \infty} F_n(x) &= \limsup_{n \to \infty} \sum_{y \leq x} P_n(\{y\}) \\
&\leq \limsup_{n \to \infty} \sum_{y} f_k(y - x) P_n(\{y\}) \\
&= \int_{-\infty}^{\infty} f_k(y - x) F_0'(y) dy \\
&\leq \int_{-\infty}^{x + \frac{1}{k}} F_0'(y) dy = F_0(x + \frac{1}{k}).
\end{aligned}$$

Hierbei haben wir zunächst verwendet, dass g auf $\mathbb{R}^-$ gleich 1 ist, dann die Voraussetzung eingesetzt und schließlich nochmals die Definition von f_k. Andererseits gilt

$$\begin{aligned}
\liminf_{n\to\infty} F_n(x) &= \liminf_{n\to\infty} \sum_{y\le x} P_n(\{y\}) \\
&\ge \liminf_{n\to\infty} \sum_{y} f_k(y-x+\frac{1}{k}) P_n(\{y\}) \\
&= \int_{-\infty}^{\infty} f_k(y-x+\frac{1}{k}) F_0'(y)dy \\
&\ge \int_{-\infty}^{x-\frac{1}{k}} F_0'(y)dy = F_0(x-\frac{1}{k}).
\end{aligned}$$

Da F_0 insbesondere überall stetig ist, folgt

$$\lim_{k\to\infty} F_0(x+\frac{1}{k}) = \lim_{x\to\infty} F_0(x-\frac{1}{k}) = F_0(x),$$

also insgesamt

$$\lim_{n\to\infty} F_n(x) = F_0(x) \quad \forall\, x \in \mathbb{R}.$$

□

Mit diesem Hilfsmittel an der Hand können wir nun die folgende, allgemeinere Version des Satzes von de Moivre/Laplace beweisen:

(4.14) Satz. (Satz von Lindeberg-Levy/Spezialfall) Es seien für alle n $X_1, X_2, \ldots, X_n$ stochastisch unabhängige Zufallsvariablen, die alle dieselbe diskrete Verteilung besitzen und deren Erwartungswerte $\mathbb{E}X_1$ und Varianzen $\mathbb{V}(X_1) > 0$ existieren. Dann gilt

$$\lim_{n\to\infty} P(-\infty < \frac{\sum_{i=1}^n (X_i - \mathbb{E}X_1)}{\sqrt{n\mathbb{V}(X_1)}} \le x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt,$$

d. h. die Variablen $\sum_{i=1}^n (X_i - \mathbb{E}X_1)/\sqrt{n\mathbb{V}(X_1)}$ sind verteiungskonvergent mit Limes

$$F_0(x) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt.$$

(4.15) Bemerkungen.

a) Da $e^{-y^2/2}$ schneller fällt als jede Potenz, ist $\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$ existent.

b) F_0 ist monoton wachsend, stetig und $\lim_{x\to-\infty} F_0(x) = 0$. Außerdem ist F_0 differenzierbar und nach dem Hauptsatz der Infinitesimalrechnung gilt

$$F_0'(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Schließlich lernt man auch in der Analysis, dass

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-t^2/2} dt = 1 \quad \text{also} \quad \lim_{x\to\infty} F_0(x) = 1$$

gilt.

c) Man beachte, dass die Aussage des Satzes unabhängig ist von der Gestalt der Verteilung von X_1.

Beweis des Satzes. Setze

$$Y_i := \frac{X_i - \mathbb{E}X_1}{\sqrt{\mathbb{V}X_1}}.$$

Die Y_i sind mit den X_i unabhängig und identisch verteilt. Es gilt

$$\mathbb{E}Y_i = 0 \quad \text{und} \quad \mathbb{V}(Y_i) = 1.$$

Setzen wir weiter

$$\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

so ist

$$\sqrt{n}\bar{Y}_n = \frac{\sum_{i=1}^{n}(X_i - \mathbb{E}X_1)}{\sqrt{n\mathbb{V}X_1}}.$$

Mit

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

wollen wir also für jede stetige Funktion $f : \mathbb{R} \to \mathbb{R}$ mit existenten Limiten $\lim_{x\to\pm\infty} f(x)$ beweisen, dass

$$\lim_{n\to\infty} \sum_y f(y) P(\sqrt{n}\bar{Y}_n = y) = \int_{-\infty}^{\infty} f(y)\varphi(y)dy =: I(f).$$

Da man von f immer die Konstante $I(f)$ subtrahieren kann, können wir o.B.d.A. $I(f) = 0$ annehmen. Betrachte

$$h(x) = \frac{1}{\varphi(x)} \int_{-\infty}^{x} f(y)\varphi(y)dy.$$

Da f konstruktionsgemäß gleichmäßig stetig und beschränkt ist, ist h wohldefiniert und als Quotient stetiger Funktionen stetig. Da

$$\varphi'(x) = -\frac{x}{\sqrt{2\pi}} e^{-x^2/2} = -x\varphi(x)$$

gilt, folgt

$$h'(x) = \frac{f(x)\varphi^2(x) - \int_{-\infty}^{x} f(y)\varphi(y)dy\varphi'(x)}{\varphi^2(x)} = f(x) + xh(x),$$

für alle $x \in \mathbb{R}$. Natürlich ist auch $xh(x)$ stetig und mit l'Hospital folgt

$$\begin{aligned}
\lim_{x\to\pm\infty} xh(x) &= \lim_{x\to\pm\infty} \frac{\int_{-\infty}^{x} f(y)\varphi(y)dy}{\frac{\varphi(x)}{x}} \\
&= \lim_{x\to\pm\infty} \frac{f(x)\varphi(x)}{\frac{-x^2\varphi(x)-\varphi(x)}{x^2}} = \lim_{x\to\pm\infty} \frac{f(x)\varphi(x)}{-\varphi(x)(1+\frac{1}{x^2})} \\
&= -\lim_{x\to\pm\infty} f(x).
\end{aligned}$$

Dies wenden wir folgendermaßen an:

$$\begin{aligned}\sum_y f(y)P(\sqrt{n}\bar{Y}_n = y) &= \mathbb{E}[f(\sqrt{n}\bar{Y}_n)] \\ &= \mathbb{E}[h'(\sqrt{n}\bar{Y}_n)] - \mathbb{E}[\sqrt{n}\bar{Y}_n h(\sqrt{n}\bar{Y}_n)] \\ &= \mathbb{E}[h'(\sqrt{n}\bar{Y}_n)] - \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\mathbb{E}[Y_j h(\sqrt{n}\bar{Y}_n)].\end{aligned}$$

Aufgrund der Unabhängigkeit und identischen Verteilung der Y_j ist dies gleich

$$= \mathbb{E}[h'(\sqrt{n}\bar{Y}_n)] - \sqrt{n}\mathbb{E}[Y_1 h(\sqrt{n}\bar{Y}_n)].$$

Nun betrachten wir die Taylor-Entwicklung von h um

$$Z_n := \frac{1}{\sqrt{n}}\sum_{j=2}^{n} Y_j.$$

Dies ergibt:

$$h(\sqrt{n}\bar{Y}_n) = h(Z_n) + h'(Z_n)\frac{Y_1}{\sqrt{n}} + \frac{Y_1}{\sqrt{n}}R_n$$

mit

$$R_n = h'(Z_n + \vartheta\frac{Y_1}{\sqrt{n}}) - h'(Z_n) \quad \text{für ein} \quad \vartheta \in [0,1).$$

Nun sind konstruktionsgemäß Y_1 und Z_n stochastisch unabhängig. Daraus folgt

$$\begin{aligned}\mathbb{E}[Y_1 h(\sqrt{n}\bar{Y}_n)] &= \mathbb{E}(Y_1)\mathbb{E}(h(Z_n)) + \mathbb{E}(Y_1^2)\frac{1}{\sqrt{n}}\mathbb{E}[h'(Z_n)] + \frac{1}{\sqrt{n}}\mathbb{E}[Y_1^2 R_n] \\ &= \frac{\mathbb{E}[h'(Z_n)]}{\sqrt{n}} + \frac{\mathbb{E}[Y_1^2 R_n]}{\sqrt{n}}.\end{aligned}$$

Insgesamt ergibt dies:

$$\mathbb{E}[f(\sqrt{n}\bar{Y}_n)] = \mathbb{E}[h'(Z_n + \frac{Y_1}{\sqrt{n}}) - h'(Z_n)] - \mathbb{E}[Y_1^2 \cdot (h'(Z_n + \frac{\vartheta Y_1}{\sqrt{n}}) - h'(Z_n)].$$

Da h' gleichmäßig stetig ist, konvergieren für festes ω wegen

$$\lim \frac{Y_1(w)}{\sqrt{n}} = 0$$

die Summanden unter beiden Erwartungswerten gegen 0. Da außerdem h' beschränkt ist, konvergieren auch die zugehörigen Erwartungswerte gegen 0. Dies ergibt

$$\lim_{n\to\infty}\mathbb{E}[f(\sqrt{n}\bar{Y}_n)] = \lim_{n\to\infty}\sum f(y)P[\sqrt{n}Y_n = y] = 0.$$

Das war zu zeigen. □

Ein Anwendungsbeispiel (Außersinnliche Wahrnehmung (ASW))
1973 machte C. Tert (Univ. California, Davis) ein Experiment zu ASW. Eine Aquarius genannte Maschine wählte zufällig ein Symbol von A,B,C,D und die Versuchsperson sollte erraten, welches. Tert nahm 15 Personen mit vermuteten "hellseherischen Fähigkeiten" und testete jede 500 Mal. Von den entstandenen 7500 Versuchen waren 2006 Treffer. Bei rein zufälligem Raten wären $7500 : 4 = 1875$ Treffer zu erwarten gewesen. Frage: Können die restlichen $2006 - 1875 = 131$ Treffer durch Zufallsschwankungen erklärt werden ?

Zur Beantwortung dieser Frage bezeichnen wir mit X die Anzahl der Treffer unter der Annahme, daß diese rein zufällig zustande kommen. Wir verwenden den Satz von de Moivre und Laplace mit

$$n = 7500; p = \frac{1}{4}; (1-p) = \frac{3}{4}$$

und erhalten

$$P(X \geq 2006) = P\left(\frac{X - 1875}{\sqrt{7500 \times \frac{1}{4} \times \frac{3}{4}}} \geq \frac{131}{\sqrt{7500 \times \frac{1}{4} \times \frac{3}{4}}}\right).$$

Nach dem Satz von de Moivre und Laplace ist die Größe auf der rechten Seite der Gleichung annähernd normalverteilt, d. h. gemäß $\mathcal{N}(0,1)$. Also

$$P(X \geq 2006) \approx P(X^* \geq 3.5) \approx 0.00023,$$

wobei X^* eine standardnormalverteilte Zufallsvariable bezeichnet. Die Wahrscheinlichkeit dafür, daß die auftretende Differenz das Produkt einer Zufallsschwankung ist, liegt also bei 2.3 Promille und ist damit extrem klein. Trotzdem beweist dieses Experiment nicht mit Sicherheit, daß es ASW gibt, da z.B. im Nachhinein festgestellt wurde, daß der Zufallsgenerator nicht besonders zuverlässig war. (Quellen: C. Tert; Learning to use extrasensory perception, Chicago Univ. Press (1976); M. Gardner; ESP at random, New York book reviews (1977))

Eine weitere Anwendungsmöglichkeit des Satzes von de Moivre und Laplace ist die, auszurechnen, wie groß eine Stichprobe sein muß, um Aussagen über den Parameter p einer Binomialverteilung mit einer gewissen Sicherheit und Genauigkeit machen zu können. Obwohl diese Fragestellung eigentlich in die Statistik gehört, wollen wir uns hierzu schon einmal ein Beispiel anschauen:

Beispiel: In einer Population will man den Anteil an Linkshändern mit 95% Sicherheit auf 1% Genauigkeit bestimmen. Wie viele Personen sollte man dazu (mit Zurücklegen) befragen?

Wir wollen die Wkeit mit Hilfe der Approximation durch die Normalverteilung berechnen. Dazu sei X die Anzahl der Linkshänder in der Stichprobe, $\frac{X}{n}$ ist dann der geschätzte Prozentsatz an Linkshändern in der Gesamtpopulation (warum das eine sinnvolle Schätzung ist, werden wir in dem Kapitel über Statistik diskutieren). Wir wollen, daß

$$|\frac{X}{n} - p| \leq \varepsilon = 0.01$$

und das mit 95% Sicherheit, also

$$P(|\frac{X}{n} - p| \leq 0.01) \geq 0.95. \tag{4.4}$$

Bringt man die Wahrscheinlichkeit auf die Form im Satz von de Moivre und Laplace so ergibt sich:

$$\begin{aligned} P(|\frac{X}{n} - p| \leq 0.01) &= P(-0.01 \leq \frac{X}{n} - p \leq 0.01) \\ &= P\left(\frac{-0.01\sqrt{n}}{\sqrt{p(1-p)}} \leq \frac{X - np}{\sqrt{np(1-p)}} \leq \frac{0.01\sqrt{n}}{\sqrt{p(1-p)}}\right). \end{aligned}$$

Nun kennen wir p dummerweise nicht; aber es gilt stets $p(1-p) \leq \frac{1}{4}$. Setzen wir dies ein, erhalten wir

$$\begin{aligned} & P\left(\frac{-0.01\sqrt{n}}{\sqrt{p(1-p)}} \leq \frac{X}{-np}\sqrt{np(1-p)} \leq \frac{-0.01\sqrt{n}}{\sqrt{p(1-p)}}\right) \\ \geq\ & P\left(-0.01 \times 2\sqrt{n} \leq \frac{X - np}{\sqrt{np(1-p)}} \leq 0.01 \times 2\sqrt{n}\right). \end{aligned}$$

Nach dem Satz von de Moivre und Laplace ergibt sich

$$P\left(-0.01 \times 2\sqrt{n} \leq \frac{X - np}{\sqrt{np(1-p)}} \leq 0.01 \times 2\sqrt{n}\right) \approx \Phi(z) - \Phi(-z) = 2\Phi(z) - 1,$$

da $\Phi(-z) = 1 - \Phi(z)$, wobei

$$z := 0.02\sqrt{n}.$$

Um nun (4.8) zu erfüllen, bestimmen wir aus einer $\mathcal{N}(0,1)$-Tafel z so, daß

$$2\Phi(z) - 1 = 0.95 \Leftrightarrow \Phi(z) = 0.975.$$

Dies ergibt (ungefähr) $z \approx 2$. Setzen wir die Definition von z wieder ein, erhalten wir

$$z = 0.02\sqrt{n} = 2, \text{ d.h.: } n = 10000.$$

Zu bemerken ist noch, daß der benötigte Umfang n der Stichprobe n quadratisch von der Approximationsgenauigkeit ε abhängt. Benötigt man beispielsweise nur eine Genauigkeit von 2% (oder 5%), so genügt eine Stichprobe vom Umfang 2500 (400), um das Ziel mit 95% Sicherheit zu erreichen.

Desweiteren bietet sich noch die Möglichkeit, den Stichprobenumfang durch eine Vorabinformation, wo ungefähr p liegen könnte, zu verkleinern.

5 Die Poisson-Approximation

Im vierten Kapitel hatten wir mit der Normalverteilung die sicherlich wichtigste und meiststudierte Verteilung der W.-Theorie kennengelernt und gesehen, daß man diese als Limes eine geeignet skalierten Binomialverteilung erhalten kann. In diesem Kapitel werden wir eine weitere zentrale Verteilung kennenlernen, die sich ebenfalls als Limes einer (natürlich anders skalierten) Binomialverteilung schreiben läßt.

Wir wollen diese Verteilung an einem Beispiel kennenlernen.
Das Experiment von Rutherford und Geiger
In einem bekannten Experiment beobachteten die Physiker Rutherford und Geiger den Zerfall einer radioaktiven Substanz. Genauer studierten sie die Emission von α-Teilchen eines radioaktiven Präparates in $n = 2608$ Zeitabschnitten von 7.5 Sekunden. Die folgende Tabelle gibt die Versuchsergebnisse wieder. Hierbei steht n_i für jedes natürliche i für die Anzahl der Zeitabschnitte, in denen genau i α-Teilchen emittiert wurden, r_i bezeichnet die relativen Häufigkeiten dieser Zeitabschnitte.

i	n_i	r_i
0	57	0.02186
1	203	0.0778
2	383	0.1469
3	525	0.2013
4	532	0.2040
5	408	0.1564
6	273	0.1047
7	139	0.0533
8	45	0.0173
9	27	0.0103
10	10	0.0038
11	4	0.0015
12	0	0
13	1	0.0004
14	1	0.0004

Offensichtlich sind diese Daten weit davon entfernt von einer Normalverteilung zu stammen. Wir benötigen vielmehr eine Verteilung, die die "Enden", d.h. die großen Zahlen mit einem sehr viel kleineren Gewicht versieht. Eine solche Verteilung ist die *Poisson-Verteilung*.

(5.1) Definition. Sei $\lambda > 0$ eine reelle Zahl. Eine Zufallsgröße X mit $X(\Omega) = \mathbb{N}_0$ und der Verteilung π_λ gegeben durch

$$\pi_\lambda(k) = \frac{e^{-\lambda}}{k!}\lambda^k, \quad k \in \mathbb{N}_0,$$

heißt *Poisson-verteilt mit Parameter* $\lambda > 0$.

Zunächst bemerken wir, daß die Poisson-Verteilung auf den natürlichen Zahlen, incl. der Null $\mathbb{N}_0$ konzentriert ist. Desweiteren überzeugt man sich rasch, daß

$$\sum_{k=0}^{\infty} \pi_\lambda(k) = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} e^{\lambda} = 1$$

ist. π_λ ist also tatsächlich eine Wahrscheinlichkeit.

Der Erwartungswert dieser Verteilung ist leicht zu berechnen:

$$\sum_{k=0}^{\infty} k\pi_\lambda(k) = e^{-\lambda} \sum_{k=0}^{\infty} k\frac{\lambda^k}{k!} = e^{-\lambda}\lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = e^{-\lambda}\lambda \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda}\lambda e^{+\lambda} = \lambda.$$

Eine Poisson-verteilte Zufallsgröße hat also Erwartungswert λ.

Als nächstes wollen wir die Varianz ausrechnen:

$$\begin{aligned} E(X^2) &= \sum_{k=0}^{\infty} k^2 \pi_\lambda(k) = e^{-\lambda} \sum_{k=1}^{\infty} k^2 \frac{\lambda^k}{k!} \\ &= e^{-\lambda} \sum_{k=1}^{\infty} (k(k-1)+k) \frac{\lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^{k+2}}{k!} + \lambda = \lambda^2 + \lambda. \end{aligned}$$

Somit gilt

$$V(X) = E(X^2) - (EX)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

Wir fassen diese beiden Feststellungen noch einmal in folgendem Lemma zusammen.

(5.2) Lemma. *Erwartungswert und Varianz einer Poisson-verteilten Zufallsgröße sind gleich dem Parameter λ.*

Wir wollen nun einmal die eingangs gezeigten Daten aus Rutherford's Experiment mit denen einer Poissonverteilung vergleichen. Dabei stellt sich die Frage, wie wir den Parameter λ am geschicktesten wählen. Vor dem Hintergrund des Gesetzes der großen Zahlen, nach dem man eine mittlere Zahl emittierter Teilchen erwarten kann, die nahe am Erwartungswert liegt und Lemma (5.2) ist eine gute Wahl die, λ als die durchschnittliche Anzahl der Emissionen zu wählen. Diese betrug im Experiment von Rutherford und Geiger

$$a = \frac{10097}{2608} \sim 3.87.$$

Die nächste Tabelle zeigt den Vergleich der relativen Häufigkeiten r_k aus dem Experiment von Rutherford und Geiger mit den Wahrscheinlichkeiten $\pi_\lambda(k)$ einer Poissonverteilung zum Parameter $\lambda = 3.87$.

k	r_k	$\pi_\lambda(k)$
0	0.0219	0.0208
1	0.0778	0.0807
2	0.1469	0.1561
3	0.2013	0.2015
4	0.2040	0.1949
5	0.1564	0.1509
6	0.1047	0.0973
7	0.0533	0.0538
8	0.0173	0.0260
9	0.0103	0.0112
10	0.0038	0.0043
11	0.0015	0.0015
12	0	0.0005
13	0.0004	0.0002
14	0.0004	4×10^{-5}

Die beobachteten relativen Häufigkeiten differieren also von den durch die entsprechende Poisson-Verteilung vorhergesagten Werten nur um wenige Tausendstel. Warum dies ein plausibles Ergebnis ist, soll am Ende dieses Kapitels in einem Satz geklärt werden, der zeigen wird, daß viele Prozesse, die einer Reihe von Anforderungen genügen, eine Poisson-Approximation erlauben. Grundlage dieses Satzes ist eine Festellung darüber, wie genau sich die Binomialverteilung $b(\cdot; n, p)$ für kleine Parameter p und große n durch die Poissonverteilung $\pi_\lambda(k)$ approximieren läßt. Wieder bleibt das Problem, λ zu wählen. Wir lösen es so, daß wir λ so bestimmen, daß die Erwartungswerte der Binomialverteilung und der Poissonverteilung übereinstimmen, daß also $\lambda = np$ ist. Wir wollen also zeigen: $b(k; n, p)$ liegt nahe bei $\pi_\lambda(k)$ für $\lambda = np$.

Um das zu präzisieren, benötigen wir ein Maß für den Abstand zweier Wahrscheinlichkeiten. Dies wird in unserem Fall gegeben sein durch

$$\Delta(n, p) := \sum_{k=0}^{\infty} |b(k; n, p) - \pi_{np}(k)|.$$

$\Delta(n, p)$ läßt sich ähnlich auf für den Abstand beliebiger anderer Wahrscheinlichkeiten definieren und heißt *Abstand der totalen Variation*.

Wir zeigen das folgende Resultat, das sogar noch wesentlich weitreichender ist als unser oben gestecktes Ziel:

(5.3) Satz. Es seien $X_1, \ldots, X_n$ unabhängige Zufallsvariablen, definiert auf einem gemeinsamen Wahrscheinlichkeitsraum, mit $P(X_i = 1) = p_i$ und $P(X_i = 0) = 1 - p_i$ mit $0 < p_i < 1$ für alle $i = 1, \ldots, n$. Sei $X = X_1 + \cdots + X_n$ und $\lambda = p_1 + \cdots + p_n$, dann gilt:

$$\sum_{k=0}^{\infty} |P(X = k) - \pi_\lambda(k)| \leq 2 \sum_{i=1}^{n} p_i^2.$$

Es folgt also im Fall $p = p_1 = \cdots = p_n$:

(5.4) Satz. Für alle $n \in \mathbb{N}$ und $p \in (0,1)$ gilt $\Delta(n,p) \leq 2np^2$.

Die Schranken in den Sätzen (5.3) und (5.4) sind natürlich nur interessant, falls $\sum_{i=1}^n p_i^2$ klein wird bzw. p^2 klein wird gegen n. Offenbar benötigt man in Satz (5.4) dazu mindestens $p \ll \frac{1}{\sqrt{n}}$, d.h. die Wahrscheinlichkeit eines Einzelerfolges wird klein mit n. Aus diesem Grund heißt die Poisson-Verteilung auch Verteilung seltener Ereignisse. Insbesondere folgt der sogenannte *Poissonsche Grenzwertsatz*, der von *Siméon Denis Poisson* (1781-1840) im Jahre 1832 entdeckt wurde:

(5.5) Satz. *(Grenzwertsatz von Poisson)* Ist $\lambda > 0$ und gilt $np_n \to \lambda > 0$ für $n \to \infty$, so gilt für jedes $k \in \mathbb{N}_0$:

$$\lim_{n\to\infty} b(k;n,p_n) = \pi_\lambda(k).$$

(5.5) folgt sofort aus (5.4): Aus $np_n \to \lambda$ folgt $p_n \to 0$ für $n \to \infty$ und $np_n^2 \to 0$. Ferner ist $|b(k;n,p) - \pi_{np}(k)| \leq \Delta(n,p)$ für jedes $k \in \mathbb{N}_0$. Demzufolge gilt

$$\lim_{n\to\infty} |b(k;n,p_n) - \pi_{np_n}(k)| = 0.$$

Wegen $\pi_{np_n}(k) \to \pi_\lambda(k)$ folgt (5.5).

Offenbar unterscheidet sich (5.4) von (5.5) dadurch, daß die Aussage von (5.4) auch im Fall, wo $np_n^2 \to 0$, $np_n \to \infty$ gilt, von Interesse ist (z.B. $p_n = 1/n^{2/3}$). Der wichtigste Vorzug von (5.3) und (5.4) im Vergleich zu (5.5) ist jedoch, daß eine ganz konkrete Approximationsschranke vorliegt. Dafür ist Satz (5.3) auch schwieriger zu beweisen als (5.5) (den wir hier allerdings nur als Korollar aus Satz (5.4) ableiten wollen).

Bevor wir den Beweis von Satz (5.3) geben, stellen wir einen wichtigen Aspekt der Poissonverteilung bereit:

(5.6) Proposition. X und Y seien unabhängig und Poisson-verteilt mit Parametern λ beziehungsweise $\mu > 0$. Dann ist $X + Y$ Poisson-verteilt mit Parameter $\lambda + \mu$.

Beweis. Für $n \in \mathbb{N}_0$ gilt:

$$\begin{aligned}
P(X+Y=n) &= \sum_{k=0}^{n} P(X=k, Y=n-k) \\
&= \sum_{k=0}^{n} P(X=k)P(Y=n-k) \quad \text{(Unabhängigkeit)} \\
&= \sum_{k=0}^{n} \frac{\lambda^k}{k!}\frac{\mu^{n-k}}{(n-k)!}e^{-\lambda}e^{-\mu} = \frac{1}{n!}\left(\sum_{k=0}^{n}\binom{n}{k}\lambda^k\mu^{n-k}\right)e^{-(\lambda+\mu)} \\
&= \frac{1}{n!}(\lambda+\mu)^n e^{-(\lambda+\mu)} = \pi_{\lambda+\mu}(n).
\end{aligned}$$

□

(5.7) Bemerkung. Per Induktion folgt sofort, daß die Summe von endlich vielen unabhängigen Poisson-verteilten Zufallsgrößen wieder Poisson-verteilt ist, wobei der Parameter sich als Summe der Einzelparameter ergibt.

Beweis von Satz 5.3.
Der Beweis des Satzes (5.3) verwendet eine Technik, die man *Kopplung (coupling)* nennt.

Dabei verwenden wir wesentlich, daß bei der Berechnung des Abstands $\sum_{k=0}^{\infty} |P(X=k) - \pi_\lambda(k)|$ die Größen $P(X=k)$ bzw. $\pi_\lambda(k)$ zwar die Verteilungen von Zufallsvariablen sind, daß aber in die Berechnung der zugrunde liegende W.-Raum nicht eingeht. Wir können also einen W.-Raum und Zufallsvariablen mit den gegebenen Verteilungen so wählen, daß sie für unsere Zwecke besonders geeignet sind und das bedeutet, daß sie sich bei gegebener Verteilung möglichst wenig unterscheiden. Konkret konstruieren wir:

Sei $\Omega_i = \{-1, 0, 1, 2, \ldots\}$, $P_i(0) = 1 - p_i$ und $P_i(k) = \frac{e^{-p_i}}{k!} p_i^k$ für $k \geq 1$ sowie $P_i(-1) = 1 - P_i(0) - \sum_{k\geq 1} P_i(k) = e^{-p_i} - (1-p_i)$. Nach Konstruktion sind somit (Ω_i, P_i) W.-Räume. Betrachte dann den Produktraum (Ω, P) der (Ω_i, P_i) im Sinne der Definition (2.13). Wir setzen für $\omega \in \Omega$

$$X_i(\omega) := \begin{cases} 0, & \text{falls } \omega_i = 0, \\ 1, & \text{sonst,} \end{cases}$$

und

$$Y_i(\omega) := \begin{cases} k, & \text{falls } \omega_i = k, k \geq 1, \\ 0, & \text{sonst.} \end{cases}$$

Dann haben nach Definition die Zufallsgrößen X_i die geforderte Verteilung: $P(X_i = 1) = p_i$ und $P(X_i = 0) = 1 - p_i$. Sie sind weiter nach Definition des Produktraumes unabhängig. Die Y_i sind nach Definition Poisson-verteilt zum Parameter p_i und ebenfalls unabhängig. Also folgt mit Proposition (5.6), daß $Y = Y_1 + \cdots + Y_n$ Poisson-verteilt ist zum Parameter λ. Nun stimmen die Zufallsgrößen in den Werten 0 und 1 überein, und es ist $P(X_i = Y_i) = P_i(0) + P_i(1) = (1 - p_i) + e^{-p_i} p_i$, und somit

$$P(X_i \neq Y_i) = p_i(1 - e^{-p_i}) \leq p_i^2,$$

denn für $x > 0$ gilt $1 - e^{-x} \leq x$. Damit folgt

$$\begin{aligned}
&\sum_{k=0}^{\infty} |P(X=k) - \pi_\lambda(k)| = \sum_{k=0}^{\infty} |P(X=k) - P(Y=k)| \\
= &\sum_{k=0}^{\infty} |P(X=k=Y) + P(X=k\neq Y) - (P(X=k=Y) + P(X\neq k=Y))| \\
\leq &\sum_{k=0}^{\infty} P(X=k\neq Y) + P(X\neq k=Y) \\
= &\; 2P(X\neq Y) \leq 2\sum_{i=1}^{n} P(X_i \neq Y_i) \leq 2\sum_{i=1}^{n} p_i^2.
\end{aligned}$$

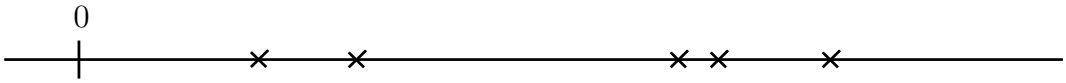
Das beweist Satz (5.3). □

Nun können wir auch klären, warum die Ergebnisse im Experiment von Rutherford und Geiger so erstaunlich nahe an den Vorhersagen einer Poisson–Verteilung lagen. Dies geschieht im Rahmen des sogenannten Poissonschen Punktprozesses.

Der Poissonsche Punktprozeß (Poisson point process)
Wir konstruieren ein mathematisches Modell für auf einer Zeitachse zufällig eintretende Vorkommnisse. Beispiele sind etwa: Ankommende Anrufe in einer Telefonzentrale, Registrierung radioaktiver Teilchen in einem Geigerzähler, Impulse in einer Nervenfaser etc.

Die Zeitachse sei $(0, \infty)$, und die ,,Vorkommnisse" seien einfach zufällige Punkte auf dieser Achse. Die Konstruktion eines unterliegenden Wahrscheinlichkeitsraumes ist leider etwas aufwendig und soll hier einfach weggelassen werden (wir glauben hier einfach mal, daß man das kann).

Ist $I = (t, t + s]$ ein halboffenes Intervall, so bezeichnen wir mit N_I die zufällige Anzahl der Punkte in I. N_I ist also eine Zufallsgröße mit Werten in $\mathbb{N}_0$. Statt $N_{(0,t]}$ schreiben wir auch einfach N_t.



zufällige Punkte

An unser Modell stellen wir eine Anzahl von Bedingungen (P1) bis (P5), die für Anwendungen oft nur teilweise realistisch sind.

(P1) Die Verteilung von N_I hängt nur von der Länge des Intervalls I ab. Anders ausgedrückt: Haben die beiden Intervalle I, I' dieselbe Länge, so haben die Zufallsgrößen N_I und $N_{I'}$ dieselbe Verteilung. Man bezeichnet das auch als (zeitliche) Homogenität des Punktprozesses.

(P2) Sind $I_1, I_2, \ldots, I_k$ paarweise disjunkte Intervalle, so sind $N_{I_1}, N_{I_2}, \ldots, N_{I_k}$ unabhängige Zufallsgrößen.

(P3) Für alle I (stets mit endlicher Länge) existiert EN_I. Um Trivialitäten zu vermeiden, fordern wir:

(P4) Es existiert ein Intervall I mit $P(N_I > 0) > 0$.

Aus (P1), (P3), (P4) lassen sich schon einige Schlüsse ziehen: Sei

$$\lambda(t) = EN_t \geq 0.$$

Offensichtlich gilt $\lambda(0) = 0$, denn N_0 setzen wir natürlich 0. Die Anzahl der Punkte in einer Vereinigung disjunkter Intervalle ist natürlich die Summe für die Einzelintervalle. Insbesondere gilt:

$$N_{t+s} = N_t + N_{(t,t+s]}.$$

Demzufolge:

$$\lambda(t + s) = \lambda(t) + EN_{(t,t+s]},$$

was wegen (P1)

$$= \lambda(t) + \lambda(s)$$

ist.

Nach einem Satz aus der Analysis, der hier nicht bewiesen werden soll, muß eine derartige Funktion linear sein, das heißt, es existiert $\lambda \geq 0$ mit $\lambda(s) = \lambda s$. $\lambda = 0$ können wir wegen (P4) sofort ausschließen. In diesem Fall müßte nach (P1) $EN_I = 0$ für jedes Intervall gelten. Dies widerspricht offensichtlich (P4).

Für kleine Intervalle ist die Wahrscheinlichkeit dafür, daß überhaupt ein Punkt in diesem Intervall liegt, klein. Es gilt nämlich:

$$P(N_I \geq 1) = \sum_{k=1}^{\infty} P(N_I = k) \leq \sum_{k=1}^{\infty} kP(N_I = k) = EN_I$$

und demzufolge

$$P(N_{(t,t+\varepsilon]} \geq 1) \leq \lambda\varepsilon \quad \text{für alle } t, \varepsilon \geq 0.$$

Unsere letzte Forderung besagt im wesentlichen, daß sich je zwei Punkte separieren lassen, es also keine Mehrfachpunkte gibt. Dazu sei für $T > 0$

$$D_T(\omega) := \inf_{t,s \leq T} \{|t - s| : |N_t - N_s| \geq 1\}$$

dann besagt unsere Forderung (P5):

(P5) $P(D_T \leq \alpha_n) \underset{n\to\infty}{\longrightarrow} 0$

für jede Nullfolge α_n und jedes endliche T.

Natürlich haben wir in keiner Weise belegt, daß eine Familie von Zufallsgrößen N_I mit den Eigenschaften (P1)–(P5) als mathematisches Objekt existiert. Wir können dies im Rahmen dieser Vorlesung nicht tun. Wir können jedoch nachweisen, daß für einen Punktprozeß, der (P1) bis (P5) erfüllt, die N_I alle Poisson-verteilt sein müssen:

(5.8) Satz. Sind (P1) bis (P5) erfüllt, so sind für alle $t, s \geq 0$ die Zufallsgrößen $N_{(t,t+s]}$ Poisson-verteilt mit Parameter λs.

Beweis. Wegen (P1) genügt es, $N_s = N_{(0,s]}$ zu betrachten. Wir halten $s > 0$ fest. Für $k \in \mathbb{N}$, $1 \leq j \leq k$, definieren wir

$$X_j^{(k)} := N_{(s(j-1)/k, sj/k]}$$

$$\bar{X}_j^{(k)} := \begin{cases} 1, & \text{falls} \quad X_j^{(k)} \geq 1, \\ 0, & \text{falls} \quad X_j^{(k)} = 0. \end{cases}$$

Für jedes feste k sind die $X_j^{(k)}$ nach (P2) unabhängig und die $\bar{X}_j^{(k)}$ damit ebenfalls.

Wir stellen einige einfach zu verifizierende Eigenschaften dieser Zufallsgrößen zusammen:

$$N_s = \sum_{j=1}^{k} X_j^{(k)}.$$

Sei $\bar{N}_s^{(k)} := \sum_{j=1}^{k} \bar{X}_j^{(k)}$. Dann gilt für jede mögliche Konfiguration der Punkte:

$$\bar{N}_s^{(k)} \leq N_s.$$

Demzufolge gilt für jedes $m \in \mathbb{N}$:

$$P(\bar{N}_s^{(k)} \geq m) \leq P(N_s \geq m). \tag{5.1}$$

Sei $p_k = P(\bar{X}_i^{(k)} = 1) = P(X_i^{(k)} \geq 1) = P(N_{s/k} \geq 1)$.

$$\bar{N}_s^{(k)} \quad \text{ist binomialverteilt mit Parameter} k, p_k. \tag{5.2}$$

Wir verwenden nun (P5), um nachzuweisen, daß sich für große k $\bar{N}_s^{(k)}$ nur wenig von N_s unterscheidet. In der Tat bedeutet ja $\bar{N}_s^{(k)} \neq N_s$, daß es mindestens ein Intervall der Länge $1/k$ gibt, in dem 2 Punkte liegen, also

$$\{\bar{N}_s^{(k)} \neq N_s\} \subseteq \{D_s \leq 1/k\}.$$

Wegen (P5) folgt

$$P(\bar{N}_s^{(k)} \neq N_s) \leq P(D_s \leq 1/k) \to 0 \tag{5.3}$$

für $k \to \infty$. Für $m \in \mathbb{N}$ und $k \in \mathbb{N}$ gilt:

$$\begin{aligned} P(N_s = m) &\geq P(\bar{N}_s^{(k)} = m,\ \bar{N}_s^{(k)} = N_s) \\ &\geq P(\bar{N}_s^{(k)} = m) - P(\bar{N}_s^{(k)} \neq N_s) \\ P(N_s = m) &\leq P(\bar{N}_s^{(k)} = m,\ \bar{N}_s^{(k)} = N_s) + P(\bar{N}_s^{(k)} \neq N_s) \\ &\leq P(\bar{N}_s^{(k)} = m) + P(\bar{N}_s^{(k)} \neq N_s). \end{aligned}$$

Unter Benutzung von (5.2) und (5.3) folgt:

$$P(N_s = m) = \lim_{k\to\infty} P(\bar{N}_s^{(k)} = m) = \lim_{k\to\infty} b(m; k, p_k) \tag{5.4}$$

und analog

$$P(N_s \geq m) = \lim_{k\to\infty} P(\bar{N}_s^{(k)} \geq m). \tag{5.5}$$

Wir zeigen nun:

$$\lim_{k\to\infty} kp_k = \lambda s. \tag{5.6}$$

$$kp_k = E\bar{N}_s^{(k)} = \sum_{j=1}^{\infty} jP(\bar{N}_s^{(k)} = j) = \sum_{l=1}^{\infty} P(\bar{N}_s^{(k)} \geq l).$$

$P(\bar{N}_s^{(k)} \geq l)$ ist nach (5.1) nicht größer als $P(N_s \geq l)$ und strebt nach (5.5) für $k \to \infty$ gegen diese obere Grenze. Nach einem Satz über reelle Zahlenfolgen (falls nicht bekannt oder vergessen: Übungsaufgabe!) folgt daraus

$$\lim_{k\to\infty} kp_k = \lim_{k\to\infty} \sum_{l=1}^{\infty} P(\bar{N}_s^{(k)} \geq l) = \sum_{l=1}^{\infty} P(N_s \geq l) = EN_s = \lambda s.$$

Damit ist (5.6) gezeigt. Unser Satz folgt nun aus (5.4), (5.6) und dem Satz (5.5). □

Der Poissonsche Punktprozeß wird oft verwendet um etwa eintreffende Anrufe in einer Telefonzentrale, ankommende Jobs in einem Computernetzwerk etc. zu modellieren. Man überlegt sich etwa, daß auch das eingangs geschilderte Rutherford-Experiment in diesen Rahmen paßt, wenn man sich die radioaktive Substanz als aus sehr vielen Atomen aufgebaut vorstellt, von denen jedes eine innere Uhr trägt. Diese Uhren laufen unabhängig voneinander und ist die Uhr eines Teilchens abgelaufen, so zerfällt es unter Emission eines α-Teilchens. Man überlegt sich schnell, das in der Regel (P1)–(P5) erfüllt sind, wobei (P2) natürlich nur dann eine Chance hat zu gelten, wenn die Halbwertzeit des Materials sehr groß ist gegenüber der Beobachtungsdauer, während (P5) bedeutet, daß keine zwei Uhren gleichzeitig ablaufen.

Allgemein sind die Annahmen (P1)–(P5) natürlich nicht immer sehr realistisch oder nur näherungsweise richtig. Problematisch in Anwendungen sind oft (P1) und (P2).

Wir wollen das Kapitel abschließen mit einem weiteren Beispiel der Poisson–Approximation in der Physik.

Das Ideale Gas
Ein Ideales Gas in einem Volumen V besteht aus einem System von N nicht–interagierenden Teilchen (den Molekülen). Wir nehmen an, daß V der d-dimensionale Würfel mit Zentrum 0 und Kantenlänge R ist. Wir wollen nun R und N gegen ∞ gehen lassen und zwar so, daß die mittlere Teilchendichte konstant bleibt, d.h. $N/R^d \to \lambda > 0$, wenn $N, R \to \infty$. Dies heißt manchmal auch *thermodynamischer Limes*. Das Hinschreiben eines zugrunde liegenden W.-Raumes bereitet ähnliche Schwierigkeiten wie im Falle des Poissonschen Punktprozesses. Eine gute Wahl für die Zustandmenge wäre beispielsweise die Menge aller Punkte, die die N Teilchen einnehmen können, also das N-fache Produkt des Würfels mit sich selbst. Dieser Raum hat allerdings für uns den Nachteil, nicht abzählbar zu sein.

Wenn wir für den Moment annehmen, daß man diese Schwierigkeiten tatsächlich überwinden kann, so ist es vernünftig anzunehmen, daß die Wahrscheinlichkeit, ein Teilchen, in einer Teilmenge $Q \subset V$ zu finden, proportional ist zum Volumen von Q. Genauer wählen wir die Wahrscheinlichkeit als $p(Q) = \frac{vol(Q)}{R^d}$. Die Annahme, die Teilchen mögen nicht interagieren, drückt sich in der Unabhängigkeit der Orte der einzelnen Teilchen aus, d. h. insbesondere, ob ein Teilchen sich im Volumen Q befindet, hängt nur von Q, nicht aber von den anderen Teilchen ab. Sei nun für festes Q die Größe $\nu^Q(\omega)$ die Anzahl der Teilchen, die sich bei einer zufälligen Verteilung der Teilchen in V in Q einfinden. Dann gilt

Satz 5.9.

$$\lim P(\nu^Q(\omega) = k) = \frac{(\lambda vol(Q))^k}{k!} e^{-\lambda vol(Q)}.$$

Beweis. Der Beweis folgt einer Überlegung, die wir schon kurz bei der Maxwell–Boltzmann–Statistik kennengelernt hatten. Es seien $i_1, \ldots, i_k$ die Indizes der Teilchen, die in Q liegen und $C_{i_1,\ldots,i_k} := \{\omega : x_{i_s} \in Q, 1 \leq s \leq k, x_j \notin Q, j \neq i_1, \ldots i_k\}$. Dann ist offenbar

$$P(\nu^Q(\omega) = k) = \sum P(C_{i_1,\ldots,i_k}).$$

Weiter gilt wegen des obigen Ansatzes für P

$$P(C_{i_1,\dots,i_k}) = \big(\frac{vol(Q)}{R^d}\big)^k \big(1 - \frac{vol(Q)}{R^d}\big)^{N-k}.$$

Und daher

$$P(\nu^Q(\omega) = k)\binom{N}{k}\big(\frac{vol(Q)}{R^d}\big)^k \big(1 - \frac{vol(Q)}{R^d}\big)^{N-k},$$

d.h. $P(\nu^Q(\omega) = k)$ ist binomialverteilt zu den Parametern N und $p_N = \frac{vol(Q)}{r^d}$. Nun ist aber $p_N N = \frac{N vol(Q)}{R^d} \to \lambda vol(Q)$ und daher folgt die Behauptung aus Satz (5.5). □

Dieses Beispiel ist gewissermaßen die d-dimensionale Verallgemeinerung des vorher vorgestellten Poissonschen Punktprozesses. Dieser ist auch in der aktuellen Forschung ein oft verwandtes Modell des Idealen Gases.

6 Große und moderate Abweichungen

In diesem Kapitel wollen wir noch einmal auf das Gesetz der großen Zahlen (Satz (3.30)) eingehen. Wir werden Verschärfungen dieses Gesetzes kennenlernen, die zum einen von theoretischem Interesse sind, zum anderen aber auch von praktischem Nutzen, da sie beispielsweise die Konvergenzgeschwindigkeit im Gesetz der großen Zahlen angeben und somit die Frage klären, wie groß eine Stichprobe, die (3.30) genügt, sein muß, damit der Mittelwert der Stichprobe eine gute Approximation für den Erwartungswert der einzelnen Zufallsvariablen ist (eine wesentliche Fragestellung in der Statistik). Desweiteren werden wir sehen, daß einer der in diesem Kapitel formulierten Sätze einem zentralen und wohlbekannten physikalischen Sachverhalt entspricht.

Wir werden uns zunächst mit der Binomialverteilung beschäftigen. Sei also S_n eine binomialverteilte Zufallsgröße zu den Parametern n und p, d.h.

$$S_n = \sum_{i=1}^{n} X_i,$$

wobei die X_i unabhängige Zufallvariablen sind, die mit Wahrscheinlichkeit p den Wert 1 annehmen und mit Wahrscheinlichkeit $1-p$ den Wert 0.

Unser erster Satz beruht auf der Beobachtung, daß wir im Gesetz der großen Zahlen gesehen hatten, daß die Zufallsvariable $S_n - np$, wenn man sie durch n dividiert, gegen 0 konvergiert (und zwar mit einer Wahrscheinlichkeit, die selber gegen 1 strebt). Normiert man hingegen $S_n - np$, indem man durch $\sqrt{n}$ dividiert, so ergibt sich nach dem Satz von de Moivre und Laplace eine Normalverteilung (die in diesem Fall nicht notwendig Varianz 1 hat). Eine berechtigte Frage ist, was eigentlich "dazwischen" geschieht, d.h., wenn wir $S_n - np$ mit n^α, $1/2 < \alpha < 1$ normieren.

(6.1) Satz. Sei S_n binomialverteilt zu den Parametern n und p. Dann gilt für jedes $1/2 < \alpha \leq 1$ und jedes $\varepsilon > 0$

$$P\left(\left|\frac{S_n - np}{n^\alpha}\right| > \varepsilon\right) \to 0$$

wenn $n \to \infty$.

Beweis. Der Beweis folgt dem Beweis des gewöhnlichen Gesetzes der großen Zahlen (Satz (3.30)). Nach der Tschebyscheff–Ungleichung ist

$$P\left(\left|\frac{S_n - np}{n^\alpha}\right| > \varepsilon\right) \leq \frac{V(\frac{S_n - np}{n^\alpha})}{\varepsilon^2} = \frac{np(1-p)}{n^{2\alpha}\varepsilon^2} \to 0.$$

da $\alpha > 1/2$. Das beweist den Satz. □

Satz (6.1) besagt also, daß das Gesetz der großen Zahlen auch dann erhalten bleibt, wenn wir statt mit n mit n^α, $1/2 < \alpha < 1$, skalieren.

Die nächste Frage, mit der wir uns beschäftigen wollen, ist die nach der Konvergenzgeschwindigkeit im Gesetz der großen Zahlen. Betrachtet man den Beweis von Satz (3.30) noch einmal, so sieht man, daß man mit der üblichen Abschätzung durch die Tchebyscheff–Ungleichung für eine $b(\cdot; n, p)$-verteilte Zufallsvariable S_n eine Schranke der Form

$$P\Big(\Big|\frac{S_n}{n} - p\Big| \geq \varepsilon\Big) \leq \frac{1}{\varepsilon^2} V\Big(\frac{S_n}{n}\Big) = \frac{p(1-p)}{n\varepsilon^2},$$

erhält. Dies ergibt zum Beispiel für den symmetrischen Münzwurf ($p = 1/2$)und $\varepsilon = 1/10$ und $n = 1000$

$$P\Big(\Big|\frac{S_{1000}}{1000} - \frac{1}{2}\Big| \geq \frac{1}{10}\Big) \leq \frac{1}{40}.$$

Diese Abschätzung liegt jedoch um Größenordnungen über der richtigen Wahrscheinlichkeit. Dies sieht man am leichtesten ein, indem man statt der üblichen Tschebyscheff–Ungleichung eine andere From der Markoff–Ungleichung anwendet. Benutzt man diese nämlich mit der monotonen Funktion $\mathbb{R} \ni x \mapsto e^{\lambda x}$, $\lambda > 0$, wobei λ zunächst beliebig ist, so erhält man

$$P\Big(S_n \geq \alpha n\Big) \leq e^{-n\alpha\lambda} E\Big(e^{\lambda S_n}\Big),$$

wobei der Erwartungswert auf der rechten Seite existiert, da S_n nur endlich viele Werte annimmt. Dieser Ansatz geht auf *S.N.Bernstein* zurück. Um diesen Erwartungswert auszuwerten, schreiben wir $\lambda S_n = \sum_{i=1}^n \lambda X_i$, wobei $X_1, \ldots, X_n$ die unabhängigen Zufallsgrößen mit $P(X_i = 1) = p$ und $P(X_i = 0) = (1-p)$ sind, die die Ergebnisse der einzelnen Würfe beschreiben. Da die X_i unabhängig sind, folgt die Unabhängigkeit der $e^{\lambda X_i}$ aus Satz (3.24). Demnach folgt aus der Bemerkung (3.26) für jedes $\lambda > 0$

$$P\Big(S_n \geq \alpha\Big) \leq e^{-n\alpha\lambda} \prod_{i=1}^n E\Big(e^{\lambda X_i}\Big) = e^{-n\alpha\lambda}\Big(E\Big(e^{\lambda X_i}\Big)\Big)^n.$$

Dies berechnen wir als

$$P\Big(S_n \geq \alpha n\Big) \leq e^{-\alpha n\lambda}\Big(pe^{\lambda} + (1-p)\Big)^n = \exp\Big(n\{-\alpha\lambda + \log M(\lambda)\}\Big),$$

wobei $M(\lambda) = pe^{\lambda} + (1-p)$ ist. Es bezeichne $f(\lambda)$ den Ausdruck in den geschweiften Klammern. Wir wollen nun $\lambda > 0$ so wählen, daß wir eine möglichst gute obere Abschätzung erhalten, d. h., wir bestimmen das Minimum von f. Zunächst bemerken wir, daß

$$f''(\lambda) = \frac{M''(\lambda)}{M(\lambda)} - \left(\frac{M'(\lambda)}{M(\lambda)}\right)^2 = \frac{p(1-p)e^{\lambda}}{M(\lambda)^2} > 0$$

für alle $\lambda > 0$ und $0 < p < 1$ ist. Demzufolge ist $f'(\lambda)$ streng monoton steigend. Es existiert also höchstens eine Nullstelle λ_0 von f', und in dieser muß die Funktion f ihr absolutes Minimum annehmen. Ist $\alpha \in (p, 1)$, so ergibt sich aus $f'(\lambda_0) = 0$ nach einer kleinen Rechnung die Nullstelle

$$\lambda_0 = \log \frac{\alpha(1-p)}{p(1-\alpha)} > 0.$$

Einsetzen in f liefert

$$f(\lambda_0) = -\alpha \log\Big(\frac{\alpha}{p}\Big) - (1-\alpha)\log\Big(\frac{1-\alpha}{1-p}\Big) =: -H(\alpha|p).$$

Die Funktion $H(\alpha|p)$ heiß *relative Entropie von α bezüglich p* und hat die folgenden schönen Eigenschaften:

(6.2) Lemma. *Für $0 < p < 1$ ist $H(\cdot|p) \geq 0$ und $H(\alpha|p) = 0$ genau dann wenn $\alpha = p$. Für ein Intervall $I = (a, b)$ ist $\inf_{\alpha \in I} H(\alpha|p) = 0$, falls $p \in \bar{I}$. $H(\cdot|p)$ ist stetig und strikt konvex.*

Beweis. Wir betrachten die folgende Hilfsfunktion $\psi(t) := t \log t - t + 1$ für $t > 0$ und $\psi(0) := 1$. Dann gilt: ψ ist nicht negativ, strikt konvex und $\psi(t) = 0$ genau dann wenn $t = 1$. Es gilt weiter

$$H(\alpha|p) = p\psi\Big(\frac{\alpha}{p}\Big) + (1-p)\psi\Big(\frac{1-\alpha}{1-p}\Big).$$

Somit folgen die Eigenschaften jeweils aus den Eigenschaften der Funktion ψ. Wir betrachten exemplarisch den Beweis der Konvexität: seien $\alpha_1, \alpha_2 \in (0,1)$ und $0 \leq \mu \leq 1$. Dann gilt mittels der Konvexität von ψ

$$\begin{aligned} H(\mu\alpha_1 + (1-\mu)\alpha_2|p) &\leq p\mu\psi\Big(\frac{\alpha_1}{p}\Big) + p(1-\mu)\psi\Big(\frac{\alpha_2}{p}\Big) \\ &+ (1-p)\mu\psi\Big(\frac{1-\alpha_1}{1-p}\Big) + (1-p)(1-\mu)\psi\Big(\frac{1-\alpha_2}{1-p}\Big) \\ &= \mu H(\alpha_1|p) + (1-\mu)H(\alpha_2|p). \end{aligned}$$

□

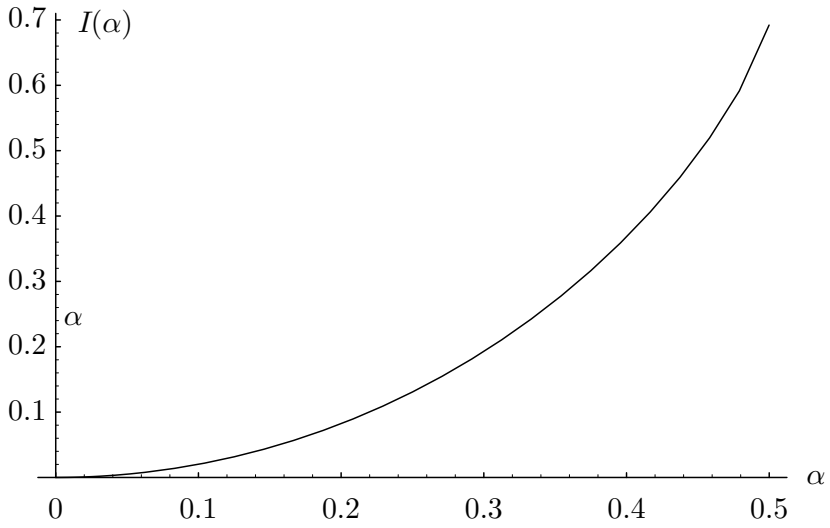
Zusammenfassend haben wir also gezeigt, daß für die Anzahl S_n der „Kopf"-Würfe die Abschätzung

$$P\Big(\bar{S}_n \geq \alpha\Big) \leq \exp(-nH(\alpha|p))$$

für alle $\alpha \in (p, 1)$ gilt (wobei wir $\bar{S}_n$ für $\frac{S_n}{n}$ schreiben). Wir wollen uns fragen, was uns diese Anstrengung gebracht hat. Für den symmetrischen Münzwurf gilt

$$P\Big(\Big|\bar{S}_n - \frac{1}{2}\Big| \geq \alpha\Big) = 2P\Big(\bar{S}_n \geq \alpha + 1/2\Big) \leq 2\exp\Big(-nH(\alpha + 1/2|1/2)\Big)$$

für alle $\alpha \in (0, 1/2)$. Der Graph von $I(\alpha) := H(\alpha + 1/2|1/2)$ ist:



Für $\alpha = 1/10$ und $n = 1000$ erhalten wir zum Beispiel

$$P\Big(\Big|\frac{S_{1000}}{1000} - \frac{1}{2}\Big| \geq \frac{1}{10}\Big) \leq 2\Big(\frac{5}{6}\Big)^{600}\Big(\frac{5}{4}\Big)^{400} \leq 3,6 \cdot 10^{-9},$$

was phantastisch viel besser ist als $1/40$ aus der Tschebyscheff-Ungleichung.

Interessanterweise ist diese Abschätzung "auf einer logarithmischen Skala" schon optimal. Genauer gilt:

(6.3) Satz. *(Prinzip großer Abweichungen von Cramér, large deviation principle)* Bezeichnet S_n die Anzahl der Erfolge bei einem Bernoulli-Experiment zu den Parametern n und p und ist $\bar{S}_n = \frac{S_n}{n}$, so gilt für alle $0 \leq a < b \leq 1$:

$$\lim_{n\to\infty} \frac{1}{n} \log P\Big(\bar{S}_n \in (a,b)\Big) = -\inf_{x\in(a,b)} H(x|p).$$

Beweis. Unser wesentliches Hilfsmittel wird wieder einmal die Stirling–Formel sein. Setzt man $(a,b) =: I$ so gilt:

$$P(\bar{S}_n \in I) = \sum_{na<k<nb} b(k;n,p),$$

wobei $b(k;n,p)$ die Binomialverteilung bezeichnet.
Mit $A_n := (na, nb)$ ist dann

$$\max_{k\in A_n} b(k;n,p) \leq P(S_n \in I) \leq (n+1) \max_{k\in A_n} b(k;n,p).$$

Das asymptotische Verhalten der Wahrscheinlichkeit ist also durch den größten Summanden bestimmt. Die Monotonie der Logarithmusfunktion liefert:

$$\begin{aligned} \max_{k\in A_n} [\frac{1}{n} \log b(k;n,p)] &\leq \frac{1}{n} \log P(\bar{S}_n \in I) \\ &\leq \frac{1}{n} \log(n+1) + \max_{k\in A_n} [\frac{1}{n} \log b(k;n,p)]. \end{aligned}$$

Wir betrachten nun den entscheidenden Term mit Hilfe der Stirlingschen Formel genauer:

$$\begin{aligned}\frac{1}{n}\log\ b(k;n,p) &= \frac{1}{n}\log\ [\binom{n}{k}p^k(1-p)^{n-k}] \\ &= \frac{1}{n}\log\binom{n}{k} + \frac{k}{n}\log p + \frac{n-k}{n}\log(1-p) \quad \text{und} \\ \frac{1}{n}\log\binom{n}{k} &= \log n - \frac{k}{n}\log k - \frac{n-k}{n}\log(n-k) + \frac{1}{n}R_n^k \\ &\text{mit} \quad \lim_{n\to\infty}\frac{1}{n}R_n^k = 0 \quad \forall k.\end{aligned}$$

Hierbei haben wir in dem Term R_n^k sowohl die Logarithmen der $\sqrt{2\pi n}$, $\sqrt{2\pi k}$ bzw. $\sqrt{2\pi(n-k)}$ als auch die Logarithmen der Quotienten aus den Fakultäten und ihren Stirling-Approximationen gesammelt. Da letztere persönlich gegen 0 konvergieren und die Konvergenz von $\frac{(\log n)^\beta}{n^\gamma} \to 0$ für alle $\beta, \gamma > 0$ die ersten Terme gegen 0 streben läßt, gilt in der Tat $\lim_{n\to\infty}\frac{1}{n}R_n^k = 0 \quad \forall k$.

Da $\log n = -\frac{k}{n}\log\frac{1}{n} - \frac{n-k}{n}\log\frac{1}{n}$, folgt insgesamt:

$$\frac{1}{n}\log b(k;n,p) = -\frac{k}{n}\log\frac{\frac{k}{n}}{p} - (1-\frac{k}{n})\log\frac{(1-\frac{k}{n})}{(1-p)} + \frac{1}{n}R_n^k.$$

Erinnern wir uns an die Definition von $H(\cdot|p)$, so erhalten wir

$$\frac{1}{n}\log b(n;k;p) = -H(\frac{k}{n}|p) + \frac{1}{n}R_n^k.$$

Nun ist $\bar{I} = [a,b]$ eine kompakte Menge, und daher nimmt $H(\cdot|p)$ als stetige Funktion sein Minimum auf $[a,b]$ an. Eine kleine Rechnung ergibt, daß die Stetigkeit von $H(\cdot|p)$ zusammen mit der Tatsache, daß sich jedes $x \in [a,b]$ durch eine Folge k/n, $k \in A_n$, approximieren läßt, dann impliziert, daß

$$\lim_{n\to\infty}\max_{k\in A_n} -H(\frac{k}{n}|p) = \max_{x\in[a,b]} -H(x|p) = -\inf_{x\in(a,b)} H(x|p).$$

Insgesamt ergibt sich also

$$\lim_{n\to\infty}\frac{1}{n}\log P(\bar{S}_n \in I) = -\inf_{x\in(a,b)} H(x|p).$$

□

In der Sprache der Wahrscheinlichkeitstheorie haben wir damit für den Münzwurf ein Prinzip großer Abweichungen mit Geschwindigkeit n und Rate $H(\cdot|p)$ bewiesen. Symbolisch schreibt man hierfür:

$$P(\bar{S}_n \in I) \approx \exp(-n\inf_{x\in I} H(x|p)).$$

Das bedeutet, daß die Wahrscheinlichkeit für ein untypisches Verhalten des empirischen Mittelwertes der Anzahl der Erfolge exponentiell klein wird. Untypisch sind hierbei offenbar alle Werte p' für die $H(p'|p) > 0$ ist und somit alle $p' \neq p$. Die Wahrscheinlichkeit

dafür, daß der empirische Mittelwert in einer Menge I liegt, wird gesteuert durch den Wert $p' \in I$ mit minimaler relativer Entropie bzgl. p und das ist aufgrund der Konvexität von $H(\cdot|p)$ dasjenige p' mit geringstem Abstand zu p. Dies ist eine deutliche Verschärfung des Gesetzes der großen Zahlen.

Man kann sich nun natürlich fragen, ob es nicht eine Aussage gibt, die sich zu der eingangs in Satz (6.1) bewiesenen Konvergenzaussage verhält wie das Prinzip großer Abweichungen zum Gesetz der großen Zahlen, eine Aussage, die die Konvergenzgeschwindigkeit in Satz (6.1) angibt. Schon bei der Betrachtung des Beweises von Satz (6.1) kann man den Verdacht hegen, daß die Konvergenzgeschwindigkeit von $\frac{S_n - np}{n^\alpha}$ gegen 0 ganz entscheidend vom gewählten $1/2 < \alpha < 1$ abhängt. Dies ist in der Tat wahr und wird durch den folgenden Satz präzisiert:

(6.4) Satz. *(Prinzip moderater Abweichungen, moderate deviation principle)*
Bezeichnet S_n die Anzahl der Erfolge bei einem Bernoulli-Experiment zu den Parametern n und p, so gilt für alle $-\infty < a < b < \infty$ und alle $1/2 < \alpha < 1$:

$$\lim_{n\to\infty} \frac{1}{n^{2\alpha-1}} \log P\Big(\frac{S_n - np}{n^\alpha} \in (a,b)\Big) = - \inf_{x\in(a,b)} \frac{x^2}{2p(1-p)}.$$

Beweis. Die Tatsache, daß die Aussage des Satzes das Verhalten der S_n "zwischen dem Satz von de Moivre und Laplace und dem Prinzip großer Abweichungen" analysiert, spiegelt sich auch im Beweis wieder. Zunächst benutzen wir die schon im Beweis des Prinzips großer Abweichungen verwendeten Abschätzungen

$$P\left(\frac{S_n - np}{n^\alpha} \in (a,b)\right) = \sum_{n^\alpha a + np < k < n^\alpha b + np} b(k;n,p),$$

und

$$\max_{k\in A_n} b(k;n,p) \le P\left(\frac{S_n - np}{n^\alpha} \in (a,b)\right) \le (n+1) \max_{k\in A_n} b(k;n,p),$$

wobei wir jetzt $A_n := (n^\alpha a + np, n^\alpha b + np)$ wählen. Folgen wir dem obigen Beweis des Prinzips der großen Abweichungen, so erhalten wir wieder

$$\log b(k;n,p) = -nH(\frac{k}{n}|p) + R_n^k.$$

wobei analog zu obigen Überlegungen $\frac{R_n^k}{n^\beta} \to 0$ für jedes $\beta > 0$. Da nun die $\frac{k}{n} \to p$ für alle $k \in A_n$ und $H(p|p) = 0$, können wir nicht ohne weiteres den Beweis des Prinzips der großen Abweichungen "fortführen". Stattdessen fahren wir fort wie im Beweis des Satzes von de Moivre/Laplace und entwickeln die Funktion $H(\cdot|p)$. Wir erhalten

$$H(x|\,p) = \frac{(x-p)^2}{2p(1-p)} + \mathcal{O}(|x-p|^3),$$

wobei wir für zwei Funktionen f, g (in diesem Fall von n) $f = \mathcal{O}(g)$ schreiben, falls es

eine Konstante C gibt, so daß $f \leq Cg$ ist. Insbesondere ist

$$\begin{aligned} nH(\frac{k}{n}|p) &= n\frac{(\frac{k}{n}-p)^2}{2p(1-p)} + n\mathcal{O}(|\frac{k}{n}-p|^3) \\ &= \frac{1}{2p(1-p)}\left(\frac{k-np}{n^\alpha}\right)^2 n^{2\alpha-1} + \mathcal{O}\left(\left|\frac{k-np}{n^\alpha}\right|^3\right)n^{3\alpha-2}. \end{aligned}$$

Somit gilt

$$\begin{aligned} \frac{1}{n^{2\alpha-1}}\log b(k;n,p) &= -\frac{1}{2p(1-p)}\left(\frac{k-np}{n^\alpha}\right)^2 + \mathcal{O}\left(\left|\frac{k-np}{n^\alpha}\right|^3\right)n^{\alpha-1} \\ &\to -\frac{1}{2p(1-p)}\left(\frac{k-np}{n^\alpha}\right)^2, \end{aligned}$$

da aufgrund der Defintion von A_n der Term $\left|\frac{k-np}{n^\alpha}\right|^3$ beschränkt ist. Dies ergibt

$$\lim_{n\to\infty}\frac{1}{n^{2\alpha-1}}P(\frac{S_n-np}{n^\alpha}\in(a,b)) = -\lim_{n\to\infty}\max_{k\in A_n}\left(\frac{k-np}{2p(1-p)n^\alpha}\right)^2.$$

Benutzt man nun wie im Beweis des Prinzips der großen Abweichungen die Tatsache, daß $[a,b]$ eine kompakte Menge ist, daß zu jedem $x\in[a,b]$ eine Folge $a_n = \frac{k-np}{n^\alpha}$ mit $k\in A_n$ existiert, die gegen x konvergiert und die Stetigkeit der Quadratfunktion, so erhält man

$$\lim_{n\to\infty}\max_{k\in A_n}\frac{1}{2p(1-p)}\left(\frac{k-np}{n^\alpha}\right)^2 = \max_{x\in[a,b]}\frac{1}{2p(1-p)}x^2 = \sup_{x\in(a,b)}\frac{1}{2p(1-p)}x^2$$

und damit folgt die Aussage des Satzes sofort. □

Als eine Anwendung und Ausweitung des Prinzips der großen Abweichungen auf die Multinomialverteilung wollen wir ein Grundprinzip der statistischen Mechanik betrachten.

Boltzmanns Gesetz
Zunächst wollen wir in wenigen Worten die Herkunft des Begriffs der Entropie in der Physik klären (dieser hatte námlich ursprünglich wenig mit userem Begriff zu tun).

In der klassischen Mechanik wird der Zustand eines Systems mehrer Teilchen durch Punkte im Phasenraum beschrieben, in dem man die Orts- und Impulskoordinaten aufgeführt hat. Die Bewegung des Systems wird durch ein System gewöhnlicher Differentialgleichungen (Lagrange, Hamilton) beschrieben. Schon Avogadro wußte, daß die Teilchenzahl pro Mol in der Größenordnung von 10^{23} Partikeln liegt. Dies führt zu einem Differentialgleichungssystem, das keiner vernünftigen Behandlung mehr zugänglich ist.

Die Thermodynamik, die in der Mittel des letzten Jahrhunderts entstand, hat das Ziel, das Verhalten eines Systems mit Hilfe makroskopischer Variablen, sogenannte Zustandsgrößen, z.B. Druck, Volumen, Temperatur, innere Energie, oder die 1865 von R. Clausius eingeführte Entropie, zu beschreiben.

Die grundlegende Beobachtung von Clausius, gestützt auf Arbeiten Carnots, war, daß für einen reversiblen, d.h. zeitlich umkehrbaren, thermodynamischen Kreisprozeß das (notwendigerweise entlang einer geschlossenen Kurve verlaufende) Integral über die Änderung

der Wärme dQ pro Temperatur T verschwindet, also in Formeln $\oint \frac{dQ}{T} = 0$. Mathematisch impliziert das (über eine Form des Hauptsatzes der Integral- und Differentialrechnung für Vektorfelder) die Existenz einer Stamm- oder Potentialfunktion für den Integranden, die eine Zustandsfunktion des zugrunde liegenden Systems ist. Diese Zustandsfunktion nannte Clausius nach dem griechischen $\varepsilon\nu\tau\varrho o\pi\eta$ (Umkehr) Entropie. Eine wesentliche Eigenschaft der Entropie ist, daß sie für nicht reversible Prozesse stets positiv ist (und für reversible Prozesse – wie oben erwähnt – verschwindet). Diese Beobachtung führte zur Formulierung des zweiten Hauptsatzes der Thermodynamik:

Prozesse in einem abgeschlossenen thermodynamischen System verlaufen stets so, daß sich die Entropie des Systems vergrößert.

Eine Begründung der thermodynamischen Gesetze auf der Basis der Atomhypothese liefert die statistische Mechanik. Deren Wurzeln wurden mit Hilfe der sich entwickelnden Wahrscheinlichkeitstheorie von L. Boltzmann und J.W. Gibbs gelegt.

Betrachten wir dazu eine Teilchenkonstellation zu einem festen Zeitpunkt, eine sogenannte Konfiguration. Boltzmann ordnete jeder Konfiguration eine Wahrscheinlichkeit zu (und zwar jeder Konfiguration die gleiche) und fragte wieviele Konfigurationen dasselbe makroskopische Bild liefern, also denselben Zustand beschreiben. Er erkannte, daß die wesentlichen Zustände, also diejenigen die man beobachtet, diejenigen mit maximaler Wahrscheinlichkeit sind, sogenannte Gleichgewichtszustände. Ein System tendiert stets zu seinem wahrscheinlichsten Zustand hin, um dann um ihn zu fluktuieren. Bereits 1872 beschrieb er das Verhältnis von Wahrscheinlichkeitstheorie und Mechanik mit den Worten:

"Lediglich dem Umstand, daß selbst die regellosesten Vorgänge, wenn sie unter denselben Verhältnissen vor sich gehen, doch jedes Mal dieselben Durchschnittswerte liefern, ist es zuzuschreiben, daß wir auch im Verhalten warmer Körper ganz bestimmte Gesetze wahrnehmen. Denn die Moleküle der Körper sind ja so zahlreich und ihre Bewegungen so rasch, daß uns nie etwas anderes als jene Durchschnittswerte wahrnehmbar sind. Die Bestimmung der Durchschnittswerte ist Aufgabe der Wahrscheinlichkeitsrechnung."

Boltzmanns wichtigster neuer Gedanke ist also die Idee, daß man für gewönlich Zustände maximaler Wahrscheinlichkeit beobachtet. Andererseits sollten dies nach dem 2. Hauptsatz auch Zustände maximaler Entropie sein. Es liegt also nahe, einen Zusammenhang zwischen Wahrscheinlichkeit und der Entropie herzustellen. Da die Entropie von zwei Systemen gleich der Summe der einzelnen Entropien ist, und die Wahrscheinlichkeit der beiden Systeme im Falle der Unabhängigkeit multiplikativ ist, sollte der Zusammenhang zwischen Entropie und Wahrscheinlichkeit logarithmisch sein:

$$S = k \log W,$$

wobei S die Entropie des Systems ist, W seine Wahrscheinlichkeit und k schließlich ein Proportionalitätsfaktor, die sogenannte Boltzmannkonstante. Boltzmann bestimmte k anhand eines idealen Gases und erhielt den Wert $k = 1,38 \cdot 10^{-23} J/K$. Die Boltzmannkonstante k ist eine fundamentale Naturkonstante.

Dieses Boltzmannsche Gesetz wollen wir im folgenden auf der Basis der großen Abweichungen für das Ideale Gas nachvollziehen. Gegeben sei also ein endliches Volumen V, das

unser Gasbehälter sein soll. In V wollen wir unabhängig n Teilchen realisieren. Wir hatten schon im vorangegangenen Kapitel gesehen, daß der mehrdimensionale Poissonsche Punktprozeß ein gutes Modell für das Ideale Gas darstellt; wir hatten aber auch gesehen, daß wir noch nicht das mathematische Werkzeug besitzen, diesen wirklich zu behandeln. Um diese Probleme zu umgehen, unterteilen wir V in r Zellen Z_1 bis Z_r mit relativen Volumina (in Bezug auf V), $\pi_1 := \frac{vol(Z_1)}{vol(V)}$ bis $\pi_r := \frac{vol(Z_r)}{vol(V)}$. Die Wahrscheinlichkeit in den Zellen $Z_1, \ldots, Z_r$ Teilchenzahlen $k_1, \ldots, k_r$ zu haben (man sagt auch man Besetzungszahlen $k_1, \ldots, k_r$), ist dann gegeben durch die *Multinomialverteilung* zu den Parametern n und $\pi_1, \ldots, \pi_r$, i.e.

$$\frac{n!}{k_1! \cdots k_r!} \pi_1^{k_1} \cdots \pi_r^{k_r}.$$

Sei nun $\mathcal{M}(X)$ die Menge der Wahrscheinlichkeiten auf $X := \{1, \ldots, r\}$, versehen mit der Supremumsnorm $\| \quad \|_{sup}$. Um dies besser zu verstehen, identifizieren wir dabei $\mathcal{M}(X)$ mit $\{(\rho_1, \ldots, \rho_r) \, ; \rho_i \geq 0; \ \sum_{i=1}^r \rho_i = 1\} \subset \mathbb{R}^r$. $\mathcal{M}(X)$ ist daher offenbar kompakt und konvex.
Auf dieser Menge definieren wir eine Entropie–Funktion in Analogie zum Fall $r = 2$:

$$H(\rho|\pi) := \sum_{i=1}^{r} \rho_i \log \frac{\rho_i}{\pi_i} \quad \text{mit } \rho, \pi \in \mathcal{M}(X) \quad \text{und } \pi_i > 0 \quad \forall i.$$

Man beachte, daß diese Defintion konsistent ist mit der Defintion von $H(\cdot|p)$ im Falle $r = 2$. Wieder heißt $H(\rho|\pi)$ die relative Entropie von ρ bezüglich π.
Ebenfalls analog zum Fall $r = 2$ zeigt man: $H(\cdot|\pi)$ ist stetig und konvex und mißt den Abstand zwischen ρ und π in dem Sinne, daß $H(\rho|\pi) \geq 0$ und $H(\rho|\pi) = 0 \Leftrightarrow \rho = \pi$ (siehe $r = 2$).

Wir wollen nun die Wahrscheinlichkeit berechnen, daß *untypische* Besetzungszahlen vorliegen. Dazu sei k_i für festes n und eine feste Beobachtung ω definiert als die Zahl der Teilchen in der Zelle Z_i, und $L_n(\omega, \cdot)$ sei der Vektor der relativen Häufigkeiten der Teilchenzahlen in den verschiedenen Zellen, also $L_n(\omega, i) = \frac{k_i}{n}, i = 1, \ldots, r$.

Weiter sei P_π das durch π gebildete n-fache Produktmaß (definiert wie in Kapitel 3). Wir interessieren uns nun für die Größenordnung von $P_\pi(L_n(\omega, \cdot) \in A)$ für ein $A \subset \mathcal{M}(X)$, das ein untypisches Verhalten beschreibt, d.h. für die Wahrscheinlichkeit in einer Zelle wesentlich mehr oder weniger Teilchen vorzufinden als erwartet. Wir wählen fortan

$$A := \{\nu \in \mathcal{M}(X) : \|\nu - \pi\|_{sup} \geq \varepsilon\}$$

mit $\varepsilon > 0$ (bemerke, daß A abgeschlossen und beschränkt und damit kompakt ist). A ist also die Menge aller Konfigurationen, bei denen die Besetzungszahl mindestens einer Zelle i um mindestens $n\varepsilon$ von der zu erwartenden Zahl $n\pi_i$ abweicht.

Aufgrund der Multinomialverteilung von L_n gilt:

$$P_\pi(L_n(\omega, \cdot) \in A) = \sum_{k=(k_1,\ldots,k_r) \in E_n} \frac{n!}{k_1! \cdots k_r!} \pi_1^{k_1} \cdots \pi_r^{k_r}$$

mit

$$E_n = \{(k_1, \ldots, k_r) : \sum_{i=1}^{r} k_i = n;\ k_i \in \{0, \ldots, n\} \text{ und } \|\frac{k}{n} - \pi\|_{\sup} \geq \varepsilon\}.$$

Wieder erhalten wir

$$\begin{aligned} \max_{k \in E_n} \frac{1}{n} \log(m(n,k)\, \pi^k) &\leq \frac{1}{n} \log P_\pi(L_n(\omega, \cdot) \in A) \\ &\leq \frac{1}{n} \log(n+1)^r + \max_{k \in E_n} \frac{1}{n} \log(m(n,k)\, \pi^k) \end{aligned}$$

mit $m(n,k) := \frac{n!}{k_1! \cdots k_r!}$ und $\pi^k := \pi_1^{k_1} \cdots \pi_r^{k_r}$. Mit der Stirlingschen Formel folgt:

$$\begin{aligned} \frac{1}{n} \log m(n,k) &= \log n - \sum_{j=1}^{r} \frac{k_j}{n} \log k_j + \mathcal{O}\left(\frac{\log n}{n}\right) \\ &= -\sum_{j=1}^{r} \frac{k_j}{n} \log \frac{k_j}{n} + \mathcal{O}\left(\frac{\log n}{n}\right), \end{aligned}$$

da $\log n = -\sum_{j=1}^{r} \frac{k_j}{n} \log \frac{1}{n}$. Somit ist

$$\begin{aligned} \frac{1}{n} \log(m(n,k)\, \pi^k) &= \sum_{j=1}^{r} \frac{k_j}{n} (\log \pi_j - \log \frac{k_j}{n}) + \mathcal{O}\left(\frac{\log n}{n}\right) \\ &= -H(\rho_{\frac{k}{n}} | \pi) + \mathcal{O}\left(\frac{\log n}{n}\right), \end{aligned}$$

wobei $\rho_{\frac{k}{n}}$ der (Wahrscheinlichkeits-)Vektor ist mit Einträgen $\frac{k_i}{n}$ ist. Eingesetzt ergibt das

$$\begin{aligned} \max_{k \in E_n} -H(\rho_{\frac{k}{n}} | \pi) - \mathcal{O}\left(\frac{\log n}{n}\right) &\leq \max_{k \in E_n} \frac{1}{n} \log(m(n,k)\, \pi^k) \\ &\leq \max_{k \in E_n} -H(\rho_{\frac{k}{n}} | \pi) + \mathcal{O}\left(\frac{\log n}{n}\right), \end{aligned}$$

Nun ist $\{\rho \in \mathcal{M}(X) : \rho = \rho_{\frac{k}{n}}, k \in E_n\}$ kompakt und steigt für wachsendes n auf gegen A. Da $H(\cdot|\pi)$ stetig ist, folgt wieder

$$\max_{k \in E_n} -H(\rho_{\frac{k}{n}} | \pi) \to_{n \to \infty} \max_{\rho \in A} -H(\rho | \pi).$$

Also erhalten wir insgesamt:

$$\lim_{n \to \infty} \frac{1}{n} \log P_\pi(L_n(\omega, \cdot) \in A) = \max_{\rho \in A} -H(\rho|\pi) = -\inf_{\rho \in A} H(\rho|\pi).$$

Formal haben wir somit ein Prinzip großer Abweichungen mit Geschwindigkeit n und Rate $H(\cdot|\pi)$ für die Multinomialverteilung gezeigt.

Inhaltlich bedeutet dies, daß in dem oben konstruierten Modell eines Idealen Gases untypische Besetzungszahlen exponentiell unwahrscheinlich sind mit einer Rate, die durch die Entropie der so entstandenen Konfiguration gebildet wird. Dies rechtfertigt die Boltzmannsche Formel, daß die Entropie der Logarithmus der Wahrscheinlichkeit ist.

7 Allgemeine Wahrscheinlichkeitsräume und Zufallsgrößen mit Dichten

In Kapitel 4 sind wir auf Wahrscheinlichkeiten gestoßen, die sich durch Integrale *approximieren* lassen. Wir hatten gesehen, daß für S_n, die Anzahl der Erfolge in einem Bernoulli-Experiment mit Erfolgswahrscheinlichkeit p,

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

gilt. Es ist naheliegend, Zufallsgrößen einzuführen, für die sich $P(a < X \le b)$ durch ein Integral ausdrücken läßt. Gibt es so etwas?

Zunächst sei bemerkt, daß diese Frage für die Ergebnisse von Kapitel 4 irrelevant ist, denn dort ist nur von (diskreten) Zufallsgrößen die Rede, für die sich die entsprechenden Wahrscheinlichkeiten durch Integrale approximieren lassen. Dennoch ist es eine bequeme mathematische Idealisierung, etwa von normalverteilten Zufallsgrößen zu sprechen, d. h. von Zufallsgrößen X mit

$$P(a < X \le b) = \int_a^b \varphi(x)\,dx, \qquad \varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Eine derartige Zufallsgröße hätte eine erstaunliche Eigenschaft: Ist $a \in \mathbb{R}$ beliebig, so gilt

$$P(X = a) \le P\Big(a - \frac{1}{n} < X \le a\Big) = \int_{a-\frac{1}{n}}^{a} \varphi(x)\,dx$$

für alle $n \in \mathbb{N}$, und die rechte Seite konvergiert gegen null für $n \to \infty$. Somit gilt $P(X = a) = 0$ für jedes $a \in \mathbb{R}$. Es ist evident, daß eine Zufallsgröße, wie sie in Kapitel 3 definiert wurde, diese Eigenschaft nicht haben kann. Ist nämlich $p(\omega) > 0$ für ein $\omega \in \Omega$, so gilt $P(X = a) \ge p(\omega) > 0$ für $a = X(\omega)$.

Um z. B. normalverteilte Zufallsgrößen exakt zu definieren, muß der Begriff des W.-Raumes erweitert werden. Offenbar funktioniert unsere bisherige Definition nicht, da $\Omega = \mathbb{R}$ überabzählbar ist. Andererseits gibt es Beispiele (die man beispielsweise in der Analysis III kennenlernt), dafür, daß man nicht mit *jedem* Maß *jede* beliebige Teilmenge eines überabzählbaren Ω messen kann. Man beschränkt sich daher auf Mengensysteme, die mit dem Begriff der Wahrscheinlichkeit konsistent sind. Und zwar ist es plausibel, daß, kennt man die Wahrscheinlichkeit zweier Ereignisse A und B, man auch an der Wahrscheinlichkeit des Eintretens von A oder B oder von A und B interessiert ist, oder auch daran, daß A nicht eintritt. Dies führt zu folgender

(7.1) Definition. Sei Ω eine Menge. Eine nichtleere Familie $\mathcal{F}$ von Teilmengen von Ω heißt *Algebra*, falls für alle $A, B \in \mathcal{F}$ auch A^c, $A \cap B$ und $A \cup B$ in $\mathcal{F}$ sind. Eine Algebra heißt *σ-Algebra*, wenn zusätzlich für jede Folge $(A_n)_{n\in\mathbb{N}}$ aus $\mathcal{F}$ auch $\bigcup_{n=1}^{\infty} A_n$ in $\mathcal{F}$ ist.

Jede Algebra enthält $\emptyset$ und Ω, weil $\emptyset = A \cap A^c$ für $A \in \mathcal{F}$ und $\Omega = \emptyset^c$ gelten. Die einfachste σ-Algebra, die man bilden kann besteht daher aus $\mathcal{F} = \{\emptyset, \Omega\}$.

(7.2) Bemerkung. Ein Mengensystem $\mathcal{F}$ ist genau dann eine σ-Algebra, wenn die folgenden drei Eigenschaften erfüllt sind:

1. $\Omega \in \mathcal{F}$,
2. $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$,
3. Ist $(A_n)_{n\in\mathbb{N}}$ eine Folge in $\mathcal{F}$, so gilt $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$.

Der Beweis ist eine einfache Übungsaufgabe.

Eine σ-Algebra $\mathcal{F}$ sollte man sich als ein hinreichend reichhaltiges Mengensystem vorstellen. Alle abzählbaren Mengenoperationen in $\mathcal{F}$ führen nicht aus $\mathcal{F}$ heraus.

(7.3) Bemerkung. Zu jedem Mengensystem $\mathcal{C}$ in Ω gibt es eine kleinste σ-Algebra $\sigma(\mathcal{C})$, die $\mathcal{C}$ enthält. Dies ist einfach der Durchschnitt aller σ-Algebren, die $\mathcal{C}$ enthalten (und dies ist als unmittelbare Folgerung aus der Definition wieder eine σ-Algebra). Mindestens eine σ-Algebra, nämlich $\mathcal{P}(\Omega)$ (die Potenzmenge), umfaßt $\mathcal{C}$.

(7.4) Beispiel. Das für uns wichtigste Beispiel ist $\Omega = \mathbb{R}^n$. Sei $\mathcal{C}$ die Familie aller nach links halboffenen Intervall. Dabei ist für $a = (a_1, \ldots, a_n), b = (b_1, \ldots, b_n) \in \mathbb{R}^n$ mit $a \le b$ (d.h. $a_i \le b_i$ für alle i) ein nach links halboffenes Intervall definiert durch

$$]a, b] = \{\, x = (x_1, \ldots, x_n) \in \mathbb{R}^n : a_i < x_i \le b_i \quad \text{für} \quad i = 1, \ldots, n\}.$$

Dann heißt $\mathcal{B}_n := \sigma(\mathcal{C})$ die *Borelsche σ-Algebra* in $\mathbb{R}^n$, und die zu $\mathcal{B}_n$ gehörigen Mengen heißen *Borelsche Mengen (Borel sets)*. Da sich jede offene Teilmenge des $\mathbb{R}^n$ als abzählbare Vereinigung von Intervallen schreiben läßt, ist jede offene Menge (und damit auch jede abgeschlossene Menge) in $\mathbb{R}^n$ Borelsch.

Wie definieren nun einen allgemeinen Wahrscheinlichkeitsraum:

(7.5) Definition. Sei Ω eine Menge und $\mathcal{F}$ eine σ-Algebra von Teilmengen von Ω. Ein *Wahrscheinlichkeitsmaß (probability measure)* ist eine auf $\mathcal{F}$ definierte Funktion P mit Werten in $[0, 1]$, welche den folgenden Bedingungen genügt:

1. $P(A) \ge 0$ für alle $A \in \mathcal{F}$,
2. $P(\Omega) = 1$,
3. P ist *σ-additiv*, d.h., für disjunkte $A_1, A_2, \ldots \in \mathcal{F}$ gilt

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

$(\Omega, \mathcal{F}, P)$ heißt dann *Wahrscheinlichkeitsraum (probability space)*, P *Wahrscheinlichkeit (probability)*.

Im diskreten Fall hatten wir jede Abbildung X von Ω nach $\mathbb{R}$ Zufallsgröße genannt. Für einen allgemeinen Wahrscheinlichkeitsraum ist dies nicht zweckmäßig. Wir wollen Wahrscheinlichkeiten von Ereignissen der Form $\{a < X \le b\}$ bestimmen. Für unsere Zwecke

genügt die folgende Definition:

(7.6) Definition. Sei $(\Omega, \mathcal{F}, P)$ ein W.-Raum und $X : \Omega \to \mathbb{R}$ eine Abbildung. X heißt *Zufallsgröße (random variable)* (oder Zufallsvariable), wenn für alle $a \in \mathbb{R}$ gilt:

$$X^{-1}(]-\infty, a]) \in \mathcal{F}.$$

(7.7) Bemerkungen. Der Begriff Zufallsgröße hat zunächst nichts mit der Wahrscheinlichkeit P zu tun. Liegt keine Wahrscheinlichkeit vor, so spricht man von einer *meßbaren (measurable)* Abbildung auf $(\Omega, \mathcal{F})$. Die Familie $\mathcal{F}_X := \{A \subset \mathbb{R} : X^{-1}(A) \in \mathcal{F}\}$ ist eine σ-Algebra. Dies ist eine einfache Übung. Ist X eine Zufallsgröße, so gilt nach Definition $]-\infty, a] \in \mathcal{F}_X$ für jedes $a \in \mathbb{R}$. Somit liegt auch jedes Intervall der Form $]a, b] =]-\infty, b] \cap (]-\infty, a])^c$ in $\mathcal{F}_X$. Da $\mathcal{B}_1$ von Intervallen dieser Form erzeugt wird, liegt somit (unmittelbare Folgerung der Definition (7.6)) das Urbild jeder Borelschen Menge in $\mathcal{F}$. Eine äquivalente Definition einer Zufallsgröße ist also durch die Forderung gegeben, daß das Urbild jeder Borelschen Menge in der vorgegebenen σ-Algebra $\mathcal{F}$,,landet".

Schließlich bemerken wir auch noch, daß unser "neuer" Begriff einer Zufallsgröße konsistent ist mit dem Begriff, den wir für diskrete Ω geprägt hatten. Dort benutzt man ja die Potenzmenge $\mathcal{P}(C)$ als $\mathcal{F}$. Somit ist

$$X^{-1}(]-\infty, a]) \in \mathcal{F}.$$

trivialerweise immer erfüllt.

Wir führen nun den Begriff der Dichte ein.

(7.8) Definition. Eine Lebesgue-integrierbare Funktion $f : \mathbb{R} \to [0, \infty)$ heißt *Dichte (density)*, wenn $\int_{-\infty}^{\infty} f(x)\, dx = 1$ gilt. ($\int \ldots dx$ bezeichne das Lebesgue-Integral.)

Falls das Lebesgue-Integral nicht bekannt ist, so setze man voraus, daß f Riemann-integrierbar ist und das uneigentliche Riemann-Integral $\int_{-\infty}^{\infty} f(x)\, dx$ existiert und gleich 1 ist.

(7.9) Beispiele.

1. Die Dichte der *Standard-Normalverteilung (standard normal distribution)* ist definiert durch
$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad x \in \mathbb{R}.$$
Wir hatten schon in (6.22) gesehen, daß $\int_{-\infty}^{\infty} \varphi(x)\, dx = 1$ ist.

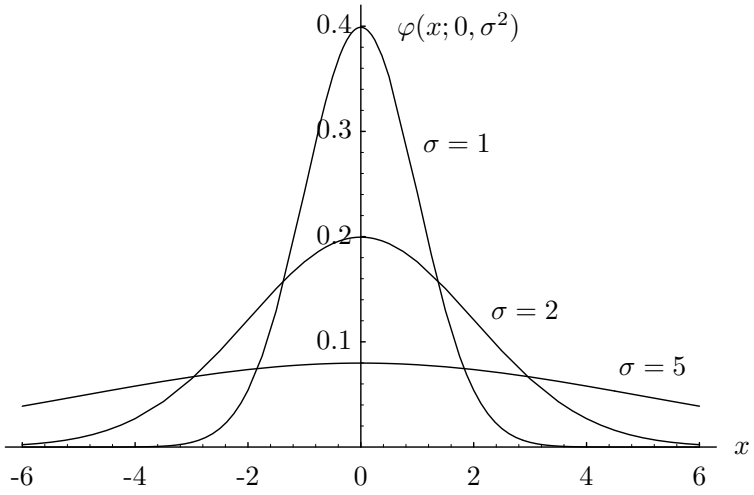
2. Die Dichte der *Normalverteilung (normal distribution) mit Mittel* $\mu \in \mathbb{R}$ *und Varianz* $\sigma^2 > 0$ ist definiert durch
$$\varphi(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}, \quad x \in \mathbb{R},$$
wobei die Namensgebung der Parameter $\mu \in \mathbb{R}$ und $\sigma > 0$ im Beispiel (7.14 (2)) klar werden wird. Durch die Transformation $y = (x - \mu)/\sigma$ geht die Dichte $\varphi(\,\cdot\,; \mu, \sigma^2)$

in die Dichte $\varphi(\,\cdot\,;0,1)$ der Standard-Normalverteilung aus Beispiel (1) über, und es gilt

$$\int_{-\infty}^{\infty} \varphi(x;\mu,\sigma^2)\,dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\,dy = 1$$

gemäß (6.22).

Normalverteilungen



3. Für $a < b$ ist die Dichte der *gleichförmigen Verteilung (uniform distribution) auf* $[a,b]$ definiert durch

$$f(x) = \begin{cases} 1/(b-a) & \text{für } x \in [a,b], \\ 0 & \text{für } x \in \mathbb{R} \setminus [a,b]. \end{cases}$$

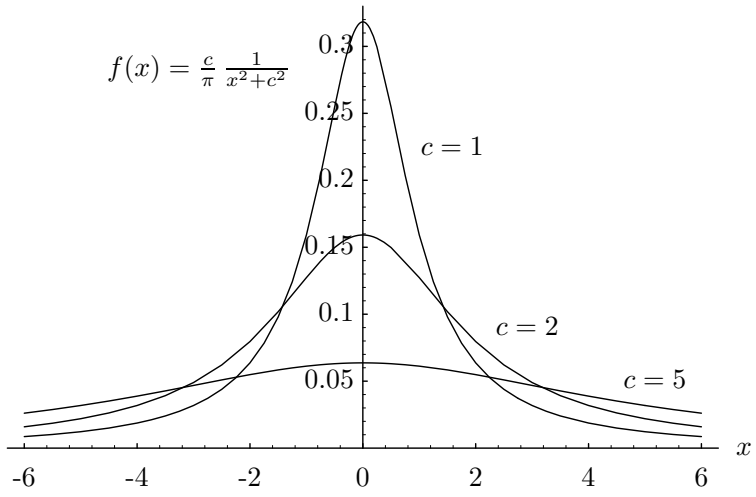
4. Die Dichte der *Exponentialverteilung (exponential distribution) zum Parameter* $\lambda > 0$ ist definiert durch

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{für } x \geq 0, \\ 0 & \text{für } x < 0. \end{cases}$$

5. Die Dichte der *Cauchy-Verteilung zum Parameter* $c > 0$ ist definiert durch

$$f(x) = \frac{c}{\pi}\,\frac{1}{x^2 + c^2}, \quad x \in \mathbb{R}.$$

Cauchy-Verteilungen



(7.10) Definition. Eine Funktion $F : \mathbb{R} \to [0,1]$ heißt *Verteilungsfunktion (distribution function),* wenn die folgenden Eigenschaften gelten: (i) F ist monoton steigend (nondecreasing), d. h. für alle $s \le t$ gilt $F(s) \le F(t)$.

(ii) F ist rechtsseitig stetig (right-continuous), d. h. für jedes $t \in \mathbb{R}$ und jede gegen t konvergente Folge $\{t_n\}_{n\in\mathbb{N}}$ mit $t_n \ge t$ für alle $n \in \mathbb{N}$ gilt $\lim_{n\to\infty} F(t_n) = F(t)$.

(iii) $\lim_{t\to\infty} F(t) = 1$ und $\lim_{t\to-\infty} F(t) = 0$.

Bemerkung. Für jede Dichte f ist natürlich $\int_{-\infty}^{t} f(s)\,ds$ eine Verteilungsfunktion, die nicht nur (ii) erfüllt, sondern sogar stetig ist. Wir nennen eine stetige Funktion $F : \mathbb{R} \to [0,1]$, die (i) und (iii) erfüllt, eine *stetige* Verteilungsfunktion. Nicht jede stetige Verteilungsfunktion hat eine Dichte, was hier nicht gezeigt wird.

(7.11) Definition. Es seien $(\Omega, \mathcal{F}, P)$ ein Wahrscheinlichkeitsraum und X eine Zufallsgröße, dann heißt die Funktion $F_X(t) := P(X \le t)$, $t \in \mathbb{R}$, die *Verteilungsfunktion* von X.

Für eine Zufallsgröße X, wie sie in Kapitel 3 definiert wurde, läßt sich die Verteilungsfunktion leicht beschreiben: In den (höchstens abzählbar vielen) Punkten $t \in X(\Omega)$ hat F_X einen Sprung der Höhe $P(X = t)$ und ist in diesem Punkt rechtsseitig stetig. Ansonsten ist sie konstant. Offensichtlich erfüllt F_X dann (i)–(iii) der Definition (7.10).

(7.12) Definition. Es seien $(\Omega, \mathcal{F}, P)$ ein Wahrscheinlichkeitsraum und f eine Dichte. Eine Zufallsgröße X heißt *absolutstetig* mit Dichte f, falls

$$F_X(t) = \int_{-\infty}^{t} f(s)\,ds$$

für alle $t \in \mathbb{R}$ gilt. Ist X absolutstetig mit Verteilungsfunktion F_X so nennt man auch F_X absolutstetig.

Eine Dichte ist nicht ganz eindeutig durch die Zufallsgröße bzw. deren Verteilungsfunktion bestimmt. Hat zum Beispiel X die in (7.9 (3)) angegebene Dichte, so ist

$$\widetilde{f}(x) = \begin{cases} 1/(b-a) & \text{für } x \in (a,b), \\ 0 & \text{für } x \in \mathbb{R} \setminus (a,b), \end{cases}$$

ebensogut eine Dichte für X. Änderungen einer Dichte in abzählbar vielen Punkten (oder allgemeiner: auf einer Nullmenge bezüglich des Lebesgue-Maßes) ändern an den Integralen nichts.

Eine absolutstetige Zufallsvariable braucht natürlich keine stetige Dichte zu besitzen. Ist jedoch eine Dichte f in einem Punkt a stetig, so gilt nach dem Fundamentalsatz der Differential- und Integralrechnung

$$f(a) = \frac{dF(x)}{dx}\Big|_{x=a};$$

also hat eine Verteilungsfunktion F genau dann eine stetige Dichte, wenn sie stetig differenzierbar ist. Diese stetige Dichte ist, wenn sie existiert, eindeutig durch F bestimmt.

Hat eine Zufallsgröße X eine Dichte f, so gilt für alle $a < b$

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = \int_a^b f(x)\,dx.$$

Mit dem zu Beginn des Kapitels vorgestellten Argument folgt, daß $P(X = x) = 0$ für alle $x \in \mathbb{R}$ ist, wenn X eine Dichte besitzt. Demzufolge gilt

$$P(a < X \leq b) = P(a \leq X \leq b) = P(a \leq X < b) = P(a < X < b).$$

Wir nennen eine Zufallsgröße *normalverteilt, gleichförmig verteilt, exponentialverteilt* bzw. *Cauchy-verteilt,* wenn sie eine Dichte gemäß Beispiel (7.9 (2)), (3), (4) bzw. (5) hat.

(7.13) Definition. Die Zufallsgröße X auf einem W.-Raum $(\Omega, \mathcal{F}, P)$ habe eine Dichte f. Sei $g : \mathbb{R} \to \mathbb{R}$ eine meßbare Abbildung bezüglich der Borelschen Mengen auf $\mathbb{R}$.

(a) Ist die Funktion $\mathbb{R} \ni x \mapsto g(x)f(x)$ Lebesgue-integrierbar, so sagen wir, daß der *Erwartungswert von* $g(X)$ existiert. Er ist dann definiert durch

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)\,dx.$$

(b) Ist $g(x) = x$ und $\mathbb{R} \ni x \mapsto xf(x)$ Lebesgue-integrierbar, so sagen wir, daß der *Erwartungswert von* X existiert. Er ist dann definiert durch

$$E(X) = \int_{-\infty}^{\infty} xf(x)\,dx.$$

(c) Es existiere $E(X)$ und es sei $g(x) = (x - E(X))^2$. Ist $\mathbb{R} \ni x \mapsto (x - E(X))^2 f(x)$ Lebesgue-integrierbar, so ist die *Varianz von* X definiert durch

$$V(X) = \int_{-\infty}^{\infty} (x - E(X))^2 f(x)\,dx.$$

Bemerkung. Die eigentliche Idee hinter dieser Konstruktion ist die folgende: Für eine

diskrete Zufallsvariable X ist wohlbekannt, daß man den Erwartungswert als

$$\sum_{a_i \in X(\Omega)} a_i P(X = a_i)$$

definiert ist. Eine beliebige Zufallsvariable X "diskretisiert" man, indem man für $k \in \mathbb{Z}$ und $n \in \mathbb{N}$ die Mengen

$$A_{nk} := \{k/n \leq X \leq (k+1)/n\}$$

und neue Zufallsvariablen

$$X_n := \sum_{k=-\infty}^{\infty} (k/n) 1_{A_{nk}}$$

definiert. Die X_n steigen gegen X auf und es ist $X_1 \leq X < X_1 + 1$. Daher definiert man den Erwartungswert von X, falls EX_1

$$\lim_n EX_n$$

existiert und setzt ihn in diesem Fall gleich dem obigen Limes.

Vor dem Hintergrund dieser Konstruktion und der entsprechenden des Lebesgue-Integrals, wird man schnell für sich klären können, daß die Definition des Erwartungswertes und der Varianz mit den Definitionen dieser Größen in Kapitel 3 im Fall diskreter W.-Räume zusammenfällt. Man muß natürlich wichtige Eigenschaften wie zum Beispiel die Linearität des Erwartungswertes erneut beweisen. Wir wollen uns diese Arbeit hier ersparen.

(7.14) Beispiele.

(1) Sei X standardnormalverteilt. Dann ist

$$\int_{-\infty}^{\infty} |x| \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = \frac{2}{\sqrt{2\pi}} \int_0^{\infty} x e^{-x^2/2}\, dx = \frac{2}{\sqrt{2\pi}} \left(-e^{-x^2/2}\right)\Big|_0^{\infty} = \sqrt{\frac{2}{\pi}} < \infty,$$

also existiert der Erwartungswert von X, und es gilt

$$E(X) = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = 0,$$

da der Integrand eine ungerade Funktion ist. Die Varianz berechnet sich wie folgt: Es gilt

$$V(X) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2}\, dx = \lim_{N\to\infty} \frac{1}{\sqrt{2\pi}} \int_{-N}^{N} x(x e^{-x^2/2})\, dx,$$

und mittels partieller Integration folgt

$$V(X) = \lim_{N\to\infty} \frac{1}{\sqrt{2\pi}} (-x e^{-x^2/2})\Big|_{-N}^{N} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2}\, dx = 0 + 1 = 1.$$

(2) Sei X normalverteilt mit den Parametern $\mu \in \mathbb{R}$ und $\sigma > 0$. Mit der Transformation $y = (x - \mu)/\sigma$ folgt unter Verwendung von Beispiel (1)

$$\begin{aligned} \int_{-\infty}^{\infty} |x| \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}\, dx &= \int_{-\infty}^{\infty} |\mu + \sigma y| \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy \\ &\leq |\mu| + \sigma \int_{-\infty}^{\infty} |y| \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy < \infty, \end{aligned}$$

also existiert der Erwartungswert, und es gilt

$$E(X) = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (y\sigma + \mu) e^{-y^2/2}\, dy = \mu.$$

Mit der gleichen Transformation und dem Ergebnis aus Beispiel (1) folgt

$$V(X) = \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} dx = \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y^2 e^{-y^2/2}\, dy = \sigma^2.$$

Eine Zufallsgröße X ist genau dann normalverteilt mit Erwartungswert μ und Varianz σ^2, wenn $(X-\mu)/\sigma$ standardnormalverteilt ist. Etwas allgemeiner: Ist X normalverteilt mit Erwartungswert μ und Varianz σ^2, und sind $a, b \in \mathbb{R}$, $a \neq 0$, so ist $aX + b$ normalverteilt mit Erwartungswert $a\mu + b$ und Varianz $a^2\sigma^2$. Dies ergibt sich im Fall $a > 0$ aus der Tatsache, daß sowohl $P(X \leq t) = P(aX+b \leq at+b)$ als auch (mittels der Transformation $y = ax + b$)

$$\int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}\, dx = \int_{-\infty}^{at+b} \frac{1}{\sqrt{2\pi}a\sigma} e^{-(y-a\mu-b)^2/2a^2\sigma^2}\, dy$$

für alle $t \in \mathbb{R}$ gelten, also $\varphi(\,\cdot\,; a\mu + b,\, a^2\sigma^2)$ eine Dichte von $aX + b$ ist.

(3) Sei X exponentialverteilt mit Parameter $\lambda > 0$. Partielle Integration ergibt

$$E(X) = \int_0^{\infty} \lambda x e^{-\lambda x}\, dx = -x e^{-\lambda x}\Big|_0^{\infty} + \int_0^{\infty} e^{-\lambda x} dx = 0 + \Big(-\frac{1}{\lambda} e^{-\lambda x}\Big)\Big|_0^{\infty} = \frac{1}{\lambda},$$

insbesondere existiert der Erwartungswert. Ausmultiplizieren von $(x - 1/\lambda)^2$, Verwenden von $E(X) = 1/\lambda$ und zweimalige partielle Integration liefern

$$V(X) = \int_0^{\infty} \Big(x - \frac{1}{\lambda}\Big)^2 \lambda e^{-\lambda x}\, dx = \int_0^{\infty} \lambda x^2 e^{-\lambda x}\, dx - \frac{2}{\lambda} E(X) + \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

Als nächstes wollen wir gemeinsame Eigenschaften von mehreren Zufallsgrößen $X_1, \ldots, X_n$, definiert auf einem gemeinsamen W.Raum $(\Omega, \mathcal{F}, P)$, betrachten.

(7.15) Definition.

a) Eine Lebesgue-integrierbare Funktion $f : \mathbb{R}^n \to [0, \infty)$ heißt *n-dimensionale Dichte*, wenn

$$\int_{\mathbb{R}^n} f(x)\, dx = 1$$

ist, wobei x ein n-Tupel $(x_1, \ldots, x_n)$ aus dem $\mathbb{R}^n$ bezeichnet.

b) Die Funktion f sei eine n-dimensionale Dichte, und $X_1, \ldots, X_n$ seien n Zufallsgrößen. Man sagt, daß sie die *gemeinsame Dichte (joint density)* f haben, wenn

$$P(X_1 \leq a_1,\, X_2 \leq a_2, \ldots, X_n \leq a_n) = \int_{(-\infty, a_1] \times \cdots \times (-\infty, a_n]} f(x)\, dx$$

für alle $a_1, \ldots, a_n \in \mathbb{R}$ gilt.

(7.16) Definition. $X_1, \ldots, X_n$ seien n Zufallsgrößen, definiert auf einem gemeinsamen W.-Raum $(\Omega, \mathcal{F}, P)$. Sie heißen *unabhängig,* wenn für alle $a_1, \ldots, a_n \in \mathbb{R}$ gilt:

$$P(X_1 \le a_1, \ldots, X_n \le a_n) = P(X_1 \le a_1) \cdots P(X_n \le a_n).$$

Bemerkung. Man prüft leicht nach, daß diese Definition für diskrete Zufallsgrößen äquivalent zu der in Kapitel 3 gegebenen ist.

(7.17) Satz. $X_1, \ldots, X_n$ seien n Zufallsgrößen, definiert auf einem gemeinsamen W.-Raum $(\Omega, \mathcal{F}, P)$. Jedes der X_j habe eine Dichte f_j. (Wir setzen nicht voraus, daß eine gemeinsame Dichte existiert.) Dann sind die Zufallsgrößen $X_1, \ldots, X_n$ genau dann unabhängig, wenn eine gemeinsame Dichte für $X_1, \ldots, X_n$ durch $\mathbb{R}^n \ni (x_1, x_2, \ldots, x_n) \mapsto f_1(x_1) f_2(x_2) \ldots f_n(x_n)$ gegeben ist.

Beweis. Ist $\mathbb{R}^n \ni (x_1, x_2, \ldots, x_n) \mapsto f_1(x_1) f_2(x_2) \ldots f_n(x_n)$ eine gemeinsame Dichte, so ergibt sich für alle $a_1, \ldots, a_n \in \mathbb{R}$

$$\begin{aligned} P(X_1 \le a_1, \ldots, X_n \le a_n) &= \int_{-\infty}^{a_1} \ldots \int_{-\infty}^{a_n} f_1(x_1) \ldots f_n(x_n)\, dx_n \ldots dx_1 \\ &= \prod_{j=1}^{n} \int_{-\infty}^{a_j} f_j(x_j)\, dx_j = \prod_{j=1}^{n} P(X_j \le a_j). \end{aligned}$$

Somit sind $X_1, \ldots, X_n$ unabhängig. Gilt umgekehrt letzteres, so folgt

$$\begin{aligned} P(X_1 \le a_1, \ldots, X_n \le a_n) &= \prod_{j=1}^{n} P(X_j \le a_j) \\ &= \prod_{j=1}^{n} \int_{-\infty}^{a_j} f_j(x_j)\, dx_j \\ &= \int_{-\infty}^{a_1} \ldots \int_{-\infty}^{a_n} f_1(x_1) \ldots f_n(x_n)\, dx_n \ldots dx_1, \end{aligned}$$

und somit ist $\mathbb{R}^n \ni (x_1, \ldots, x_n) \mapsto f_1(x_1) \ldots f_n(x_n)$ eine gemeinsame Dichte. □

Wir wollen nun die Dichte von $X + Y$ berechnen, wenn X und Y unabhängig sind, und ihre Verteilungen durch die Dichten f und g gegeben sind. Wir bemerken zunächst, daß $X + Y$ nach einer Übung wieder eine Zufallsgröße ist. Wir wollen $P(X + Y \le a)$ für alle $a \in \mathbb{R}$ bestimmen. Mit $C_a := \{(x, y) \in \mathbb{R}^2 : x + y \le a\}$ können wir dies als $P((X, Y) \in C_a)$ schreiben. Wichtig ist die Tatsache, daß aus der definierenden Eigenschaft (7.15(b)) folgt, daß für Teilmengen $C \subset \mathbb{R}^n$, für die die Funktion $\mathbb{R}^n \ni x \mapsto 1_C(x) f(x)$ Lebesgue-integrierbar ist,

$$P((X_1, \ldots, X_n) \in C) = \int_C f(x)\, dx$$

gilt. Wir wollen dies hier nicht beweisen. Es sei auf eine Vorlesung ,,Wahrscheinlichkeitstheorie“ verwiesen. Es gilt mit der Substitution $u = x + y$ und $v = y$ nach Satz (7.17):

$$\begin{aligned} P(X + Y \le a) &= \int_{C_a} f(x)g(y)\,dx\,dy \\ &= \int_{-\infty}^{a}\int_{-\infty}^{\infty} f(u - v)g(v)\,dv\,du. \end{aligned}$$

Somit gilt:

(7.18) Satz. Es seien X und Y unabhängige Zufallsgrößen. X habe die Dichte f und Y die Dichte g. Dann hat $X + Y$ die Dichte

$$h(x) = \int_{-\infty}^{\infty} f(x - y)g(y)\,dy, \quad x \in \mathbb{R}. \tag{7.1}$$

Sind f und g zwei Dichten, so definiert (*) eine neue Dichte h, die man als die *Faltung (convolution)* von f und g bezeichnet und meist als $f * g$ schreibt.

Als Anwendung von (7.18) können wir eine wichtige Eigenschaft von normalverteilten Zufallsgrößen zeigen:

(7.19) Satz. Es seien X_i, $1 \le i \le n$, unabhängige und normalverteilte Zufallsgrößen mit Erwartungswerten μ_i und Varianzen σ_i^2. Dann ist $\sum_{i=1}^n X_i$ normalverteilt mit Erwartungswert $\sum_{i=1}^n \mu_i$ und Varianz $\sum_{i=1}^n \sigma_i^2$.

Beweis. Sind $X_1, \ldots, X_n$ unabhängig, so sind $X_1 + \cdots + X_{n-1}$ und X_n ebenfalls unabhängig (warum?). Der Satz folgt also mit Induktion nach n aus dem Fall $n = 2$.

Die Zufallsgrößen $Y_1 = X_1 - \mu_1$ und $Y_2 = X_2 - \mu_2$ sind normalverteilt mit Erwartungswert 0. Nach (7.18) ist die Dichte h von $Y_1 + Y_2$ gegeben durch

$$h(x) = \frac{1}{2\pi\sigma_1\sigma_2}\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\left[\frac{(x-y)^2}{\sigma_1^2} + \frac{y^2}{\sigma_2^2}\right]\right) dy$$

für alle $x \in \mathbb{R}$. Schreibt man den Term in der eckigen Klammer in der Form

$$\frac{(x-y)^2}{\sigma_1^2} + \frac{y^2}{\sigma_2^2} = \left(\frac{\sqrt{\sigma_1^2 + \sigma_2^2}}{\sigma_1\sigma_2}y - \frac{\sigma_2}{\sigma_1\sqrt{\sigma_1^2 + \sigma_2^2}}x\right)^2 + \frac{x^2}{\sigma_1^2 + \sigma_2^2}.$$

und benutzt die Transformation

$$z(y) = \frac{\sqrt{\sigma_1^2 + \sigma_2^2}}{\sigma_1\sigma_2}y - \frac{\sigma_2}{\sigma_1\sqrt{\sigma_1^2 + \sigma_2^2}}x,$$

so ergibt sich

$$h(x) = \frac{1}{\sqrt{2\pi(\sigma_1^2 + \sigma_2^2)}}\exp\left(-\frac{1}{2}\,\frac{x^2}{\sigma_1^2 + \sigma_2^2}\right)\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-z^2/2}\,dz = \varphi(x; 0, \sigma_1^2 + \sigma_2^2).$$

Also ist $Y_1 + Y_2$ normalverteilt mit Erwartungswert 0 und Varianz $\sigma_1^2 + \sigma_2^2$. Demzufolge ist $X_1 + X_2$ normalverteilt mit Erwartungswert $\mu_1 + \mu_2$ und Varianz $\sigma_1^2 + \sigma_2^2$. □

8 Grundlagen der Statistik

In diesem Kapitel wollen wir einen kurzen Einblick in die mathematische Statistik geben. Die Statistik ist ein sehr reichhaltiges Teilgebiet der Stochastik, die oft in mehreren Vorlesungen gesondert behandelt wird; wir können daher hier nur einige zentrale Ideen und Konzepte betrachten. Man unterteilt die mathematische Statistik in die *beschreibende Statistik* und die *schließende Statistik*. Die beschreibende Statistik faßt Datensätze zusammen und macht deren Besonderheiten mit Hilfe von Kennzahlen und Grafiken sichtbar. Wir wollen uns damit hier nicht befassen.

Die Fragestellung der schließenden Statistik ist gewissermaßen dual zu der Fragestellung der Wahrscheinlichkeitstheorie. Während man in der Wahrscheinlichkeitstheorie von einem festen Modell ausgeht und analysiert, was man an Beobachtungen erwarten kann, sind in der schließenden Statistik die Beobachtungen gegeben und man versucht Rückschlüsse auf das zugrunde liegende Modell zu ziehen. Man hat also in der mathematischen Statistik a priori eine ganze Klasse von möglichen Modellen und das bedeutet von möglichen Wahrscheinlichkeits–Verteilungen zur Verfügung. In den einfacheren Fällen – und das sind u.a. alle hier behandelten – lassen sich diese Wahrscheinlichkeits–Verteilungen durch einen *strukturellen Parameter* klassifizieren, der meist reellwertig ist und direkt mit der ursprünglichen Fragestellung zusammenhängt. Beispielsweise kann von einer Beobachtung bekannt sein (woher auch immer), daß sie von einer Poisson-Verteilung zum Parameter $\lambda > 0$ stammt, bloß kennt man das λ nicht und möchte wissen, welches λ am besten zu der Beobachtung paßt. In diesem Fall wäre also die Klasse der möglichen Modelle, die Menge aller Poisson–Verteilungen $\{\pi_\lambda : \lambda > 0\}$. In solchen Fällen spricht man auch von *parametrischen Modellen*. Die schwierigere (aber auch interessantere) Problematik der sogenannten *nichtparametrische Modelle*, die wesentlich reichhaltigere Klassen von Wahrscheinlichkeits–Verteilungen zuläßt, kann hier nicht behandelt werden.

Man unterscheidet drei verschiedene Problemstellungen: man möchte den Parameter durch einen *Schätzwert* beschreiben, man möchte ein Prüfverfahren entwickeln, mit dem getestet werden kann, ob vorgegebene Hypothesen über den Parameterwert mit den Daten verträglich sind (*statistische Tests*), und man möchte Schranken berechnen, die einen unbekannten Parameter mit vorgegebener Wahrscheinlichkeit einfangen (*Konfidenzintervalle*).

Der allgemeine Rahmen dieser Probleme hat immer folgende Zutaten:

1. eine nichtleere Menge $\mathcal{X}$, der sogenannte Stichprobenraum (so heißt häufig der zugrunde liegende Raum in der Statistik, im Gegensatz zu Ω in der Wahrscheinlichkeitstheorie) versehen mit einer σ-Algebra $\mathcal{F}$

2. eine Familie $\{P_\theta : \theta \in \Theta\}$ von Wahrscheinlichkeiten auf $\mathcal{X}$; hierbei nehmen wir an, daß $\Theta \subseteq \mathbb{R}^d$ für ein d ist und daß Θ ein verallgemeinertes Intervall ist.

Schätzprobleme
Das Problem in diesem Abschnitt ist das folgende: Man möchte aus vorliegenden Beobachtungen (Realisierungen von Zufallsgrößen), die nach P_θ verteilt sind, den tatsächlich zugrunde liegenden Parameter θ schätzen, oder allgemeiner eine Funktion $g : \Theta \to \mathbb{R}^k$

(ist der Parameter selbst zu schätzen, so ist $g(\theta) = \theta$). Der Vorteil der allgemeineren Formulierung liegt darin, daß auch einfache Fälle, in denen g etwas komplizierter aussieht, eingeschlossen sind. So könnte man die Varianz $np(1-p)$ einer Binomialverteilung schätzen wollen. Dann ist $\theta = p$ und $g(p) = np(1-p)$. Im Falle der Normalverteilung ist der Parameterbereich zweidimensional, also $\theta = (\mu, \sigma^2)$, eine zu schätzende Funktion ist zum Beispiel $g(\theta) = \mu$.

Zunächst müssen wir natürlich klären, was ein Schätzer überhaupt sein soll. Hierzu nehmen wir an, daß wir n Beobachtungen $X_1, \ldots, X_n \in \mathcal{X}$ gegeben haben. Die Großbuchstaben sollen hierbei andeuten, daß es sich bei $X_1, \ldots, X_n$ um Zufallveriablen handelt, von denen wir annehmen wollen, daß sie unabhängig sind und alle nach P_θ verteilt. Ein Schätzer für $g(\theta)$ ist dann sinnvollerweise eine Funktion der Zufallsgrößen $X_1, \ldots, X_n$, d.h. eine Funktion $\hat{g} : \mathcal{X}^n \to \mathbb{R}^k$ (im dem Falle, daß $g(\theta) = \theta$ ist, werden anstatt $\hat{g}$ oft auch $\hat{\theta}$ schreiben). Damit ist natürlich $\hat{g}$ selbst wieder eine Zufallsvariable.

Nun, da wir wissen, was Schätzer eigentlich sind, stellt sich die Frage nach der Güte von Schätzern. Wir wollen hier zwei Kriterien vorstellen:

(8.1) Definition. Es sei $X = (X_1, \ldots, X_n) \in \mathcal{X}^n$ eine Beobachtung und $\hat{g}$ ein Schätzer für das unbekannte g. $\hat{g}$ heißt *erwartungstreu (unbiased)* für $g(\theta)$, wenn für alle $\theta \in \Theta$ die Gleichung

$$E_\theta \hat{g} = g(\theta)$$

gilt.
$\hat{g}$ heißt *konsistent (consistent)* für $g(\theta)$, wenn für alle $\theta \in \Theta$ und alle $\delta > 0$

$$\lim_{n\to\infty} P_\theta\Big(|\hat{g}(X_1, \ldots, X_n) - g(\theta)| > \delta\Big) = 0.$$

Ein konsistenter Schätzer genügt also dem Gesetz der großen Zahlen und wird somit für große Datenmengen immer besser.

Nun haben wir zwar zwei sinnvolle Kriterien zur Beurteilung von Schätzern aufgestellt. Eine für die Praxis relevante Frage ist allerdings die, wie man eigentlich solche Schätzer findet. Wir werden in der folgenden Definition das Konzept eines *Maximum-Likelihood-Schätzers* vorstellen, der in vielen Fällen obigen Güte–Kriterien genügt. Die Idee hinter der Konstruktion ist die, daß man – kennt man θ nicht – am plausibelsten annimmt, daß man einen für P_θ typischen Wert beobachtet hat. Typisch soll hier der $P_\theta(x)$ bzw. eine Dichte $f(x|\theta)$ maximierende Wert für θ sein. Da Θ ein Intervall in $\mathbb{R}^d$ ist, kann dann ein Maximum-Likelihood-Schätzer durch Differentiation gefunden werden.

(8.2) Definition

(a) Ist $\mathcal{X}$ eine endliche oder abzählbare Menge, so heißt die Funktion $\theta \mapsto L_x(\theta) = P_\theta(x)$ mit $x \in \mathcal{X}^n$ *Likelihood-Funktion*. Es seien X eine Zufallsvariable, definiert auf einem allgemeinen W.-Raum, mit Werten in $\mathcal{X}^n = \mathbb{R}^n$ und $\{P_\theta : \theta \in \Theta\}$ eine Familie von Verteilungen von X. Ist P_θ verteilt mit einer n-dimensionalen Dichte $f(\cdot|\theta)$, so heißt hier die Funktion $\theta \mapsto L_x(\theta) = f(x|\theta)$ die *Likelihood-Funktion*.

(b) Nimmt $L_x(\cdot)$ einen Maximalwert in $\hat{\theta}(x)$ an, ist also

$$L_x(\hat{\theta}(x)) = \sup\{L_x(\theta) : \theta \in \Theta\},$$

so nennen wir $\hat{\theta}(x)$ eine *Maximum-Likelihood-Schätzung (Schätzer, estimator)* von θ und $g(\hat{\theta}(x))$ eine Maximum-Likelihood-Schätzung von $g(\theta)$.

(8.3) Bemerkung. $L_x(\theta)$ gibt also an, wie wahrscheinlich die gemachte Beobachtung x ist, wenn die zugrunde liegende Verteilung P_θ ist.

Wir wollen nun den Maximum-Likelihood-Schätzer in einigen gut bekannten Situationen kennenlernen.

(8.4) Beispiele. (a) *Bernoulli-Experiment:*
In einem Bernoulli-Experiment zu den Parametern n und p soll p aus der Anzahl x der Erfolge geschätzt werden. Es ist also $\Theta = [0,1]$ und $L_x(p) = b(x;n,p)$. Aufgrund der Monotonie der Logarithmus–Funktion hat $\log L_x(p)$ dieselben Maxima wie $L_x(p)$. Es ist $(\log L_x(p))' = \frac{x}{p} - \frac{n-x}{1-p}$, womit man $(\log L_x(p))' = 0$ bei $\hat{p}(x) = \frac{x}{n}$ findet. Es ist leicht zu sehen, daß es sich bei $\hat{p}(x)$ um ein Maximum von $\log L_x(p)$ handelt. $\frac{x}{n}$ ist also der Maximum-Likelihood-Schätzer für p. Dies entspricht der naiven Mittelwertbildung, die man üblicherweise durchführen würde (es ist nämlich $\hat{p} = \frac{\sum_{i=1}^n X_i}{n}$).

Aufgrund der Linearität des Erwartungswertes gilt außerdem $E_p(\hat{p}) = E_p(\frac{\sum_{i=1}^n X_i}{n}) = \frac{np}{n} = p$, $\hat{p}$ ist also erwartungstreu. Schließlich liefert das Gesetz der großen Zahlen (Satz 3.30) die Konsistenz von $\hat{p}$.

Wir haben also im Falle der Binomialverteilung gesehen, daß der Maximum-Likelihood-Schätzer der naiven Vorgehensweise entspricht und diesem Falle auch unsere Gütekriterien an eine Schätzung erfüllt. Interessanterweise ist der Maximum-Likelihood-Schätzer im Falle der Binomialverteilung sogar der einzige Schätzer, der dies tut.

(8.5) Satz. Ist in obiger Situation S ein erwartungstreuer Schätzer für p so gilt $S = \hat{p}$.

Beweis. Sei $T := S - \hat{p}$. Da sowohl S also auch $\hat{p}$ erwartungstreu sind gilt für alle p

$$E_p(T) = E_p(S - \hat{p}) = E_p(S) - E_p(\hat{p}) = p - p = 0.$$

Also ist für alle p

$$0 = E_p(T) = \sum_{k=0}^{n} T(k)\binom{n}{k}p^k(1-p)^{n-k} = (1-p)^n \sum_{k=0}^{n} T(k)\binom{n}{k}\left(\frac{p}{1-p}\right)^k.$$

Setzt man $s := \frac{p}{1-p}$, so läuft s von 0 bis ∞, wenn p das Einheitsintervall durchläuft und es ist

$$f(s) := (1-p)^n \sum_{k=0}^{n} T(k)\binom{n}{k}s^k$$

konstant gleich 0. Andererseits ist $f(s)$ ein Polynom, so daß alle Koeffizienten von f schon 0 sein müssen, was impliziert, daß $T(k) = 0$ für alle k und somit $S = \hat{p}$. □

Um ein weiteres Qualitätsmerkmal des oben gewonnenen Schätzers zu diskutieren, definieren wir den Abstand eines Schätzers zu seiner erwartungstreuen Variante.

(8.6) Definition. Der *Bias* eines Schätzers T eines Parameters p ist als

$$b(p, T) = E_p[T - p]$$

definiert.

Die Qualität eines Schätzers messen wir durch seinen quadratischen Abstand zum zu schätzenden Wert.

(8.7) Definition. Das *quadratische Risiko* eines Schätzers T eines Parameters p ist definiert als

$$R(p, T) = E_p[(T - p)^2].$$

Den Zusammenhang dieser beiden Definitionen klärt die folgende Proposition.

(8.8) Proposition. Sei T ein Schätzer für p. Dann gilt

$$R(p, T) = V_p(T) + b^2(p, T).$$

Beweis. Es gilt

$$\begin{aligned} R(p,T) &= E_p[(T-p)^2] \\ &= E_p[(T - E_pT) + (E_pT - p))^2] \\ &= E_p[(T - E_pT)^2] + 2E_p(T - E_pT)(E_pT - p) + (E_p[T - p])^2 \\ &= E_p[(T - E_pT)^2] + (E_p[T - p])^2 \\ &= V_p(T) + b^2(p, T). \end{aligned}$$

□

Wir werden nun die Qualität des oben hergeleiteten Schätzers $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ für p untersuchen. Sei $\mathcal{U}$ die Klasse aller erwartungstreuen Schätzer $\tilde{p}$ von p, d. h. die Klasse aller Schätzer, für die $b(p, \tilde{p}) = 0$ gilt.

(8.9) Satz. Für den Schätzer $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ und jeden erwartungstreuen Schätzer $\tilde{p} \in \mathcal{U}$ gilt

$$V_p(\tilde{p}) \geq \frac{p(1-p)}{n} = V_p(\hat{p})$$

für alle $p \in (0, 1)$. Da $b(p, \tilde{p}) = b(p, \hat{p}) = 0$ impliziert dies

$$R(p, \tilde{p}) \geq R(p, \hat{p})$$

für alle $p \in (0,1)$ und alle $\tilde{p} \in \mathcal{U}$.

Für den Beweis benötigen wir einen neuen Begriff.

(8.10) Definition. Für $x \in \{0,1\}^n$ sei

$$L_x(p) = P_p(x)$$

und

$$\mathcal{L}_x(p) = \log P_p(x).$$

Dann ist

$$\mathcal{L}'_x(p) = \frac{L'_x(p)}{L_x(p)}.$$

Mit

$$I(p) := E_p[(\mathcal{L}'_x(p))^2]$$

bezeichnen wir die *Fisher Information* von p.

Beweis von Satz 8.9. Sei $\tilde{p} \in \mathcal{U}$. Da $\tilde{p}$ erwartungstreu ist, gilt

$$p = E_p\tilde{p} = \sum_{x\in\{0,1\}^n} \tilde{p}(x)P_p(x).$$

Dies impliziert

$$1 = p' = \sum_{x\in\{0,1\}^n} \tilde{p}(x)\frac{d}{dp}P_p(x) = \sum_{x\in\{0,1\}^n} \tilde{p}(x)\mathcal{L}'_x(p)P_p(x) = E_p[\tilde{p}\mathcal{L}'_x(p)].$$

Andererseits gilt

$$E_p\mathcal{L}'_x(p) = \sum_{x\in\{0,1\}^n} \mathcal{L}'_x(p)P_p(x) = \sum_{x\in\{0,1\}^n} \frac{d}{dp}P_p(x) = \frac{d}{dp}1 = 0.$$

Daher gilt auch

$$E_p\tilde{p}E_p\mathcal{L}'_x(p) = 0.$$

Subtrahiert man die vorherige Gleichung, erhält man

$$1 = E_p((\tilde{p} - E_p\tilde{p})\mathcal{L}'_x(p)).$$

Mit Cauchy-Schwarz folgt

$$1 = 1^2 \leq E(\tilde{p} - E_p\tilde{p})^2 \, E_p[(\mathcal{L}'_x(p))^2] = V_p(\tilde{p})I(p)$$

nach Definition von $I(\cdot)$. Da $I(p) > 0$ (was wir im Anschluss beweisen), folgt

$$V_p(\tilde{p}) \geq \frac{1}{I(p)}.$$

Dies heißt auch die *Cramér-Rao Ungleichung*. In einem letzten Schritt berechnen wir $I(\cdot)$.

$$\begin{aligned} I(p) &= E_p\left[\left(\frac{d}{dp}\log P_p(x)\right)^2\right] = E_p\left[\left(\frac{d}{dp}\log\, p^{\sum_i x_i}(1-p)^{n-\sum_i x_i}\right)^2\right] \\ &= E_p\left[\left(\frac{d}{dp}\sum_{i=1}^{n}\log\, P_p(x_i)\right)^2\right]. \end{aligned}$$

Nun gilt für jedes i

$$E_p\left[\frac{d}{dp}\log P_p(x_i)\right] = \frac{d}{dp}\left[\frac{1}{p}p - (1-p)\frac{1}{1-p}(1-p)\right] = 0.$$

Daher erhalten wir

$$\begin{aligned} I(p) &= E_p\left[\left(\frac{d}{dp}\sum_{i=1}\neg n\log P_p(x_i)\right)^2\right] \\ &= \sum_{i\neq j} EW_p\frac{d}{dp}\log P_p(x_i)E_p\frac{d}{dp}\log P_p(x_j) + \sum_{i=1}^{n} E_p\left[\left(\frac{d}{dp}\log P_p(x_i)\right)^2\right] \\ &= 0 + np\frac{1}{p^2} + n(1-p)\frac{1}{(1-p)^2} \\ &= n\left(\frac{1}{p}+\frac{1}{1-p}\right) \\ &= \frac{n}{p(1-p)}. \end{aligned}$$

Daher gilt für $\tilde{p}\in\mathcal{U}$

$$R(p,\tilde{p}) = V_p(\tilde{p}) \geq \frac{p(1-p)}{n}.$$

Wählen wir für $\tilde{p} = \hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$

$$R(p,\hat{p}) = V_p(\hat{p}) = V_p\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}p(1-p).$$

Dies beweist den Satz. □

Nun wenden wir die Konzepte auf den Fall normalverteilter Zufallsvariablen an.

(8.4) Beispiele fortgesetzt. (b) *Normalverteilung:*
Hier leiten wir nur den Maximum–Likelihood–Schätzer für die verschiedenen Fälle normalverteilter Zufallsvariablen her; ihre Güte zu diskutieren erfordert ein wenig Extraarbeit, die wir im Anschluß erledigen werden. Seien $X_1, X_2, \ldots, X_n$ unabhängig und normalverteilt zu den Parametern μ und σ^2 (wir schreiben im folgenden $\mathcal{N}(\mu,\sigma^2)$-verteilt). Dann ist $\theta = (\mu,\sigma^2)$. Die Dichte von $X = (X_1,\ldots,X_n)$ an der Stelle $x = (x_1,\ldots,x_n)$ ergibt sich nach Satz (7.17) zu

$$f(x|\theta) = \Big(\frac{1}{\sqrt{2\pi}\sigma}\Big)^n \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\Big).$$

Wir betrachten wieder

$$\log f(x|\theta) = -n\log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

und unterscheiden die folgenden Fälle:

(1) *(Varianz bekannt, Schätzung des Erwartungswertes)*
Sei μ unbekannt und $\sigma^2 = \sigma_0^2$ bekannt. Dann ist $\Theta = \{(\mu,\sigma^2) : \mu \in \mathbb{R}, \sigma^2 = \sigma_0^2\}$. Nun ist $\frac{d}{d\mu}\log f(x|\theta) = 0$ genau dann, wenn $\sum_{i=1}^{n}(x_i-\mu) = 0$ ist. Daraus ergibt sich der Maximun-Likelihood-Schätzer zu

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Dies ist erneut die naive Mittelwertbildung. Man muß natürlich noch durch Bildung der zweiten Ableitung überprüfen, daß wirklich ein Maximum in $\hat{\mu}$ vorliegt. Dies sei dem Leser überlassen.

(2) *(Erwartungswert bekannt, Schätzung der Varianz)*
Sei $\mu = \mu_0$ bekannt und $\sigma^2 > 0$ unbekannt. Hier ist $\Theta = \{(\mu,\sigma^2) : \mu = \mu_0, \sigma^2 > 0\}$. Nun ist $\frac{d}{d\sigma}\log f(x|\theta) = 0$ genau dann, wenn

$$-\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu_0)^2 = 0$$

ist. Daraus ergibt sich für σ^2 der Maximun-Likelihood-Schätzer zu

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_0)^2.$$

Auch dieser Schätzer entspricht dem naiven Ansatz, aus den Daten die mittlere quadratische Abweichung zu bestimmen.

(3) *(beide Parameter unbekannt)*
Seien nun beide Parameter μ und σ^2 unbekannt. Die Gleichungen

$$\frac{d}{d\mu}\log f(x|\theta) = 0 \quad \text{und} \quad \frac{d}{d(\sigma^2)}\log f(x|\theta) = 0$$

liefern (simultan gelöst) die Maximum-Likelihood-Schätzer $\hat{\mu}$ für μ und

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2$$

für σ^2. Hier muß man allerdings mit Hilfe der Hesseschen Matrix überprüfen, ob es sich um ein Maximum handelt. Dazu beachte, daß

$$\frac{d^2}{d\mu^2}\log f(x|\theta) = -\frac{n}{\sigma^2} \quad \text{und} \quad \frac{d^2}{d(\sigma^2)^2}\log f(x|\theta) = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i-\mu)^2$$

sowie

$$\frac{d^2}{d\mu\, d(\sigma^2)} \log f(x|\theta) = -\frac{1}{\sigma^4} \sum_{i=1}^{n} (x_i - \mu).$$

Somit ist die Determinante der Hesseschen Matrix an der Stelle $(\hat{\mu}, \hat{\sigma}^2)$ identisch gleich $\frac{n^2}{2(\sigma^2)^3} > 0$ und $\frac{d^2}{d\mu^2} \log f(x|\theta) < 0$, also ist die Hessesche Matrix an dieser Stelle negativ definit, und somit liegt an der Stelle $(\hat{\mu}, \hat{\sigma}^2)$ ein isoliertes Maximum vor.

Die Diskussion der Güte obiger Maximum–Likelihood–Schätzer ist ein wenig aufwendig und muß durch ein Lemma vorbereitet werden.

(8.11) Lemma. *Die Verteilung der Summe der Quadrate von n unabhängigen $\mathcal{N}(0,1)$-verteilten Zufallsgrößen nennt man eine χ^2_n-Verteilung (χ^2-Verteilung mit n Freiheitsgraden, χ^2-distribution with n degrees of freedom). Ihre Dichte ist gegeben durch*

$$g_n(x) = \frac{1}{2^{n/2}\Gamma(n/2)}\, x^{(n/2)-1}\, e^{-x/2}, \quad x > 0.$$

Hierbei ist

$$\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx$$

die Γ-Funktion.
Der Erwartungswert einer χ^2_n-verteilten Zufallsgröße ist n, die Varianz $2n$.

Beweis. Wir beweisen den ersten Teil des Satzes via Induktion über die Anzahl der Variablen n. Der zweite Teil ist eine Übung.
$n = 1$: Es ist für eine $\mathcal{N}(0,1)$-verteilte Zufallsgröße X_1

$$\begin{aligned} P(X_1^2 \le x) = P(-\sqrt{x} < X_1 < \sqrt{x}) &= 2\int_0^x \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \\ &= \int_0^x \frac{1}{\sqrt{2\pi}} z^{-\frac{1}{2}} e^{-\frac{z}{2}} dz \end{aligned}$$

mittles der Substitution $t = \sqrt{z}$. Das – zusammen mit $\Gamma(1/2) = \sqrt{\pi}$ – beweist den Induktionsanfang.

Aufgrund der Unabhängigkeit der gegebenen Zufallsvariablen und gemäß der Definition der Faltung ist nach Induktionsvoraussetzung

$$\begin{aligned} g_n(z) &= \int_{-\infty}^{\infty} g_{n-1}(x) g_1(z-x) dx \\ &= \int_0^z \frac{1}{2^{(n-1)/2}\Gamma((n-1)/2)}\, x^{((n-1)/2)-1}\, e^{-x/2} \frac{1}{\sqrt{2\pi}} (z-x)^{-1/2} \exp\left(-\frac{z-x}{2}\right) dx. \end{aligned}$$

Substituiert man $y = \frac{z}{x}$ ergibt sich

$$\begin{aligned} g_n(z) &= \frac{e^{-z/2}}{\sqrt{2\pi}2^{(n-1)/2}\Gamma((n-1)/2)} \int_0^1 z^{(n-1)/2-1} y^{(n-1)/2-1} z^{-1/2}(1-y)^{-1/2} z dy \\ &= \frac{z^{n/2-1}e^{-z/2}}{\sqrt{2\pi}2^{(n-1)/2}\Gamma((n-1)/2)} \int_0^1 y^{(n-1)/2-1}(1-y)^{-1/2} dy \\ &= \frac{z^{n/2-1}e^{-z/2}}{\Gamma(1/2)2^{(n-1)/2}\Gamma((n-1)/2)} \frac{\Gamma((n-1)/2)\Gamma(1/2)}{\Gamma(n/2)} \\ &= \frac{z^{n/2-1}e^{-z/2}}{2^{(n-1)/2}\Gamma(n/2)}. \end{aligned}$$

□

Dieses Lemma hilft uns die Frage nach der Güte der oben vorgeschlagenen Maximum–Likelihood-Schätzer im wesentlichen zu klären.

(8.4) Beispiel, 2. Fortsetzung. (b) Zunächst wollen wir hier wieder die Güte von $\hat{\mu}$ überprüfen. Nach Satz (7.19) ist $\sum_{i=1}^n X_i$ normalverteilt mit Erwartungswert $n\mu$ und Varianz $n\sigma_0^2$. Dann ist nach den Ausführungen in Beispiel (7.14)(2) der Erwartungswert von $S_1 = \frac{1}{n}\sum_{i=1}^n X_i$ gleich μ, also ist S_1 erwartungstreu für $g(\theta) = \mu$. Das schwache Gesetz der großen Zahlen war in Kapitel 3 nur für diskrete W.-Räume formuliert worden. Aber die Markoff-Ungleichung (Satz (3.28)) erhalten wir analog für absolutstetig verteilte Zufallsgrößen X mit Dichte f (verwende $E(|X|) = \int_{\mathbb{R}} |x| f(x) dx$; dieser Erwartungswert existiert, hier nach Beispiel (7.14)(2)). Da nun nach (7.14)(2) die Varianz von S_1 gleich σ_0^2/n ist, erhalten wir hier ebenfalls ein schwaches Gesetz und damit die Konsistenz des Schätzers S_1 für μ.

Die Güte des Schätzers für σ^2 bei bekanntem μ diskutieren wir ähnlich wie oben. Wir schreiben dazu $\tilde{\sigma}^2$ als:

$$\tilde{\sigma}^2 = \frac{\sigma^2}{n} \sum_{i=1}^n \Big(\frac{X_i - \mu_0}{\sigma}\Big)^2.$$

Nach Beispiel (7.14)(2) sind die Zufallsgrößen $X_i^* := (X_i - \mu_0)/\sigma$ standardnormalverteilt. Nach Lemma (8.6) ist dann $\sum_{i=1}^n (X_i^*)^2$ χ_n^2-verteilt mit Erwartungswert n und Varianz $2n$. Also ist nach Definition (7.13) $E(\tilde{\sigma}^2) = \sigma^2$ und $V(S_2) = \frac{2\sigma^4}{n}$. Damit ist $\tilde{\sigma}^2$ erwartungstreu für σ^2, und wir erhalten entsprechend der Diskussion im Fall (1) die Konsistenz von S_2 für σ^2.

Schließlich wollen wir noch verstehen, daß der Schätzers für σ^2 bei unbekanntem μ **nicht** erwartungstreu ist. Dazu betrachten wir $\hat{\sigma}^2 + \hat{\mu}^2$ und berechnen zum einen

$$\begin{aligned} E_{\mu,\sigma^2}(\hat{\sigma}^2 + \hat{\mu}^2) &= E_{\mu,\sigma^2}(\frac{1}{n}\sum_{i=1}^n X_i^2 - \hat{\mu}^2 + \hat{\mu}^2) \\ &= \frac{1}{n}\sum_{i=1}^n E_{\mu,\sigma^2} X_i^2 = V_{\mu,\sigma^2}(X_i^2) + (E_{\mu,\sigma^2} X_i)^2 = \sigma^2 + \mu^2. \end{aligned}$$

Andererseits ist

$$\begin{aligned} E_{\mu,\sigma^2}(\hat{\sigma}^2 + \hat{\mu}^2) &= E_{\mu,\sigma^2}(\hat{\sigma}^2) + E_{\mu,\sigma^2}(\hat{\mu}^2) \\ &= E_{\mu,\sigma^2}(\hat{\sigma}^2) + V_{\mu,\sigma^2}(\hat{\mu}^2) + E_{\mu,\sigma^2}(\hat{\mu})^2 = E_{\mu,\sigma^2}(\hat{\sigma}^2) + \frac{\sigma^2}{n} + \mu^2. \end{aligned}$$

Löst man diese beiden Gleichungen nach $E_{\mu,\sigma^2}(\hat{\sigma}^2)$ auf, so ergibt sich

$$E_{\mu,\sigma^2}(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2.$$

Somit ist $\hat{\sigma}^2$ nicht erwartungstreu für σ^2, wohl aber

$$S^2 := \frac{n}{n-1}\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2.$$

Ob S^2 auch konsistent ist für σ^2, kann erst geklärt werden, wenn wir die Verteilung von S^2 kennen. Da wir diese im Abschnitt über das Testen sowieso berechnen müssen, verschieben wir den Beweis der Konsistenz nach dort. Hier sei nur vorab bemerkt, daß S^2 in der Tat konsistent ist und daß man diese Konsistenz im wesentlichen wie oben zeigt.

Statistische Tests
Die Zutaten in diesem Abschitt sind die gleichen wie die im vorangegangenen mit der zusätzlichen Ingrediens, daß nun noch eine Teilmenge $H \subset \Theta$ gegeben ist. H nennen wir Hypothese, $K := \Theta \setminus H$ heißt Alternative. Das Problem besteht nun darin, festzustellen, ob ein gegebener Datensatz nahelegt, daß der zugrunde liegende Parameter in H ist oder in K. Betrachten wir ein Beispiel, das wir weniger wegen seiner praktischen Relevanz als aufgrund dessen gewählt haben, daß es Grundlage einer historischen Diskussion zwischen den Statistikern *R.A. Fisher* und *J. Neyman* über die Konstruktion von Tests war.

(8.12) Beispiel. Eine englische Lady trinkt ihren Tee stets mit etwas Milch. Eines Tages verblüfft sie ihre Teerunde mit der Behauptung, sie könne allein am Geschmack unterscheiden, ob zuerst die Milch oder zuerst der Tee eingegossen wurde. Dabei sei ihr Geschmack zwar nicht unfehlbar, aber sie würde häufiger die richtige Eingieß-Reihenfolge erschmecken, als dies durch blindes Raten möglich wäre.

Um der Lady eine Chance zu geben, ihre Behauptung unter Beweis zu stellen, könnte man sich folgenden Versuch vorstellen: der Lady werden jeweils n mal 2 Tassen gereicht, von denen jeweils eine vom Typ "Milch vor Tee", die andere vom Typ "Tee vor Milch" ist; ihre Reihenfolge wird jeweils zufällig ausgewürfelt. Die Lady soll nun durch Schmecken erkennen, welche Tasse von welchem Typ ist.

Aufgrund dieses Experiments modellieren wir die n Geschmacksproben als unabhängige Erfolg/Mißerfolg-Experiment mit Erfolgswahrscheinlichkeit p, also als n-stufiges Bernoulli-Experiment. Der Parameter p variiert dabei im Intervall $[1/2, 1]$ (da $p = 1/2$ schon die Erfolgswahrscheinlichkeit bei purem Raten ist und daher $p < 1/2$ unrealistisch ist). Es liegt nun nahe $H = 1/2$ und $K = (1/2, 1]$ zu wählen, d.h. wir testen die Hypothese "die Lady rät" gegen die Alternative "die Lady schmeckt den Unterschied". Natürlich

könnten wir auch $K = 1/2$ und $H = (1/2, 1]$ wählen, denn bislang scheint die Situation zwischen H und K komplett symmetrisch zu sein. Daß es tatsächlich einen Unterschied macht, was man als welche Hypothese wählt, versteht man, wenn man die möglichen Fehler betrachtet, die eine Entscheidung $\theta \in H$ bzw. $\theta \in K$ mit sich bringen kann.

Offenbar gibt es zwei mögliche Fehler: Ist $\theta \in H$ und wird die Hypothese verworfen, so spricht man von einem *Fehler erster Art (type I error)*, ist $\theta \in K$ und wird die Hypothese angenommen, so spricht man von einem *Fehler zweiter Art (type II error)*. Ein Test ist beschrieben durch die Angabe der Menge R derjenigen x, für die die Hypothese verworfen wird. R heißt auch *Verwerfungsbereich (rejection region)*. Um sich den Unterschied zwischen den beiden Fehlern deutlich zu machen, stelle man sich vor ein Angeklagter solle verurteilt werden. Offenbar gibt es auch hier zwei Möglichkeiten: den Angeklagten unschuldig zu verurteilen oder einen Schuldigen freizusprechen. Dem Rechtsgrundsatz "in dubio pro reo" würde es dann entsprechen, den ersten Fehler so klein wie möglich zu halten.

Bei einem statistischen Test beide Fehler gleichzeitig zu minimieren ist offenbar schwer möglich (es sei denn man verschafft sich über eine große Stichprobe eine große Sicherheit über den zugrunde liegenden Parameter); eine Minimierung des Fehlers erster Art würde in letzter Konsequenz bedeuten, die Hypothese stets zu akzeptieren, während man den Fehler zweiter Art dadurch klein halten könnte, indem man stets die Hypothese verwirft. Man hat sich darauf geeinigt, bei einem Test immer den Fehler erster Art zu kontrollieren, indem man gewährleistet, daß er kleiner ist als eine vorgegebene Irrtumswahrscheinlichkeit α. Unter dieser Randbedingung versucht man den Fehler 2. Art möglichst klein zu halten (trotzdem kann es passieren, daß dieser besonders bei sehr kleinen Stichproben sehr groß wird). Diese Konstruktion beeinflußt auch die Wahl von H bzw. K.

Wie soll man nun zu einem gegebenen Testproblem einen Test konstruieren? In der mathematischen Statistik gibt es verschiedene Ansätze optimale Tests theoretisch zu konstruieren. Wir werden an dieser Stelle darauf verzichten ein solch theoretisches Fundament zu legen und nur eine heuristisch sinnvolle Konstruktion anführen, die in den nachfolgend diskutierten Beispielen in der Tat zu Tests führt, die in gewissem Sinne optimal sind (was wir allerdings nicht beweisen werden).

Nehmen wir also an, wir wollen die Hypothese H mit einer *Fehlerwahrscheinlichkeit (error probability)* (man sagt oft auch zum *Signifikanzniveau (level of significance)*) $\alpha > 0$ testen. Es ist sinnvoll, dazu zunächst eine Stichprobe vom (möglichst großen) Umfang n aufzunehmen. Aufgrund dieser Stichprobe schätzen wir dann θ möglichst gut durch $\hat{\theta}_n$. Sieht $\hat{\theta}_n$ nicht signifikant anders aus als man es unter Vorliegen von H erwarten würde, so entscheidet man sich für H ansonsten für K. Genauer bedeutet das, man wählt ein Intervall $\hat{H} = \hat{H}(\alpha, n)$ möglichst klein, so daß

$$P_\theta(\hat{\theta}_n \notin \hat{H}) \leq \alpha \quad \forall \theta \in H$$

und entscheidet sich H zu akzeptieren, falls $\hat{\theta}_n \in \hat{H}$ und ansonsten H abzulehnen. Die Wahrscheinlichkeit eines Fehlers erster Art ist also maximal α.

Da wir im vergangenen Abschnitt schon gesehen haben, wie man in einigen Fällen gute Schätzer konstruiert, können wir uns jetzt einmal die obige Methode in Aktion betrachten.

(8.12 a) Beispiel. Zunächst behandeln wir das eingangs gestellte Problem der Teatesting Lady. Wie wir sehen werden, ist dies der allgemeine Fall des Testens im Falle der Binomialverteilung. Nehmen wir an, die Lady testet 20 Mal, wir führen also 20 unabhängige 0-1 Experimente mit unbekanntem Erfolgsparameter p durch. Wir hatten $\Theta = [1/2, 1]$ angenommen; wir werden aber gleich sehen, daß wir ebenso gut $\Theta = [0, 1]$ nehmen können, ohne das Testergebnis zu beeinflussen. Die Anzahl der Erfolge X ist $b(k; 20, p)$-verteilt. Sei die Hypothese $H = \{1/2\}$ (bzw. $H = [0, 1/2]$ im Falle von $\Theta = [0, 1]$), d.h. die Lady rät. Wir suchen den Verwerfungsbereich $R = \{c, c+1, \ldots, n = 20\}$ in Abhängigkeit vom Niveau α (entsprechend versuchen wir den Fehler zu entscheiden, die Lady habe die behauptete geschmackliche Fähigkeit, obwohl sie in Wahrheit rät, kleiner als das gegebene α zu bekommen). Aus dem vorigen Abschnitt ist bekannt, daß $\hat{p} = \frac{X}{n}$ ein guter Schätzer für p ist. Wir werden also $\hat{H}$ so wählen, daß

$$P_p(\frac{X}{n} \in \hat{H}) \leq \alpha \qquad \forall p \in H$$

und so, daß dabei $\hat{H}$ dabei möglichst klein ist. $R := n\hat{H}$ ist dann der Verwerfungsbereich. Um diesen zu berechnen bemerken wir, daß

$$P_p(X \in R) = \sum_{k=c}^{20} \binom{20}{k} p^k (1-p)^{20-k}$$

gilt. Da dies in p monoton wachsend ist, das Supremum über alle $p \in H$ also bei $p = 1/2$ angenommen wird, ist es offenbar egal, ob wir $\Theta = [1/2, 1]$ und $H = \{1/2\}$ oder $\Theta = [0, 1]$ und $H = [0, 1/2]$ wählen. Wir können nun c als Funktion von α einfach als Lösung der folgenden Ungleichung bestimmen: $2^{-20} \sum_{k=c}^{20} \binom{20}{k} \leq \alpha < 2^{-20} \sum_{k=c-1}^{20} \binom{20}{k}$. Insbesondere ist für $\alpha \in [0.021, 0.058]$ das entsprechende $c = 15$ (also für zulässige Fehler zwischen 2% und 5%). Wir können noch den Fehler 2. Art diskutieren. Es gilt:

p	0.6	0.7	0.8	0.9
Fehler 2. Art	0.874	0.584	0.196	0.011

Dies bedeutet zum Beispiel für den Wert $p = 0.7$, daß die Wahrscheinlichkeit einer Annahme der Hypothese, obwohl sie falsch ist, bei 0.6 liegt. Eine Verkleinerung des Fehlers 2.Art, ohne dabei das Niveau des Tests zu vergrößern, ist also hier allein durch eine größer gewählte Stichprobe möglich.

Wir wollen nun die Qualität der benutzten Testmethode theoretisch untersuchen. Dazu sollten wir zunächst eine mathematische Definition des Begriffs "Test" geben.

8.13 Definition. Zu testen sei die Hypothese $H \subset \Theta$ gegen die Alternative $K \neq \emptyset$. Ein Test ist eine Abbildung

$$\phi : \{0, 1\}^n \to \{0, 1\}.$$

$\phi(x) = 0$ soll bedeuten, daß wir uns für H entscheiden, während $\phi(x) = 1$ bedeutet, wir entscheiden uns für K (wir lehnen die Hypothese ab). Ein Test ist vollständig festgelegt durch das Gebiet $R \subseteq \Theta$, auf dem wir die Hypothese verwerfen (R ist das Verwerfungsgebiet von ϕ), d. h. $\phi(x) = 1 \Leftrightarrow x \in R$.

Neben gewöhnlichen Tests betrachten wir auch randomisierte Tests: Ein randomisierter Test ist eine Abbildung.

$$\phi : \{0,1\}^n \to [0,1].$$

$\phi(x)$ ist die Wahrscheinlichkeit H abzulehnen.

Natürlich will man zwei gegebene Tests der gleichen Hypothese und Alternative vergleichen. Dies geht einerseits über das Niveau des Tests

$$\max_{p \in H} P_p(x \in R)$$

(dies möchte man i. a. durch die gegebene Schranke $\alpha > 0$ kontrollieren). Sind zwei Tests zu einem Niveau α vorgelegt, so bietet sich der Fehler zweiter Art als Vergleichskriterium an. Man definiert daher (äquivalent)

$$\beta(p) = P_p(x \in R)$$

als die Macht eines Tests mit Verwerfungsbereich R in $p \in K$.

Wir werden nun eine Untersuchung der Güte der oben diskutierten Tests im einfachst möglichen Fall präsentieren, dem Fall, in dem sowohl die Hypothese H als auch die Alternative K aus einem einzigen Punkt bestehen. Im Fall einer Folge von i.i.d. Bernoulli-Variablen mit unbekanntem Erfolgsparameter p testen wir also die einfache Hypothese

$$H : \{p = p_0\}$$

gegen die einfache Alternative

$$K : \{p = p_1\}.$$

Da wir uns in diesem Fall auch mit randomisierten Tests befassen wollen, verallgemeinern wir die Begriffe des Niveaus und der Macht rasch auf diesen Fall: Für einen randomisierten ϕ ist

$$E_H(\phi) = \sum_x \phi(x) P_H(x)$$

das Niveau des Tests.

$$E_K(\phi) = \sum_x \phi(x) P_K(x)$$

ist seine Macht. Bemerke, daß diese Definitionen konsistent sind mit den Definitionen für nicht-randomisierte Tests.

Wir interessieren uns nun dafür, unter allen Tests $\{\phi : E_H(\phi) \leq \alpha\}$ denjenigen Test ϕ^* mit maximaler Macht zu finden.

(8.14) Definition. Ein Test ϕ^* heißt Neyman-Pearson Test, falls es eine Konstante c^*, $0 \leq c^* \leq \infty$ gibt, so daß $\phi^*(x) = 1$ falls $P_K(x) > c^* P_H(x)$ und $\phi^*(x) = 0$ falls $P_K(x) < c^* P_H(x)$. Auf $\{P_K(x) = c^* P_H(x)\}$ darf der Test ϕ^* beliebige Werte $0 \leq \gamma(x) \leq 1$ annehmen.

Wir werden im folgenden einen Test ϕ_1 *schärfer* nennen als einen Test ϕ_2, falls

$$E_K(\phi_1) > E_K(\phi_2)$$

gilt, die Chance H zu verwerfen, wenn K vorliegt bei ϕ_1 somit größer ist als bei ϕ_2.

Wir wenden uns nun der speziellen Situation des n-fachen Münzwurfs zu. Offensichtlich gilt für alle $x \in \{0,1\}^n$

$$P_H(x) + K_K(x) > 0.$$

Der folgende Satz ist zentral für das gesamte Gebiet der Test-Theorie.

(8.15) Satz [Neyman-Pearson-Lemma]. In der Situation des n-fachen Münzwurfs mit Parameter p sei die Hypothese

$$H : \{p = p_0\}$$

gegen die Alternative

$$K : \{p = p_1\}$$

zu testen. Dann gilt

- Falls ϕ^* ein Neyman-Pearson Test ist, dann ist ϕ^* schärfer als jeder andere Test ϕ mit
$$E_H(\phi) \leq E_H(\phi^*).$$
- Für jedes $0 \leq \alpha \leq 1$ gibt es einen (randomisierten) Neyman-Pearson Test ϕ^* zum Niveau α, also mit $E_H(\phi^*) = \alpha$.

Beweis. Sei ϕ^* ein Neyman-Pearson Test und ϕ ein beliebiger Test zum Niveau kleiner oder gleich $E_H(\phi^*)$. Auf

$$A := \{x \in \{0,1\}^n : \phi^*(x) > \phi(x)\}$$

gilt $\phi^*(x) > 0$ und daher

$$P_K(x) \geq c^* P_H(x).$$

Umgekehrt ist auf

$$B := \{x \in \{0,1\}^n : \phi^*(x) < \phi(x)\}$$

$\phi^*(x) < 1$ und daher

$$P_K(x) \leq c^* P_H(x).$$

Dies bedeutet

$$\begin{aligned}
E_K(\phi^*) - E_K(\phi) &= \sum_{x \in \{0,1\}^n} (\phi^*(x) - \phi(x)) P_K(x) \\
&= \sum_{x \in A} (\phi^*(x) - \phi(x)) P_K(x) + \sum_{x \in B} (\phi^*(x) - \phi(x)) P_K(x) \\
&\geq \sum_{x \in A} (\phi^*(x) - \phi(x)) c^* P_H(x) + \sum_{x \in B} (\phi^*(x) - \phi(x)) c^* P_H(x) \\
&= c^* \sum_{x \in \{0,1\}^n} (\phi^*(x) - \phi(x)) P_H(x) \\
&= c^* (E_H(\phi^*) - E_H(\phi)) \geq 0.
\end{aligned}$$

Nun beweisen wir den zweiten Teil des Satzes und konstruieren den zugehörigen Neyman-Pearson Test.

Für $\alpha = 0$ setzen wir $c^* = \infty$. Dann gilt immer

$$P_K(x) < c^* P_H(x).$$

Daher ist $\phi \equiv 0$ und somit $E_H \phi = 0$.

Nun sei $\alpha > 0$. Für $c \geq 0$ setzen wir $q(x) := P_K(x)/P_H(x)$ und

$$\begin{aligned} \alpha(c) &= P_H(q(X) > c) \qquad \text{und} \\ \alpha(c-0) &= P_H(q(X) \geq c). \end{aligned}$$

Offensichtlich gilt

$$\alpha(0-0) = P_H(q(X) \geq 0) = P_H\left(\frac{P_K(X)}{P_H(X)} \geq 0\right) = 1.$$

Desweiteren setzen wir

$$C_n = \{x \in \{0,1\}^n : q(x) > c_n\}$$

für eine strikt wachsende Folge (c_n). (C_n) ist fallend, d. h. $C_{n+1} \subseteq C_n$ für alle n und falls $c_n \uparrow\uparrow \infty$ erhalten wir

$$C := \bigcap_{n \geq 0} C_n = \emptyset.$$

Daher gilt

$$\alpha(c_n) = P_H(q(X) > c_n) = P_H(C_n) \to 0$$

(da P stetig ist). Falls umgekehrt $c_n \uparrow\uparrow c > 0$, definieren wir

$$C := \bigcap_{n \geq 0} C_N = \{x : q(x) \geq c\}.$$

Daher konvergiert $\alpha(c_n) \to P_H(C) = \alpha(c-0)$. Umgekehrt, falls $b_n \downarrow\downarrow b$, so ist

$$B_n\{x : q(x) > b_n\}$$

wachsend und

$$B := \bigcup_{n \geq 0} B_n = \{x : q(x) > b\}.$$

Das bedeutet, daß unser $\alpha(\cdot)$ eine rechts-stetige Funktion ist. Definieren wir

$$c^* = \inf\{c : \alpha(c) \leq \alpha\},$$

so erhalten wir

$$\alpha(c^*) \leq \alpha \leq \alpha(c^* - 0).$$

Falls $a(c^*) < a(c^* - 0)$ gilt, so setzen wir

$$\gamma^* = \frac{\alpha - \alpha(c^*)}{\alpha(c^* - 0) - \alpha(c^*)}$$

und schließlich $\phi(x) = 1$ falls $P_K(x) > c^* P_H(x)$, $\phi^*(x) = 0$, falls $P_K(x) < c^* P_H(x)$ und $\phi^*(x) = \gamma^*$, falls $P_K(x) = c^* P_H(x)$. Offenbar ist dies ein Neyman-Pearson Test mit

$$\begin{aligned} E_H(\phi^*) &= P_H(q(X) > c^*) + \gamma^* P_H(q(X) = c^*) \\ &= \alpha(c^*) + \gamma^*(\alpha(c^* - 0) - \alpha(c^*)) \\ &= \alpha(c^*) + \alpha - \alpha(c^*) \\ &= \alpha. \end{aligned}$$

□

Dies zeigt die Optimalität des im Beispiel der Tea-testing lady eingeführten Test-Verfahrens für den Fall einfacher Hypothesen und Alternativen. Für einseitige Tests, d. h. Tests einer Hypothese H die komplett links (oder komplett rechts) von der Alternative K liegt, überträgt sich die Optimalität mit Hilfe des folgenden

(8.16) Lemma. Sei X Binomial-verteilt zu den Parametern n und p und $x < n$. Dann ist

$$p \mapsto P_p(X \leq x)$$

stetig und strikt fallend in p und

$$P_0(X \leq x) = 1 \qquad \text{und} \qquad P_1(X \leq x) = 0.$$

Beweis. Alles bis auf die Monotonie ist trivial. Sei daher $p_1 > p_2$. Wir müssen zeigen, dass

$$P_{p^1}(X \leq x) > P_{p^2}(X \leq x).$$

Dies machen wir wieder mit Hilfe eines Kopplungsarguments. Wir wählen $p_3 \in (0, 1)$ als $p_3 := \frac{p_1}{p_2}$ und (X_i) als i.i.d. Bernoulli Variablen zum Parameter p_2. Desgleichen seien (Y_i) i.i.d. Bernoulli Variablen zum Parameter p_3 die auch unabhängig von den (X_i) sind. Definiert man $Z_i = X_i Y_i$, so nimmt auch Z_i nur die Werte 0 und 1 an und die Z_i sind unabhängig. Daher sind die (Z_i) i.i.d. Bernoulli Variablen mit Erfolgswahrscheinlichkeit

$$P(Z_i = 1) = P(X_i = 1)P(Y_i = 1) = p_2 p_3 = p_1.$$

Offensichtlich gilt $\{X_i = 0\} \subseteq \{Z_i = 0\}$ woraus wir

$$\{X_1 + \ldots + X_n \leq x\} \subseteq \{Z_1 + \ldots + Z_n \leq x\}$$

erhalten. Dies bedeutet

$$P_{p^2}(X \leq x) = P(\{X_1 + \ldots + X_n \leq x\}) < P(\{Z_1 + \ldots + Z_n \leq x\}) = P_{p^1}(X \leq x).$$

Die Ungleichung ist strikt, da die Inklusion strikt ist und die Differenz der beiden Mengen positive Wahrscheinlichkeit besitzt. □

Es soll hier erwähnt werden, dass neben einseitigen Tests auch zweiseitige Tests der Form

$$H : \{p = p_0\}$$

gegen

$$K : \{p \neq p_0\}$$

(aber nie umgekehrt – warum?) existieren. Das Testverfahren ist analog zum einseitigen Fall. Man konstruiert sich ein Intervall $I \in p_0$, so dass $P_H(\frac{\sum_{i=1}^n X_i}{n} \in I) \geq 1-\alpha$ und I dabei möglichst klein und entschließt sich H anzunehmen, falls $\frac{\sum_{i=1}^n X_i}{n} \in I$ und entscheidet sich für K, falls $\frac{\sum_{i=1}^n X_i}{n} \notin I$.

(b) *Normalverteilung:*
(i) Testen auf μ bei bekanntem σ^2
Es seien $X_1, X_2, \ldots, X_n$ unabhängig und $\mathcal{N}(\mu, \sigma^2)$-verteilt. σ^2 sei bekannt und es sei die Hypothese $H : \mu \leq \mu_0$ gegen die Alternative $K : \mu > \mu_0$ zu testen. Aus dem Abschnitt über das Schätzen wissen wir schon, daß $\hat{\mu} = \frac{\sum_{i=1}^n X_i}{n}$ das unbekannte μ gut schätzt. Wir konstruieren unseren Test also so, daß wir H verwerfen, falls $\hat{\mu} \geq \eta$ für ein noch zu bestimmendes η, das von der gegebenen Irrtumswahrscheinlichkeit α abhängt. Um η zu bestimmen, bedenken wir daß

$$P_{\mu,\sigma^2}(\hat{\mu} \geq \eta) \leq \alpha \qquad \forall \mu \leq \mu_0$$

gelten soll. Wir wissen schon, daß $\hat{\mu}$ als normierte Summe von normalverteilten Zufallsgrößen $\mathcal{N}(\mu, \sigma^2/n)$-verteilt ist. Also ist

$$P_{\mu,\sigma^2}(\hat{\mu} \geq \eta) = P_{\mu,\sigma^2}\left(\frac{\sqrt{n}(\hat{\mu}-\mu)}{\sigma} \geq \frac{\sqrt{n}(\eta-\mu)}{\sigma}\right) = 1 - \Phi\left(\frac{\sqrt{n}(\eta-\mu)}{\sigma}\right).$$

Da die rechte Seite dieser Gleichung wiederum monoton wachsend in μ ist genügt es η aus einer $\mathcal{N}(0,1)$-Tafel so zu bestimmen, daß

$$1 - \Phi\left(\frac{\sqrt{n}(\eta-\mu_0)}{\sigma}\right) = \alpha,$$

um den gewünschten Test zu konstruieren.

(ii) Testen auf σ^2 bei bekanntem μ
Wieder seien $X_1, X_2, \ldots, X_n$ unabhängig und $\mathcal{N}(\mu, \sigma^2)$-verteilt. Diesmal sei μ bekannt und wir testen $H : \sigma \geq \sigma_0$ gegen die Alternative $K : \sigma < \sigma_0$. Aus dem Abschnitt über das Schätzen wissen wir, daß $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2$ ein guter Schätzer für σ^2 ist. Nach Lemma 8.6 wissen wir schon, daß $\frac{n}{\sigma^2}\tilde{\sigma}^2$ verteilt ist gemäß der χ^2_n-Verteilung. Wollen wir also z.B. auf dem Signifikanzniveau $\alpha > 0$ die Hypothese $H : \sigma^2 \geq \sigma_0^2$ gegen $K : \sigma^2 < \sigma_0^2$ testen (der umgekehrte Test $H : \sigma^2 < \sigma_0^2$ gegen $K : \sigma^2 \geq \sigma_0^2$ geht analog), so müssen wir also ein η so finden, daß

$$P_{\mu,\sigma^2}(\tilde{\sigma}^2 < \eta) \leq \alpha$$

für alle $\sigma^2 \geq \sigma_0^2$ gilt (und dabei η möglichst klein). Wir werden dann H annehmen, falls $\tilde{\sigma}^2 \geq \eta$ und andernfalls werden wir H ablehnen. Nun ist

$$\begin{aligned} P_{\mu,\sigma^2}(\tilde{\sigma}^2 < \eta) &= P_{\mu,\sigma^2}(\frac{n}{\sigma^2}\tilde{\sigma}^2 < \frac{n}{\sigma^2}\eta) \\ &= \int_0^{n\eta/\sigma^2} g_n(z)dz \leq \int_0^{n\eta/\sigma_0^2} g_n(z)dz, \end{aligned}$$

wobei $g_n(z)$ die Dichte der χ^2_n-Verteilung bezeichnet. Wir bestimmen also η so aus einer χ^2_n-Tabelle, daß

$$\int_0^{n\eta/\sigma_0^2} g_n(z)dz = \alpha,$$

haben wir unseren Test konstruiert.

Prinzipiell underscheiden sich im Falle der Normalverteilung die Tests auf μ bzw. σ^2 bei *unbekanntem* anderen Parameter nicht von den oben gezeigten Verfahren, wenn der andere Parameter bekannt ist. Es gibt allerdings ein technisches Problem. Seien wieder $X_1, \ldots, X_n$ unabhängige, identisch nach $\mathcal{N}(\mu, \sigma^2)$ verteilte Zufallsvariablen und seien μ und σ^2 unbekannt. Wie wir schon gesehen haben, sind $\hat{\mu} = \frac{1}{n}\sum X_i$ bzw. $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \hat{\mu})^2$ dann gute Schätzer für die unbekannten Parameter, auf die getestet werden soll, aber das nüzt uns relativ wenig, denn wir haben Probleme, die Verteilung der Schätzer zu bestimmen. Zwar wissen wir, daß $\hat{\mu}$ $\mathcal{N}(\mu, \sigma^2/n)$-verteilt ist, doch kennen wir σ^2 nicht (daß wir μ nicht kennen, können wir ggf. verschmerzen, da wir ja gerade auf μ testen wollen) und von S^2 kennen wir die Verteilung überhaupt nicht. Diese Problem wollen wir im folgenden lösen. Aus technischen Gründen werden wir uns dazu zunächst um die Verteilung von S^2 kümmern. Dies geschieht mit Hilfe des folgenden Lemma:

(8.17) Lemma. *Es seien A eine orthogonale $n \times n$-Matrix, $Y = (Y_1, \ldots, Y_n)$ ein Vektor aus unabhängigen $\mathcal{N}(0,1)$-verteilten Zufallsvariablen und $Z = (Z_1, \ldots, Z_n)$ der Vektor $A(Y)$. Dann sind $Z_1, \ldots, Z_n$ unabhängig und $\mathcal{N}(0,1)$-verteilt.*

Beweis. Es bezeichne $g(y_1, \ldots, y_n)$ die Dichte von Y. Für jedes n-dimensionale Rechteck $[a, b[$ gilt nach der Transformationsformel für orthogonale Transformationen:

$$\begin{aligned} P\big(A(Y) \in [a,b[\big) &= P\big(Y \in A^{-1}([a,b[)\big) = \int_{A^{-1}([a,b[)} g(y_1,\ldots,y_n)dy_1\cdots dy_n \\ &= \int_{[a,b[} g(y_1,\ldots,y_n)dy_1\cdots dy_n = P\big(Y \in [a,b[\big). \end{aligned}$$

□

Wir wenden dieses Lemma auf die spezielle orthogonale Matrix A an, die in der ersten Zeile den Vektor $(1/\sqrt{n}, \ldots, 1/\sqrt{n})$ als Eintrag hat. Diese Vorgabe kann nach dem Gram-Schmidtschen Orthonormalisierungs-Verfahren zu einer orthogonalen Matrix aufgefüllt werden. Weiter sei $Y_i = \frac{X_i - \mu}{\sigma}$ und $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$ Hier ist dann

$$Z_1 = \frac{1}{\sqrt{n}}(Y_1 + \cdots Y_n) = \sqrt{n}\bar{Y} = \sqrt{n}\left(\frac{\hat{\mu} - \mu}{\sigma}\right).$$

Es bezeichne $\langle \cdot, \cdot \rangle$ das gewöhnliche Skalarprodukt in $\mathbb{R}^n$. Dann gilt wegen der Orthogonalität von A

$$\begin{aligned} Z_2^2 + \cdots + Z_n^2 &= \langle Z, Z\rangle - Z_1^2 = \langle Y, Y\rangle - n(\bar{Y})^2 \\ &= \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n \left(\frac{X_i - \hat{\mu}}{\sigma}\right)^2 = \frac{(n-1)}{\sigma^2} S^2 \end{aligned}$$

Da die Z_i unabhängig sind, ist Z_1 von $Z_2^2 + \cdots + Z_n^2$ unabhängig, und somit $\hat{\mu}$ von S^2. Ferner ist $\frac{(n-1)}{\sigma^2}S^2$ verteilt wie $Z_1^2 + \ldots + Z_n^2$, was nach Lemma 8.6 χ^2_{n-1}-verteilt ist. (Wir bemerken hier, daß dies insbesondere die Konsistenz von S^2 impliziert, da $V(S^2) = \frac{\sigma^4}{n-1}$ folgt.) Damit ist das Problem der Verteilung von S^2 geklärt.

Für das Problem des Testens auf μ bei unbekanntem σ^2 erinnern wir noch einmal, daß $\hat{\mu}$ eine gute Schätzung von μ war und, daß $\sqrt{n}\frac{\hat{\mu}-\mu}{\sigma}$ gemäß $\mathcal{N}(0,1)$ verteilt war. Da uns das aufgrund des unbekannten σ^2 nicht weiterhilft, ersetzen wir einfach das unbekannte σ^2 durch seine gute Schätzung S^2. Dies fürt zu folgender Statistik:

$$T := T(X) := \sqrt{n}\frac{\hat{\mu}-\mu}{S} = \frac{\sqrt{n}\frac{\hat{\mu}\mu}{\sigma}}{\sqrt{\frac{S^2(n-1)}{\sigma^2}}}\sqrt{n-1},$$

wobei wir die 2. Schreibweise gewählt haben, um anzudeuten, daß T der Quotient aus einer $\mathcal{N}(0,1)$-verteilten und einer (nach dem vorigen Schritt davon unabhängigen) χ^2_{n-1}-verteilten Variablen ist. Welche Verteilung hat nun T ? Dazu beweisen wir folgenden Satz:

(8.18) Satz. *Sind W und U_n unabhängige Zufallsvariable, und ist W $\mathcal{N}(0,1)$-verteilt und U_n χ^2_n-verteilt, so nennt man die Verteilung von*

$$T_n = \frac{W}{\sqrt{U_n/n}}$$

eine t_n-Verteilung oder auch eine t-Verteilung mit n Freiheitsgraden (t-distribution with n degrees of freedom). Die Dichte von T_n berechnet sich zu

$$h_n(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{1}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-(n+1)/2}.$$

Die t_1-Verteilung ist uns schon begegnet: Hier ist die Dichte $h_1(x) = 1/(\pi(1+x^2))$. Dies ist die Cauchy-Verteilung zu $c = 1$, siehe Beispiel (7.9)(5). Man spricht auch von der *Standard-Cauchy-Verteilung*. Die allgemeine t-Verteilung stammt von *William Sealy Gosset* (1876–1937), der unter der Pseudonym „Student“ publizierte. Dies tat er, da er als Angestellter der Guiness-Brauerei nicht publizieren durfte. Die t-Verteilung heißt daher auch *Studentsche Verteilung*.

Beweis. Da U_n χ^2_n-verteilt ist, ist $P(U_n > 0) = 1$, also ist T_n mit Wahrscheinlichkeit 1 wohldefiniert. Weiter sei $\lambda > 0$. Dann ist nach Satz (7.17)

$$\begin{aligned} P(T_n < \lambda) &= P(\sqrt{n}W < \lambda\sqrt{U_n}) \\ &= \int_0^\infty \int_{-\infty}^{\lambda\sqrt{y/n}} \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) g_n(y)\, dx\, dy. \end{aligned}$$

Wir substituieren mit $\varphi(t) = t\sqrt{y/n}$ und verwenden $\Gamma(1/2) = \sqrt{\pi}$:

$$P(T_n < \lambda) = \int_0^\infty \int_{-\infty}^{\lambda} \frac{1}{\sqrt{n}2^{\frac{n+1}{2}}\Gamma(n/2)\Gamma(1/2)} \exp\left(-\frac{1}{2}\left(y+\frac{y+t^2}{n}\right)\right) y^{\frac{n+1}{2}-1}\, dt\, dy.$$

Eine erneute Substitution $\varphi(z) = \frac{2z}{1+t^2/n}$ liefert

$$\begin{aligned} P(T_n < \lambda) &= \int_0^\infty \int_{-\infty}^{\lambda} \frac{1}{\sqrt{n}\Gamma(n/2)\Gamma(1/2)} \exp(-z) z^{\frac{n+1}{2}-1} (1+t^2/n)^{-\frac{n+1}{2}} \, dz \, dt \\ &= \int_{-\infty}^{\lambda} \frac{1}{\sqrt{n}\Gamma(n/2)\Gamma(1/2)} (1+t^2/n)^{-\frac{n+1}{2}} \Big(\int_0^\infty \exp(-z) z^{\frac{n+1}{2}-1} \, dz \Big) \, dt. \end{aligned}$$

Mit der Definition der Gammafunktion ist das innere Integral nach z gleich $\Gamma(\frac{n+1}{2})$. Da nun noch $h_n(\lambda) = h_n(-\lambda)$ gilt, ist das Lemma bewiesen. □

Mit $W = \sqrt{n}\,\bar{Y}$ und $U_{n-1} = Z_2^2 + \cdots + Z_n^2$ ist somit $T(X)$ t_{n-1} verteilt. Wir fassen also für unsere Situation zusammen:

(8.19) Satz. Sind $X_1, \ldots, X_n$ unabhängige $\mathcal{N}(\mu_0, \sigma^2)$-verteilte Zufallsgrößen, dann ist $T(X)$ t_{n-1}-verteilt.

Das Testen im Falle der Normalverteilung auf μ bzw. σ^2 bei jeweils unbekanntem anderen Parameter gestaltet sich nun genauso wie in Beispiel (8.8) (b). Die dort verwendeten Schätzer $\hat{\mu}$ und $\tilde{\sigma}^2$ für die unbekannten μ und σ^2 ersetzt man – wie oben gesehen – durch T bzw. S^2. Analog muß man die in den Tests unter Beispiel (8.8) (b) die Normalverteilung für $\hat{\mu}$ durch die t_{n-1}-Verteilung für T bzw. die χ_n^2-Verteilung von $\tilde{\sigma}^2$ durch die χ_{n-1}^2-Verteilung von S^2 ersetzen. Mit diesen Veränderungen bleiben alle weiteren Rechenschritte dieselben.

Bei genauerem Hinsehen haben wir bislang nur sogenannte *einseitige* Tests studiert; das sind solche Tests, bei denen der Parameterbereich in zwei Teilintervalle zerfällt, von denen einer die Hypothese und der andere die Alternative ist. Dies führt dazu, daß die Hypothese entweder verworfen wird, wenn der Schätzer des Parameters, auf den getestet werden soll, zu groß ist, oder wenn er zu klein ist (aber nicht beides), je nachdem, ob die Hypothese nun das "linke" oder das "rechte" Teilintervall von Θ ist. Dem gegenüber stehen *zweiseitige* Tests, bei denen Θ in drei Intervalle zerfällt. Dabei steht das mittlere Intervall für die Hypothese, die beiden anderen Intervalle bilden die Alternative. Dementsprechend wird H verworfen, wenn der Schätzer des zu testenden Parameters zu klein ist *und* dann, wenn er zu groß ist (natürlich nicht gleichzeitig !).

Die prinzipielle Testidee ändert sich nicht. Wieder approximiert man den zu testenden Parameter durch seinen guten Schätzer (die wir in den von uns betrachteten Situationen nun schon hinlänglich kennengelernt haben) und konstruiert zu gegebener Signifikanz α den Verwerfungsbereich des Tests. Wir wollen das an einem Beispiel studieren.

(8.20) Beispiel. Es seien $X_1, \ldots, X_n$ n gemäß $\mathcal{N}(\mu, \sigma^2)$ verteilte Zufallsvariablen und es seien μ und σ^2 unbekannt. Für ein gegebenes festes μ_0 wollen wir die Hypothese $\mu = \mu_0$ gegen die Alternative $\mu \neq \mu_0$ testen. Es sei also

$$H = \{(\mu, \sigma^2) : \mu = \mu_0, \sigma^2 > 0\}$$

und

$$K = \{(\mu, \sigma^2) : \mu \neq \mu_0, \sigma^2 > 0\}.$$

Schließlich sei $\Theta = H \cup K$. Da σ^2 unbekannt ist, arbeiten wir mit der Statistik T (und nicht mit $\hat{\mu}$). Wir bemerken, daß unter H die Statistik $T = \sqrt{n}\frac{\hat{\mu}-\mu_0}{S}$ t-verteilt ist, also insbesondere Erwartungswert 0 hat. Wir werden also H akzeptieren, wenn T betragsmäßig nicht zu groß ist, ansonsten lehnen wir H ab.
Sei also $\alpha > 0$ gegeben. Gesucht ist ein k (möglichst klein), so daß

$$P_{\mu_0,\sigma^2}(|T| > k) \leq \alpha.$$

Man nennt den Wert $t_{n-1,\beta}$ mit $P(T \leq t_{n-1,\beta}) = \beta$ das *β-Quantil* der t_{n-1}-Verteilung. Um einen Test zum Niveau α zu erhalten, bestimmt man aus Tabellen der t_{n-1}-Verteilung die Zahl $k = t_{n-1,1-\alpha/2}$ (das $1-\alpha/2$-Quantil). Wegen der Symmetrie der t_{n-1}-Verteilung ist dann $P(|T(X)| > k) = \alpha$. Es folgt die Entscheidungsregel: die Hypothese wird verworfen, wenn

$$|\hat{\mu} - \mu_0| > t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}.$$

Ein Beispiel: es mögen 15 unabhängige zufällige Variable mit derselben Normalverteilung $N(\mu, \sigma^2)$ die folgenden Werte angenommen haben: 0.78, 0.78, 1.27, 1.21, 0.78, 0.71, 0.68, 0.64, 0.63, 1.10, 0.62, 0.55, 0.55, 1.08, 0.52. Teste $H : \mu = \mu_0 = 0.9$ gegen $K : \mu \neq 0.9$. Bei welchem Niveau α wird H verworfen? Aus den Daten ermittelt man $\hat{\mu}$ zu 0.7934 und S zu 0.2409. Dann muß man mit Hilfe einer Tabelle der t_{14}-Verteilung α so bestimmen, daß das $1-\alpha/2$-Quantil unterhalb des Wertes $(0.9-0.7934)\sqrt{15}/0.2409$ liegt. Dies liefert den kritischen Wert $\alpha \approx 0.1$, womit für dieses Niveau und alle besseren Niveaus die Hypothese H verworfen wird.

Anschließend wollen wir noch ein Schätzproblem diskutieren, das von großer praktischer Relevanz ist. Die dort verwendeten Schätzer bzw. deren Verteilung sollen hier nicht hergeleitet werden. Die auftretenden Verteilungen haben wir allerdings schon kennengelernt. Die Testsituation tritt häufig beim Vergleich zweier Verfahren auf. Beispielsweise stelle man sich vor, ein neu den Markt kommendes Medikament B soll getestet werden. Man möchte freilich wissen, ob dieses Medikament besser ist als das bisher übliche A. Dazu wird man sinnvollerweise zwei (möglichst große) Versuchsgruppen bilden, von denen eine Medikament A, die andere Medikament B nimmt. Ist die durschnittliche Krankheitsdauer der Gruppe die Medikament B genommen hat, signifikant kürzer als die der anderen Gruppe, wird man die Hypothese, das neue Medikament bringt keine Verbesserung verwerfen, ansonsten wird man die Hypothese akzeptieren.

Prinzipiell unterscheidet man bei dieser Art Problemen zwei verschiedene Modelle. Zum einen hat man die Fälle, in denen man Medikament A und Medikament B an der gleichen Person ausprobieren kann (hir ist das Beispiel von Medikamenten auch eher schlecht gewählt; solche Tests findet man beispielsweise beim Vergleich zweier Typen Schuhsohlen, in dem man jedem Probanden unter jeden Schuh eine der beiden Sohlen nageln kann). Hat man nun n Versuchspersonen, so bekommt man für Medikament A eine Versuchsreihe $Y_1, \ldots, Y_n$, von denen wir annehmen wollen, dass die $Y_i \sim \mathcal{N}(\mu_2, \sigma^2)$ unabhängig und unabhängig von den X_i sind. Das Paar (X_i, Y_i) beschreibt somit die Messung der Wirkungen von A und B an Person i. Man spricht daher in diesem Fall auch von sogenannten gepaarten Stichproben. Wir betrachten nun die Differenz der beiden Wirkungen

$$D_i := X_i - Y_i.$$

Unter der Hypothese

$$H : \mu_1 = \mu_2$$

sind die D_i unabhängig und $\mathcal{N}(0, \sigma^2)$-verteilt. Bei unbekanntem σ^2 können wir also den oben hergeleiteten t-Test benutzen, um H gegen die Alternative

$$K : \mu_1 \neq \mu_2$$

zu testen.

(8.21) Beispiel. Wir verabreichen 10 Patienten zunächst eine Nacht Schlafmittel A und dann eine Nacht Schlafmittel B. Die folgende Tabelle gibt die Werte für D_i wieder.

Patient	1	2	3	4	5	6	7	8	9	10
D_i	1.2	2.4	1.3	1.3	0.0	1.0	1.8	0.8	4.6	1.4

Wir wollen nun H : die beiden Mittel sind gleich wirksam, also $\mu_1 = \mu_2$ auf dem Niveau $\alpha = 0.01$ testen. Aus dem Datenvektor ermitteln wir $T = 1.58\sqrt{10/1.513} = 4.06$. Da wir einen zweiseitigen Test betrachten, müssen wir dies vergleichen mit dem 0.995-Quantil der t_9-Verteilung. Dieses ist 3.25. Somit verferfen wir H aufgrund der Daten. Wir entscheiden uns also dafür, das eine Schlafmittel wir wirksamer zu erklären.

Die Modellierung durch verbundene Stichproben ist nicht immer realistisch. Meist hat man eine Gruppe, die mit Methode A behandelt wird, und eine andere, der man Methode B angedeihen läßt. Um dies zu mathematisieren seien zwei Folgen von Zufallsvariablen $X_1, \ldots, X_n$ und $Y_1, \ldots, Y_m$ gegeben, die alle unabhängig seien. Die Zufallsvariablen X_i seien alle gemäß $\mathcal{N}(\mu_1, \sigma_1^2)$, die Y_i gemäß $\mathcal{N}(\mu_2, \sigma_2^2)$ verteilt. Getestet werden soll die Hypothese $H : \mu_1 = \mu_2$ gegen die Alternative $K : \mu_1 \neq \mu_2$. Ist $\sigma_1^2 \neq \sigma_2^2$ so bekommen wir ziemlich große Probleme, die hier nicht behandelt werden sollen. Im Falle von $\sigma_1^2 = \sigma_2^2 =: \sigma^2$ schätzen wir dieses (unbekannte) σ^2 durch

$$\tilde{S}^2 := \frac{1}{m+n-2}\left(\sum_{i=1}^{m}(X_i - \bar{X})^2 + \sum_{i=1}^{n}(Y_i - \bar{Y})^2\right),$$

wobei die $\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i$ und $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ die Mittelwerte der Stichproben sind. Unter der Hypothese $H : \mu_1 = \mu_2$ ist dann die Statistik

$$\tilde{T} := \frac{\bar{X} - \bar{Y}}{\tilde{S}\sqrt{\frac{1}{m} + \frac{1}{n}}}$$

verteilt gemäß einer t_{m+n-2}-Verteilung (insbesondere hat sie den Erwartungswert 0). Wir werden also H verwerfen, wenn $\tilde{T}$ betragsmäßig zu groß wird. Ist genauer eine Signifikanz $\alpha > 0$ gegeben, so verwerfen wir H, falls

$$|\tilde{T}| > t_{m+n-1,1-\alpha/2},$$

sonst akzeptieren wir H. Das Problem, das wir somit gelöst haben, heißt auch *Zweistichprobeproblem mit unverbundenen Stichproben.*

Es sei abschließend noch erwähnt, daß man für die entsprechende nicht-parametrische Fragestellung, wenn man also nicht voraussetzt, daß die X_i und Y_j eine bestimmte Verteilung haben (z. B. eine Normalverteilung), einen anderen Test entwickelt hat. Dieser ist unter dem Namen Mann-Whinteny U-test oder Wilkoxon Zweistichproben Rangsummentest in der Literatur bekannt. Er basiert auf der ganz einfachen Idee, daß, wenn beide Stichproben $X_1, \ldots, X_n$ und $Y_1, \ldots, Y_n$ den gleichen Mittelwert haben und man die Stichprobenelemente der X_i und Y_i der Größe nach ordnet, dann auch die Summe der "Plätze" (der Statistiker sagt Ränge), an denen ein X-Stichprobenwert vorkommt, ungefähr gleich der Summe der Ränge der Y-Stichprobe sein sollte. Um so einen Test allerdings sinnvoll durchzuführen, muss man mehr über die Verteilung der Rangstatistik wissen. Das würde uns an dieser Stelle zu weit führen, wird aber in einer Vorlesung über Statistik behandelt.

Konfidenzintervalle
Das dritte Problem der Statistik (das hier nur kurz angerissen werden soll) ist das der sogenannten Konfidenzintervalle. Hierbei geht es darum Intervalle anzugeben, die den unbekannten vorgegebenen Parameter mit einer vorgegebenen Wahrscheinlichkeit einfangen. Um genau zu sein, sind wir natürlich weniger an einem einzigen Intervall interessiert, als an der Prozedur ein solches zu finden. Unter den allgemeinen Rahmenbedingugen der Statistik definieren wir

(8.22) Definition. Ein Konfidenzintervall für den unbekannten Parameter $\theta \in \Theta$ basierend auf dem Schätzer $\hat{\theta}$ ist ein Berechnungsschema, daß aus $\hat{\theta}$ ein Intervall $I(\hat{\theta})$ konstruiert, so daß $\hat{\theta} \in I(\hat{\theta})$ ist. Ein Konfidenzintervall heißt γ-Konfidenzintervall (wobei $0 \leq \gamma \leq 1$), falls

$$P_\theta(\theta \in I(\hat{\theta})) \geq \gamma \tag{8.1}$$

für alle $\theta \in \Theta$.

Es ist im allgemeinen natürlich nicht schwer einen Bereich zu finden, in dem das unbekannte θ mit großer Wahrscheinlichkeit liegt, nämlich Θ persönlich. Offenbar kann die Angabe von Θ nicht der Sinn der Konstruktion eines Konfidenzintervalls sein. Wir schließen dies aus, in dem wir fordern, daß das Konfidenzintervall in einem geeigneten Sinne möglichst klein ist (genauer, daß es kein Intervall gibt, daß echt in $I(\hat{\theta})$ enthalten ist und das auch noch die Bedingung (8.1) erfüllt).

Hat man nun einen guten Schätzer $\hat{\theta}$ für θ so bedeutet die Defintion von Konfidenzintervall, daß dies ein Intervall der Form $[\hat{\theta} - \kappa_1, \hat{\theta} + \kappa_2]$ für $\kappa_1, \kappa_2 > 0$ ist. Kennt man zudem die Verteilung, so läßt sich damit prinzipiell (nicht immer leicht) κ_1 und κ_2 berechnen. Wir werden dies an drei Beispielen sehen.

(8.23) Beispiele. *a) Binomialverteilung*
Hier wollen wir ein Konfidenzintervall für den unbekannten Parameter p konstruieren. Seien $X_1, \ldots, X_n$ n Beobachtungen. Dann wissen wir, daß $\bar{X} = \frac{X_1 + \ldots + X_n}{n}$ ein guter Schätzer für p ist. Damit ist

$$I(\bar{X}) := [\bar{X} - \kappa_1, \bar{X} + \kappa_2]$$

ein Konfidenzintervall für p. Um κ_1 und κ_2 zu berechnen, erinnern wir, daß $n\bar{X}$ binomi-

alverteilt ist zu den Parametern n und p. Um also zu gegegenem $\gamma > 0$ zu garantieren, daß

$$P_p(p \in I(\bar{X})) \geq \gamma \text{ d.h. } P_p(p \notin I(\bar{X})) \leq 1 - \gamma$$

überlegen wir, daß

$$P_p(p \notin I(\bar{X})) = P_p(\bar{X} > p + \kappa_1) + P_p(\bar{X} < p - \kappa_2).$$

Wir wollen nun κ_1 und κ_2 so wählen, daß die beiden Summanden auf der rechten Seite jeweils $\frac{1-\gamma}{2}$ sind. Unbefriedigenderweise lassen sich die entsprechenden Werte für κ_1 und κ_2 allgemein nicht gut ausrechnen. Für den Fall, daß man berechtigterweise annehmen kann, daß $\bar{X}$ dem Satz von de Moivre und Laplace genügt hatten wir schon in Kapitel 4 gesehen, wie man die entsprechenden Konfidenzintervalle konstruiert.

b) Normalverteilung
i) Konfidenzintervall für μ, σ^2 bekannt
Wieder sei aufgrund von n Beobachtungen $X_1, \ldots, X_n$, die diesmal $\mathcal{N}(\mu, \sigma^2)$-verteilt seien, ein Konfidenzintervall $I(\hat{\mu})$ für das unbekannte μ zu konstruieren, wobei $\hat{\mu} = \frac{X_1 + \ldots + X_n}{n}$ der gute Schätzer für μ ist. Bekanntlich ist $\hat{\mu}$ $\mathcal{N}(\mu, \sigma^2/n)$-verteilt. Unser Ansatz für das Konfidenzintervall ist aufgrund der Symmetrie der Normalverteilung

$$I(\hat{\mu}) = [\hat{\mu} - \eta\frac{\sigma}{\sqrt{n}}, \hat{\mu} + \eta\frac{\sigma}{\sqrt{n}}]$$

für ein zu berechnendes η. Wie oben rechnen wir für gegebenes $\gamma < 1$

$$\begin{aligned} P_{\mu,\sigma^2}(\mu \notin I(\hat{\mu})) &= P_{\mu,\sigma^2}(\hat{\mu} > \mu + \eta\frac{\sigma}{\sqrt{n}}) + P_{\mu,\sigma^2}(\hat{\mu} < \mu - \eta\frac{\sigma}{\sqrt{n}}) \\ &= P_{\mu,\sigma^2}(\sqrt{n}\frac{\hat{\mu} - \mu}{\sigma} > \eta) + P_{\mu,\sigma^2}(\sqrt{n}\frac{\hat{\mu} - \mu}{\sigma} < -\eta) \\ &= 1 - \Phi(\eta) + \Phi(-\eta) = 2(1 - \Phi(\eta)). \end{aligned}$$

Bestimmt man daher η aus einer $\mathcal{N}(0,1)$-Tafel so, daß

$$\Phi(\eta) = \frac{\gamma}{2} + \frac{1}{2}$$

so ist

$$1 - \Phi(\eta) = -\frac{\gamma}{2} + \frac{1}{2}$$

und somit

$$P_{\mu,\sigma^2}(\mu \notin I(\hat{\mu})) = 1 - \gamma;$$

das Konfidenzintervall ist also konstruiert.

ii) Konfidenzintervall für σ^2, μ bekannt
Wir erinnern daran, daß in dieser Situation $\tilde{\sigma}^2 := \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ ein guter Schätzer für σ^2 ist. Weiter ist $\frac{n}{\sigma^2}\hat{\sigma}^2$ verteilt gemäß der χ^2_n-Verteilung. Wir machen folgenden Ansatz für das Konfidenzintervall

$$I(\tilde{\sigma}^2) = [\frac{n\tilde{\sigma}^2}{\eta_1}, \frac{n\tilde{\sigma}^2}{\eta_2}].$$

Ist nun wieder γ gegeben, so ist

$$\begin{aligned} P_{\mu,\sigma^2}(\tilde{\sigma}^2 \in I(\tilde{\sigma}^2) &= P_{\mu,\sigma^2}(\frac{n\tilde{\sigma}^2}{\eta_1} \leq \sigma^2 \leq \frac{n\tilde{\sigma}^2}{\eta_2}) \\ &= P_{\mu,\sigma^2}(\eta_2 \leq \frac{n}{\sigma^2}\tilde{\sigma}^2 \leq \eta_1) \\ &= 1 - P_{\mu,\sigma^2}(\frac{n}{\sigma^2}\tilde{\sigma}^2 < \eta_2) - (1 - P_{\mu,\sigma^2}(\frac{n}{\sigma^2}\tilde{\sigma}^2 \leq \eta_1). \end{aligned}$$

Wir wählen nun η_1 und η_2 so aus einer χ^2_n-Tabelle, daß

$$P_{\mu,\sigma^2}(\frac{n}{\sigma^2}\tilde{\sigma}^2 \leq \eta_1) - P_{\mu,\sigma^2}(\frac{n}{\sigma^2}\tilde{\sigma}^2 < \eta_2) = \gamma.$$

Hierbei gibt es viele Möglichkeiten – und anders als im Falle i) ist eine symmetrische Lage des Intervalls um den unbekannten Parameter nicht besonders nahliegend, da auch die χ^2_n-Verteilung nicht symmetrisch ist. Man könnnte z.B. $\eta_1 = \infty$ wählen und erhielte ein einseitges Konfidenzintervall der Form $[0, \frac{n\tilde{\sigma}^2}{\eta_2}]$, wobei η_2 so gewählt ist, daß

$$P_{\mu,\sigma^2}(\frac{n}{\sigma^2}\tilde{\sigma}^2 < \eta_2) = 1 - \gamma$$

gilt; anderseits kann man auch $\eta_2 = 0$ setzen und erhält ein Intervall der Form $[\frac{n\tilde{\sigma}^2}{\eta_1}, \infty)$, wobei dann η_1 die Gleichung

$$P_{\mu,\sigma^2}(\frac{n}{\sigma^2}\tilde{\sigma}^2 \leq \eta_1) = \gamma$$

erfüllt. Und selbstverständlich sind auch viele Wahlen von 2-seitigen Konfidenzintervallen denkbar, beispielsweise η_1, η_2 so, daß

$$P_{\mu,\sigma^2}(\frac{n}{\sigma^2}\tilde{\sigma}^2 \leq \eta_1) = \frac{1}{2} + \frac{\gamma}{2}$$

und

$$P_{\mu,\sigma^2}(\frac{n}{\sigma^2}\tilde{\sigma}^2 < \eta_2) = \frac{1}{2} - \frac{\gamma}{2}$$

(in gewisser Weise eine symmetrische Wahl).

Konfidenzintervalle für den Fall, daß beide Parameter unbekannt sind lassen sich in ähnlicher Weise konstruieren.

9 Markoff-Ketten

Bisher haben wir uns hauptsächlich mit unabhängigen Ereignissen und unabhängigen Zufallsgrößen beschäftigt. *Andrej Andrejewitsch Markoff* (1856–1922) hat erstmalig in einer Arbeit 1906 Zufallsexperimente analysiert, bei denen die einfachste Verallgemeinerung der unabhängigen Versuchsfolge betrachtet wurde. Man spricht bei diesen Versuchsfolgen heute von Markoff-Ketten. Wir werden sehen, daß sehr viele Modelle Markoff-Ketten sind. Man kann sie anschaulich wie folgt beschreiben: Ein Teilchen bewegt sich in diskreter Zeit auf einer höchstens abzählbaren Menge I. Befindet es sich auf einem Platz $i \in I$, so wechselt es mit gewissen Wahrscheinlichkeiten (die von i abhängen) zu einem anderen Platz $j \in I$. Diese Übergangswahrscheinlichkeiten hängen aber nicht weiter von der „Vorgeschichte“ ab, das heißt von dem Weg, auf dem das Teilchen zum Platz i gekommen ist.

(9.1) Definition. Es sei I eine nichtleere, höchstens abzählbare Menge. Eine Matrix $\mathbb{P} = (p_{ij})_{i,j\in I}$ heißt *stochastische Matrix (stochastic matrix)*, wenn $p_{ij} \in [0,1]$ für alle $i, j \in I$ und $\sum_{j\in I} p_{ij} = 1$ für alle $i \in I$ gelten. Die Komponenten p_{ij} heißen *Übergangswahrscheinlichkeiten (transition probabilities)*. Eine stochastische Matrix wird im Zusammenhang mit Markoff-Ketten auch *Übergangsmatrix (transition matrix)* genannt. Eine auf einem Grundraum $(\Omega, \mathcal{F}, P)$ definierte Zufallsgröße $X : \Omega \to I$ nennt man *I-wertige Zufallsgröße*.

(9.2) Definition. Eine endlich oder unendlich lange Folge $X_0, X_1, X_2, \ldots$ I-wertiger Zufallsgrößen heißt (zeitlich homogene, time homogeneous) *Markoff-Kette (Markov chain)* mit stochastischer Matrix $\mathbb{P}$, wenn für alle $n \geq 0$ und alle $i_0, i_1, \ldots, i_n, i_{n+1} \in I$ mit $P(X_0 = i_0, \ldots, X_n = i_n) > 0$

$$P(\, X_{n+1} = i_{n+1} \mid X_0 = i_0,\, X_1 = i_1, \ldots,\, X_n = i_n \,) = p_{i_n i_{n+1}}$$

gilt. Die *Startverteilung (initial distribution)* ν einer Markoff-Kette ist definiert durch $\nu(i) = P(X_0 = i)$ für alle $i \in I$. Oft schreibt man P_ν, um die Startverteilung zu betonen. Ist die Startverteilung auf einen Punkt konzentriert, d. h. gilt $\nu(i) = 1$ für ein $i \in I$, so schreiben wir meist P_i anstelle von P_ν.

(9.3) Satz. Sei $\{X_n\}_{n\in\mathbb{N}_0}$ eine Markoff-Kette mit Startverteilung ν.

a) Für alle $n \in \mathbb{N}_0$ und $i_0, i_1, \ldots, i_n \in I$ gilt

$$P(X_0 = i_0,\, X_1 = i_1, \ldots, X_n = i_n) = \nu(i_0) p_{i_0 i_1} p_{i_1 i_2} \ldots p_{i_{n-1} i_n}.$$

b) Es seien $n < m$ und $i_n \in I$ sowie $A \subset I^{\{0,1,\ldots,n-1\}}$ und $B \subset I^{\{n+1,\ldots,m\}}$. Falls $P((X_0, X_1, \ldots, X_{n-1}) \in A,\, X_n = i_n) > 0$ ist, so gilt

$$\begin{aligned} & P(\, (X_{n+1}, \ldots, X_m) \in B \mid (X_0, \ldots, X_{n-1}) \in A,\, X_n = i_n \,) \\ = \;& P(\, (X_{n+1}, \ldots, X_m) \in B \mid X_n = i_n \,). \end{aligned}$$

Beweis. (a) folgt durch Induktion nach n: Definitionsgemäß gilt die Behauptung für $n = 0$. Gelte die Behauptung für ein $n \in \mathbb{N}_0$ und seien $i_0, i_1, \ldots, i_{n+1} \in I$. Ist $P(X_0 = i_0, \ldots, X_n = i_n) = 0$, so gilt die behauptete Formel ebenfalls für $n+1$: Ist $P(X_0 = i_0, \ldots, X_n = i_n) > 0$, so folgt aus Definition 9.2

$$\begin{aligned} P(X_0 = i_0, \ldots, X_n = i_n,\, X_{n+1} = i_{n+1}) &= P(X_{n+1} = i_{n+1} \mid X_0 = i_0, \ldots, X_n = i_n) \\ &\quad \times P(X_0 = i_0, \ldots, X_n = i_n) \\ &= \nu(i_0) p_{i_0 i_1} \ldots p_{i_{n-1} i_n} p_{i_n i_{n+1}}. \end{aligned}$$

(b) Sei $P((X_0, X_1, \ldots, X_{n-1}) \in A,\, X_n = i_n) > 0$. Mit der Definition der bedingten Wahrscheinlichkeit und Teil (a) folgt

$$\begin{aligned} & P(\,(X_{n+1}, \ldots, X_m) \in B \mid (X_0, \ldots, X_{n-1}) \in A,\; X_n = i_n\,) \\ = \; & \frac{P(\,(X_{n+1}, \ldots, X_m) \in B,\; X_n = i_n,\; (X_0, \ldots, X_{n-1}) \in A\,)}{P(\,(X_0, \ldots, X_{n-1}) \in A,\; X_n = i_n\,)} \\ = \; & \frac{\sum_{(i_{n+1}, \ldots, i_m) \in B} \sum_{(i_0, \ldots, i_{n-1}) \in A} \nu(i_0) p_{i_0 i_1} \ldots p_{i_{m-1} i_m}}{\sum_{(i_0, \ldots, i_{n-1}) \in A} \nu(i_0) p_{i_0 i_1} \ldots p_{i_{n-1} i_n}} \\ = \; & \sum_{(i_{n+1}, \ldots, i_m) \in B} p_{i_n i_{n+1}} p_{i_{n+1} i_{n+2}} \ldots p_{i_{m-1} i_m}. \end{aligned}$$

Dieser Ausdruck hängt nicht von A ab, insbesondere führt also die obige Rechnung für $A = I^{\{0,1,\ldots,n-1\}}$ zum gleichen Resultat. Aber für $A = I^{\{0,1,\ldots,n-1\}}$ gilt die in (b) behauptete Formel. □

(9.4) Bemerkung. Die Aussage von (b) heißt *Markoff-Eigenschaft (Markov property)*. Sie spiegelt genau die eingangs erwähnte Eigenschaft wieder, daß in einer Markoff-Kette die Wahrscheinlichkeit, zur Zeit $n + 1$ in einen beliebigen Zustand zu gelangen, nur vom Zustand zur Zeit n abhängt, aber nicht davon, in welchem Zustand die Kette früher war. Nicht jede Folge von I-wertigen Zufallsgrößen mit dieser Eigenschaft ist eine homogene Markoff-Kette in unserem Sinn: Die Übergangswahrscheinlichkeiten können nämlich noch von der Zeit abhängen. Genauer: Sei $X_0, X_1, \ldots$ eine Folge I-wertiger Zufallsgrößen, die die Eigenschaft aus Satz (9.3 (b)) hat. Dann existiert eine Folge $\{\mathbb{P}_n\}_{n \in \mathbb{N}_0}$ von stochastischen Matrizen $\mathbb{P}_n = (p_n(i,j))_{i,j \in I}$ mit

$$P(X_0 = i_0, \ldots, X_n = i_n) = \nu(i_0) p_0(i_0, i_1) \ldots p_{n-1}(i_{n-1}, i_n)$$

für alle $n \in \mathbb{N}_0$ und $i_0, \ldots, i_n \in I$. Der Beweis sei dem Leser überlassen. Man spricht dann von einer (zeitlich) inhomogenen Markoff-Kette. Wir werden jedoch nur (zeitlich) homogene Ketten betrachten, ohne dies jedesmal besonders zu betonen.

(9.5) Satz. Es seien $\mathbb{P} = (p_{ij})_{i,j \in I}$ eine stochastische Matrix, ν eine Verteilung auf I und $N \in \mathbb{N}_0$. Dann gibt es eine abzählbare Menge Ω, eine Wahrscheinlichkeitsverteilung p auf Ω und Abbildungen $X_i : \Omega \to I$ für alle $i \in \{0, 1, \ldots, N\}$, so daß $X_0, \ldots, X_N$ eine homogene Markoff-Kette mit Startverteilung ν und Übergangsmatrix $\mathbb{P}$ ist.

Beweis. Es sei $\Omega := I^{\{0,\ldots,N\}}$ und $p(i_0, \ldots, i_N) := \nu(i_0) p_{i_0 i_1} \ldots p_{i_{N-1} i_N}$ sowie $X_n(i_0, \ldots, i_N) = i_n$ für alle $n \in \{0, 1, \ldots, N\}$ und $(i_0, \ldots, i_N) \in \Omega$. Da die Summe der Komponenten der

stochastischen Matrix $\mathbb{P}$ in jeder Zeile gleich eins ist, gilt für alle $n \in \{0, 1, \ldots, N\}$ und $(i_0, \ldots, i_n) \in I^{\{0,\ldots,n\}}$

$$\begin{aligned} P(X_0 = i_0, \ldots, X_n = i_n) &= \sum_{(i_{n+1},\ldots,i_N) \in I^{\{n+1,\ldots,N\}}} P(X_0 = i_0, \ldots, X_N = i_N) \\ &= \sum_{(i_{n+1},\ldots,i_N) \in I^{\{n+1,\ldots,N\}}} \nu(i_0) p_{i_0 i_1} \ldots p_{i_{N-1} i_N} \\ &= \nu(i_0) p_{i_0 i_1} \ldots p_{i_{n-1} i_n}. \end{aligned}$$

Dieses Produkt ist größer als Null genau dann, wenn jeder Faktor größer als Null ist. Ist dies der Fall, so ist offenbar

$$P(\, X_{n+1} = i_{n+1} \mid X_0 = i_0, \ldots, X_n = i_n \,) = p_{i_n i_{n+1}}.$$

□

Bemerkung. Nachfolgend soll stets von einer unendlich langen Markoff-Kette ausgegangen werden, dies jedoch nur wegen einer bequemeren Notation. Alle nachfolgenden Überlegungen benötigen die Konstruktion einer unendlichen Markoff-Kette nicht, sondern kommen damit aus, daß für jedes N eine Kette gemäß Satz (9.5) konstruiert werden kann.

(9.6) Beispiele.

a) Sei $p_{ij} = q_j$ für alle $i, j \in I$, wobei $\sum_{j \in I} q_j = 1$ ist. Dann gilt

$$P(X_0 = i_0,\ X_1 = i_1, \ldots, X_n = i_n) = \nu(i_0) q_{i_1} \ldots q_{i_n}.$$

Man sieht leicht, daß $q_j = P(X_m = j)$ für $m \geq 1$ ist. Somit gilt

$$P(X_0 = i_0, \ldots, X_n = i_n) = P(X_0 = i_0) P(X_1 = i_1) \ldots P(X_n = i_n),$$

d. h., die $X_0, X_1, \ldots, X_n$ sind unabhängig. Satz (9.5) liefert also als Spezialfall die Konstruktion von unabhängigen, I-wertigen Zufallsgrößen.

b) *Irrfahrt auf $\mathbb{Z}$:* Es sei $Y_1, Y_2, \ldots$ eine Folge unabhängiger, $\{1, -1\}$-wertiger Zufallsgrößen mit $P(Y_j = 1) = p$ und $P(Y_j = -1) = 1 - p$, wobei $p \in [0, 1]$ ist. Sei $X_0 := 0$ und $X_n := \sum_{j=1}^{n} Y_j$ für $n \geq 1$. Dann ist $X_0, X_1, \ldots$ eine Markoff-Kette auf $\mathbb{Z}$. Die Übergangsmatrix $\mathbb{P} = (p_{ij})_{i,j \in \mathbb{Z}}$ ist durch $p_{i,i+1} = p$ und $p_{i,i-1} = 1 - p$ eindeutig festgelegt, und die Startverteilung ist in 0 konzentriert.

c) *Symmetrische Irrfahrt auf $\mathbb{Z}^d$:* Hier ist $I = \mathbb{Z}^d$ und $p_{(i_1,\ldots,i_d),(j_1,\ldots,j_d)} = 1/(2d)$, falls $i_k = j_k$ für alle bis auf genau ein $k \in \{1, 2, \ldots, d\}$, für das $|i_k - j_k| = 1$ ist. Alle anderen Übergangswahrscheinlichkeiten müssen dann gleich Null sein.

d) *Ehrenfests Modell der Wärmebewegung:* Es seien n Kugeln auf zwei Schachteln verteilt. Zu einem bestimmten Zeitpunkt seien r Kugeln in der rechten Schachtel und $l := n - r$ in der linken. Mit Wahrscheinlichkeit $1/2$ tun wir nun überhaupt nichts (daß diese auf den ersten Blick unsinnige Annahme begründet ist, werden

wir zu einem späteren erkennen). Im anderen Fall wird mit Wahrscheinlichkeit 1/2 eine der n Kugeln nun zufällig ausgewählt, wobei jede dieselbe Chance hat, und in die andere Schachtel gelegt. Wir können für I die Anzahl der Kugeln in der rechten Schachtel nehmen, also $I = \{0, \ldots, n\}$. Die Übergangswahrscheinlichkeiten sind gegeben durch

$$p_{r,r-1} = r/2n, \quad r \in \{1, 2, \ldots, n\},$$
$$p_{r,r+1} = 1/2 - r/2n, \quad r \in \{0, 1, \ldots, n-1\}.$$

e) *Irrfahrt auf $I = \{0, \ldots, n\}$ mit Absorption (random walk with absorbing barriers):* 0 und n seien absorbierend, also $p_{00} = 1$ und $p_{nn} = 1$. Für $i \in \{1, 2, \ldots, n-1\}$ geschehe ein Schritt nach rechts mit Wahrscheinlichkeit $p \in (0, 1)$ und ein Schritt nach links mit Wahrscheinlichkeit $q := 1 - p$, also $p_{i,i+1} = p$ und $p_{i,i-1} = q$. Die stochastische Matrix hat somit die Form

$$\mathbb{P} = \begin{pmatrix} 1 & 0 & 0 & & \\ q & 0 & p & & \\ & \ddots & \ddots & \ddots & \\ & & q & 0 & p \\ & & 0 & 0 & 1 \end{pmatrix}.$$

f) *Irrfahrt mit Reflexion (reflecting barriers):* Das gleiche Modell wie in Beispiel (e) mit der Änderung, daß $p_{01} = p_{n,n-1} = 1$ sein soll.

g) *Wettervorhersage:* Wenn wir annehmen, daß die Wahrscheinlichkeit für Regen am folgenden Tag nur von Bedingungen von heute abhängt und unbeeinflußt ist vom Wetter der vergangenen Tage, so liefert dies eine ganz einfache Markoff-Kette. Ist α die Wahrscheinlichkeit, daß es morgen regnet, wenn es heute geregnet hat, und β die Wahrscheinlichkeit, daß es morgen regnet, wenn es heute nicht geregnet hat, so hat die stochastische Matrix die Form

$$\mathbb{P} = \begin{pmatrix} \alpha & 1-\alpha \\ \beta & 1-\beta \end{pmatrix}.$$

Auf Grund der Vielzahl von Beispielen für Markoff-Ketten könnte man vermuten, daß Markoff selbst aus angewandten Fragestellungen heraus die Ketten analysiert hat. Markoff hatte jedoch bei seinen Untersuchungen primär im Sinn, Gesetze der großen Zahlen und zentrale Grenzwertsätze für die Ketten zu studieren. Er hatte nur ein Beispiel vor Augen: er analysierte die möglichen Zustände „Konsonant“ und „Vokal“ bei der Buchstabenfolge des Romans „Eugen Onegin“ von Puschkin. Die Zufallsgröße X_n soll hier den n-ten Buchstaben des Textes angeben.

Eine stochastische Matrix $\mathbb{P} = (p_{ij})_{i,j \in I}$ kann man stets ohne Probleme potenzieren: Für $n \in \mathbb{N}_0$ definiert man die n-te Potenz $\mathbb{P}^n = (p_{ij}^{(n)})_{i,j \in I}$ rekursiv durch $p_{ij}^{(0)} = \delta_{ij}$ und

$$p_{ij}^{(n+1)} = \sum_{k \in I} p_{ik}^{(n)} p_{kj}$$

für alle $i, j \in I$, das heißt, $\mathbb{P}^n$ ist das n-fache Matrixprodukt von $\mathbb{P}$ mit sich selbst. Aus der rekursiven Definition folgt, daß $\mathbb{P}^n$ selbst eine stochastische Matrix ist. Es gelten die aus der linearen Algebra bekannten Rechenregeln für Matrizen, insbesondere gilt $\mathbb{P}^m\mathbb{P}^n = \mathbb{P}^{m+n}$, das heißt

$$\sum_{k\in I} p_{ik}^{(m)} p_{kj}^{(n)} = p_{ij}^{(m+n)}, \quad i, j \in I.$$

Diese Gleichungen nennt man auch *Chapman-Kolmogoroff-Gleichungen.*

(9.7) Definition. Die Komponenten $p_{ij}^{(n)}$ der Übergangsmatrix $\mathbb{P}^n = (p_{ij}^{(n)})_{i,j\in I}$ heißen *n-stufige Übergangswahrscheinlichkeiten (n th order transition probabilities).*

(9.8) Bemerkung. Sei $X_0, X_1, X_2, \ldots$ eine Markoff-Kette mit stochastischer Matrix $\mathbb{P} = (p_{ij})_{i,j\in I}$. Sind $m, n \in \mathbb{N}_0$ und $i, j \in I$ mit $P(X_m = i) > 0$, so gilt

$$P(X_{m+n} = j \mid X_m = i) = p_{ij}^{(n)}.$$

Beweis. Es gilt

$$\begin{aligned} & P(\, X_{m+n} = j \mid X_m = i \,) \\ = & \sum_{i_{m+1},\ldots,i_{m+n-1}\in I} P(\, X_{m+1} = i_{m+1}, \ldots, \\ & X_{m+n-1} = i_{m+n-1},\ X_{m+n} = j \mid X_m = i \,) \end{aligned}$$

und mit der Definition (9.2) folgt

$$\begin{aligned} & P(\, X_{m+1} = i_{m+1}, \ldots, X_{m+n-1} = i_{m+n-1},\ X_{m+n} = j \mid X_m = i \,) \\ = \ & P(\, X_{m+n} = j \mid X_m = i,\ X_{m+1} = i_{m+1}, \ldots, X_{m+n-1} = i_{m+n-1} \,) \\ & \times \prod_{k=1}^{n-1} P(X_{m+k} = i_{m+k} \mid X_m = i,\ X_{m+1} = i_{m+1}, \ldots, X_{m+k-1} = i_{m+k-1}) \\ = \ & p_{ii_{m+1}} p_{i_{m+1}i_{m+2}} \cdots p_{i_{m+n-1}j}. \end{aligned}$$

Somit gilt

$$P(\, X_{m+n} = j \mid X_m = i \,) = \sum_{i_{m+1},\ldots,i_{m+n-1}\in I} p_{ii_{m+1}} \cdots p_{i_{m+n-1}j} = p_{ij}^{(n)}.$$

□

(9.9) Lemma. *Für alle $m, n \in \mathbb{N}_0$ und $i, j, k \in I$ gilt $p_{ij}^{(m+n)} \geq p_{ik}^{(m)} p_{kj}^{(n)}$.*

Beweis. Dies ergibt sich sofort aus den Chapman-Kolmogoroff-Gleichungen. □

(9.10) Lemma. *Es sei $X_0, X_1, X_2, \ldots$ eine Markoff-Kette mit Startverteilung ν und Übergangsmatrix $\mathbb{P}$. Dann gilt*

$$P_\nu(X_n = j) = \sum_{i\in I} \nu(i) p_{ij}^{(n)}$$

für alle $n \in \mathbb{N}_0$ und $j \in I$. Ist die Startverteilung ν auf $i \in I$ konzentriert, so gilt $P_i(X_n = j) = p_{ij}^{(n)}$.

Beweis. Aus Satz (9.3 (a)) folgt

$$\begin{aligned} P_\nu(X_n = j) &= \sum_{i_0,\ldots,i_{n-1} \in I} P_\nu(X_0 = i_0, \ldots, X_{n-1} = i_{n-1},\ X_n = j) \\ &= \sum_{i_0,\ldots,i_{n-1} \in I} \nu(i_0) p_{i_0 i_1} \ldots p_{i_{n-1} j} = \sum_{i \in I} \nu(i) p_{ij}^{(n)}. \end{aligned}$$

□

(9.11) Definition. Es sei $\mathbb{P} = (p_{ij})_{i,j \in I}$ eine stochastische Matrix. Man sagt, $j \in I$ sei von $i \in I$ aus *erreichbar (can be reached from)*, wenn ein $n \in \mathbb{N}_0$ existiert mit $p_{ij}^{(n)} > 0$. *Notation:* $i \rightsquigarrow j$.

Die in (9.11) definierte Relation auf I ist reflexiv und transitiv. Wegen $p_{ii}^{(0)} = 1 > 0$ gilt $i \rightsquigarrow i$ für alle $i \in I$. Falls $i \rightsquigarrow j$ und $j \rightsquigarrow k$ gelten, so gibt es $m, n \in \mathbb{N}_0$ mit $p_{ij}^{(m)} > 0$ und $p_{jk}^{(n)} > 0$, und dann ist $p_{ik}^{(m+n)} \geq p_{ij}^{(m)} p_{jk}^{(n)} > 0$ nach Lemma (9.9).
Die durch $i \sim j \Leftrightarrow (i \rightsquigarrow j$ und $j \rightsquigarrow i)$ für alle $i, j \in I$ definierte Relation ist offenbar eine Äquivalenzrelation auf I. Wir werden $i \sim j$ für den Rest dieses Kapitels stets in diesem Sinne verwenden.

Sind $A, B \subset I$ zwei Äquivalenzklassen der obigen Äquivalenzrelation, so sagen wir, B ist von A aus erreichbar und schreiben $A \rightsquigarrow B$, wenn $i \in A$ und $j \in B$ existieren mit $i \rightsquigarrow j$. Offensichtlich hängt dies nicht von den gewählten Repräsentanten in A und B ab.

(9.12) Definition. Es sei $\mathbb{P}$ eine stochastische Matrix.

a) Eine Teilmenge I' von I heißt *abgeschlossen (closed)*, wenn keine $i \in I'$ und $j \in I \setminus I'$ existieren mit $i \rightsquigarrow j$.

b) Die Matrix $\mathbb{P}$ und auch eine Markoff-Kette mit Übergangsmatrix $\mathbb{P}$ heißen *irreduzibel (irreducible)*, wenn je zwei Elemente aus I äquivalent sind.

Bemerkung. Es sei $\mathbb{P} = (p_{ij})_{i,j \in I}$ eine stochastische Matrix.

a) Ist $I' \subset I$ abgeschlossen, so ist die zu I' gehörige Untermatrix $\mathbb{P}' := (p_{ij})_{i,j \in I'}$ eine stochastische Matrix für I'.

b) Ist $\mathbb{P}$ irreduzibel, so existieren keine abgeschlossenen echten Teilmengen von I.

(9.13) Beispiele.

a) Die symmetrische Irrfahrt auf $\mathbb{Z}^d$ ist irreduzibel.

b) Bei der Irrfahrt auf $\{0,\dots,n\}$ mit absorbierenden Rändern gibt es drei Äquivalenzklassen, nämlich $\{0\}$, $\{1,\dots,n-1\}$ und $\{n\}$. Die Mengen $\{0\}$ und $\{n\}$ sind abgeschlossen, und es gelten $\{1,\dots,n-1\} \rightsquigarrow \{n\}$ und $\{1,\dots,n-1\} \rightsquigarrow \{0\}$.

c) Es sei $I = \{0,1,2\}$ und die stochastische Matrix gegeben durch

$$\mathbb{P} = \begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1/4 & 1/4 \\ 0 & 1/3 & 2/3 \end{pmatrix}.$$

Dann ist die Markoff-Kette irreduzibel.

d) Es sei $I = \{0,1,2,3\}$ und die stochastische Matrix gegeben durch

$$\mathbb{P} = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Dann gibt es drei Äquivalenzklassen: $\{0,1\}, \{2\}$ und $\{3\}$. Der Wert 0 ist von 2 aus erreichbar, aber nicht umgekehrt. Der Wert 3 hat absorbierendes Verhalten; kein anderer Wert ist von 3 aus erreichbar.

Es sei $X_0, X_1, X_2, \dots$ eine Markoff-Kette mit Übergangsmatrix $\mathbb{P} = (p_{ij})_{i,j\in I}$ und Startverteilung ν. Die wichtigste Frage, die uns für einen Großteils des Kapitels beschäftigen wird, ist die Diskussion der Verteilung von X_n für große n, also

$$P_\nu(X_n = j) = \sum_{i\in I} \nu(i) p_{ij}^{(n)}, \quad j \in I.$$

Zu diesem Zwecke werden wir annehmen, daß der Zustandsraum I *endlich* ist. Aus obigen Überlegungen erhält man dann, daß die Frage der asymptotischen Verteilung von X_n äquivalent ist zur Frage, wie sich große Potenzen von stochastischen Matrizen verhalten. Im dem Falle, in dem I nur aus zwei Elementen besteht, kann man sich das noch recht leicht überlegen.

(9.14) Beispiel. Sei $|I| = 2$ und

$$\mathbb{P} = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}.$$

Dann ist für $\alpha = \beta = 0$ $\mathbb{P}^n = Id$ für jedes n (wobei Id bei uns immer die Identität bezeichnet, egal auf welchem Raum sie lebt). Im Falle von $\alpha = \beta = 1$ ist offenbar $\mathbb{P}^n = \mathbb{P}$ für jedes ungerade n und $\mathbb{P}^n = Id$ für alle geraden n.

Im Falle von $0 < \alpha + \beta < 2$ (dem interessanten Fall) diagonalisieren wir $\mathbb{P}$, um seine Potenzen zu berechnen. Es ist

$$\mathbb{P} = RDR^{-1},$$

wobei

$$R = \begin{pmatrix} 1 & \alpha \\ 1 & -\beta \end{pmatrix}$$

und

$$D = \begin{pmatrix} 1 & 0 \\ 0 & 1-\alpha-\beta \end{pmatrix}$$

ist. Daher ist

$$\mathbb{P}^n = RD^nR^{-1}.$$

Nun konvergiert aber

$$D^n = \begin{pmatrix} 1 & 0 \\ 0 & (1-\alpha-\beta)^n \end{pmatrix} \underset{n\to\infty}{\longrightarrow} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

Eingesetzt ergibt das

$$\lim_{n\to\infty} \mathbb{P}^n = R \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} R^{-1} = \begin{pmatrix} \pi_1 & \pi_2 \\ \pi_1 & \pi_2 \end{pmatrix},$$

mit

$$\pi_1 = \frac{\beta}{\alpha+\beta} \qquad \pi_2 = \frac{\alpha}{\alpha+\beta}.$$

Im allgemeinen, d.h. für $|I| > 2$ sind wir leider ziemlich schnell am Ende unserer Weisheit, wenn es um die Berechnung der Eigenwerte von $\mathbb{P}$ und damit um das Diagonalisieren von $\mathbb{P}$ geht. Die obige Methode taugt also nicht, um allgemein Erkenntnisse über das Langzeitverhalten von Markoff-Ketten zu gewinnen. Der Effekt, den wir aber im Beispiel (9.14) gesehen haben, daß nämlich die Limesmatrix aus lauter identischen Zeilen besteht – und das bedeutet, daß die Markoff-Kette asymptotisch ihren Startort "vergißt" – werden wir in dem allgemeinen Limesresultat wiederfinden. Um dieses zu beweisen, müssen wir zunächst den Begriff der Entropie, den wir schon in Kapitel 4 und 6 für zweielementige Grundräume kennengelernt haben, auf größere Räume übertragen.

(9.15) Definition. Es sei I eine endliche, mindestens zweielementige Menge und ν, ϱ seinen Wahrscheinlichkeiten auf I mit $\varrho(i) > 0$ für alle $\imath \in I$. Dann heißt

$$H(\nu|\varrho) := \sum_{i\in I} \nu(i) \log\left(\frac{\nu(i)}{\varrho(i)}\right)$$

die *relative Entropie (relative entropy)* von ν bezüglich ϱ. Hierbei setzen wir $0 \log 0 = 0$.

Wir sammeln ein paar Eigenschaften der Entropiefunktion

(9.16) Proposition. *In der Situation von Definition (9.15) ist $H(\cdot|\varrho)$ positiv und strikt konvex und es ist $H(\nu|\varrho) = 0 \Leftrightarrow \nu = \varrho$.*

Beweis. Der Beweis folgt dem Beweis von Lemma (6.2). Sei also wieder die nicht–negative, strikt–konvexe Funktion $\psi(t)$ gegeben durch $\psi(t) = t \log t - t + 1$ (und wieder ist $\psi(t) = 0 \Leftrightarrow t = 1$). Dann ist

$$\begin{aligned} H(\nu|\varrho) &= \sum_{i\in I} \varrho(i) \left(\frac{\nu(i)}{\varrho(i)} \log\left(\frac{\nu(i)}{\varrho(i)}\right) - \frac{\nu(i)}{\varrho(i)} + 1 \right) \\ &= \sum_{i\in I} \varrho(i) \psi\left(\frac{\nu(i)}{\varrho(i)}\right), \end{aligned}$$

woraus die Behauptungen folgen. □

Wir kommen nun zu einem Satz, der das aymptotische Verhalten einer großen Gruppe von Markoff-Ketten klärt. Dieser Satz ist gewissermaßen ein Gesetz der großen Zahlen für Markoff-Ketten; er wird in der Literatur häufig auch als *Ergodensatz* für Markoff-Ketten bezeichnet.

(9.17) Satz. *Ergodensatz (ergodic theorem)* Sei $\mathbb{P}$ eine stochastische Matrix über einem endlichen Zustandsraum I und ν irgendeine Anfangsverteilung. Weiter existiere ein N, so daß $\mathbb{P}^N$ nur strikt positive Einträge hat. Dann konvergiert

$$\nu\mathbb{P}^n \to_{n\to\infty} \varrho,$$

wobei ϱ eine Wahrscheinlichkeit auf I ist, die der Gleichung

$$\varrho\mathbb{P} = \varrho$$

genügt.

(9.18) Bemerkung. Die Bedingung "es existiere ein N, so daß $\mathbb{P}^N$ nur strikt positive Einträge hat" impliziert natürlich, daß $\mathbb{P}$ irreduzibel ist (man kann nach spätestens N Schritten jeden Punkt von jedem anderen aus erreichen). Umgekehrt ist die Bedingung aber nicht äquivalent zur Irreduzibilität von $\mathbb{P}$. Beispielsweise ist die Matrix

$$\mathbb{P} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

irreduzibel, aber natürlich ist keine ihrer Potenzen strikt positiv. Man kann sich überlegen, daß obige Bedingung äquivalent ist zur Irreduzibilität von $\mathbb{P}$ plus einer weiteren Bedingung, die Aperiodizität von $\mathbb{P}$ heißt. Unter letzterem wollen wir verstehen, daß der ggT über sämtliche Zeiten, zu denen man mit positiver Wahrscheinlichkeit in den Punkt i zurückkehren kann, wenn man in i gestartet ist, und über sämtliche Startpunkte i eins ist. Wir werden diese Äquivalenz hier nicht beweisen und nur bemerken, daß irreduzible und aperiodische Markoff-Ketten manchmal auch *ergodisch (ergodic)* heißen.

Satz (9.17) enthält offenbar unter anderem eine unbewiesene Existenzaussage. Diese werden wir getrennt beweisen. Wir zeigen also zunächst, daß es eine Wahrscheinlichkeit ϱ mit

$$\varrho\mathbb{P} = \varrho$$

gibt. Die Existenz eines beliebigen ϱ, das obiger Gleichung genügt, ist ziemlich offensichtlich, denn offenbar ist 1 Eigenwert jeder stochastischen Matrix (die konstanten Funktionen sind rechte Eigenvektoren) – also muß es auch linke Eigenvektoren zum Eigenwert 1 geben; ein solcher ist ϱ. Auch ist es nicht schwierig, ein solches ϱ so zu normieren, daß die Summe seiner Einträge 1 ist. Was aber a priori überhaupt nicht klar ist, ist, warum ein solches ϱ eigentlich nicht-negativ sein sollte. Wer in der linearen Algebra ein wenig Perron-Froebenius Theorie betrieben hat, wird dies schon wissen. Wir werden es hier mit

Hilfe eines anderen, mehr stochastischen Arguments herleiten.

(9.19) Satz. Sei Q eine stochastische $r \times r$ Matrix. Dann existiert

$$\lim_{k\to\infty} \frac{1}{k} \sum_{j=1}^{k} Q^j =: H$$

und es gilt

$$HQ = QH = H \qquad H^2 = H.$$

Beweis. Zunächst bemerken wir, daß mit Q auch Q^n stochastisch ist (es ist z.B.

$$\sum_{f=1}^{r} Q^2(e,f) = \sum_{f=1}^{r} \sum_{d=1}^{r} Q(e,d)Q(d,f) = 1;$$

für beliebiges n geht das analog.) Damit ist dann auch

$$P_k := \frac{1}{k} \sum_{j=1}^{k} Q^j$$

stochastisch. Darüber sind die $P_k \in \mathbb{R}^{r^2}$ und als solche beschränkt. Nach dem Satz von Bolzano–Weierstraß besitzt somit die Folge der P_k einen Häufungspunkt H. Wir wollen im folgenden sehen, daß es genau einen Häufungspunkt dieser Folge gibt. Dazu betrachten wir eine Teilfolge (H_l) der Folge (P_k), die gegen H konvergiert. Damit erhalten wir

$$\begin{aligned} QH_l = H_l Q &= \frac{1}{l} \sum_{j=1}^{l} Q^{j+1} \\ = \; H_l - \frac{1}{l} Q &+ \frac{1}{l} Q^{l+1}. \end{aligned}$$

Da die letzten beiden Terme für $l \to \infty$ verschwinden, ergibt sich

$$QH = HQ = H. \tag{9.1}$$

Ist nun H' ein weiterer Häufungspunkt und (H_m) eine Folge die gegen H' konvergiert, dann erhalten wir aus (9.20) einerseits

$$H'H = HH' = H.$$

Andererseits folgert man analog zu oben

$$H'P_k = P_k H' = H'$$

für alle k und somit

$$H'H = HH' = H'.$$

Daher ist $H' = H$ und $H^2 = H$. □

Was haben wir nun damit gewonnen? Nun, die Gleichung $HQ = H$ impliziert doch, daß für jede Zeile ϱ von H gilt, daß

$$\varrho Q = \varrho,$$

jede Zeile (und jede konvexe Kombination von Zeilen) von H ist also ein linker Eigenvektor von H zum Eigenwert eins. Darüber hinaus ist die Menge der stochastischen Matrizen abgeschlossen in $\mathbb{R}^{r^2}$. Das sieht man, indem man einerseits die Abgeschlossenheit aller nicht-negativen Matrizen erkennt (das ist nicht schwer) und andererseits sieht, daß die Menge aller Matrizen mit Zeilensumme eins für alle Zeilen abgeschlossen ist (die Menge der stochastischen Matrizen ist dann der Durchschnitt dieser beiden abgeschlossenen Mengen). Letzteres ist wahr, denn die Funktionen f_i, die die i'te Zeilensumme bilden sind stetig, und die Menge der Matrizen mit Zeilensumme 1 ist dann das Urbild der (abgeschlossenen) Menge $(1, \ldots, 1)$ unter der stetigen Abbildung $f = (f_1, \ldots, f_r)$.

Somit ist H als Limes stochastischer Matrizen wieder stochastisch, seine Zeilen sind also Wahrscheinlichkeiten auf dem Grundraum. Dies beweist die Existenz einer Wahrscheinlichkeit ϱ mit

$$\varrho Q = \varrho.$$

Solche Wahrscheinlichkeiten heißen auch *stationär (stationary)* bzgl. Q. Nun sind wir in der Lage Satz (9.17) zu beweisen.

Beweis von (9.17) Wie wir eben gesehen haben, existiert eine stationäre Verteilung ϱ bzgl. $\mathbb{P}$, nämlich beispielsweise eine Zeile des entsprechend Satz (9.19) gebildeten Cesaro-Limes der Potenzen von $\mathbb{P}$ (der der Einfachhheit halber auch H heißen soll). Ein solches ϱ besitzt nur strikt positive Einträge. Wäre z.B. $\varrho(i) = 0$, so ergäbe das

$$0 = \varrho(i) = \sum_{j \in I} \varrho(j) \mathbb{P}^N(j, i)$$

im Widerspruch dazu, daß $\mathbb{P}^N$ strikt positiv ist und $\sum \varrho(j) = 1$ ist.

Darüber hinaus gibt es nur eine Verteilung ϱ, die stationär zu $\mathbb{P}$ ist (insbesondere besteht H aus lauter identischen Zeilen). Gäbe es nämlich ϱ, ϱ', die beide stationär bzgl. $\mathbb{P}$ wären, so gälte für jedes $a \in \mathbb{R}$ und $n \in \mathbb{N}$

$$\varrho - a\varrho' = (\varrho - a\varrho')\mathbb{P}^n.$$

Wir wählen

$$a = \min_{i \in I} \frac{\varrho(i)}{\varrho'(i)} =: \frac{\varrho(i_0)}{\varrho'(i_0)}.$$

Damit ist

$$0 = (\varrho - a\varrho')(i_0) = \sum_{j \in I} (\varrho - a\varrho')(j) \mathbb{P}^N(j, i_0).$$

Aus der strikten Positivität von $\mathbb{P}^N$ folgt somit, daß $\varrho(j) = a\varrho'(j)$ für alle $j \in I$ gelten muß. Da ϱ und ϱ' Wahrscheinlichkeiten sind, impliziert das, daß $a = 1$ ist und folglich $\varrho = \varrho'$. Die im Satz behauptete Konvergenz ist also die Konvergenz gegen *einen* Punkt im klassichen Sinne.

Um diese Konvergenz schließlich zu zeigen, verwenden wir die Entropiefunktion aus Definition (9.15) in der Schreibweise

$$H(\nu|\varrho) = \sum_{i\in I} \varrho(i)\psi\left(\frac{\nu(i)}{\varrho(i)}\right),$$

wobei ψ wieder die strikt konvexe Funktion

$$\psi(t) = t\log t - t + 1$$

ist. Daher ist

$$\begin{aligned}
H(\nu\mathbb{P}|\varrho) &= \sum_{i\in I} \varrho(i)\psi\left(\frac{\nu\mathbb{P}(i)}{\varrho(i)}\right) \\
&= \sum_{i\in I} \varrho(i)\psi\left(\frac{\sum_{j\in I}\nu(j)\mathbb{P}(j,i)}{\varrho(i)}\right) \\
&= \sum_{i\in I} \varrho(i)\psi\left(\frac{\sum_{j\in I}\varrho(j)\mathbb{P}(j,i)}{\varrho(i)}\frac{\nu(j)}{\varrho(j)}\right) \\
&\leq \sum_{i\in I}\sum_{j\in I} \varrho(j)\mathbb{P}(j,i)\psi\left(\frac{\nu(j)}{\varrho(j)}\right) \\
&= \sum_{j\in I} \varrho(j))\psi\left(\frac{\nu(j)}{\varrho(j)}\right) \\
&= H(\nu|\varrho),
\end{aligned}$$

wobei das "$\leq$"-Zeichen aus der Tatsache, daß $\frac{\sum_{j\in I}\varrho(j)\mathbb{P}(j,i)}{\varrho(i)}\frac{\nu(j)}{\varrho(j)}$ eine konvexe Kombination der $\frac{\nu(j)}{\varrho(j)}$ ist, folgt, zusammen mit der Konvexität von ψ und das vorletzte Gleichheitszeichen eine Konsequenz der Stochastizität von $\mathbb{P}$ ist. Somit ist

$$H(\nu\mathbb{P}|\varrho) \leq H(\nu|\varrho)$$

mit Gleichheit genau dann, wenn $\nu\mathbb{P} = \nu$, also $\nu = \varrho$ ist. Anwenden von $\mathbb{P}$ verkleinert also die Entropie und damit eine Art Distanz zum invarianten Maß.

Somit ist insbesondere die Folge $\left(H(\nu\mathbb{P}^n|\varrho)\right)_n$ monoton fallend und zwar strikt, solange $\nu\mathbb{P}^n \neq \varrho$ ist.

Wir wollen abschließend sehen, daß dies schon impliziert, daß die Folge $\varrho_n := \nu\mathbb{P}^n$ gegen ϱ konvergiert. Da ϱ_n beschränkt ist, besitzt die Folge zumindest im $\mathbb{R}^{|I|}$ einen Häufungspunkt ϱ' und es existiert eine Teilfolge $(\varrho_{n_l})_l$, die gegen ϱ' konvergiert. Wir zeigen, daß $\varrho' = \varrho$ ist (und sind dann fertig, da die Argumentation für jeden Häufungspunkt gilt und die Folge ϱ_n damit gegen ϱ konvergiert).

Nun ist einerseits

$$H(\varrho'|\varrho) \geq H(\varrho'\mathbb{P}|\varrho).$$

Andererseits haben wir

$$\begin{aligned} H(\varrho'\mathbb{P}|\varrho) &= \sum_{j\in I}\varrho(j)\psi\left(\frac{(\varrho'\mathbb{P})(j)}{\varrho(j)}\right) \\ &= \lim_{l\to\infty}\sum_{j\in I}\varrho(j)\psi\left(\frac{(\nu\mathbb{P}^{n_l})\mathbb{P}(j)}{\varrho(j)}\right) \\ &= \lim_{l\to\infty}\sum_{j\in I}\varrho(j)\psi\left(\frac{(\nu\mathbb{P}^{n_l+1})(j)}{\varrho(j)}\right). \end{aligned}$$

Nun ist $(n_l)_l$ eine Teilfolge und daher $n_l + 1 \leq n_{l+1}$. Dies ergibt mit der vorher gezeigten Monotonie

$$\begin{aligned} &\lim_{l\to\infty}\sum_{j\in I}\varrho(j)\psi\left(\frac{(\nu\mathbb{P}^{n_l+1})(j)}{\varrho(j)}\right) \\ \geq\ &\lim_{l\to\infty}\sum_{j\in I}\varrho(j)\psi\left(\frac{(\nu\mathbb{P}^{n_{l+1}})(j)}{\varrho(j)}\right) = H(\varrho'|\varrho). \end{aligned}$$

Insgesamt ist also

$$H(\varrho'|\varrho) = H(\varrho'\mathbb{P}|\varrho)$$

und daher

$$\varrho' = \varrho.$$

□

Beispiele:

1. *Irrfahrt auf dem Kreis*

   Für $n \in \mathbb{N}$ sei C_n der n-Kreis, d.h. der Graph, der entsteht, wenn man n Punkte durchnummeriert und den Punkt k mit den Punkten $k-1$ und $k+1$ verbindet (Punkt 1 wird mit 2 und n verbunden). Auf C_n definiert man eine Markoff-Kette vermöge der Übergangsvorschrift $p_{ii} = 1/2$ und $p_{i,i+1} = p_{i,i-1} = 1/4$ (dabei ist die Addition modulo n zu verstehen). Offenbar ist für die zugehörige stochastische Matrix $\mathbb{P}$ und jedes $r > n/2 + 1$, $\mathbb{P}^r$ strikt positiv. Also sind die Voraussetzungen des Ergodensatzes erfüllt und für jede beliebige Startverteilung ν konvergiert $\nu\mathbb{P}^n$ gegen das invariante Maß der Kette, was offensichtlich die Gleichverteilung auf allen Zuständen ist.

2. *Ehrenfests Urnenmodell*

   In der Situation von Beispiel (9.6 (d)) rechnet man wieder nach, daß die Bedingungen des Ergodensatzes erfüllt sind. Die Kette konvergiert daher gegen ihre Gleichgewichtsverteilung, d.h. die Binomialverteilung.

Das Arcussinusgesetz

Wir werden uns im folgenden auf eine besondere Markoff-Kette konzentrieren. Dazu bemerken wir zunächst, daß – hat man eine Folge (X_i) von unabhängigen, identisch verteilten

Zufallsvariablen mit endlich vielen Werten gegeben (daß es so eine Folge gibt, können wir allerdings hier nicht zeigen) – man daraus eine Markoffkette S_n bilden kann, indem man

$$S_n = \sum_{i=1}^{n} X_i$$

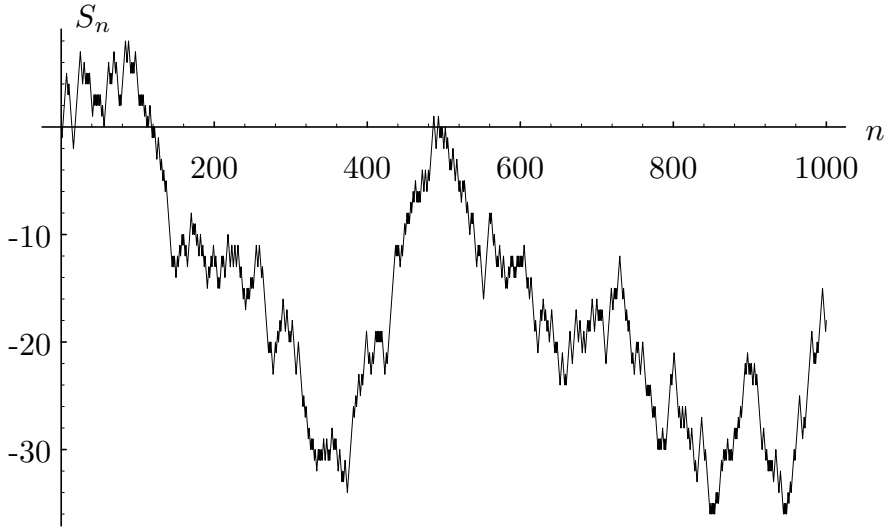
und $S_0 = 0$ setzt. In der Tat rechnet man schnell nach, daß für jedes Ereignis $\{S_{n-1} = a_{n-1}, \ldots, S_1 = a_1, S_0 = a_0\}$ mit $P(\{S_{n-1} = a_{n-1}, \ldots, S_1 = a_1, S_0 = a_0\}) > 0$ gilt

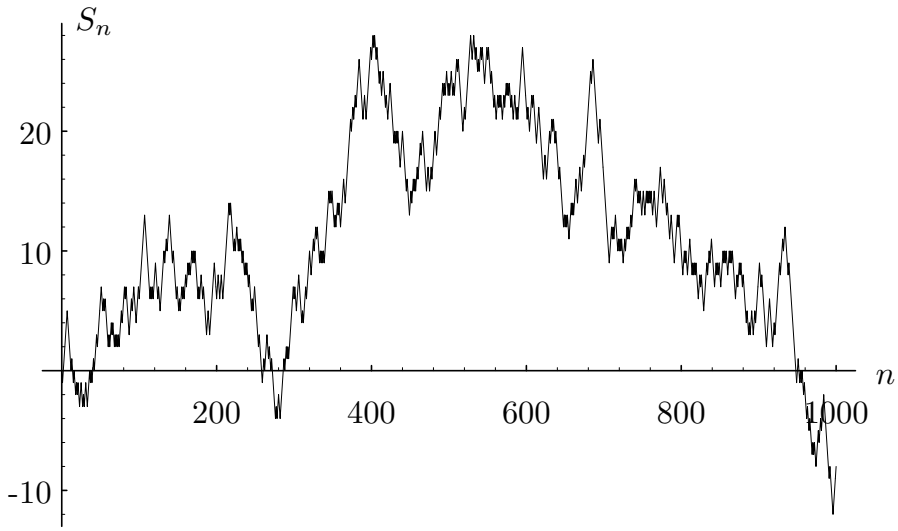
$$P(S_n = a_n | S_{n-1} = a_{n-1}, \ldots, S_1 = a_1, S_0 = a_0) = P(X_n = a_n - a_{n-1}),$$

also die Markoff-Eigenschaft erfüllt ist. Wir werden im folgenden genau eine solche Markoff-Kette betrachten, wobei die X_i unabhängige Zufallsvariablen mit Werten in $\{-1, 1\}$ und $P(X_i = 1) = P(X_i = -1) = 1/2$ sind. Anschaulich entpricht das einer Art Pfad, der in der 0 startet und in jedem Punkt $n \in \mathbb{N}$ entscheidet, ob er einen Schritt nach oben oder einen Schritt nach unten geht. Die Menge aller solcher Pfade der Länge n sei Ω_n. Aus naheliegenden Gründen bezeichnet man die Folge $S_0 = 0, S_1, \ldots, S_n$ auch als *Irrfahrt (random walk)* auf $\mathbb{Z}$. Den Index dieser Zufallsgrößen bezeichnet man meist als die „Zeit“. Wir sagen also etwa „die Wahrscheinlichkeit, daß zum Zeitpunkt 100 die Irrfahrt erstmals in 20 ist, ist…“ und meinen damit die Wahrscheinlichkeit des Ereignisses

$$A = \{S_1 \neq 20, S_2 \neq 20, \ldots, S_{99} \neq 20, S_{100} = 20\}.$$

Nachfolgend sind zwei Simulationen einer derartigen Irrfahrt mit $n = 1000$ abgebildet. Aus dem Gesetz der großen Zahlen folgt, daß zum Beispiel $S_{1000}/1000$ mit großer Wahrscheinlichkeit nahe bei 0 liegt. Um etwas zu „sehen“ müssen wir die y-Achse gegenüber der x-Achse strecken. Eine genauere theoretische Diskussion des richtigen Streckungsmaßstabs kann hier nicht gegeben werden.





Wir werden uns im folgenden also mit dem Verhalten solcher "Streckenzüge" S_n befassen. Um nicht in Konflikt zu der Tatsache zu geraten, daß wir gar nicht wissen, daß es unendlich viele unabhängige Zufallsvariablen gibt, werden wir nur Aussagen über S_n für endliche n treffen. Dazu benötigen wir nur die Existenz von unabhängigen $X_1, \ldots, X_n$, die wir schon kennen.

Zunächst betrachten wir für $k \leq n$ das Ereignis $A_k = \{S_k = 0\}$. A_k ist das unmögliche Ereignis, falls k ungerade ist. Wir betrachten also A_{2k}, $2k \leq n$. Offensichtlich gilt

$$P(A_{2k}) = \binom{2k}{k} 2^{-2k} = b(k; 2k, 1/2).$$

Wir kürzen diese Größe auch mit u_{2k} ab ($u_0 = 1$). Wir bemerken zunächst, daß $P(A_{2k})$ nicht von n, der Gesamtlänge des Experiments, abhängt, sofern nur $n \geq 2k$ gilt. Dies ist nicht weiter erstaunlich, denn die X_i sind ja unabhängig.

Wir werden diesem Phänomen noch mehrmals begegnen und wollen es deshalb genau ausformulieren: Sei $k < n$ und A ein Ereignis in Ω_k. Wir können ihm das Ereignis

$$\bar{A} = \{\, \omega = (s_0, \ldots, s_n) \in \Omega_n : (s_0, \ldots, s_k) \in A \,\}$$

in Ω_n zuordnen. Dann gilt

$$P^{(k)}(A) = P^{(n)}(\bar{A}),$$

wobei $P^{(n)}$ die durch die Gleichverteilung auf den Teilmengen von Ω_n definierte Wahrscheinlichkeit ist. Der Leser möge dies selbst verifizieren. Für ein derartiges Ereignis ist es deshalb gleichgültig, in welchem Pfadraum Ω_n die Wahrscheinlichkeit berechnet wird, sofern nur $n \geq k$ ist. Wir werden im weiteren stillschweigend auch endlich viele Ereignisse miteinander kombinieren (z.BDurchschnitte bilden), die zunächst für Pfade unterschiedlicher Länge definiert sind. Dies bedeutet einfach, daß diese Ereignisse im obigen Sinne als Ereignisse in einem gemeinsamen Raum Ω_n interpretiert werden, wobei nur n genügend groß gewählt werden muß.

Um die Größenordnung von $u_{2k} = P(A_{2k})$ für große k zu bestimmen, erinnern wir uns an den lokalen Grenzwertsatz (Satz (4.2)). Dieser liefert sofort:

(9.21) Satz.

$$u_{2k} \sim \frac{1}{\sqrt{\pi k}},$$

d.h.

$$\lim_{k\to\infty} u_{2k}\sqrt{\pi k} = 1.$$

Interessanterweise lassen sich die Wahrscheinlichkeiten einer Reihe anderer Ereignisse in Beziehung zu u_{2k} setzen. Es sei zunächst für $k \in \mathbb{N}$ f_{2k} die Wahrscheinlichkeit, daß die erste Nullstelle der Irrfahrt nach dem Zeitpunkt 0 die Zeitkoordinate $2k$ hat, das heißt

$$f_{2k} = P(S_1 \neq 0,\, S_2 \neq 0, \ldots, S_{2k-1} \neq 0,\, S_{2k} = 0).$$

Dann gilt

(9.22) Satz.

1. $f_{2k} = \frac{1}{2k}u_{2k-2} = P(S_1 \geq 0,\, S_2 \geq 0, \ldots, S_{2k-2} \geq 0,\, S_{2k-1} < 0)$
   $= u_{2k-2} - u_{2k}.$
2. $u_{2k} = P(S_1 \neq 0,\, S_2 \neq 0, \ldots, S_{2k} \neq 0) = P(S_1 \geq 0,\, S_2 \geq 0, \ldots, S_{2k} \geq 0).$
3. $u_{2k} = \sum_{j=1}^{k} f_{2j}u_{2k-2j}.$

Zum Beweis dieses Satzes müssen wir ein wenig ausholen. Insbesondere stellen wir einen eleganten Trick vor, mit dem sich die Mächtigkeit gewisser Pfadmengen bestimmen läßt. Dieser beruht auf einer teilweisen Spiegelung der Pfade an der x-Achse.

Wir sagen, daß ein Pfad $(s_i, s_{i+1}, \ldots, s_j)$ *die x-Achse berührt*, falls ein k mit $i \leq k \leq j$ existiert, für das $s_k = 0$ ist.

(9.23) Lemma. *(Reflektionsprinzip, reflection principle) Es seien $a, b \in \mathbb{N}$ und $i, j \in \mathbb{Z}$ mit $i < j$. Die Anzahl der Pfade von (i, a) nach (j, b), welche die x-Achse berühren, ist gleich der Anzahl der Pfade von $(i, -a)$ nach (j, b).*

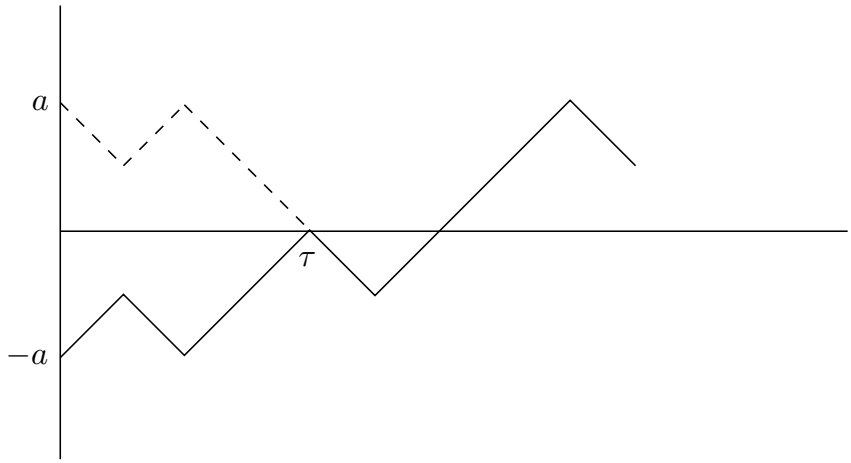
Beweis. Wir geben eine bijektive Abbildung an, die die Menge der Pfade von $(i, -a)$ nach (j, b) auf die Menge der Pfade von (i, a) nach (j, b), welche die x-Achse berühren, abbildet. Sei

$$(s_i = -a, s_{i+1}, \ldots, s_{j-1},\, s_j = b)$$

ein Pfad von $(i, -a)$ nach (j, b). Dieser Pfad muß notwendigerweise die x-Achse berühren. Sei τ die kleinste Zahl $> i$, für welche $s_\tau = 0$ gilt. Offensichtlich ist dann

$$(-s_i, -s_{i+1}, \ldots, -s_{\tau-1},\, s_\tau = 0,\, s_{\tau+1}, \ldots, s_j = b)$$

ein Pfad von (i, a) nach (j, b), der die x-Achse berührt, und die Zuordnung ist bijektiv. □



Das Spiegelungsprinzip werden wir nun verwenden, um die Menge der Pfade, die nach $2k$ Schritten zum ersten Mal wieder die x-Achse berühren abzuzählen.

(9.24) Satz.

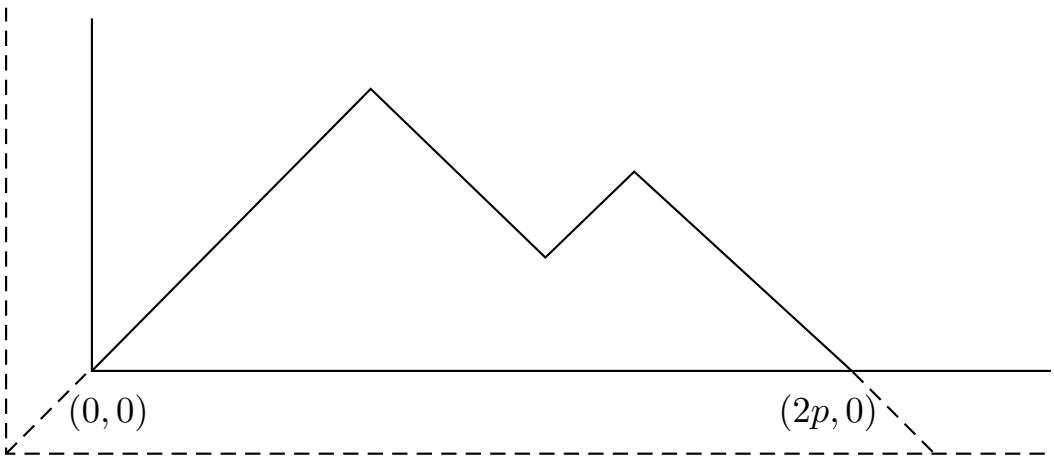
1. Es gibt $\frac{1}{p}\binom{2p-2}{p-1}$ Pfade von $(0,0)$ nach $(2p,0)$ mit $s_1 > 0,\ s_2 > 0, \ldots, s_{2p-1} > 0$.
2. Es gibt $\frac{1}{p+1}\binom{2p}{p}$ Pfade von $(0,0)$ nach $(2p,0)$ mit $s_1 \geq 0,\ s_2 \geq 0, \ldots, s_{2p-1} \geq 0$.

Beweis. (1) Es ist notwendigerweise $s_1 = 1$ und $s_{2p-1} = 1$. Wir suchen somit nach der Anzahl der Pfade von $(1,1)$ nach $(2p-1,1)$ mit $s_1 > 0,\ s_2 > 0, \ldots, s_{2p-1} = 1$. Diese ist gleich der Anzahl aller Pfade von $(1,1)$ nach $(2p-1,1)$ minus der Anzahl der Pfade, die die x-Achse berühren. Dies ist nach dem Spiegelungsprinzip gleich der Anzahl aller Pfade von $(1,1)$ nach $(2p-1,1)$ minus der Anzahl der Pfade von $(-1,1)$ nach $(2p-1,1)$. Nach ein bißchen elementarer Kombinatorik erhält man daher

$$\binom{2p-2}{p-1} - \binom{2p-2}{p} = \frac{1}{2p-1}\binom{2p-1}{p} = \frac{1}{p}\binom{2p-2}{p-1}$$

als die gesuchte Anzahl der Pfade.

(2) Wir verlängern jeden Pfad, der die Bedingung erfüllt, indem wir noch die beiden Punkte $(-1,-1)$ und $(2p+1,-1)$ anfügen und mit $(0,0)$ bzw. $(2p,0)$ verbinden.



Auf diese Weise wird eine bijektive Abbildung von der gesuchten Menge von Pfaden auf die Menge der Pfade von $(-1,-1)$ nach $(2p+1,-1)$, welche die Bedingung $s_0 > -1$, $s_1 > -1, \ldots, s_{2p} > -1$ erfüllen, hergestellt. Die Anzahl der Pfade in dieser Menge ist gleich der Anzahl der Pfade von $(0,0)$ nach $(2p+2,0)$ mit $s_1 > 0$, $s_2 > 0, \ldots, s_{2p+1} > 0$ (Verschiebung des Ursprungs). (2) folgt dann aus (1). □

Nun sind wir in der Lage Satz (9.22) zu beweisen:

Beweis von Satz (9.22). (1) Nach (9.24 (1)) gibt es $\frac{1}{k}\binom{2k-2}{k-1}$ Pfade von $(0,0)$ nach $(2k,0)$ mit $s_1 > 0, \ldots, s_{2k-1} > 0$ und natürlich genauso viele mit $s_1 < 0, \ldots, s_{2k-1} < 0$. Es folgt

$$f_{2k} = \frac{2}{k}\binom{2k-2}{k-1}2^{-2k} = \frac{1}{2k}\binom{2k-2}{k-1}2^{-2(k-1)} = \frac{1}{2k}u_{2k-2}.$$

Wir beweisen die nächste Gleichung: Falls $s_{2k-2} \geq 0$ und $s_{2k-1} < 0$ sind, so gelten $s_{2k-2} = 0$ und $s_{2k-1} = -1$. Die Anzahl der Pfade von $(0,0)$ nach $(2k-1,-1)$ mit $s_1 \geq 0, \ldots, s_{2k-3} \geq 0$, $s_{2k-2} = 0$ ist gleich der Anzahl der Pfade von $(0,0)$ nach $(2k-2,0)$ mit allen y-Koordinaten ≥ 0. Die zweite Gleichung in (1) folgt dann mit Hilfe von (9.24 (2)). Die dritte ergibt sich aus

$$u_{2k} = \binom{2k}{k}2^{-2k} = \frac{2k(2k-1)}{k \cdot k}\binom{2k-2}{k-1} \cdot \frac{1}{4} \cdot 2^{-2k+2} = \left(1 - \frac{1}{2k}\right)u_{2k-2}.$$

(2) C_{2j} sei das Ereignis $\{S_1 \neq 0, S_2 \neq 0, \ldots, S_{2j-1} \neq 0, S_{2j} = 0\}$. Diese Ereignisse schließen sich gegenseitig aus und haben Wahrscheinlichkeiten $f_{2j} = u_{2j-2} - u_{2j}$. Somit ist mit $u_0 = 1$

$$P(S_1 \neq 0, S_2 \neq 0, \ldots, S_{2k} \neq 0) = 1 - P\left(\bigcup_{j=1}^{k} C_{2j}\right) = 1 - \sum_{j=1}^{k}(u_{2j-2} - u_{2j}) = u_{2k}.$$

Die zweite Gleichung folgt analog aus der dritten Identität in (1).

(3) Für $1 \leq j \leq k$ sei $B_j = \{S_1 \neq 0, S_2 \neq 0, \ldots, S_{2j-1} \neq 0, S_{2j} = 0, S_{2k} = 0\}$. Diese Ereignisse sind paarweise disjunkt, und ihre Vereinigung ist $\{S_{2k} = 0\}$. $|B_j|$ ist offenbar gleich der Anzahl der Pfade von $(0,0)$ nach $(2j,0)$, die die x-Achse dazwischen

nicht berühren, multipliziert mit der Anzahl aller Pfade von $(2j,0)$ nach $(2k,0)$, das heißt $|B_j| = 2^{2j} f_{2j} 2^{2k-2j} u_{2k-2j}$. Somit gilt $P(B_j) = f_{2j} u_{2k-2j}$, das heißt

$$u_{2k} = \sum_{j=1}^{k} P(B_j) = \sum_{j=1}^{k} f_{2j} u_{2k-2j}.$$

□

Eine interessante Folgerung ergibt sich aus der ersten Gleichung in (2). Da nach (9.21) $\lim_{k\to\infty} u_{2k} = 0$ gilt, folgt, daß die Wahrscheinlichkeit für keine Rückkehr der Irrfahrt bis zum Zeitpunkt $2k$ mit $k \to \infty$ gegen 0 konvergiert. Man kann das folgendermaßen ausdrücken: ,,Mit Wahrscheinlichkeit 1 findet irgendwann eine Rückkehr statt.“ Man sagt auch, die Irrfahrt sei rekurrent. Wir wollen das noch etwas genauer anschauen und bezeichnen mit T den Zeitpunkt der ersten Nullstelle nach dem Zeitpunkt 0. T muß gerade sein, und es gilt $P(T = 2k) = f_{2k}$. Aus (1) und $u_{2k} \to 0$ folgt

$$\begin{aligned}
\sum_{k=1}^{\infty} f_{2k} &= \lim_{N\to\infty} \sum_{k=1}^{N} f_{2k} \\
&= \lim_{N\to\infty} \sum_{k=1}^{N} (u_{2k-2} - u_{2k}) \\
&= \lim_{N\to\infty} (u_0 - u_{2N}) = 1.
\end{aligned}$$

Wir sehen also, daß $(f_{2k})_{k\in\mathbb{N}}$ eine Wahrscheinlichkeitsverteilung auf den geraden natürlichen Zahlen definiert, die Verteilung von T. Daraus läßt sich der Erwartungswert von T berechnen:

$$ET = \sum_{k=1}^{\infty} 2k f_{2k} = \sum_{k=1}^{\infty} u_{2k-2},$$

wobei wir die Gleichung (9.22 (1)) anwenden. Nach (9.21) divergiert jedoch diese Reihe! Man kann auch sagen, daß ET gleich ∞ ist. Mit Wahrscheinlichkeit 1 findet also ein Ausgleich statt; man muß jedoch im Schnitt unendlich lange darauf warten.

Obgleich $P(S_1 \neq 0, \ldots, S_{2k} \neq 0) = P(S_1 \geq 0, \ldots, S_{2k} \geq 0) \sim 1/\sqrt{\pi k}$ gegen 0 konvergiert, ist diese Wahrscheinlichkeit erstaunlich groß. Wieso erstaunlich? Wir betrachten das Ereignis $F_j^{(k)}$, daß die Irrfahrt während genau $2j$ Zeiteinheiten bis $2k$ positiv ist. Aus formalen Gründen präzisieren wir ,,positiv sein“ wie folgt: Die Irrfahrt ist positiv im Zeitintervall von l bis $l+1$, falls S_l *oder* $S_{l+1} > 0$ ist. Es kann also auch $S_l = 0$, $S_{l+1} > 0$ oder $S_l > 0$, $S_{l+1} = 0$ sein. Man überzeugt sich leicht davon, daß die Anzahl der Intervalle, wo dieses der Fall ist, gerade ist. $F_k^{(k)}$ ist natürlich gerade das Ereignis $\{S_1 \geq 0, S_2 \geq 0, \ldots, S_{2k} \geq 0\}$. Aus Gründen der Symmetrie ist $P(F_0^{(k)}) = P(F_k^{(k)})$, was nach (9.24 (2)) gleich $u_{2k} \sim 1/\sqrt{\pi k}$ ist.

Die $F_j^{(k)}$ sind für $0 \leq j \leq k$ paarweise disjunkt, und es gilt

$$\sum_{j=0}^{k} P(F_j^{(k)}) = 1.$$

Mithin können nicht allzuviele der $P(F_j^{(k)})$ von derselben Größenordnung wie $P(F_k^{(k)})$ sein, denn sonst müßte die obige Summe > 1 werden. Anderseits ist wenig plausibel, daß unter diesen Wahrscheinlichkeiten gerade $P(F_k^{(k)})$ und $P(F_0^{(k)})$ besonders groß sind. Genau dies ist jedoch der Fall, wie aus dem folgenden bemerkenswerten Resultat hervorgehen wird.

(9.25) Satz. Für $0 \leq j \leq k$ gilt

$$P(F_j^{(k)}) = u_{2j}u_{2k-2j}.$$

Beweis. Wir führen einen Induktionsschluß nach k. Für $k = 1$ gilt

$$P(F_0^{(1)}) = P(F_1^{(1)}) = \frac{1}{2} = u_2.$$

Wir nehmen nun an, die Aussage des Satzes sei bewiesen für alle $k \leq n-1$, und beweisen sie für $k = n$.

Wir hatten in (9.22 (2)) schon gesehen, daß $P(F_0^{(n)}) = P(F_n^{(n)}) = u_{2n}$ ist (u_0 ist $= 1$). Wir brauchen deshalb nur noch $1 \leq j \leq n-1$ zu betrachten. Zunächst führen wir einige spezielle Mengen von Pfaden ein.

Für $1 \leq l \leq n$, $0 \leq m \leq n-l$ sei $G_{l,m}^+$ die Menge der Pfade der Länge $2n$ mit: $s_0 = 0$, $s_1 > 0$, $s_2 > 0, \ldots, s_{2l-1} > 0$, $s_{2l} = 0$ und $2m$ Strecken des Pfades zwischen den x-Koordinaten $2l$ und $2n$ sind positiv.

Analog bezeichne $G_{l,m}^-$ für $1 \leq l \leq n$, $0 \leq m \leq n-l$, die Menge der Pfade mit: $s_0 = 0$, $s_1 < 0$, $s_2 < 0, \ldots, s_{2l-1} < 0$, $s_{2l} = 0$ und $2m$ Strecken des Pfades zwischen den x-Koordinaten $2l$ und $2n$ sind positiv.

Die $G_{l,m}^+$, $G_{l,m}^-$ sind offensichtlich alle paarweise disjunkt. Ferner gilt

$$G_{l,m}^+ \subset F_{l+m}^{(n)},\ G_{l,m}^- \subset F_m^{(n)}.$$

Man beachte, daß für $1 \leq j \leq n-1$ jeder Pfad aus $F_j^{(n)}$ zu genau einer der Mengen $G_{l,m}^+$, $G_{l,m}^-$ gehört. Dies folgt daraus, daß ein solcher Pfad mindestens einmal das Vorzeichen wechseln, also auch die 0 passieren muß. Ist $2l$ die x-Koordinate der kleinsten Nullstelle > 0, so gehört der Pfad zu $G_{l,j-l}^+$, falls der Pfad vor $2l$ positiv, und zu $G_{l,j}^-$, falls er vor $2l$ negativ ist. Demzufolge ist

$$P(F_j^{(n)}) = \sum_{l=1}^{j} P(G_{l,j-l}^+) + \sum_{l=1}^{n-j} P(G_{l,j}^-).$$

Es bleibt noch die Aufgabe, die Summanden auf der rechten Seite dieser Gleichung zu berechnen.

Offensichtlich enthalten $G_{l,m}^+$ und $G_{l,m}^-$ gleich viele Pfade. $|G_{l,m}^+|$ ist gleich der Anzahl der Pfade von $(0,0)$ nach $(2l,0)$ mit $s_1 > 0$, $s_1 > 0, \ldots, s_{2l-1} > 0$ multipliziert mit der Anzahl der Pfade der Länge $2n-2l$ mit Start in $(2l,0)$ und $2m$ positiven Strecken, das heißt

$$\begin{aligned} |G_{l,m}^+| &= |G_{l,m}^-| = \frac{1}{2} f_{2l} 2^{2l} P(F_m^{(n-l)}) 2^{2n-2l}, \\ P(G_{l,m}^+) &= P(G_{l,m}^-) = \frac{1}{2} f_{2l} P(F_m^{(n-l)}). \end{aligned}$$

Nach der weiter oben stehenden Gleichung ist also

$$P(F_j^{(n)}) = \frac{1}{2}\sum_{l=1}^{j} f_{2l} P(F_{j-l}^{(n-l)}) + \frac{1}{2}\sum_{l=1}^{n-j} f_{2l} P(F_j^{(n-l)}).$$

Nach Induktionsvoraussetzung ist das

$$= \frac{1}{2}\sum_{l=1}^{j} f_{2l}\, u_{2j-2l}\, u_{2n-2j} + \frac{1}{2}\sum_{l=1}^{n-j} f_{2l}\, u_{2n-2j-2l}\, u_{2j} = u_{2j}\, u_{2n-2j} \quad \text{nach (9.22 (3)).}$$

□

Um das Verhalten von $P(F_j^{(k)})$ für festes k als Funktion von j zu untersuchen, betrachten wir für $1 \le j \le k-1$ die Quotienten

$$\begin{aligned}\frac{P(F_j^{(k)})}{P(F_{j+1}^{(k)})} &= \frac{\binom{2j}{j}\binom{2k-2j}{k-j}}{\binom{2j+2}{j+1}\binom{2k-2j-2}{k-j-1}} = \frac{(2j)!(2k-2j)!((j+1)!)^2((k-j-1)!)^2}{(j!)^2((k-j)!)^2(2j+2)!(2k-2j-2)!} \\ &= \frac{(2k-2j-1)(j+1)}{(2j+1)(k-j)}.\end{aligned}$$

Dieser Quotient ist > 1, $= 1$ oder < 1, je nachdem, ob $j < \frac{k-1}{2}$, $j = \frac{k-1}{2}$ oder $j > \frac{k-1}{2}$ ist. Als Funktion von j fällt also $P(F_j^{(k)})$ für $j < \frac{k-1}{2}$ und steigt an für $j > \frac{k-1}{2}$.

$P(F_0^{(k)}) = P(F_k^{(k)})$ ist also der größte vorkommende Wert und $P(F_{\lceil \frac{k-1}{2} \rceil})$ der kleinste. Es ist bedeutend wahrscheinlicher, daß die Irrfahrt über das ganze betrachtete Zeitintervall positiv ist, als daß sich positive und negative Zahlen ausgleichen. Dies scheint im Widerspruch zum Gesetz der großen Zahlen zu stehen. Ohne dies genauer diskutieren zu können, sei aber daran erinnert, daß die Rückkehrzeit T nach 0 keinen endlichen Erwartungswert hat, wie wir oben gezeigt haben.

Mit Hilfe von (9.23) läßt sich eine einfach Approximation für $P(F_j^{(k)})$ für große j und $k-j$ gewinnen:

(9.26) Satz. Für $j \to \infty$, $k - j \to \infty$ gilt $P(F_j^{(k)}) \sim \frac{1}{\pi}\frac{1}{\sqrt{j(k-j)}}$, das heißt

$$\lim_{\substack{j\to\infty \\ k-j\to\infty}} \sqrt{j(k-j)}\, P(F_j^{(k)}) = \frac{1}{\pi}.$$

□

Betrachten wir speziell $x \in (0,1)$ so gilt für $j, k \to \infty$ mit $j/k \sim x$

$$P(F_j^{(k)}) \sim \frac{1}{\pi k}\frac{1}{\sqrt{x(1-x)}}.$$

Diese Wahrscheinlichkeiten sind also von der Größenordnung $1/k$, das heißt asymptotisch viel kleiner als

$$P(F_0^{(k)}) = P(F_k^{(k)}) \sim \frac{1}{\sqrt{\pi k}}.$$

Die Funktion $(x(1-x))^{-1/2}$ hat für $x = 0$ und 1 Pole. Das steht in Übereinstimmung damit, daß für $j/k \sim 0$ und $j/k \sim 1$ die Wahrscheinlichkeiten $P(F_j^{(k)})$ von einer anderen Größenordnung als $1/k$ sind.

Eine Aussage wie (9.26) ist gewissermaßen auch ein lokaler Grenzwertsatz, da wir damit Informationen über die Wahrscheinlichkeit, daß der Zeitraum der Führung exakt $= 2j$ ist, erhalten. Da diese Wahrscheinlichkeiten jedoch alle für große k klein werden, interessiert man sich eher zum Beispiel für die Wahrscheinlichkeit, daß der relative Anteil der Zeit, wo die Irrfahrt positiv ist, $\geq \alpha$ ist.

Es seien $0 < \alpha < \beta < 1$. $\gamma_k(\alpha, \beta)$ sei die Wahrscheinlichkeit, daß dieser relative Anteil der Zeit zwischen α und β liegt. Genauer: T_k sei (die auf Ω_{2k} definierte) Zufallsgröße, die die Dauer der Führung zählt:

$$T_k := \sum_{j=1}^{2k} 1_{\{S_{j-1} \geq 0,\, S_j \geq 0\}}.$$

Dann ist

$$\gamma_k(\alpha, \beta) := P\Big(\alpha \leq \frac{T_k}{2k} \leq \beta\Big) = \sum_{j: \alpha \leq \frac{j}{k} \leq \beta} P(F_j^{(k)}).$$

Wir wollen nun aus (9.26) für $k \to \infty$ folgern:

$$\gamma_k(\alpha, \beta) \sim \frac{1}{\pi} \sum_{j: \alpha \leq \frac{j}{k} \leq \beta} \frac{1}{k} \frac{1}{\sqrt{\frac{j}{k}(1 - \frac{j}{k})}}. \tag{9.2}$$

Die rechte Seite ist nichts anderes als die Riemann-Approximation für

$$\int_\alpha^\beta \frac{1}{\pi} \frac{1}{\sqrt{x(1-x)}} dx = \frac{2}{\pi}(\arcsin\sqrt{\beta} - \arcsin\sqrt{\alpha}).$$

Es folgt damit:

(9.28) Satz. *(Arcus-Sinus-Gesetz)*

$$\lim_{k \to \infty} \gamma_k(\alpha, \beta) = \frac{2}{\pi}(\arcsin\sqrt{\beta} - \arcsin\sqrt{\alpha}).$$

Beweis. Wir müssen (9.27) zeigen. Wir schreiben die Stirling-Approximation als $n! = \sqrt{2\pi n}\big(\frac{n}{e}\big)^n F(n)$ mit $\lim_{n\to\infty} F(n) = 1$. Es folgt

$$P(F_j^{(k)}) = \binom{2j}{j}\binom{2k-2j}{k-j}\frac{1}{2^{2k}} = \frac{1}{\pi}\frac{1}{\sqrt{(\frac{j}{k})(1-(\frac{j}{k}))}}\frac{1}{k}\frac{F(2j)\,F(2(k-j))}{F(j)\,F(j)\,F(k-j)\,F(k-j)}.$$

Wir wählen nun ein $\delta > 0$ mit $0 < \delta < 1/2$ und betrachten für jedes k nur die Werte j für die gilt

$$\delta \leq \frac{j}{k} \leq 1 - \delta,$$

womit $k\delta \leq j$ und $k\delta \leq k - j$ folgt. Für $k \to \infty$ konvergiert nun jedes $F(j), F(k-j), F(2j)$ gleichmäßig für alle obigen Werte von j. Somit existiert für $\delta \leq \alpha < \beta \leq 1 - \delta$ ein $G_{\alpha,\beta}(k)$ für jedes $k = 1, 2, \ldots$, so daß für jedes obige $\delta > 0$ gilt:

$$\lim_{k\to\infty} G_{\alpha,\beta}(k) = 1 \quad \text{gleichmäßig für} \quad \delta \leq \alpha < \beta \leq 1 - \delta$$

und

$$\sum_{\alpha\le\frac{j}{k}\le\beta} P(F_j^{(k)}) = \left(\frac{1}{k}\sum_{\alpha\le\frac{j}{k}\le\beta}\frac{1}{\pi}\frac{1}{\sqrt{(j/k)(1-(j/k))}}\right) G_{\alpha,\beta}(k).$$

Nun folgt die Behauptung gleichmäßig für $\delta \le \alpha < \beta \le 1-\delta$, wie auch immer $0 < \delta < 1/2$ gewählt war. Damit folgt die Behauptung. □

(9.29) Bemerkung. Die Aussage von (9.28) ist auch richtig für $\alpha = 0$ oder $\beta = 1$. Das heißt etwa, daß $\gamma_k(0,\beta)$ — die Wahrscheinlichkeit dafür, daß der relative Anteil der Zeit, in der K_1 führt, $\le \beta$ ist — gegen $\frac{2}{\pi}\arcsin\sqrt{\beta}$ konvergiert.

Beweis Offensichtlich gilt $\lim_{k\to\infty}\gamma_k(0,\frac{1}{2}) = 1/2$. Ist $\beta \in (0,1/2)$, so folgt

$$\lim_{k\to\infty}\gamma_k(0,\beta) = \lim_{k\to\infty}(\gamma_k(0,1/2) - \gamma_k(\beta,1/2)) = \frac{2}{\pi}\arcsin\sqrt{\beta},$$

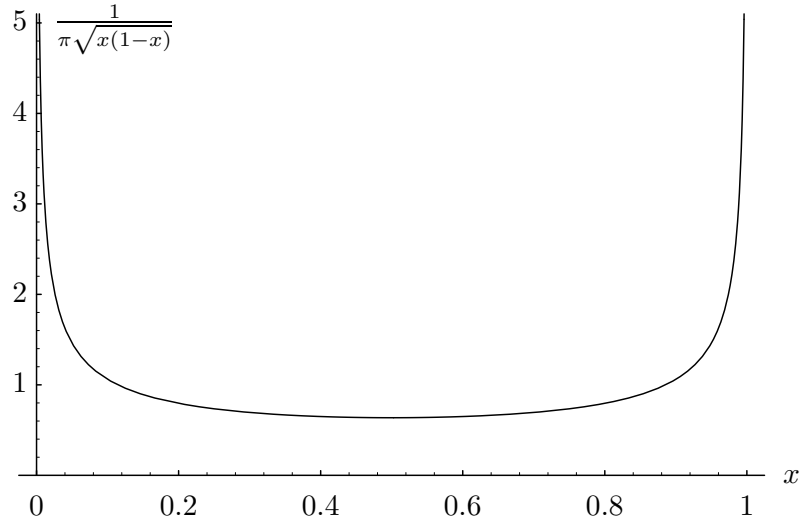
für $\beta > 1/2$

$$\lim_{k\to\infty}\gamma_k(0,\beta) = \lim_{k\to\infty}(\gamma_k(0,1/2) + \gamma_k(1/2,\beta)) = \frac{2}{\pi}\arcsin\sqrt{\beta}.$$

Für $\gamma_k(\alpha,1)$ führt dasselbe Argument zum Ziel. □

Der Beweis des Arcus-Sinus-Gesetzes wurde in einer allgemeineren Form zuerst von *Paul Pierre Lévy* (1886-1971) im Jahre 1939 gegeben.

Die Funktion $\frac{1}{\pi}\frac{1}{\sqrt{x(1-x)}}$ hat das folgende Aussehen:



10 Informationstheorie

Die mathematische Disziplin, die heutzutage Informationstheorie heißt, wurde durch den amerikanischen Ingenieur C. E. Shannon begründet. Shannon nannte seine bahnbrechende Arbeit "A mathematical theory of communication". Erst später hat die Bezeichnung ,,Informationstheorie" Eingang gefunden. Die Bezeichnung kann höhere Erwartungen wecken als die Theorie erfüllen kann. Es ist wichtig, darauf hinzuweisen, daß die Theorie nichts über die Bedeutung, den Inhalt oder den Wert einer Mitteilung (einer ,,Information") aussagt.

a) Optimale Quellenkodierung nach Huffman, Entropie.
Wir betrachten ein Zufallsexperiment mit n möglichen Ausgängen, das wir einfach durch einen endlichen Wahrscheinlichkeitsraum (Ω, p) mit $\Omega = \{\omega_1, \ldots, \omega_n\}$ beschreiben können. Die Wahrscheinlichkeiten $p(\omega_i)$ kürzen wir mit p_i ab, und p sei der Wahrscheinlichkeitsvektor $(p_1, \ldots, p_n)$. Bevor das Zufallsexperiment ausgeführt ist, herrscht Unsicherheit, Ungewißheit über den möglichen Ausgang. Wir möchten eine Zahl $H(p)$, die Entropie des Experimentes, definieren, die ein Maß für die Unbestimmtheit sein soll. Das Wort ,,Entropie" ist aus dem griechischen $\grave{\varepsilon}\nu\tau\rho\acute{\varepsilon}\pi\varepsilon\omega$ (umwenden) abgeleitet. Es wurde 1876 von Clausius in die Thermodynamik eingeführt. Auf die Beziehungen zwischen Informationstheorie und statistischer Mechanik kann hier nicht eingegangen werden. Das am wenigsten unbestimmte Experiment, das wir uns vorstellen können, ist das deterministische, dessen Ausgang von vornherein vorausgesagt werden kann. Ein solches muß die Entropie 0 haben. Im allgemeinen haben wir lediglich $H(p) \geq 0$.

Wir haben die Funktion H eingeführt, ohne zu sagen, wie sie genau definiert ist. Um zu einer vernünftigen Definition zu gelangen, gehen wir vom oben eingeführten, wahrscheinlichkeitstheoretischen Modell mit dem Wahrscheinlichkeitsvektor $p = (p_1, \ldots, p_n)$ aus. Bezeichnet man mit $\log_2$ den Logarithmus zur Basis 2 und verwendet man die Konvention $0 \log_2 0 = 0$, so kann man die Entropie einfach durch

$$H(p) = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{10.1}$$

definieren, wie es in vielen Lehrbüchern geschieht. Die Gründe, die zu dieser Definition der Entropie führen, bleiben dann aber rätselhaft.

Wir wollen versuchen, zu einer Herleitung der Entropie zu kommen, die deren Interpretation als ,,Maß der Unbestimmtheit" Rechnung trägt. Dazu stellen wir uns vor, daß das Experiment ausgeführt wurde und daß eine Person A weiß, wie es ausgegangen ist, während eine Person B nicht über dieses Wissen verfügt. Wieviel ist nun das Wissen von A wert, verglichen mit dem Mangel an Wissen von B? Anders gesagt: Wieviel Anstrengung wird es B kosten, um sein Wissen auf dasselbe Niveau wie das von A zu bringen? Wir können versuchen, diese Anstrengung z. B. durch die Zeit zu messen, die B braucht, um den Ausgang zu erfahren. Das Problem ist, eine vernünftige und wohldefinierte Handlungsweise, der B folgen soll, zu finden. Eine erste Annäherung an eine Definition der Entropie wäre, die Anzahl der Fragen an A zu zählen, die B stellen muß, um den tatsächlichen Ausgang zu finden. Wir denken dabei an Fragen mit möglicher Antwort ,,ja" oder ,,nein". Man darf natürlich nicht fragen: ,,Welches der ω_i ist es¿', sondern etwa: ,,Ist es ω_1 oder ω_5¿'

Die Anzahl der benötigten Fragen hängt natürlich vom Geschick des Fragestellers ab, ferner im allgemeinen vom Ausgang des Zufallsexperimentes. Wir wollen deshalb die mittlere Anzahl der benötigten Fragen betrachten, wenn der Fragesteller optimal fragt. Leider ist auch dies, selbst wenn wir das genau präzisiert haben, noch nicht die übliche Definition von H, d. h. der Ausdruck in (10.1). Wir werden diesen Punkt noch ausführlich diskutieren. Die Größe, zu der wir nach einigen Präzisierungen gelangen werden, nennen wir die *wahre Entropie* und bezeichnen sie mit H_0. Zur klaren Unterscheidung nennen wir H aus (10.1) die *ideelle Entropie*.

Wir fassen die bisherige Diskussion in der nachfolgenden Definition (10.2) zusammen, wir werden sie später durch die Definition (10.7) präzisieren.

(10.2) Definition. Für ein Zufallsexperiment (Ω, p) ist die *wahre Entropie* $H_0(p)$ definiert als der Erwartungswert der Anzahl benötigter Fragen bei optimaler Fragestrategie.

(10.3) Beispiele.

1. Beim Münzwurf, also bei $p = (1/2, 1/2)$, fragt man etwa: „Ist es ω_1?“ Das ist offensichtlich optimal. Somit ist $H_0(1/2, 1/2) = 1$.

2. Auch für $p = (1/2, 1/4, 1/4)$ kann man die optimale Fragestrategie leicht erraten: Man fragt natürlich: „Ist es ω_1?“ Falls die Antwort „nein“ ist, so fragt man nach ω_2. Die mittlere Anzahl der Fragen ist

$$\frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{4} \cdot 2 = \frac{3}{2}.$$

   Fragt man zuerst nach ω_2 und dann, falls nötig, nach ω_1, so beträgt die mittlere Anzahl der benötigten Fragen

$$\frac{1}{4} \cdot 1 + \frac{1}{2} \cdot 2 + \frac{1}{4} \cdot 2 = \frac{7}{4},$$

   was offenbar schlechter ist.

3. Bei $p = (1/4, 1/4, 1/4, 1/4)$ fragt man am besten zunächst: „Ist es ω_1 oder ω_2?“ und dann nach ω_1 bzw. ω_3. Man braucht also bei jedem Versuchsausgang zwei Fragen. Fragt man jedoch der Reihe nach „Ist es ω_1?“, „Ist es ω_2?“ und „Ist es ω_3?“, so benötigt man zwar nur eine Frage, wenn ω_1 der Ausgang ist, im Mittel aber mehr, nämlich

$$\frac{1}{4} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{4} \cdot 3 + \frac{1}{4} \cdot 3 = \frac{9}{4}.$$

Um zu präzisieren, was eine Fragestrategie ist, führen wir den Begriff *Code* ein. Statt „ja“ und „nein“ verwenden wir die Zeichen 1 und 0. Ein *Wort* sei eine endliche Folge von Nullen und Einsen. Ist μ ein Wort, so bezeichnen wir mit $|\mu|$ die Länge von μ, zum Beispiel hat $\mu = 001101$ die Länge $|\mu| = 6$. Die leere Folge nennen wir das leere Wort. Es hat die Länge 0.

Ein Wort μ_1 heißt *Präfix* eines Wortes μ_2, wenn $|\mu_1| < |\mu_2|$ ist und die ersten $|\mu_1|$ Stellen von μ_2 mit μ_1 identisch sind. Zum Beispiel ist 01 ein Präfix von 010010 aber nicht von

000. Das leere Wort ist natürlich Präfix von jedem anderen Wort. Die Idee, die hinter der folgenden Definition eines Codes steht, kann leicht aus der folgenden Betrachtung eingesehen werden: Gegeben sei eine Fragestrategie für (Ω, p). Wir nehmen an, daß wir zum Beispiel fünf Fragen brauchen, um den Ausgang zu bestimmen, der ω_1 sein möge; die Antworten auf die fünf Fragen in diesem Fall seien etwa „ja“, „ja“, „nein“, „ja“, „nein“. Da 0 dem „nein“ und 1 dem „ja“ entspricht, ist es natürlich, ω_1 das Codewort 11010 zuzuordnen.

(10.4) Definition. Ein Code κ für (Ω, p) ist eine injektive Abbildung, die jedem Versuchsausgang ω_i in Ω ein Codewort $\kappa(\omega_i)$ zuordnet. Dabei darf keines der Wörter $\kappa(\omega_i)$ Präfix eines anderen Wortes $\kappa(\omega_j)$ sein.

Wir können einen Code durch eine Tabelle, ein sogenanntes *Codebuch,* darstellen, d. h. durch ein Schema, in dem in einer Spalte die möglichen Versuchsausgänge (Nachrichten) und in einer anderen Spalte die zugehörigen Codewörter stehen. Als Beispiel eines Codebuchs können wir das Morsealphabet nehmen, wo wir 0 statt „Punkt“ und 1 statt „Strich“ lesen. Allerdings ist das Morsealphabet kein Code im Sinne unserer Definition, da die Präfixeigenschaft nicht erfüllt ist: Es ist „Punkt“ das Codewort für „e“ und „Punkt–Strich“ das Codewort für „a“.

Von den fünf Vorschlägen für ein Codebuch in der nachfolgenden Tabelle sind nur κ_3, κ_4 und κ_5 als Codes brauchbar, denn κ_1 ist nicht injektiv und κ_2 hat nicht die Präfixeigenschaft.

(10.5) Beispiel.

	κ_1	κ_2	κ_3	κ_4	κ_5
ω_1	00	0	00	010	1
ω_2	10	01	01	011	01
ω_3	110	001	10	101	001
ω_4	00	000	11	11	000

Es ist nun nicht schwer, den Zusammenhang zwischen Fragestrategien für (Ω, p) und Codes zu erörtern. Wenn wir ein Verfahren haben, um Fragen zu stellen, so konstruieren wir den zugehörigen Code κ folgendermaßen: Die erste Ziffer von $\kappa(\omega_i)$ setzen wir gleich 1 bzw. 0, je nachdem ob die Antwort auf die erste Frage „ja“ bzw. „nein“ ist, falls das Ereignis ω_i eintritt. Wenn man, falls ω_i eintritt, nur eine Frage zu stellen braucht, so haben wir das Codewort $\kappa(\omega_i)$ bereits gefunden. Benötigt man dagegen mehrere Fragen, so setzen wir die zweite Ziffer in $\kappa(\omega_i)$ gleich 1 bzw. 0, je nachdem ob die Antwort auf die zweite Frage „ja“ bzw. „nein“ lautet, falls ω_i eintritt. Auf diese Weise fahren wir fort, bis der ganze Code festgelegt ist.

Wenn uns umgekehrt ein Code gegeben ist, so lautet die erste Frage der zugehörigen Fragestrategie: „Ist die erste Ziffer des Codeworts für das eingetretene Ereignis gleich 1?” Als nächstes die Frage: „Ist die zweite Ziffer des Codewortes des eingetretenen Ereignisses eine 1?”, etc. Da der Code die Präfixeigenschaft hat, ist jederzeit klar, ob man mit den Fragen aufhören kann.

(10.6) Beispiele.

1. Die erste Strategie in Beispiel (10.3 (2)) ergibt den untenstehenden Code κ_1; die zweite führt auf κ_2.

	κ_1	κ_2
ω_1	1	01
ω_2	01	1
ω_3	00	00

2. Für das Beispiel (10.3 (3)) ergeben sich die beiden folgenden Codes:

	κ_1	κ_2
ω_1	11	1
ω_2	10	01
ω_3	01	001
ω_4	00	000

Unsere Codes haben eine zusätzliche angenehme Eigenschaft. Wir stellen uns vor, daß das Experiment mehrfach hintereinander ausgeführt wird und daß wir laufend eine Mitteilung über den Ausgang jedes einzelnen in Codeform erhalten vermöge eines bestimmten Codes κ. Da kein Codewort Präfix eines anderen ist, sind wir nie im Zweifel darüber, wo ein Codewort aufhört und das nächste anfängt. Jede mit Hilfe des Codes gegebene Mitteilung kann daher auf eindeutige Weise decodiert oder entziffert werden. Wenn wir z. B. den Code κ_1 aus Beispiel (10.6 (1)) benutzen und die Folge 11100101100 empfangen, so entspricht dies eindeutig den Versuchsausgängen ω_1, ω_1, ω_1, ω_3, ω_1, ω_2, ω_1, ω_3.

Welcher Code, d. h. welche Fragestrategie, optimal ist, hängt natürlich vom Wahrscheinlichkeitsvektor $p = (p_1, \ldots, p_n)$ ab. Der Erwartungswert der Länge eines Codes κ ist

$$E(|\kappa|) = \sum_{i=1}^{n} p_i |\kappa(\omega_i)|,$$

dies ist gleichzeitig der Erwartungswert der Anzahl der Fragen bei Verwendung der zu κ gehörigen Fragestrategie. Wir können nun unsere Definition (10.2) präzisieren:

(10.7) Definition. Für ein Zufallsexperiment (Ω, p) ist die *wahre Entropie* $H_0(p)$ definiert durch

$$H_0(p) = \min\{E(|\kappa|) : \kappa \text{ ist Code für } (\Omega, p)\}.$$

Man müßte korrekterweise zunächst das Infimum betrachten. Wir werden jedoch sehen, daß stets ein optimaler Code existiert, d. h. ein Code κ_0 mit $E(|\kappa|) \geq E(|\kappa_0|)$ für jeden anderen Code κ für (Ω, p). Natürlich ist die obige Definition von H_0 jetzt noch unhandlich, denn wir haben noch keinen optimalen Code und damit noch keine Möglichkeit, $H_0(p)$ zu berechnen.

Manchmal ist es nützlich, Codes als binäre Bäume zu veranschaulichen. Dabei ist die Knotenmenge des Baumes die Menge aller Codewörter und aller ihrer Präfixe. Wir bezeichnen diese Knotenmenge mit $K(\kappa)$. Wir ziehen eine Verbindung zwischen μ und μa,

$a \in \{0,1\}$, sofern μ und μa zu $K(\kappa)$ gehören. Die Menge dieser Verbindungen bezeichnen wir mit $V(\kappa)$. $(K(\kappa), V(\kappa))$ ist dann ein Graph, der offensichtlich zusammenhängend ist und keine Kreise aufweist. (Ein Kreis in einem Graphen (K, V) ist eine Folge $(e_1, \ldots, e_n)$ von verschiedenen Knoten mit $n \geq 3$, $\{e_i, e_{i+1}\}$, $\{e_n, e_1\} \in V$ für $1 \leq i \leq n-1$.)

Wir ordnen die Elemente von $K(\kappa)$ aufsteigend der Länge nach. Auf der untersten Ebene das leere Wort, sozusagen die ,,Wurzel“ des Baumes, und dann aufsteigend die Wörter der Länge $1, 2, \ldots$ Dabei zeichnen wir eine Verbindung nach rechts oben von μ nach $\mu 1$ und nach links oben von μ nach $\mu 0$, sofern $\mu 1$ beziehungsweise $\mu 0 \in K(\kappa)$ sind.

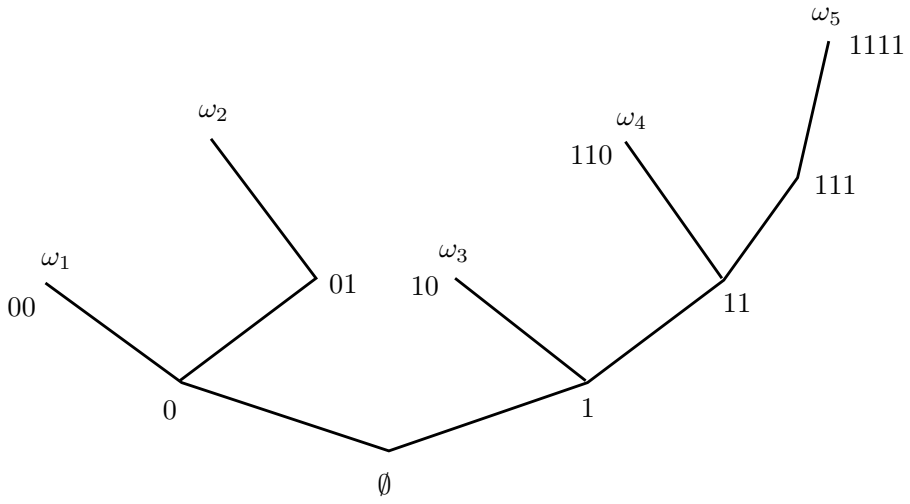
(10.8) Beispiel. $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}$.

$$\begin{aligned} \kappa(\omega_1) &= 00, \qquad \kappa(\omega_2) = 010, \qquad \kappa(\omega_3) = 10, \\ \kappa(\omega_4) &= 110, \qquad \kappa(\omega_5) = 1111. \end{aligned}$$

Dann ist

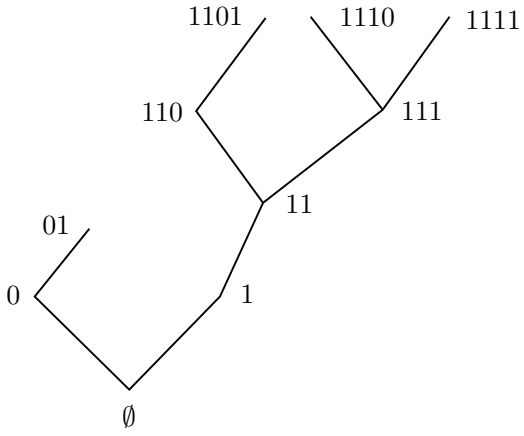
$$K(\kappa) = \{\emptyset, 0, 1, 00, 01, 10, 11, 010, 110, 111, 1111\}.$$

Der Baum:

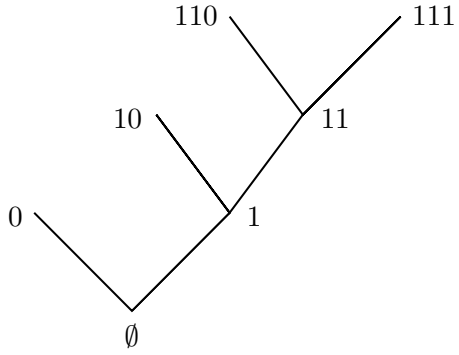


Aus dem Baum eines Codes läßt sich die zugehörige Fragestrategie sofort ablesen. Im obigen Beispiel fragt man zuerst: ,,Ist es ω_3, ω_4 oder ω_5.“ Falls ,,ja“ so befindet man sich im Knoten 1 und falls ,,nein“ im Knoten 0, und dann fährt man entsprechend weiter. Wir nennen einen derartigen binären Baum *vollständig,* falls für jedes Wort $\mu \in K(\kappa)$, das kein Blatt ist, das heißt, das nicht zu den Codewörtern des Codes gehört, sowohl $\mu 0$ wie $\mu 1$ zu $K(\kappa)$ gehören. Es ist evident, daß man sich bei der Suche nach einem optimalen Code auf solche beschränken kann, die zu vollständigen Bäumen führen. Fragestrategien mit unvollständigen Bäumen enthalten überflüssige Fragen. Wir nennen einen Code *vollständig,* falls der zugehörige Baum es ist. Unvollständige Bäume lassen sich durch Weglassen der überflüssigen Knoten zu vollständigen verkürzen und entsprechend lassen sich unvollständige Codes verbessern.

(10.9) Beispiel. Wir betrachten den Code mit den Codewörtern 01, 1101, 1110, 1111. Daraus ergibt sich der Baum



Wir können ihn zu folgendem Baum verkürzen



und erhalten den besseren Code mit den Codewörtern 0, 10, 110, 111.

Ein Verfahren für einen optimalen Code ist von Huffman angegeben worden. Man bezeichnet diesen Code als *Huffman-Code*. Die Konstruktion des Codes erfolgt rekursiv nach der Anzahl n der möglichen Versuchsausgänge. Wir setzen dabei stets $p_i > 0$ für alle $i \in \{1, \ldots, n\}$ voraus, denn gilt $p_i = 0$ für ein i, so lassen wir ω_i aus der Betrachtung weg. Ferner genügt es, nur den Grundraum $\Omega = \{1, 2, \ldots, n\}$ zu betrachten, wodurch die Notation einfacher wird. Für $n = 2$ ist $\kappa(1) = 0$ und $\kappa(2) = 1$ offensichtlich eine optimale Codierung von (p_1, p_2).

Sei also $n > 2$. Wir nehmen an, daß wir den Huffman-Code für alle Wahrscheinlichkeitsvektoren der Länge $n-1$ schon konstruiert haben und geben nun den Code für $(p_1, \ldots, p_n)$ an.

Zunächst bemerkt man, daß die Reihenfolge der p_i für die Codierung keine Rolle spielt, denn wenn $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$ eine Permutation und κ ein Code für $(p_1, \ldots, p_n)$ mit den Codewörtern $\kappa(1), \ldots, \kappa(n)$ ist, so ist $\kappa(\pi(1)), \ldots, \kappa(\pi(n))$ natürlich ein Code für $(p_{\pi(1)}, \ldots, p_{\pi(n)})$ mit derselben mittleren Länge.

Wir können daher voraussetzen, daß $p_1 \geq p_2 \geq \cdots \geq p_n$ gilt. Nun faßt man die beiden kleinsten Wahrscheinlichkeiten zusammen und betrachtet den Wahrscheinlichkeitsvektor $(p_1, p_2, \ldots, p_{n-2}, p_{n-1}+p_n)$ mit $n-1$ Komponenten. Natürlich braucht $p_{n-1}+p_n$ nicht mehr die kleinste Komponente dieses Vektors zu sein. Bezeichnet gemäß Induktionsvoraussetzung $\kappa(1), \ldots, \kappa(n-1)$ den Huffman-Code für diesen Vektor, so ist $\kappa(1), \kappa(2), \ldots, \kappa(n-2), \kappa(n-1)0, \kappa(n-1)1$ der Huffman-Code für $(p_1, \ldots, p_n)$. Es ist offensichtlich, daß der

Huffman-Code stets zu einem vollständigen Baum führt. Das beweist natürlich noch lange nicht, daß er optimal ist.

Bevor wir zeigen, daß der Huffman-Code optimal ist, betrachten wir ein Beispiel:

(10.10) Beispiel. In der untenstehenden Tabelle ist der zu codierende Wahrscheinlichkeitsvektor $(p_1, \ldots, p_8)$ die erste Spalte:

$p_1 = 0,36$	0,36	0,36	0,36	0,36	<u>0,37</u>	<u>0,63</u>	<u>1</u>
$p_2 = 0,21$	0,21	0,21	0,21	<u>0,27</u>	0,36	0,37	
$p_3 = 0,15$	0,15	0,15	<u>0,16</u>	0,21	0,27		
$p_4 = 0,12$	0,12	0,12	0,15	0,16			
$p_5 = 0,07$	0,07	<u>0,09</u>	0,12				
$p_6 = 0,06$	0,06	0,07					
$p_7 = 0,02$	<u>0,03</u>						
$p_8 = 0,01$							

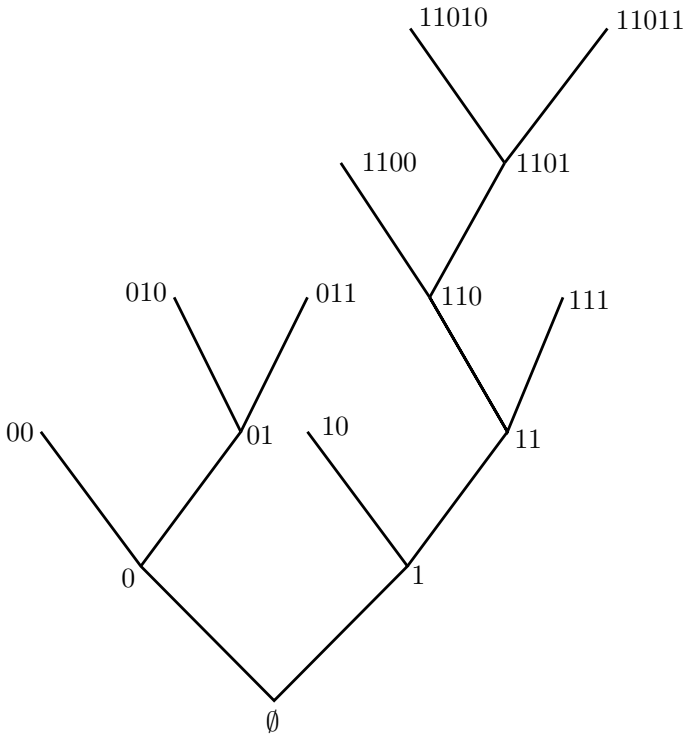
Die Spalten sind die Wahrscheinlichkeitsvektoren. Die erste ist der ursprüngliche, der codiert werden soll. Den nächsten gewinnt man jeweils, indem man die beiden kleinsten Wahrscheinlichkeiten zusammenzählt und gleich richtig einordnet. Diese Summe ist im neuen Wahrscheinlichkeitsvektor jeweils unterstrichen.

Den Huffman-Code gewinnt man rückwärts. Für den Vektor der Länge zwei besteht der zugehörige Code aus den Wörtern 0 und 1. Danach wird jeweils das Codewort, das zur unterstrichenen Wahrscheinlichkeit in der vorangehenden Tabelle gehört, durch Anhängen der Ziffer 0 bzw. 1 aufgespaltet, um die beiden neuen Codewörter für die beiden letzten Wahrscheinlichkeiten zu erhalten. In der folgenden Tabelle sind die aufgespaltenen Codewörter jeweils unterstrichen:

$\kappa(1) = 00$	00	00	00	00	<u>1</u>	<u>0</u>
$\kappa(2) = 10$	10	10	10	<u>01</u>	00	1
$\kappa(3) = 010$	010	010	<u>11</u>	10	01	
$\kappa(4) = 011$	011	011	010	11		
$\kappa(5) = 111$	111	<u>110</u>	011			
$\kappa(6) = 1100$	1100	111				
$\kappa(7) = 11010$	<u>1101</u>					
$\kappa(8) = 11011$						

Sowohl aus der Definition von $E(|\kappa|)$ als auch durch Addition der unterstrichenen Wahrscheinlichkeiten in der ersten Tabelle ergibt sich, daß die mittlere Länge dieses Codes 2,55 ist. Benötigt man also für einen Wahrscheinlichkeitsvektor nur die mittlere Länge des zugehörigen Huffman-Codes, so genügt die erste Tabelle.

Der zugehörige Baum sieht wie folgt aus:



Der Huffman-Code ist offenbar nicht immer eindeutig definiert. Es kann nämlich passieren, daß die Summe der beiden kleinsten Wahrscheinlichkeiten gleich einer der anderen ist, so daß die Einordnung nicht eindeutig ist. Dies ist jedoch ohne Belang, denn offensichtlich haben die entstehenden Huffman-Codes dieselbe mittlere Länge.

(10.11) Satz. Jeder Huffman-Code ist optimal.

Beweis. Der Beweis verläuft mit Induktion nach n, der Länge des Wahrscheinlichkeitsvektors. Der Fall $n = 2$ ist trivial.

Induktionsschluß von $n-1$ auf n:
Wir nehmen an, daß der Satz für Vektoren der Länge $n-1 \geq 2$ gezeigt ist. Sei $(p_1, \ldots, p_n)$ ein beliebiger Wahrscheinlichkeitsvektor der Länge n mit $p_i > 0$ für alle $i \in \{1, \ldots, n\}$. Wir können annehmen, daß $p_1 \geq p_2 \geq \cdots \geq p_n > 0$ gilt, denn dies läßt sich durch Vertauschen stets erreichen.

Sei $\kappa^{(n)}_{\text{Huff}}$ ein Huffman-Code für diesen Vektor. Sei κ ein beliebiger anderer Code mit den Codewörtern $\mu_1, \ldots, \mu_n$. Wir zeigen nun

$$E(|\kappa|) \geq E(|\kappa^{(n)}_{\text{Huff}}|). \tag{10.2}$$

Zunächst ordnen wir die Codewörter von κ nach aufsteigender Länge. Den geordneten Code nennen wir $\kappa' = (\mu'_1, \ldots, \mu'_n)$; für die Codewörter gilt $|\mu'_1| \leq |\mu'_2| \leq \cdots \leq |\mu'_n|$. Die Menge der Codewörter ist dieselbe geblieben. Es ist ziemlich offensichtlich, daß $E(|\kappa|) \geq E(|\kappa'|)$ ist (Nachprüfen!).

Falls $|\mu'_n| > |\mu'_{n-1}|$ ist, so stutzen wir das Wort μ'_n, indem wir von μ'_n die letzten $|\mu'_n| - |\mu'_{n-1}|$ Binärzeichen weglassen. Dieses Wort sei μ''_n. Wegen der Präfix-Eigenschaft unterscheidet sich dieses Wort von $\mu'_1, \ldots, \mu'_{n-1}$. Das neue Wort μ''_n kann aber auch nicht Präfix eines

der anderen Wörter sein, denn seine Länge ist zumindest die der anderen. Also ist $\kappa'' = (\mu'_1, \ldots, \mu'_{n-1}, \mu''_n)$ ein Code.

Gilt $|\mu'_{n-1}| = |\mu'_n|$, so setzen wir $\kappa'' = \kappa'$. In jedem Fall gilt $E(|\kappa'|) \geq E(|\kappa''|)$.

Mindestens zwei Wörter von κ'' haben die Länge $m := |\mu''_n|$. Sei α das aus den ersten $m-1$ Zeichen von μ''_n bestehende Wort. Dann gilt $\mu''_n = \alpha 0$ oder $\mu''_n = \alpha 1$. Wir nehmen das letztere an, der andere Fall geht genau gleich. Wir betrachten nun zwei Fälle:

(i) Eines der anderen Wörter von κ'' der Länge m ist das Wort $\alpha 0$. Falls $\alpha 0$ nicht bereits das zweitletzte Wort ist, so vertauschen wir $\alpha 0$ mit dem zweitletzten Wort. Diesen (eventuell neuen) Code nennen wir κ'''.

(ii) Keines der anderen Wörter der Länge m ist $\alpha 0$. Dann ersetzen wir μ'_{n-1} durch $\alpha 0$ und nennen den neuen Code κ'''. Die Präfixeigenschaft wird dadurch nicht zerstört, denn $\alpha 1$ war ja schon Codewort.

Es gilt offenbar $E(|\kappa''|) = E(|\kappa'''|)$, denn die Längen sind gleichgeblieben. Wir schreiben $\kappa''' = (\nu_1, \ldots, \nu_n)$ mit $\nu_{n-1} = \alpha 0$ und $\nu_n = \alpha 1$. Dann ist $(\nu_1, \ldots, \nu_{n-2}, \alpha)$ ein Code für $(p_1, \ldots, p_{n-2}, p_{n-1} + p_n)$. Um dies einzusehen, müssen wir nur die Präfixeigenschaft nachprüfen. Das Wort α kann aber kein Präfix von $\nu_1, \ldots, \nu_{n-2}$ sein, denn die Längen dieser Codewörter sind kleiner oder gleich $|\alpha| + 1$, und $\alpha 0$, $\alpha 1$ waren verschieden von $\nu_1, \ldots, \nu_{n-2}$.

Nach Induktionsvoraussetzung ist die mittlere Länge des Codes $(\nu_1, \ldots, \nu_{n-2}, \alpha)$ größer oder gleich der mittleren Länge des zugehörigen Huffman-Codes, also

$$\sum_{i=1}^{n-2} p_i |\nu_i| + (p_{n-1} + p_n)|\alpha| \geq E(|\kappa_{\text{Huff}}^{(n-1)}|),$$

wobei $\kappa_{\text{Huff}}^{(n-1)}$ ein Huffman-Code für $(p_1, \ldots, p_{n-2}, p_{n-1} + p_n)$ ist. Nach der rekursiven Konstruktion des Huffman-Codes $\kappa_{\text{Huff}}^{(n)}$ aus $\kappa_{\text{Huff}}^{(n-1)}$ ist

$$E|\kappa_{\text{Huff}}^{(n)}| = E|\kappa_{\text{Huff}}^{(n-1)}| + p_{n-1} + p_n.$$

Somit gilt

$$\begin{aligned} E|\kappa'''| = \sum_{i=1}^{n} p_i |\nu_i| &= \sum_{i=1}^{n-2} p_i |\nu_i| + (p_{n-1} + p_n)|\alpha| + (p_{n-1} + p_n) \\ &\geq E|\kappa_{\text{Huff}}^{(n-1)}| + p_{n-1} + p_n = E|\kappa_{\text{Huff}}^{(n)}|. \end{aligned}$$

Damit ist (10.2) gezeigt. □

Wegen der Optimalität des Huffman-Codes haben wir natürlich auch ein effektives Berechnungsverfahren für $H_0(p)$ gewonnen. Wir wollen nun noch die Beziehung zwischen $H_0(p)$ und dem bereits in (10.1) angegebenen Ausdruck für die ideelle Entropie $H(p)$ diskutieren.

Im allgemeinen stimmen $H_0(p)$ und $H(p)$ nicht überein. Das sieht man schon bei $n = 2$, wo stets $H_0(p) = 1$ ist. Der folgende Satz zeigt, daß die wahre Entropie $H_0(p)$ nur wenig oberhalb der ideellen Entropie $H(p)$ liegen kann. Man beachte, daß wegen $p_i \leq 1$ stets $\log_2 p_i \leq 0$ und somit $H(p) = -\sum_{i=1}^{n} p_i \log_2 p_i \geq 0$ ist.

(10.12) Satz. Für jeden Wahrscheinlichkeitsvektor $p = (p_1, \ldots, p_n)$ gilt

$$H(p) \leq H_0(p) < H(p) + 1.$$

Da ein Versuchsausgang $\omega_i \in \Omega$ mit $p_i = p(\omega_i) = 0$ bei den Definitionen der ideellen und der wahren Entropie in (10.1) bzw. (10.7) keinen Beitrag liefert, können wir für den Beweis des Satzes $p_i > 0$ für alle $i \in \{1, \ldots, n\}$ voraussetzen. Wir benötigen einige einfache Aussagen über die Längen der Codewörter eines Codes.

(10.13) Proposition. (a) *$l_1, \ldots, l_n$ seien die Längen der Codewörter eines Codes. Dann gilt $\sum_{i=1}^{n} 2^{-l_i} \leq 1$ und Gleichheit gilt genau dann, wenn der Code vollständig ist.*

(b) *Seien $l_1, \ldots, l_n \in \mathbb{N}$ mit $\sum_{i=1}^{n} 2^{-l_i} \leq 1$. Dann existiert ein Code mit den Wortlängen $l_1, \ldots, l_n$.*

Beweis. (a) Wir zeigen zunächst mit Induktion nach n, daß für einen vollständigen Code $\sum_{i=1}^{n} 2^{-l_i} = 1$ gilt. Für $n = 2$ ist die Aussage trivial, denn dann muß $l_1 = l_2 = 1$ gelten. Sei $n \geq 3$. O.E.d.A. können wir annehmen, daß $l_1 \leq l_2 \leq \cdots \leq l_n$ gilt. Aus der Vollständigkeit folgt, daß $l_{n-1} = l_n \geq 2$ gilt. Die letzten beiden Codewörter sind dann von der Form $\mu 0$ und $\mu 1$. Ersetzen wir diese beiden Codewörter durch das eine μ, so erhalten wir einen vollständigen Code mit $n-1$ Codewörtern, wobei das letzte die Länge $l_n - 1$ hat. Wenden wir nun die Induktionsvoraussetzung an, so folgt $\sum_{i=1}^{n} 2^{-l_i} = \sum_{i=1}^{n-2} 2^{-l_i} + 2^{-l_n+1} = 1$.

Ein unvollständiger Code läßt sich zu einem vollständigen verkürzen. Damit folgt sofort $\sum_{i=1}^{n} 2^{-l_i} \leq 1$ für jeden Code, wobei das Gleichheitszeichen nur für vollständige gilt.

(b) Wir wenden wieder Induktion nach n an. Für $n = 2$ ist die Sache trivial. Sei $n \geq 3$. Wir können wieder annehmen, daß $l_1 \leq l_2 \leq \cdots \leq l_n$ gilt. Wegen $\sum_{i=1}^{n} 2^{-l_i} \leq 1$ folgt $\sum_{i=1}^{n-1} 2^{-l_i} < 1$. Per Induktionsvoraussetzung existiert ein Code mit Wortlängen $l_1, \ldots, l_{n-1}$, der jedoch nach (a) nicht vollständig ist. Der zugehörige Baum hat also einen Knoten μ, der kein Codewort ist und so, daß entweder $\mu 0$ oder $\mu 1$ keine Knoten sind. Da l_n mindestens so groß wie die anderen sind, ergibt sich, daß wir den Baum mit einem neuen Blatt ergänzen können, das μ als Präfix hat und das die Länge l_n hat. □

Wir benötigen noch das folgende elementare analytische Ergebnis:

(10.14) Lemma. *Für alle $i \in \{1, \ldots, n\}$ seien s_i und r_i positive reelle Zahlen mit $\sum_{i=1}^{n} s_i \geq \sum_{i=1}^{n} r_i$. Dann gilt $\sum_{i=1}^{n} s_i \log_2(s_i/r_i) \geq 0$.*

Beweis. Es gilt $\log x \leq x - 1$ für alle $x > 0$, wobei log den Logarithmus zur Basis e

bezeichnet. Somit folgt

$$\sum_{i=1}^{n} s_i \log \frac{r_i}{s_i} \leq \sum_{i=1}^{n} s_i \left(\frac{r_i}{s_i} - 1 \right) = \sum_{i=1}^{n} r_i - \sum_{i=1}^{n} s_i \leq 0,$$

also $\sum_{i=1}^{n} s_i \log(s_i/r_i) \geq 0$. Die $\log_2$-Funktion ist jedoch proportional zur log-Funktion. Damit ist (10.14) gezeigt. □

Beweis von $H(p) \leq H_0(p)$:
Es seien $l_1, \ldots, l_n$ die Wortlängen des Huffman-Codes κ für $p = (p_1, \ldots, p_n)$. Da dieser vollständig ist, folgt nach (10.13 (a)) $\sum_{i=1}^{n} 2^{-l_i} = 1 = \sum_{i=1}^{n} p_i$. Nach (10.14) ist dann $\sum_{i=1}^{n} p_i \log_2(p_i/2^{-l_i}) = \sum_{i=1}^{n} p_i \log_2 p_i + \sum_{i=1}^{n} l_i p_i \geq 0$. Das bedeutet, daß $E(|\kappa|) \geq H(p)$ gilt. □

Beweis von $H_0(p) < H(p) + 1$:
Zu vorgegebenen p_i können wir natürliche Zahlen l_i wählen mit $-\log_2 p_i \leq l_i < -\log_2 p_i + 1$. Aus der ersten Ungleichung folgt $\sum_{i=1}^{n} 2^{-l_i} \leq \sum_{i=1}^{n} p_i = 1$. Nach (10.13 (b)) existiert ein Code mit diesen l_i als Wortlängen. Wegen der zweiten Ungleichung für die l_i folgt $\sum_{i=1}^{n} p_i l_i < -\sum_{i=1}^{n} p_i \log_2 p_i + 1$. Der optimale Code hat jedoch höchstens die mittlere Länge $\sum_{i=1}^{n} p_i l_i$. □

Bemerkung. Der letzte Beweisteil von (10.12) deutet darauf hin, daß bei einem optimalen Code die Länge des i-ten Codewortes ungefähr gleich $-\log_2 p_i$ sein wird.

Es ist klar, daß die wahre Entropie H_0 als Maß für die Ungewißheit in einigen Situation etwas unbefriedigend ist. Am deutlichsten sieht man das bei einem Experiment mit zwei möglichen Ausgängen, die mit den Wahrscheinlichkeiten p_1 und $p_2 = 1 - p_1$ auftreten, denn dann gilt $H_0(p_1, 1 - p_1) = 1$ für jedes $p_1 \in (0, 1)$.

Wir können noch eine andere Beziehung zwischen H_0 und H herleiten, indem wir unabhängige Repetitionen des Zufallsexperimentes (Ω, p) betrachten. Nach Kapitel 2 ist der geeignete W.-Raum für eine k-fache Repetition der Produktraum (Ω^k, p^k), mit $p^k(\omega_1, \ldots, \omega_k) = p(\omega_1) \ldots p(\omega_k)$ für alle $\omega_1, \ldots, \omega_k \in \Omega$.

Es ist klar, wie aus einer Fragestrategie (d. h. einem Code) für p eine für p^k gewonnen werden kann: Man fragt zunächst nach dem Ausgang des ersten Experimentes, dann nach dem zweiten etc. bis nach dem k-ten. Die gesamte Anzahl der benötigten Fragen ergibt sich als Summe der benötigten Fragen für die einzelnen Experimente; somit summieren sich auch die Erwartungswerte. Ist κ ein optimaler Code für p, so ist der optimale Code für p^k natürlich mindestens so gut wie dieser „Repetitionscode“, der die mittlere Länge $kE(|\kappa|)$ hat. Somit folgt:

$$H_0(p^k) \leq kH_0(p).$$

Es zeigt sich jedoch, daß die oben beschriebene k-fache Repetition der optimalen Fragestrategie für p im allgemeinen nicht die optimale Fragestrategie für p^k ist.

(10.15) Beispiel. Sei $(p_1, p_2) = (3/4,\ 1/4)$. Dann ist $H_0(p) = 1$. Der Huffman-Algorithmus für p^2 wird durch das folgende Schema gegeben:

$$\begin{array}{llll} 9 & 9 & 9 & \underline{16} \\ 3 & \underline{4} & \underline{7} & \\ 3 & 3 & & \\ 1 & & & \end{array}$$

wobei die einzelnen Zahlen mit $1/16$ zu multiplizieren sind. Die mittlere Länge des zugehörigen Huffman-Codes ist also $27/16$, was deutlich kleiner als 2 ist.

(10.16) Satz. Sei $p = (p_1, \ldots, p_n)$ ein Wahrscheinlichkeitsvektor. Dann gilt

$$\lim_{k\to\infty} \frac{1}{k} H_0(p^k) = H(p).$$

Beweis. Einsetzen in die Definition (10.1) ergibt $H(p^k) = kH(p)$. Aus Satz (10.12) folgt dann $H(p) \le H_0(p^k)/k < H(p) + 1/k$, woraus sich Satz (10.16) ergibt. □

Die ideelle Entropie $H(p)$ ist also die pro Versuch benötigte mittlere Anzahl von Fragen bei vielen unabhängigen Repetitionen des Versuchs. In der Regel liegt $H_0(p^k)/k$ bereits für kleine k sehr nahe an der ideellen Entropie $H(p)$.

Die ideelle Entropie H hat einige interessante Eigenschaften. Zu vorgegebenem $n \in \mathbb{N}$ ist sie definiert auf der Menge von Wahrscheinlichkeitsvektoren

$$\triangle_n = \left\{ (p_1, \ldots, p_n) \in \mathbb{R}^n \;\middle|\; p_1 \ge 0,\ \ldots,\ p_n \ge 0,\ \sum_{j=1}^{n} p_j = 1 \right\}.$$

Als Durchschnitt von n Halbräumen und einer Hyperebene ist $\triangle_n$ eine konvexe Teilmenge des $\mathbb{R}^n$. Benutzt man die Konvention $0 \log_2 0 = 0$, so wird durch (10.1) eine stetige Funktion $H : \triangle_n \to [0, \infty)$ definiert. Der Beweis des folgenden Satzes sei dem Leser überlassen, für Teil (b) ist Lemma (10.14) hilfreich:

(10.17) Satz.

a) Die Funktion H ist streng konkav auf $\triangle_n$, das heißt für $\lambda \in (0, 1)$ und $p, p' \in \triangle_n$ mit $p \ne p'$, gilt $H(\lambda p + (1-\lambda)p') > \lambda H(p) + (1-\lambda)H(p')$.

b) Für alle $p \in \triangle_n$ gilt $H(p) \le H(1/n, \ldots, 1/n)$.