§1 Diskrete Wahrscheinlichkeitsräume

Auch wenn dem Eindruck, die Stochastik diene hauptsächlich der Lösung von Fragen im Zusammenhang mit Glücksspielen, in dieser Vorlesung entschieden entgegengewirkt werden soll, sei zu Beginn ein Spiel erwähnt, welches 1991 in der Presse für ein wenig Aufsehen sorgte:

Das 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.

Ziel dieses Kapitels ist es, die mathematischen Begriffe bereitzustellen, die es uns ermöglichen, unter anderem dieses Spiel in ein geeignetes Modell zu übersetzen. In diesem Kapitel wird der Begriff „Wahrscheinlichkeit" mathematisch präzisiert. Zufallsereignissen werden dabei Wahrscheinlichkeiten zugeordnet. Man spricht von der „Wahrscheinlichkeit eines Ereignisses".

Zufälligen Phänomenen begegnet man in *Experimenten*, die unter gleichbleibenden, wohldefinierten Bedingungen viele Male wiederholt werden (können), wobei jede Ausführung des Experiments ein wohlbestimmtes Resultat liefern möge. Kennzeichnend für diese Experimente ist, daß die Versuchsausgänge in einer irregulären, nicht vorhersehbaren Weise von Versuch zu Versuch variieren, auch wenn die Bedingungen, unter denen die Versuche stattfinden, so gleichbleibend wie nur möglich gehalten werden. Man nennt solche Experimente *zufällige Experimente*.

Zunächst soll der Begriff „Ereignis" präzisiert werden. Am besten zerlegt man die Ereignisse gewissermaßen in Atome, in die sogenannten Elementarereignisse: die kleinsten Ereignisse, die in einer bestimmten Situation interessant oder von Bedeutung sind. Die Festlegung, was in einer Situation die Elementarereignisse sind, ist eher willkürlich.

Formal sind die *Elementarereignisse* (*sample point*) einfach die Elemente einer (zunächst) endlichen oder abzählbaren Menge, die meist mit Ω bezeichnet wird. Ω ist dann die Menge der möglichen Versuchsausgänge eines zufälligen Experiments und heißt auch *Ergebnisraum* oder *Stichprobenraum* (*sample space*).

Die *Wahrscheinlichkeiten* (*probability*) der Elementarereignisse $\omega \in \Omega$ sind Zahlen $p(\omega)$ zwischen 0 und 1, die sich zu 1 aufsummieren.

(1.1) Definition. Ein *diskreter Wahrscheinlichkeitsraum* (Ω, p) (kurz W.-Raum, *discrete probability space*) besteht aus einer endlichen oder abzählbar unendlichen Menge Ω und einer Abbildung $p\colon \Omega \to [0,1]$, für die $\sum_{\omega\in\Omega} p(\omega) = 1$ gilt.

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.

Soweit ist mathematisch alles sehr simpel. Mit Wahrscheinlichkeitsräumen sollen jedoch konkrete Zufallssituationen modelliert werden. In der Regel gibt es dann mehr als eine vernünftige Wahl für einen W.-Raum. Man wählt 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:

(1.2) 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. 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 $\frac{1}{n!}$.

Natürlich sollen nicht nur den Elementarereignissen Wahrscheinlichkeiten zugeordnet werden, sondern auch zusammengesetzten Ereignissen, etwa in Beispiel (2) oben dem Ereignis, daß die Augenzahlen der beiden Würfel gleich sind. Ereignisse sind einfach Zusammensetzungen von Elementarereignissen. In mathematischer Formulierung:

(1.3) Definition. (Ω, p) sei ein W.-Raum. *Ereignisse* (*events*) sind Teilmengen von Ω. Für ein Ereignis $A \subset \Omega$ ist die *Wahrscheinlichkeit von A* definiert durch $P(A) = \sum_{\omega \in A} p(\omega)$. Die leere Menge $\emptyset$ ist das sogenannte *unmögliche Ereignis*, es hat Wahrscheinlichkeit $P(\emptyset) = 0$. Die Grundmenge Ω ist das *sichere Ereignis*.

Bemerkung. Es hat sich eingebürgert, Ereignisse mit großen lateinischen Buchstaben vom Anfang des Alphabets zu bezeichnen: $A, B, C, \ldots$. Die Wahrscheinlichkeit wird meist mit einem großen P (probability) bezeichnet.

Es mag etwas verwirren, daß Ereignisse Teilmengen sind. Am anschaulichsten ist vielleicht die folgende Vorstellung: Das zufällige Geschehen besteht in der zufälligen Auswahl eines Elementarereignisses. Eine Teilmenge A von Ω entspricht dann dem Ereignis, daß dieses zufällig gewählte Elementarereignis in A liegt.

Mengenoperationen entsprechen aussagenlogischen Operationen gemäß der folgenden Übersetzungstabelle:

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 schließen sich aus	$A \cap B = \emptyset$
A impliziert B	$A \subset B$

Bemerkung. Für jedes Elementarereignis ω ist die Menge $\{\omega\}$ offenbar ein Ereignis, das sich formal mathematisch von ω unterscheidet. Elementarereignisse sind formal nach unserer Definition keine Ereignisse. Sowohl $p(\omega)$ als auch $P(\{\omega\})$ bezeichnen die Wahrscheinlichkeit von $\omega \in \Omega$.

Wahrscheinlichkeiten genügen einigen einfachen Regeln, die im nächsten Satz aufgelistet sind.

(1.4) 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 $P\left(\bigcup_{i\in\mathbb{N}} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$ (*abzählbar additiv, countable additive*).
(4) In (3) ohne die Voraussetzung, daß die A_i paarweise disjunkt sind, gilt noch $P\left(\bigcup_{i\in\mathbb{N}} A_i\right) \leq \sum_{i=1}^{\infty} P(A_i)$ (*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) folgen sofort aus der Definition.

(3), (4): Jedes $\omega \in \bigcup_{i=1}^{\infty} A_i$ gehört zu mindestens einem der A_i und zu genau einem, wenn die A_i paarweise disjunkt sind. Demzufolge gilt

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

wenn die A_i paarweise disjunkt sind. Im Fall (4) ist das mittlere Gleichheitszeichen durch „$\leq$“ zu ersetzen, denn die $p(\omega)$'s werden in der Summe auf der rechten Seite eventuell mehrfach gezählt, nämlich einmal für jede Menge A_i, die das entsprechende ω enthält.

(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). □

Es bezeichne $\mathcal{P}(\Omega)$ die Potenzmenge einer endlichen oder abzählbar unendlichen Menge Ω. Dann ist P eine Abbildung von $\mathcal{P}(\Omega)$ nach $[0,1]$, die gemäß (1.4) den folgenden *Kolmogoroffschen Axiomen* genügt.

(K1) $P(\Omega) = 1$.

(K2) Es sei I eine höchstens abzählbare Indexmenge und $A_i \in P(\Omega)$ für jedes $i \in I$. Sind die A_i paarweise disjunkt, so gilt $P(\bigcup_{i \in I} A_i) = \sum_{i \in I} P(A_i)$.

A.N. Kolmogoroff (1903-1987) gab 1933 die heute übliche Definition von W.-Räumen via der Axiome (K1) und (K2) an. Eine Abbildung $P\colon \mathcal{P}(\Omega) \to [0,1]$, die (K1) und (K2) erfüllt, legt einen W.-Raum im Sinne von (1.1) eindeutig fest, wie die folgende Überlegung zeigt: Wegen $P(\Omega) = P(\Omega \cup \emptyset) = P(\Omega) + P(\emptyset)$ (nach (K2)) folgt $P(\emptyset) = 0$. Für $\omega \in \Omega$ sei $p(\omega)$ definiert durch $P(\{\omega\})$. Wegen (K2) gilt dann $P(A) = \sum_{\omega \in A} p(\omega)$ für alle $A \subset \Omega$. Mit $A = \Omega$ folgt insbesondere $\sum_{\omega \in \Omega} p(\omega) = 1$. Also ist (Ω, p) ein W.-Raum entsprechend der Definition (1.1), und $P(A)$ berechnet sich gemäß der Definition (1.3). Unsere Definition (1.1) ist also gleichbedeutend damit, daß $P\colon \mathcal{P}(\Omega) \to [0,1]$ mit (K1) und (K2) gegeben ist.

Die Bedeutung des Kolmogoroffschen Aufbaus liegt darin, daß er sich auf überabzählbare Räume verallgemeinern läßt.

(1.5) Beispiele.

(1) In Beispiel (1.2 (2)) wird man jedem Elementarereignis die Wahrscheinlichkeit $1/36$ zuordnen. Für jedes Ereignis A ist $P(A) = |A|/36$, wobei $|A|$ die Anzahl der Elemente in A ist. Sei z.B. $A = \{(1,1), (2,2), \dots, (6,6)\}$ 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 $\Omega = \{\,(i,j) \in \{1,2,\dots,2n\}^2 : i \neq j\}$ als Menge der Elementarereignisse. Hierbei ist $(i,j) \in \Omega$ das Elementarereignis, daß sich der erste

Joker am Platz i und der zweite am Platz j befindet. Nach guter Mischung hat jedes dieser Elementarereignisse 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. 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 Elementarereignis 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) *Würfelproblem von de Méré:* Man wirft einen idealen Würfel (alle sechs Seiten sind gleichwahrscheinlich) vier mal und fragt nach der Wahrscheinlichkeit dafür, daß dabei mindestens einmal eine Sechs auftritt. Wir verzichten hier leichtsinnigerweise auf eine genaue Modellierung des zugrundeliegenden Wahrscheinlichkeitsraums. Die Anzahl der möglichen Ergebnisse bei vier Würfen ist 6^4. In 5^4 Fällen kommt keine sechs vor. Also ist die gesuchte Wahrscheinlichkeit $(6^4 - 5^4)/6^4 \geq 0.5$. Wenn man mit zwei Würfeln 24 Würfe durchführt, so ist die Wahrscheinlichkeit dafür, einen doppelten Sechser (mindestens einmal) zu erhalten, aber kleiner als 0.5. Sie berechnet sich analog zu $1 - (35/36)^{24}$. Dies ist überraschend, wenn man bedenkt, daß die Chance, eine Sechs zu erhalten, sechsmal so groß ist wie die Chance, einen doppelten Sechser zu werfen, und 24 ja gerade sechsmal vier ist. Dieses von dem französischen Glücksspieler *A.G. Chevalier de Méré* (1610-1685) gestellte Problem wurde in einem Briefwechsel zwischen *Blaise Pascal* (1623-1662) und *Pierre de Fermat* (1601-1665) am 29.7.1654 diskutiert (siehe hierzu auch der historische Anhang).

(5) *Auto-Ziege Problem:* 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.

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: $\Omega = \{2,3,4,\dots,12\}$. Es ist offensichtlich, daß diese Elementarereignisse nicht gleichwertig sind. Deshalb nimmt man besser die Elementarereignisse aus (1.2 (2)).

In vielen Fällen wäre die Festlegung, daß alle Elementarereignisse gleich wahrscheinlich sind, aber ganz unsinnig.

Als Beispiel betrachten wir das Problem festzulegen, wie groß die Wahrscheinlichkeit ist, mit der etwa ein produziertes Werkstück defekt ist. In Fällen, wo man auf lange Produktionsreihen zurückgreifen kann, setzt man die Wahrscheinlichkeit als die relative Häufigkeit des Defekts an. Eine gewisse theoretische Begründung für diesen Ansatz gibt das Gesetz der großen Zahlen (siehe Kapitel 3). Sind etwa bei der Produktion von 10 000 Werkstücken 200 defekt gewesen, so wird man die Wahrscheinlichkeit als 0,02 annehmen. Dabei handelt es sich nicht um eine „Naturkonstante", sondern lediglich um eine Arbeitshypothese, die gegebenenfalls wieder revidiert werden muß. Das Vertrauen, das man zu einem über relative Häufigkeiten ermittelten Wert für eine Wahrscheinlichkeit hat, hängt natürlich auch von der Anzahl der Versuche ab. Es ist z.B. klar, daß 200 Defekte auf 10 000 aussagekräftiger ist, als 2 auf 100. Eine genauere Diskussion derartiger Probleme gehört in die Statistik.

Nun ein Beispiel mit einem unendlichen W.-Raum:

(1.6) 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 (1.5 (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 Definition (1.1): $\sum_{i\in\mathbb{N}} p(i) = 1$. 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 (1.4 (7)) sieht wie folgt aus: $A_1, \dots, A_n$ seien n Ereignisse. $A_1 \cup \dots \cup A_n$ ist das Ereignis, daß mindestens eines der A_i eintritt.

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

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

Beweis. Induktion nach n: Für $n = 2$ ist dies (1.4 (7)).
Induktionsschluß:

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

nach (1.4 (7))

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

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

Exkurs zu Abzählmethoden

Zur Berechnung der Wahrscheinlichkeiten in Laplace-Experimenten sind die folgenden kombinatorischen Ergebnisse von Nutzen. In einer Urne seinen n Kugeln mit $1, 2, \dots, n$ 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 $(\omega_1, \dots, \omega_k)$ 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 $\{\omega_1, \dots, \omega_k\}$.

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

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

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

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

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

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

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:

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

§2 Bedingte Wahrscheinlichkeiten, Unabhängigkeit

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

Sei Ω ist die Menge der Einwohner Bochums. Ein Reporter des WDR befragt einen rein zufällig herausgegriffenen Bochumer. Wir nehmen an, daß jeder Einwohner die gleiche Chance hat, befragt zu werden. Ist N die Anzahl der Einwohner, so ist die Wahrscheinlichkeit für jedes Elementarereignis $1/N$. Der Reporter möchte nach der Meinung zur Einführung von Studiengebühren fragen. Natürlich ist es keineswegs sicher, daß der Befragte einem Reporter des WDR eine Antwort gibt. Sei B das Ereignis, daß eine Antwort gegeben wird. In unserem Modell ist B einfach die Menge der Bochumer, die eine Antwort geben werden, und es gilt $P(B) = |B|/N$. Sei A das Ereignis, daß der Angesprochene die Einführung insgeheim befürwortet. Es gilt dann $P(A) = |A|/N$. Natürlich ist es denkbar, daß die Meinung derjenigen, die eine Antwort geben, im Schnitt eine andere als die der Gesamtbevölkerung ist. Der relative Anteil der Antwortgeber, die die Studiengebühren befürworten, ist $|A \cap B|/|B| = P(A \cap B)/P(B)$. Man bezeichnet dies als bedingte Wahrscheinlichkeit von A gegeben B.

Ganz allgemein können wir wie folgt definieren:

(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:

(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

$$P\Big(\bigcup_{i=1}^{\infty} A_i \Big| B\Big) = \sum_{i=1}^{\infty} P(A_i|B).$$

(4) $P(A^c|B) = 1 - P(A|B)$.

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

(3)

$$P\Big(\bigcup_{i=1}^{\infty} A_i \Big| B\Big) = \frac{1}{P(B)} P\Big(\Big(\bigcup_{i=1}^{\infty} A_i\Big) \cap B\Big) = \frac{1}{P(B)} P\Big(\bigcup_{i=1}^{\infty} (A_i \cap B)\Big)$$
$$= \sum_{i=1}^{\infty} \frac{P(A_i \cap B)}{P(B)} = \sum_{i=1}^{\infty} P(A_i|B).$$

(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. 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 nummerieren 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 $P(A|B) = (1/6)/(1/2) = 1/3$.

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. Ein Beispiel dazu:

(2.5) Beispiel. Eine Leitung überträgt die zwei Signale „0“ und „1“. Dabei können Übertragungsfehler auftreten, wobei die Wahrscheinlichkeit dafür davon abhängt, welches Signal gesendet wird. Unser mathematisches Modell für die Übertragung eines Zeichens ist ein W.-Raum Ω mit den vier Elementen $(0,0)$, $(0,1)$, $(1,0)$, $(1,1)$, wobei an der ersten Stelle des Paares das gesendete und an der zweiten Stelle das empfangene Zeichen steht. $S_i := \{(i,0),(i,1)\}$ ist das Ereignis, daß i gesendet wird, und $E_i := \{(0,i),(1,i)\}$, daß i empfangen wird. $F := \{(0,1),(1,0)\}$ ist das Ereignis, daß ein Übertragungsfehler auftritt. Oft kennt man die Wahrscheinlichkeit für einen Übertragungsfehler in Abhängigkeit von den gesendeten Zeichen (d.h. unter der entsprechenden Bedingung). Sei $f_i = P(F|S_i)$, also

$$f_0 = P(\{(0,1),(1,0)\}|S_0) = P(\{(0,1)\}|S_0)$$

und

$$f_1 = P(\{(0,1),(1,0)\}|S_1) = P(\{(1,0)\}|S_1).$$

Die Angabe dieser Größen statt der totalen (d.h. unbedingten) Fehlerwahrscheinlichkeit ist deshalb angebracht, weil die f_i im allgemeinen nur vom Übertragungssystem und nicht von der relativen Häufigkeit der Nullen und Einsen in der gesendeten Nachricht, d.h. von $P(S_i)$ abhängen. Es ist einleuchtend, daß die totale Fehlerwahrscheinlichkeit sich aus den f_i und $P(S_i)$ mittels $P(F) = f_0 P(S_0) + f_1 P(S_1)$ berechnen läßt. 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:

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

$\square$

Als weitere Anwendung von (2.6) betrachten wir im Beispiel (2.5) das in der Praxis wichtige Problem, die bedingte Wahrscheinlichkeit für eine richtige Übertragung, gegeben das empfangene Zeichen, etwa $P(S_1|E_1)$ zu berechnen. Das läßt sich zunächst mittels $P(S_1|E_1) = P(S_1 \cap E_1)/P(E_1)$ umschreiben. Per Definition gilt

$$P(S_1 \cap E_1) = P(E_1|S_1)P(S_1) = (1 - f_1)P(S_1).$$

Nach (2.6) gilt

$$P(E_1) = P(E_1|S_1)P(S_1) + P(E_1|S_0)P(S_0) = (1 - f_1)P(S_1) + f_0 P(S_0),$$

also

$$P(S_1|E_1) = \frac{(1 - f_1)P(S_1)}{(1 - f_1)P(S_1) + f_0 P(S_0)}.$$

(2.7) Bemerkung. Das obige Beispiel ist ein Spezialfall der sogenannten *Bayes-Formel:* 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). $\square$

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 $A_1, \dots, A_n$ 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 $A_1, \dots, A_n$ sind genau dann unabhängig, wenn für alle $(k_1, \dots, k_n) \in \{1, c\}^n$*

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

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 $A_1, \dots, A_{n+1}$ seien unabhängig. Wir beweisen die obige Gleichung (für $n + 1$) mit Induktion nach der Anzahl m der Komplementzeichen in $(k_1, \dots, k_{n+1})$. Für $m = 0$ folgt sie aus der Unabhängigkeit. Induktionsschluß $m \to m + 1$ für $0 \leq m < n + 1$: Es seien $m + 1 \geq 1$ Komplementzeichen in $(k_1, \dots, k_{n+1})$. 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 $(k_1, \dots, k_n) \in \{1, c\}^n$. Wir zeigen die Unabhängigkeit von $A_1, \dots, A_n$.

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

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

□

Die Notationen mögen etwas verwirren. Man schreibe die Argumente für $n = 2$ und $n = 3$ aus; dann wird der Beweisgang klar. 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. □

Unabhängigkeit hängt eng mit den sogenannten *Produkträumen* zusammen. Es 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 $\sum_{\omega\in\Omega} p(\omega) = 1$.

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

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

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

Beweis. Es gilt $A_i^{(i)c} = \{\omega \in \Omega \colon \omega_i \in A_i^c\} = A_i^{c(i)}$. 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 $A_i \subset \Omega_i$, $1 \leq i \leq n$, gilt. Die linke Seite dieser Gleichung ist gleich

$$\sum_{\omega\in A_1^{(1)}\cap\cdots\cap A_n^{(n)}} p(\omega) = \sum_{\omega_1\in A_1} \cdots \sum_{\omega_n\in A_n} p_1(\omega_1) \cdots p_n(\omega_n)$$

$$= \prod_{j=1}^n \sum_{\omega_j\in A_j} p_j(\omega_j) = \prod_{j=1}^n \sum_{\omega\in A_j^{(j)}} p(\omega) = \prod_{j=1}^n P(A_j^{(j)}).$$

□

Der Produktraum liefert somit ein Modell für eine unabhängige Hintereinanderreihung von n einzelnen Zufallsexperimenten. Offenbar ist unser Modell für einen n-fachen Münzwurf das n-fache Produkt des W.-Raumes für einen Münzwurf. Wir können das gleich etwas verallgemeinern: Zunächst betrachten wir ein Zufallsexperiment mit zwei möglichen Ausgängen, die wir mit E (für „Erfolg") und M (für „Mißerfolg") bezeichnen. Man denke etwa an ein Spiel, das darin besteht, eine Münze zu werfen, und bei dem der eine Spieler eine Einheit gewinnt, wenn „Kopf" fällt. Wir wollen nicht voraussetzen, daß E und M gleich wahrscheinlich sind. Der W.-Raum ist also die zweielementige Menge $\{E, M\}$ mit den entsprechenden Wahrscheinlichkeiten.

Der n-fache Produktraum, das Modell für die unabhängige, n-malige Wiederholung des Spiels, ist also der W.-Raum $\Omega = \{E, M\}^n$, 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.

(2.15) 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 große Klasse von Beispielen nennt man Urnenmodelle:

(2.16) Beispiel.

(1) Ziehung 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 scheint 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.

Man kann sich das auch wie folgt überlegen: Wir denken uns die einzelnen Kugeln wieder von 1 bis $r + s$ durchnumeriert; 1 bis r sind rot, $r + 1$ bis $r + s$ schwarz. In der Beschreibung des W.-Raums unterscheiden wir nun zwischen den einzelnen Kugeln, d.h. $\Omega = \{1, \ldots, r + s\}^n$. Die Elementarereignisse sind also die Folgen $\omega =$

$(\omega_1, \ldots, \omega_n)$ mit $\omega_k \in \{1, \ldots, r+s\}$. Unter idealen Bedingungen sind diese Elementarereignisse alle gleich wahrscheinlich, haben also die Wahrscheinlichkeit $(r+s)^{-n}$. Das Ereignis einer speziellen Rot-Schwarz-Folge ist hier kein Elementarereignis; man kann die Anzahl der Elementarereignisse darin jedoch leicht abzählen: Eine spezielle Rot-Schwarz-Folge mit k-mal Rot und $(n-k)$-mal Schwarz wird durch $r^k s^{n-k}$ Elementarereignisse repräsentiert, hat also die Wahrscheinlichkeit $\left(\frac{r}{r+s}\right)^k \left(\frac{s}{r+s}\right)^{n-k}$. Die Wahrscheinlichkeit des Ereignisses A_k, genau k-mal Rot zu ziehen, ist somit

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

(2) Ziehung ohne Zurücklegen (sampling without replacement)
Wir betrachten dieselbe Situation 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 distributuion).*

In unserem W.-Raum können wir jedoch das Ereignis, daß die erste Kugel rot ist, nicht betrachten, denn wir unterscheiden die Reihenfolge der Ziehungen nicht. Um dieses Ereignis zu untersuchen, brauchen wir einen anderen, größeren Wahrscheinlichkeitsraum. Wir betrachten dazu analog wie beim Modell mit Zurücklegen die Menge Ω' der Folgen $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$ mit $1 \leq \omega_i \leq r+s$ aber mit der Einschränkung $\omega_i \neq \omega_j$ für $i \neq j$. Dann bedeutet $1 \leq \omega_i \leq r$, daß die i-te Kugel rot ist, $r+1 \leq \omega_i \leq r+s$, daß sie schwarz ist. Ω' enthält offenbar $(r+s)(r+s-1)\cdots(r+s-n+1)$ Elemente. Betrachtet man diese Elementarereignisse als gleich wahrscheinlich, so hat unser obiges Ereignis A_k (entsprechend als Teilmenge von Ω' formuliert) dieselbe Wahrscheinlichkeit wie oben (nachprüfen!). Im Gegensatz zu der Situation in Ω können wir nun jedoch die einzelnen Ziehungen unterscheiden. Sei R_i das Ereignis, daß die i-te Kugel rot ist. Jedes der R_i enthält gleich viele Elementarereignisse, nämlich $r(r+s-1)(r+s-2)\cdots(r+s-n+1)$.

Somit ist $P(R_i) = r/(r+s)$ der gleiche Wert wie beim Modell mit Zurücklegen. Dennoch sind die Wahrscheinlichkeiten für A_k in beiden Modellen verschieden. Dies liegt daran, daß hier $R_1, \ldots, R_n$ abhängig sind: Das Ereignis $R_1 \cap R_2$ enthält $r \times (r-1)(r+s-2)\cdots(r+s-n+1)$ Elementarereignisse und somit ist

$$P(R_1 \cap R_2) = \frac{r(r-1)}{(r+s)(r+s-1)} \neq P(R_1)P(R_2),$$

der Unterschied ist aber klein, sofern r und s groß sind. Dies ist plausibel, denn wenn die Gesamtzahl $r+s$ der Kugeln sehr groß ist, so beeinflussen sich die einzelnen Ziehungen wenig. $P(A_k)$ 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:

(2.17) 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

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

Wir kehren noch einmal zur Bayes-Formel zurück und diskutieren ein Beispiel, das im Zusammenhang mit sogenannten Expertensystemen wichtig ist.

Ein Patient kommt zu einem Arzt und klagt, daß er zwei Beschwerden hat, nennen wir sie B_1 und B_2. Der Arzt hat den Verdacht, daß der Patient die Krankheit A hat. Er untersucht noch auf ein weiteres Symptom B_3. Die Untersuchung fällt jedoch negativ aus. Mit welcher Wahrscheinlichkeit hat der Patient A? Gesucht ist also die bedingte Wahrscheinlichkeit $P(A|B_1\cap B_2\cap B_3^c)$.

Um diese Größe zu berechnen, muß einiges bekannt sein. Wir nehmen an, daß die bedingten Wahrscheinlichkeiten $P(B_i|A)$ und $P(B_i|A^c)$ bekannt sind, und daß der Arzt weiß, mit welcher Häufigkeit A eintritt, daß er also $P(A)$ kennt. Damit läßt sich immer noch wenig anfangen: Nach der Bayes-Formel ist

$$P(A|B_1\cap B_2\cap B_3^c) = \frac{P(B_1\cap B_2\cap B_3^c|A)P(A)}{P(B_1\cap B_2\cap B_3^c|A)P(A) + P(B_1\cap B_2\cap B_3^c|A^c)P(A^c)}.$$

Die hier auftretenden Größen sind aber aus unseren Angaben nicht berechenbar.

Unter einer (in vielen Fällen unrealistischen) Annahme kann man die Rechnung zu Ende führen: Die Annahme besagt, daß die Ereignisse B_1, B_2, B_3 bedingt unabhängig sind gegeben A und gegeben A^c. Das bedeutet, daß für die bedingten Wahrscheinlichkeiten $P(\cdot|A)$ und $P(\cdot|A^c)$ die Eigenschaften aus Definition (2.9) (mit B_1, B_2, B_3) gelten. In diesem Fall ist

$$P(B_1\cap B_2\cap B_3^c|A) = P(B_1|A)P(B_2|A)(1-P(B_3|A))$$

und genauso mit A^c, womit nun die gewünschte Größe aus unseren Angaben berechenbar ist.

Rechnungen dieser Art spielen bei sogenannten Expertensystemen — computerisierten Diagnosesystemen — eine große Rolle.

Interessant sind auch Anwendungsmöglichkeiten 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 Häufigkeit $2v > 0$, 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 Wahrscheinlichlkeit, daß beide Genotyp AA haben, dann u^2. Die folgende Tabell 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.

Zum Schluß des Kapitels soll ein Beispiel diskutiert werden, das wir an einigen Stellen mathematisch noch nicht ganz exakt durchführen können, zu dessen Lösung uns jedoch die inzwischen gewonnene Intuition im Umgang mit Wahrscheinlichkeiten befähigt:

A schlägt B das folgende Spiel vor: Sie werfen solange eine symmetrische Münze, bis zum erstenmal eine von zwei Dreiersequenzen auftaucht. A gewinnt, wenn zuerst die Sequenz „Zahl-Kopf-Zahl“ (abgekürzt ZKZ) auftritt; B gewinnt, wenn zuerst die Sequenz KKK vorkommt. Wie groß sind die Erfolgswahrscheinlichkeiten? Es ist nicht ganz einfach, einen geeigneten Wahrscheinlichkeitsraum zur Beschreibung des Experimentes zu finden. Da diese Aufgabe uns (noch) zu große Schwierigkeiten macht, lassen wir sie ganz weg und versuchen, die gewünschten Wahrscheinlichkeiten anders zu bestimmen.

Zunächst bemerkt man, daß von einer Folge von Münzwürfen nur der Teil am Ende wichtig ist, der als Anfangsstück einer der Gewinnsequenzen vorkommt, etwa bei der Folge $KKZZKKKZK$ die letzten zwei. Wir schreiben alle diese relevanten Sequenzen (Anfänge der Gewinnsequenzen und die Gewinnsequenzen selbst) auf:

$$\begin{tikzcd}
K \arrow[r] \arrow[d] & KK \arrow[r] \arrow[dl] & KKK \\
Z \arrow[r] \arrow[loop below] & ZK \arrow[u] \arrow[r] & ZKZ
\end{tikzcd}$$

Die Pfeile geben an, wie diese Sequenzen nach einem weiteren Wurf verändert werden können; z.B. wenn in der obigen Folge als nächstes Z geworfen wird, so hat B gewonnen, und sonst ist man bei KK.

Mit $q(K), q(KK), \ldots$ bezeichnen wir die bedingte Wahrscheinlichkeit, daß A gewinnt, gegeben das Spiel ist in dem entsprechenden Zustand. Sei q die unbedingte Wahrscheinlichkeit, daß A gewinnt. Zunächst gilt natürlich $q(KKK) = 0$ und $q(ZKZ) = 1$. Ist das Spiel in KK, so gelangt es mit Wahrscheinlichkeit $1/2$ nach KKK und mit Wahrscheinlichkeit $1/2$ nach Z. Eine ungenierte Anwendung der Formel über die totale Wahrscheinlichkeit liefert:

$$q(KK) = \frac{1}{2}q(KKK) + \frac{1}{2}q(Z) = \frac{1}{2}q(Z).$$

Analog

$$\begin{aligned}
q(ZK) &= \frac{1}{2}q(KK) + \frac{1}{2}, \\
q(K) &= \frac{1}{2}q(KK) + \frac{1}{2}q(Z), \\
q(Z) &= \frac{1}{2}q(Z) + \frac{1}{2}q(ZK)
\end{aligned}$$

und schließlich

$$q = \frac{1}{2}q(Z) + \frac{1}{2}q(K).$$

Die 4. Gleichung ergibt $q(Z) = q(ZK)$, also nach der zweiten $q(KK) = 2q(Z) - 1$. In die erste eingesetzt folgt $2q(Z) - 1 = q(Z)/2$ oder $q(Z) = 2/3$, also $q(KK) = 1/3$. Nach der dritten gilt $q(K) = 1/2$ und nach der letzten Gleichung schließlich $q = 7/12$, also um $1/12$ mehr als $1/2$! (Wer es nicht glaubt, soll es ausprobieren!)

Man kann versuchen, eine optimale Dreiersequenz zu finden, d.h. eine, die gegenüber jeder anderen eine Erfolgswahrscheinlichkeit $\geq 1/2$ hat. Es stellt sich jedoch heraus, daß es eine solche Sequenz nicht gibt: Zu jeder Sequenz gibt es mindestens eine, die sie schlägt!

§3 Zufallsgrössen, Gesetz der grossen Zahlen

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

Statt Zufallsgröße wird oft auch der Begriff *Zufallsvariable* benutzt. Für die formale Definition ist p zunächst völlig belanglos. Eine Zufallsgröße ist einfach eine Abbildung und keine „zufällige" Abbildung. Natürlich werden wir jedoch nun die Eigenschaften von X im Zusammenhang mit p untersuchen.

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:

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

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

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

(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.

(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): $X(\Omega) = \{2, 3, 4, \ldots, 12\}$,

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

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.

Geometrisch verteilte Zufallsgrößen:
In einem Bernoulli-Experiment mit Erfolgswahrscheinlichkeit p führen wir das Experiment so lange fort, bis zum erstenmal „Erfolg“ eintritt. X sei der Zeitpunkt des ersten Erfolges. Die Festlegung eines geeigneten W.-Raumes macht uns hier jedoch etwas Schwierigkeiten, da die Länge des Bernoulli-Experimentes nicht von vornherein festliegt. Am liebsten würden wir Ω als die Menge aller abzählbaren E-M-Folgen definieren, $\Omega := \{E, M\}^{\mathbb{N}}$ und $X(\omega) = \min\{\, k \in \mathbb{N} \colon \omega_k = E \,\}$. Außer der (eher belanglosen) Schwierigkeit, daß X auf der unendlich langen Pechsträhne $(M, M, M, \ldots)$ nicht definiert ist, besteht die Hauptschwierigkeit darin, daß Ω nicht abzählbar ist. Wir behelfen uns mit einer etwas unnatürlich anmutenden Notlösung und setzen $\Omega = \mathbb{N}$, wobei $n \in \mathbb{N}$ das Elementarereignis ist, daß der erste Erfolg

zum Zeitpunkt n vorkommt. Dieses Ereignis ist auch ein Elementarereignis im Bernoulli-Experiment der festen Länge n, nämlich das Ereignis, daß nach $n-1$ Mißerfolgen ein Erfolg vorkommt. Somit gilt $p(n) = (1-p)^{n-1}p$. Tatsächlich ist $\sum_{n=1}^{\infty} p(n) = p\sum_{n=0}^{\infty}(1-p)^n = 1$, womit wir nachgeprüft haben, daß (Ω, p) ein W. -Raum ist. Wir setzen dann einfach $X(n) = n$ für alle $n \in \Omega$.

(3.5) Definition. Eine Zufallsgröße, die diese Verteilung hat, heißt *geometrisch verteilt.*

Geometrisch verteilte Zufallsgrößen haben die folgende interessante Eigenschaft:

(3.6) Satz. Sei X geometrisch verteilt. Für $k \in \mathbb{N}$ ist die bedingte Wahrscheinlichkeit $P(\, X = n-1+k \mid X \geq n\,)$ gleich $P(X = k)$, also insbesondere unabhängig von $n \in \mathbb{N}$.

Beweis. Für alle $k, n \in \mathbb{N}$ gilt

$$P(\, X = n-1+k \mid X \geq n\,) = \frac{P(X = n-1+k)}{P(X \geq n)} = \frac{p(n-1+k)}{\sum_{m=n}^{\infty} p(m)}$$

und

$$\sum_{m=n}^{\infty} p(m) = \sum_{m=n}^{\infty} (1-p)^{m-1}p = p(1-p)^{n-1}\sum_{m=0}^{\infty}(1-p)^m = (1-p)^{n-1}.$$

Somit folgt $p(n-1+k)/\sum_{m=n}^{\infty} p(m) = (1-p)^{k-1}p$. $\square$

Der Satz hat die folgende intuitive Interpretation: Die Tatsache, daß bis zu einem Zeitpunkt kein Erfolg eingetreten ist, verändert nicht die bedingte Verteilung des Moments des ersten Erfolges, gerechnet von diesem Zeitpunkt an. (Viele Menschen sind anderer Ansicht, da sie, geleitet von der Beobachtung, daß sich Erfolge und Mißerfolge zum Beispiel bei einem symmetrischen Bernoulli-Experiment ungefähr ausgleichen, dem Trugschluß erliegen, daß nach einer langen Pechsträhne die Wahrscheinlichkeit für einen Erfolg steigt.)

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.

(3.7) 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 Verteilungen abhängt. Zwei verschiedene Zufallsgrößen mit derselben Verteilung haben also denselben Erwartungswert. Wir lassen die Klammern oft weg und schreiben EX statt $E(X)$. *Physikalische Interpretation:* Die Punkte in $X(\Omega)$ seien Massepunkte auf der reellen Achse. $z \in X(\Omega)$ habe die Masse $P(X = z)$. Dann ist EX der Schwerpunkt dieser Masseverteilung.

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

(3.8) 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. □

(3.9) 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) $X_1, \dots, X_n$ seien (auf einem gemeinsamen W.-Raum definierte) Zufallsgrößen, deren Erwartungswerte existieren, und $a_1, \dots, a_n$ 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 $a_1EX_1 + \cdots + a_nEX_n$. *(Man sagt, der Erwartungswert sei linear.)*

(3) X, Y seien Zufallsgrößen. Gilt $X \le Y$ und existiert der Erwartungswert von Y, so gilt $EX \le EY$. *(Man sagt, der Erwartungswert sei monoton.)*

Beweis.
(1) und (3) sind nach der Definition des Erwartungswertes evident.
(2) Wir benutzen (3.8):

$$\begin{aligned}&\sum_{\omega} p(\omega)|a_1X_1(\omega) + \cdots + a_nX_n(\omega)|\\ &\qquad\le |a_1|\sum_{\omega} p(\omega)|X_1(\omega)| + \cdots + |a_n|\sum_{\omega} p(\omega)|X_n(\omega)| < \infty.\end{aligned}$$

Somit existiert der Erwartungswert und es gilt

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

□

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 (3.9 (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 $L_1(\Omega, p)$.

(3.10) 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 $E(X) = np$.

(3) X *geometrisch verteilt* mit Parameter $p > 0$:
Es gilt $E(X) = \sum_{k=1}^\infty k(1-p)^{k-1}p$. Eine Anwendung des Quotientenkriteriums zeigt, daß die Reihe konvergiert. Zur Berechnung verwenden wir den folgenden Trick: Sei $f(s) := \sum_{k=1}^\infty s^k = s/(1-s)$ für alle $|s| < 1$ (geometrische Reihe). Dann gilt

$$f'(s) = \sum_{k=1}^\infty k s^{k-1} = \frac{(1-s) - s(-1)}{(1-s)^2} = \frac{1}{(1-s)^2}.$$

Setzt man $s = 1 - p$ ein, so ergibt sich

$$\sum_{k=1}^\infty k(1-p)^{k-1} = \frac{1}{p^2}, \quad \text{also} \quad E(X) = \frac{1}{p}.$$

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 (3.10 (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 „Massekonzentrationsphänomene" gäbe es keine Anwendungen der Wahrscheinlichkeitstheorie.

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

(3.11) 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 (3.8) nützlich:

(3.12) 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 $g \colon \Omega' \to \mathbb{R}$. 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 (3.8). $\square$

(3.13) 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 gilt $V(X) = E(X^2) - (EX)^2$.*
(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 (3.12) 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 (3.12) existiert dann $E(X^2)$.

Falls $E(X^2)$ existiert, so folgt

$$\sum_{z\in X(\Omega)} |z|P(X=z) = \sum_{\substack{z\in X(\Omega)\\ |z|\le 1}} |z|P(X=z) + \sum_{\substack{z\in X(\Omega)\\ |z|>1}} |z|P(X=z)$$
$$\le 1 + \sum_{z\in X(\Omega)} z^2 P(X=z) < \infty.$$

Somit existiert EX. Wegen $(z-EX)^2 \le 2(EX)^2 + 2z^2$ 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 $(X(\omega)+Y(\omega))^2 \le 2X(\omega)^2 + 2Y(\omega)^2$ für alle $\omega \in \Omega$. Nach (2) folgt dann die Existenz von $V(X+Y)$. $\square$

(3.14) Beispiel. Wir berechnen die Varianz einer geometrisch verteilten Zufallsgröße X und verwenden dazu denselben Trick wie bei der Berechnung des Erwartungswertes in (3.10 (3)). Sei also $f(s) := \sum_{k=1}^{\infty} s^k = s/(1-s)$ für alle $|s|<1$. Dann gilt

$$f''(s) = \sum_{k=1}^{\infty} k(k-1)s^{k-2} = \frac{2}{(1-s)^3}, \quad |s|<1.$$

Da $E(X^2) = E(X + X(X-1)) = E(X) + E(X(X-1))$ ist, folgt mit Lemma (3.12) und der obigen Formel mit $s = 1-p$

$$E(X(X-1)) = \sum_{k=1}^{\infty} k(k-1)p(1-p)^{k-1} = p(1-p)\sum_{k=1}^{\infty} k(k-1)(1-p)^{k-2} = 2\frac{1-p}{p^2},$$

also $E(X^2) = 1/p + 2(1-p)/p^2 = (2-p)/p^2$, wobei wir $E(X) = 1/p$ gemäß (3.10 (3)) benützt haben. Aus $V(X) = E(X^2) - (EX)^2$ nach (3.13 (3)), folgt $V(X) = (1-p)/p^2$.

Im allgemeinen gilt $V(X+Y) \ne V(X) + V(Y)$. Eine einfache Rechnung ergibt nämlich

$$\textbf{(3.15)} \qquad \begin{aligned} V(X+Y) &= E\big(((X+Y) - E(X+Y))^2\big) \\ &= E\big((X-EX)^2\big) + E\big((Y-EY)^2\big) \\ &\quad + 2E\big((X-EX)(Y-EY)\big) \\ &= V(X) + V(Y) + 2E\big((X-EX)(Y-EY)\big), \end{aligned}$$

und der letzte Summand ist in vielen Fällen ungleich Null, z. B. für $X = Y$, $V(X) \ne 0$. 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.

(3.16) 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.

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

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

(3.18) 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. $\square$

Die Gleichung (3.15) kann wie folgt verallgemeinert werden:

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

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

Beweis.

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

$\square$

(3.20) 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 (3.13 (2)) folgt die Existenz von $E(XY)$ und nach der Bemerkung (3.17) auch die von $\operatorname{cov}(X,Y)$. Für $\lambda, \mu \in \mathbb{R}$ folgt aus (3.18) und (3.19):

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

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. $\square$

(3.21) Bemerkung. Der Vollständigkeit halber sei auf den folgenden Sachverhalt hingewiesen. Die Existenz von $\operatorname{cov}(X,Y)$ setzt nach (3.17) 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.

(3.22) 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 (3.19)

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

(*Gleichheit nach Irénée Jules Bienaymé* (1796-1878)).

Die für uns zunächst wichtigste Klasse von unkorrelierten Zufallsgrößen sind unabhängige:

(3.23) 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.

(3.24) 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)$\Rightarrow$(b): Summation der Gleichung in (3.23) über $(z_1,\ldots,z_n) \in A_1 \times A_2 \times \cdots \times A_n$.

(b)$\Rightarrow$(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)$\Rightarrow$(d) ist trivial und (d)$\Rightarrow$(a) ergibt sich aus der Definition. □

(3.25) 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 $A_i = \{x_i \in \mathbb{R}\colon f_i(x_i) = y_i\}$. Dann ist $\{Y_i = y_i\} = \{X_i \in A_i\}$. Die Aussage folgt somit aus Satz (3.24). $\square$

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

Beweis. Sind X und Y unabhängig, so folgt

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

Nach (3.12) 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 (3.17) folgt daraus die Unkorreliertheit von X und Y. $\square$

(3.27) 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.

(3.28) 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 $(1 \leq i \leq n)$. 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 (3.26) sind die X_i paarweise unkorreliert. Nach (3.19) 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 (3.26) 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 (3.17) 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 $X_j X_k$ 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 $E(X_j X_k) = 1/(n(n-1))$. Nach Bemerkung (3.17) ist $\operatorname{cov}(X_j, X_k) = 1/(n(n-1)) - 1/n^2 = 1/(n^2(n-1))$. Nach Satz (3.19) 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 *Tschebyschev-Ungleichung* präzisiert. *Pafnuty Lwowitsch Tschebyschev* (1821-1894) bewies diese Ungleichung 1867. Wir beweisen zunächst eine andere Ungleichung, die später noch nützlich sein wird:

(3.29) Satz *(Markov-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.

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

□

(3.30) Satz *(Tschebyschev-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 (3.29)

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

□

Beispiel: Sei $a > 1$ 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 Tschebyschev-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 Tschebyschev-Ungleichung abgeschätzt werden. Wir werden dies im nächsten Kapitel diskutieren.

Die Tschebyschev-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 1713 in Basel erschien:

(3.31) 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 (3.30), (3.13 (4)) und (3.19) 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.

(3.32) 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 (3.31) setzen wir jedoch nur voraus, daß für jedes n ein W.-Raum $\Omega^{(n)}$ existiert, auf dem die $X_1,\dots,X_n$ existieren. Wenn wir ganz pedantisch wären, sollten wir deshalb $X_1^{(n)},\dots,X_n^{(n)}$ schreiben. Es macht keine Schwierigkeiten, eine solche Folge von W.-Räumen und die dazugehörenden Zufallsgrößen als mathematisch präzis definierte Objekte zu konstruieren: Es seien $f_1,\dots,f_n$ beliebige W.-Verteilungen auf abzählbaren Teilmengen A_i von $\mathbb{R}$ (d. h. $f_i\colon A_i\to[0,1]$ 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 $\omega=(\omega_1,\dots,\omega_n)\in\Omega$ setzen wir $X_i(\omega)=\omega_i$ für alle i in $\{1,\dots,n\}$ und $p(\omega)=f_1(\omega_1)f_2(\omega_2)\cdots f_n(\omega_n)$. Per Konstruktion sind $X_1,\dots,X_n$ 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 (3.31) läßt sich natürlich auf binomialverteilte Zufallsgrößen anwenden, denn diese lassen sich ja in der Form $X_1+\cdots+X_n$ schreiben, wobei die $X_1,\dots,X_n$ 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 (3.31) folgt, 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 (3.31) heißt schwaches Gesetz der großen Zahlen, um es vom sogenannten *starken Gesetz der großen Zahlen (strong law of large numbers)* zu unterscheiden. Dieses besagt

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

(*) 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:

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

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 Tschebyschev-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 $(B_n^f)_{n\in\mathbb{N}}$.

Im Weiteren wollen wir mit Hilfe der entwickelten Begriffe Aussagen über die Laufzeit von *rekursiven Algorithmen* herleiten. Die mittlere Laufzeit wird hierbei der Erwartungswert eines einfachen stochastischen Modells sein.

Wir betrachten den Sortieralgorithmus QUICKSORT: Der Algorithmus sortiert eine Liste von n Zahlen der Größe nach. Der Bequemlichkeit halber wollen wir annehmen, daß alle Elemente der Liste verschieden sind. Es gibt verschiedene Versionen dieses Algorithmus; wir betrachten hier die folgende, die für die Praxis nicht die optimale ist. Im 1. Schritt wird das erste Element der Liste mit den $n-1$ anderen verglichen und dann an die richtige Stelle gebracht. Das heißt, die Elemente, die kleiner sind, werden vor dieses 1. Elemente der ursprünglichen Liste gebracht, und die größeren werden hinter ihm gelassen. Dabei wird jedoch zunächst die interne Reihenfolge der größeren und der kleineren Elemente nicht angetastet.

Zum Beispiel wird aus $\underline{6}$ 8 3 5 1 7 nach dem ersten Schritt 3 5 1 $\underline{6}$ 8 7. Die kleineren Elemente (im Bsp. 3 5 1) und die größeren (im Bsp. 8 7) bilden nun zwei kürzere Teillisten. Die Prozedur ruft sich nun rekursiv auf, um diese zu ordnen. Listen der Länge 0 und 1 brauchen nicht mehr geordnet zu werden. Dies ist das Abbruchkriterium für den Algorithmus. Wir definieren den Aufwand für diesen Algorithmus als die Anzahl der Vergleiche zweier Zahlen, die bis zum Schluß benötigt werden. Natürlich ist dies eine Vereinfachung der realen Situation. Der tatsächliche Aufwand hängt auch sehr von der Programmiersprache ab. (Der Algorithmus ist am einfachsten in einer Sprache zu programmieren, in der man Prozeduren rekursiv aufrufen kann, er benötigt dann aber eher etwas mehr Rechenzeit. Auf programmiertechnische Fragen soll natürlich hier nicht eingegangen werden.) Im 1. Schritt werden stets $n-1$ Vergleiche durchgeführt. Wie viele jedoch nachher gebraucht werden, hängt davon ab, wie die Einteilung in die Teillisten erfolgt. Im obigen numerischen Beispiel:

1 Schritt			5 Vergleiche.
Ordnen von	$\underline{3}\ 5\ 1$	$1\ \underline{3}\ 5$	2 Vergleiche.
Ordnen von	$\underline{8}\ 7$	$7\ \underline{8}$	1 Vergleich.
Zusammen also			8 Vergleiche.

Man kann sich leicht überlegen, daß der Algorithmus im ungünstigsten Fall insgesamt $(n-1)+(n-2)+\cdots+1 = n(n-1)/2$ Vergleiche benötigt (z. B. wenn die Liste schon geordnet ist!). In der Regel braucht man jedoch bedeutend weniger. Tatsächlich gehört er zu den schnellsten Sortieralgorithmen.

Was heißt „in der Regel"? Wir machen dazu ein wahrscheinlichkeitstheoretisches Modell: Als gleich wahrscheinliche Elementarereignisse nehmen wir die möglichen Reihenfolgen einer Menge von n verschiedenen Elementen. Wir haben also $n!$ Elementarereignisse. X_n sei die Anzahl der benötigten Vergleiche bei QUICKSORT, etwa $X_1 = 0$ für jede einelementige Liste. Dann ist z. B. $X_6((6,8,3,5,1,7)) = 8$, wie oben berechnet, oder $X_6((1,3,5,6,7,8)) = 15$.

Der Erwartungswert $E(X_n)$ ist gleich $\frac{1}{n!}\sum_{\omega\in\Omega} X_n(\omega)$, da Ω $n!$ gleich wahrscheinliche Elementarereignisse enthält. Dieser mittlere Aufwand soll nun berechnet werden.

Offenbar können wir annehmen, daß die zu ordnende Liste genau die Zahlen 1 bis n enthält. Die Elementarereignisse sind die Permutationen von 1 bis n, d. h. die bijektiven Abbildungen $\omega\colon \{1,\ldots,n\} \to \{1,\ldots,n\}$. Ist $\omega(1) = k$, so hat nach dem ersten Durchgang des Algorithmus die Liste die Gestalt

$$(\tau(1),\ldots,\tau(k-1),k,\tau'(k+1),\ldots,\tau'(n)) \tag{*}$$

Dabei sind τ und τ' Permutationen der Zahlen 1 bis $k-1$ bzw. $k+1$ bis n. Die vordere Liste entfällt für $k=1$ und die hintere für $k=n$. Es gibt im allgemeinen mehrere Elementarereignisse, die nach dem ersten Durchgang gleich aussehen, z. B. $(3,2,4,1),(3,2,1,4),(3,4,2,1)$. Wir bezeichnen mit $\Omega_{k,\tau,\tau'}$ die Menge der Elementarereignisse, die nach dem ersten Durchgang die obige Liste (*) ergeben.

Der erste Durchgang des Algorithmus benötigt $n-1$ Vergleiche. Ist $\omega(1) = k$, so ist demzufolge die gesamte Anzahl $X_n(\omega)$ der benötigten Vergleiche

$$X_n(\omega) = (n-1) + X_{k-1}(\tau) + X'_{n-k}(\tau'),$$

wobei $X_{k-1}(\tau)$ und $X'_{n-k}(\tau')$ die Anzahl der benötigten Vergleiche für das Ordnen der Listen $(\tau(1), \ldots, \tau(k-1))$ bzw. $(\tau'(k+1), \ldots, \tau'(n))$ bezeichnen. Somit gilt

$$\begin{aligned} E(X_n) &= \frac{1}{n!} \sum_{\omega \in \Omega} X_n(\omega) = \frac{1}{n!} \sum_{k=1}^{n} \sum_{\omega\,:\,\omega(1)=k} X_n(\omega) \\ &= (n-1) + \frac{1}{n!} \sum_{k=1}^{n} \sum_{\tau} \sum_{\tau'} \sum_{\omega \in \Omega_{k,\tau,\tau'}} (X_{k-1}(\tau) + X'_{n-k}(\tau')). \end{aligned}$$

Die Summation über τ geht über alle Permutationen der Zahlen 1 bis $k-1$, und diejenige über τ' geht über alle Permutationen der Zahlen $k+1$ bis n.

Zunächst müssen wir abzählen, wie viele Elemente $\Omega_{k,\tau,\tau'}$ enthält, d. h. wieviele Möglichkeiten es gibt, die Elemente $\tau(1), \ldots, \tau(k-1)$ unter Erhaltung ihrer Ordnung in den $n-1$ Elementen auf den Plätzen 2 bis n der ursprünglichen Liste einzuordnen. Dies ist einfach die Anzahl der Möglichkeiten, $k-1$ Elemente aus $\{2, \ldots, n\}$ auszuwählen, also $\binom{n-1}{k-1}$. Somit gilt

$$\begin{aligned} E(X_n) &= (n-1) + \frac{1}{n!} \sum_{k=1}^{n} \sum_{\tau} \sum_{\tau'} \binom{n-1}{k-1} \big(X_{k-1}(\tau) + X'_{n-k}(\tau')\big) \\ &= (n-1) + \frac{1}{n} \sum_{k=1}^{n} \left(\sum_{\tau} \frac{1}{(k-1)!} X_{k-1}(\tau) + \sum_{\tau'} \frac{1}{(n-k)!} X'_{n-k}(\tau') \right) \\ &= (n-1) + \frac{1}{n} \sum_{k=1}^{n} (E(X_{k-1}) + E(X_{n-k})) \\ &= (n-1) + \frac{2}{n} \sum_{k=1}^{n} E(X_{k-1}). \end{aligned}$$

Da $E(X_0)$ und $E(X_1)$ gleich 0 sind, können wir die obige Gleichung wie folgt umschreiben:

$$nE(X_n) = n(n-1) + 2 \sum_{k=2}^{n-1} E(X_k).$$

Dasselbe mit $n-1$ anstelle von n:

$$(n-1)E(X_{n-1}) = (n-1)(n-2) + 2 \sum_{k=2}^{n-2} E(X_k).$$

Subtrahieren wir die zweite Gleichung von der ersten, so ergibt sich

$$nE(X_n) - (n-1)E(X_{n-1}) = 2(n-1) + 2E(X_{n-1})$$

d. h.

$$nE(X_n) - (n+1)E(X_{n-1}) = 2(n-1).$$

Dividiert man durch $n(n+1)$, so ergibt sich:

$$\frac{E(X_n)}{n+1} - \frac{E(X_{n-1})}{n} = \frac{2(n-1)}{n(n+1)} = 2\left[\frac{2}{n+1} - \frac{1}{n}\right],$$

also, da $E(X_1) = 0$ ist,

$$\begin{aligned}\frac{E(X_n)}{n+1} &= \sum_{j=2}^{n}\left(\frac{E(X_j)}{j+1} - \frac{E(X_{j-1})}{j}\right) = 2\sum_{j=2}^{n}\left(\frac{2}{j+1} - \frac{1}{j}\right) \\ &= 2\left(\sum_{j=2}^{n}\frac{2}{j+1}\sum_{j=1}^{n-1}\frac{1}{j+1}\right) \\ &= 2\left(\sum_{j=2}^{n-1}\frac{1}{j+1} + \frac{2}{n+1} - \frac{1}{2}\right) = 2\left(\sum_{j=1}^{n}\frac{1}{j+1} + \frac{1}{n+1} - 1\right).\end{aligned}$$

Damit haben wir $E(X_n)$ berechnet, allerdings etwas unhandlich, da wir $\sum_{j=1}^{n}\frac{1}{j+1}$ nicht explizit hinschreiben können. Es gelten aber die folgenden Abschätzungen, bei denen log den Logarithmus zur Basis e bezeichnet:

$$\log(n+1) - 1 \le \log(n+2) - \log(2) = \int_2^{n+2}\frac{dx}{x} = \sum_{j=1}^{n}\int_{j+1}^{j+2}\frac{dx}{x} \le \sum_{j=1}^{n}\frac{1}{j+1};$$

die letzte Ungleichung gilt, da $\frac{1}{x} \ge \frac{1}{j+1}$ für $x \in [j+1, j+2]$ ist. Ferner ist $\frac{1}{j+1} \le \frac{1}{x}$ für $x \in [j, j+1]$, also ist

$$\sum_{j=1}^{n}\frac{1}{j+1} \le \int_1^{n+1}\frac{dx}{x} = \log(n+1).$$

Für den hergeleiteten Ausdruck für $E(X_n)$ bedeutet das

$$2(n+1)(\log(n+1) - 2) \le E(X_n) \le 2(n+1)\log(n+1).$$

Wegen $\lim_{n\to\infty}\frac{n}{n+1} = 1$ und $\lim_{n\to\infty}\frac{\log n}{\log(n+1)} = 1$ folgt

(3.33) Satz.

$$\lim_{n\to\infty}\frac{E(X_n)}{n\log n} = 2.$$

Der Aufwand für Quicksort ist also im Mittel etwa $2n\log n$. Man weiß, daß es keinen Sortieralgorithmus geben kann mit einem Aufwand, dessen Größenordnung unter $n\log n$ ist. Es gibt allerdings Algorithmen, die jede vorgegebene Liste in weniger als const. $n\log n$ Schritten ordnen (z. B. Mergesort), während Quicksort in ungünstigen Fällen wesentlich mehr braucht. Der mittlere Aufwand ist jedoch bei Quicksort günstiger als bei Mergesort. Quicksort ist wohl der am meisten verwendete Sortieralgorithmus. (Meistens wird er in einer leicht modifizierten Form angewandt, wo vermieden wird, daß die ungünstigsten Listen gerade die schon weitgehend geordneten sind, wie dies in unserer Version der Fall ist.)

Der obige Satz ist ein Beispiel für eine in der Algorithmik sehr wichtige Analyse. Viele Algorithmen (z. B. bei Optimierungsproblemen) haben ein sehr schlechtes Verhalten in den ungünstigsten Fällen, jedoch eine gute mittlere Laufzeit.

Für Quicksort kann ein „Gesetz der großen Zahlen“ bewiesen werden. Dazu muß die Varianz der Zufallsgröße X_n, der Anzahl der benötigten Vergleiche, abgeschätzt werden. Dazu ohne Beweis das folgende Ergebnis:

(3.34) Lemma. *Sei X_n die Anzahl der benötigten Vergleiche bei* QUICKSORT *für eine Liste der Länge n. Dann existiert $c = \lim_{n\to\infty} V(X_n)/n^2$ und ist größer als 0.*

Der Beweis hierzu ist eine harte kombinatorische Nuß. Zu zeigen, daß $\limsup_{n\to\infty} V(X_n)/(n^2 \log n) < \infty$ gilt, ist übrigens wesentlich einfacher und reicht für (3.35) unten aus.

(3.35) Satz. Für jedes $\varepsilon > 0$ gilt

$$\lim_{n\to\infty} P\Big(\Big|\frac{X_n}{n\log n} - 2\Big| \geq \varepsilon\Big) = 0.$$

Beweis. Nach (3.33) existiert $N_\varepsilon \in \mathbb{N}$, so daß

$$\Big|\frac{E(X_n)}{n\log n} - 2\Big| \leq \varepsilon/2$$

für alle $n \geq N_\varepsilon$ gilt. Für jedes derartige n gilt

$$\Big\{\Big|\frac{X_n}{n\log n} - 2\Big| \geq \varepsilon\Big\} \subset \Big\{\Big|\frac{X_n}{n\log n} - \frac{E(X_n)}{n\log n}\Big| \geq \frac{\varepsilon}{2}\Big\}$$

und mittels der Tschebyschev-Ungleichung folgt

$$P\Big(\Big|\frac{X_n}{n\log n} - 2\Big| \geq \varepsilon\Big) \leq P\Big(\Big|\frac{X_n}{n\log n} - \frac{E(X_n)}{n\log n}\Big| \geq \frac{\varepsilon}{2}\Big) \leq \frac{4}{\varepsilon^2}\,\frac{V(X_n)}{n^2\log^2 n}.$$

Nach (3.34) konvergiert dies gegen 0 für $n \to \infty$. $\square$

§4 Grosse Abweichungen

In diesem Kapitel widmen wir uns ausschließlich dem Bernoulli-Experiment der Länge n mit Erfolgswahrscheinlichkeit p. Wir sind bereits jetzt in der Lage, in diesem Modell ein paar genauere Analysen vorzunehmen. Wir wollen betrachten, wie schnell der Limes im schwachen Gesetz der großen Zahlen gegen null konvergiert. Das Resultat ermöglicht eine genauere Untersuchung langer Erfolgs-Ketten im Bernoulli-Experiment.

Wir bezeichnen wieder mit S_n die Anzahl der „Kopf"-Würfe nach n Würfen. Wie wir aus den Beispielen (3.10 (2)) und (3.28 (1)) schon wissen, gilt $E(S_n) = np$ und $V(S_n) = np(1-p)$. Wir wollen die Wahrscheinlichkeit untersuchen, daß der relative Anteil S_n/n der Anzahl der „Kopf"-Würfe in den ersten n Würfen eine feste Mindestabweichung vom Erwartungswert hat, d. h. für $\alpha > 0$ untersuchen wir

$$P\Big(\Big|\frac{S_n}{n} - p\Big| \geq \alpha\Big).$$

Zu welcher Erkenntnis verhilft uns die Tschebyschev-Ungleichung? Antwort:

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

also zum Beispiel für den symmetrischen Münzwurf und $\alpha = 1/10$ 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. Eine bessere Abschätzung erhält man auf die folgende Weise:

Wir wenden die Markov-Ungleichung für die monotone Funktion $\mathbb{R} \ni x \mapsto e^{\lambda x}$, $\lambda > 0$, an, wobei λ zunächst beliebig gewählt sei. Es gilt

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

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 $\lambda(S_n - n\,p) = \sum_{i=1}^n \lambda(X_i - p)$, 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 $e^{\lambda(X_i - p)}$ aus Satz (3.25). Demnach folgt aus der Bemerkung (3.27) für jedes $\lambda > 0$

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

Nun ist $|X_i - p| \leq 1$ und $E(X_i - p) = 0$ für jedes $i = 1, \ldots, n$. Da die Funktion $e^{\lambda x}$ konvex ist, ist sie im Intervall $[-1, 1]$ durch die Sekante, die die Punkte $(-1, e^{-\lambda})$ und $(1, e^{\lambda})$ verbindet, nach oben beschränkt:

$$e^{\lambda x} \leq \frac{e^{\lambda} + e^{-\lambda}}{2} + \frac{e^{\lambda} - e^{-\lambda}}{2} x.$$

Aus der Monotonie des Erwartungswertes folgt also

$$E\left(e^{\lambda(X_i-p)}\right) \le \frac{e^{\lambda}+e^{-\lambda}}{2} + \frac{e^{\lambda}-e^{-\lambda}}{2} E(X_i-p) = \cosh(\lambda).$$

Jetzt nutzen wir noch die Ungleichung $\cosh(\lambda) \le e^{\lambda^2/2}$ aus. Sie ergibt sich unmittelbar aus den Potenzreihenentwicklungen

$$\exp(x) = \sum_{k=0}^{\infty} \frac{x^k}{k!} \quad \text{und} \quad \cosh(x) = \sum_{k=0}^{\infty} \frac{x^{2k}}{(2k)!}.$$

Wir erhalten

$$P\Big(S_n - n\,p \ge \alpha\Big) \le e^{-\alpha\lambda} e^{n\lambda^2/2}$$

für jedes $\lambda > 0$. Wir wollen nun $\lambda > 0$ so wählen, daß wir eine möglichst gute obere Abschätzung erhalten, d.h., wir bestimmen das Minimum der Funktion $f(\lambda) = -\alpha\lambda + n\lambda^2/2$. Eine einfache Rechnung ergibt, daß das Minimum bei $\lambda = \frac{\alpha}{n}$ angenommen wird. Einsetzen liefert

$$P\Big(S_n - n\,p \ge \alpha\Big) \le e^{-\alpha^2/(2n)}.$$

Wir haben also eine Abschätzung erhalten, die nicht vom Parameter p abhängt. Dies ist, wie wir weiter unten sehen, nicht unbedingt von Vorteil. Halten wir dieses Ergebnis, welches *Wassily Hoeffding* im Jahre 1963 in einer grundlegenden Arbeit lieferte, fest:

(4.1) Satz *(1.Hoeffding-Ungleichung)***.** Bezeichnet S_n die Anzahl der „Kopf"-Würfe nach n Würfen bei einem Bernoulli-Experiment zu den Parametern n und p, so gilt:

$$P\Big(S_n - n\,p \ge \alpha\Big) \le e^{-\alpha^2/(2n)}.$$

Was hat uns die Anstrengung gebracht? In der Notation $\bar{S}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$ folgt aus der 1.Hoeffding-Ungleichung:

$$P\Big(\bar{S}_n - p \ge \alpha\Big) \le e^{-n\alpha^2/2}.$$

Für den symmetrischen Münzwurf sind „Kopf" und „Zahl" gleich wahrscheinlich, also gilt

$$P\Big(S_n - \frac{n}{2} \ge \alpha n\Big) = P\Big(S_n - \frac{n}{2} \le -\alpha n\Big)$$

und somit

$$P\Big(\Big|\bar{S}_n - \frac{1}{2}\Big| \ge \alpha\Big) = P\Big(\Big|S_n - \frac{n}{2}\Big| \ge \alpha n\Big) = 2P\Big(S_n - \frac{n}{2} \ge \alpha n\Big) \le 2e^{-n\alpha^2/2}.$$

Für $\alpha = 1/10$ und $n = 1000$ erhalten wir dann als Schranke $2e^{-5} < 1/40$.

Eine genauere Analyse des Beweises führt aber recht schnell zu einer wesentlichen Verbesserung unserer ersten *exponentiellen Ungleichung (exponential inequality)*. In der obigen Rechnung haben wir zunächst den Erwartungswert $E(e^{\lambda X})$ einer durch 1 beschränkten Zufallsgröße X nach oben abgeschätzt und anschließend das Resultat im „freien“ Parameter λ optimiert. Nun optimieren wir einfach direkt die aus dem Bernstein-Ansatz gewonnene Abschätzung. Es gilt zunächst ganz analog:

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

Da erneut mit den X_i auch die $e^{\lambda X_i}$ unabhängig sind, folgt

$$\begin{aligned} P\Big(S_n \geq \alpha n\Big) &\leq e^{-\alpha n\lambda} \prod_{i=1}^{n} E\Big(e^{\lambda X_i}\Big) = e^{-\alpha n\lambda}\Big(pe^{\lambda} + (1-p)\Big)^n \\ &= \exp\Big(n\{-\alpha\lambda + \log M(\lambda)\}\Big), \end{aligned}$$

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

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

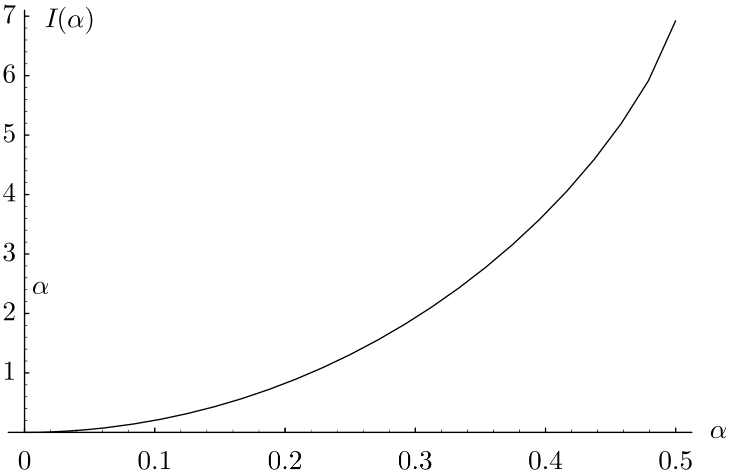
Zusammenfassend haben wir also gezeigt, daß für die Anzahl S_n der „Kopf“-Würfe die Abschätzung

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

für alle $\alpha \in (p, 1)$ gilt. Erneut stellen wir die Frage, 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 Tschebyschev-Ungleichung oder $2e^{-5}$ aus der 1.Hoeffding-Ungleichung. Wir halten unser Resultat fest:

(4.2) Satz *(2.Hoeffding-Ungleichung)***.** Bezeichnet S_n die Anzahl der „Kopf"-Würfe nach n Würfen bei einem Bernoulli-Experiment zu den Parametern n und p, so gilt:

$$P\Big(\bar{S}_n - p \geq \alpha\Big) \leq e^{-nH(\alpha+p|p)}.$$

Genaugenommen ist dieses Resultat nur für die Werte $0 < \alpha < 1-p$ interessant, denn für $\alpha > 1-p$ ist $P(\bar{S}_n - p \geq \alpha) = 0$ und im Fall $\alpha = 1-p$ ist $P(\bar{S}_n - p \geq 1-p) = P(\bar{S}_n = 1) = P(S_n = n) = p^n = b(n; n, p)$. Tatsächlich konvergiert die rechte Seite der 2.Hoeffding-Ungleichung für $\alpha \to (1-p)$ gegen p^n, wie man sich leicht klar machen kann.

Bemerkung. Natürlich ist die Abschätzung in Satz (4.2) schärfer als die in Satz (4.1). Dazu zeigen wir $H(\alpha+p|p) \geq \alpha^2/2$ für alle $\alpha > 0$. Da $H(0+p|p) = 0$, folgt dies zum Beispiel aus der Tatsache, daß die zweite Ableitung von $H(\alpha+p|p)$ nach α immer größer als 1 ist. Dies überlassen wir dem Leser.

Eine natürliche Frage ist, ob wir die Güte der Abschätzung der 2.Hoeffding-Ungleichung noch verbessern können, also durch geschickte Techniken das Resultat noch verbessern können. Die Antwort auf diese Frage ist im Fall des Bernoulli-Experimentes, aber auch in vielen anderen Fällen, in denen ein schwaches Gesetz der großen Zahlen zugrundeliegt, nein. Dies wollen wir im folgenden präzisieren. Wir stellen zunächst ein paar Eigenschaften der Funktion $H(\alpha|p)\colon (0,1) \to \mathbb{R}$ zusammen.

(4.3) 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 $p \in I$. $H(\cdot|p)$ ist stetig, streng monoton wachsend für $\alpha > p$ und streng monoton fallend für $\alpha < p$.*

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. $\square$

Mit Hilfe der Eigenschaften der Funktion $H(\cdot|p)$ stellen wir das Resultat aus Satz (4.2) in ein anderes Licht. Das schwache Gesetz der großen Zahlen besagt, daß die Wahrscheinlichkeit, das $\bar{S}_n$ von p um mehr als ein $\alpha > 0$ abweicht, für jedes beliebige $\alpha > 0$ in n gegen null konvergiert. Da $H(\alpha + p|p) > 0$ für alle $\alpha > 0$, sagt uns Satz (4.2), daß die Konvergenz sogar exponentiell schnell ist. In dieser Skala spricht man von *großen Abweichungen*. Zu jedem $\alpha > 0$ finden wir einen Zahlenwert $H(\alpha + p|p)$, und dieser bestimmt neben der *Geschwindigkeit* n, die immer auftaucht, das genaue Konvergenzverhalten. Das schwache Gesetz schildert uns quasi das typische Verhalten des Bernoulli-Experimentes, die Abweichungen von diesem typischen Verhalten werden in ihrer Größenordnung mit Hilfe der 2.Hoeffding-Ungleichung bestimmt. Wir formulieren nun diese Ungleichung ein wenig um: für $\alpha \in (p, 1)$ gilt

$$P\Big(\bar{S}_n \in (\alpha, \infty)\Big) \leq P\Big(\bar{S}_n \in [\alpha, \infty)\Big) \leq \exp\Big(-n \inf_{x \in (\alpha, 1)} H(x|p)\Big),$$

wobei wir die Monotonie und die Stetigkeit von $H(\cdot|p)$ nutzen. Für $\alpha \in (0, p)$ gilt nun analog:

$$P\Big(\bar{S}_n \in (-\infty, \alpha)\Big) \leq \exp\Big(-n \inf_{x \in (0, \alpha)} H(x|p)\Big).$$

Hierzu muß man wirklich nur den Beweis einmal durchsehen, und man sieht schnell, daß alles analog berechnet werden kann. Man kann nun wegen des Monotonieverhaltens von $H(\cdot|p)$ in beiden Ungleichungen das Intervall (α, ∞) (bzw. $(-\infty, \alpha)$) durch ein Intervall $I = (a, b)$ mit $a > p$ (bzw. $b < p$) ersetzen. Interessant sind dann noch Intervalle der Form $I = (a, b)$ mit $p \in (a, b)$. Nach Lemma (4.3) wissen wir, daß $\inf_{x \in (a,b)} H(x|p) = 0$. Mit $m := \min(p - a, b - p)$ gilt $P(|\bar{S}_n - p| \leq m) \leq P(\bar{S}_n \in (a, b))$, woraus nach dem schwachen Gesetz der großen Zahlen folgt:

$$\lim_{n \to \infty} P\Big(\bar{S}_n \in (a, b)\Big) = 1.$$

Also haben wir die obige Ungleichung auch im Grenzübergang für diese Intervalle gezeigt. Wir erhalten somit für ein beliebiges Intervall $I = (a, b)$:

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

Wir müssen hier etwas vorsichtig sein. Tatsächlich haben wir keine Konvergenz bewiesen, aber der Limessuperior existiert immer. Entlang der obigen Überlegungen folgt sofort die analoge Abschätzung für ein abgeschlossenes Intervall $I = [a, b]$, wobei hier das Infimum über $x \in [a, b]$ gebildet wird. Wir werden nun zeigen, daß

$$\liminf_{n \to \infty} \frac{1}{n} \log P\Big(\bar{S}_n \in I\Big) \geq -\inf_{x \in I} H(x|p)$$

gilt mit $I = (a, b)$ oder $I = [a, b]$. Es folgt daraus insgesamt der folgende Grenzwertsatz:

(4.4) Satz *(Prinzip großer Abweichungen von Cramér, large deviation principle)***.** Bezeichnet S_n die Anzahl der „Kopf"-Würfe nach n Würfen bei einem Bernoulli-Experiment zu den Parametern n und p, so gilt für alle $0 \leq a < b \leq 1$:

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

wobei $I = (a, b)$ oder $I = [a, b]$.

Beweis. Auf Grund der obigen Diskussion können wir uns erneut auf ein Intervall der Form $I = (\alpha, \infty)$ (bzw. $I = [\alpha, \infty)$) mit $\alpha > p$ beschränken. Es gilt dann für ein beliebiges $c > \alpha$

$$\begin{aligned} P\Big(\bar{S}_n > \alpha\Big) &= \sum_{k>n\alpha} b(k; n, p) \\ &= \sum_{k>n\alpha} \exp\Bigg(- \log\Big(\frac{b(k;n,c)}{b(k;n,p)}\Big)\Bigg)\, b(k; n, c). \end{aligned}$$

Da nun

$$\log\Big(\frac{b(k;n,c)}{b(k;n,p)}\Big) = n(k/n)\log(c/p) + n(1 - (k/n))\log((1 - c)/(1 - p)) =: nH(k/n)$$

mit $H(x) = x\log(c/p) + (1 - x)\log((1 - c)/(1 - p))$, erhalten wir für jedes $\varepsilon > 0$

$$\begin{aligned} P\Big(\bar{S}_n > \alpha\Big) &= \sum_{k\,:\,k>n\alpha} \exp(-nH(k/n))\, b(k; n, c) \\ &\geq \sum_{k\,:\,k>n\alpha,\ H(k/n)\leq H(c)+\varepsilon} \exp(-n(H(c) + \varepsilon))\, b(k; n, c). \end{aligned}$$

Für $c > p$ ist $H(\cdot)$ eine lineare Funktion. Daher gilt $\{k\colon k > n\alpha, H(k/n) - H(c) \leq \varepsilon\} \supseteq \{k\colon k > n\alpha, |k/n - c| \leq H^{-1}(\varepsilon)\}$. Nach dem schwachen Gesetz der großen Zahlen gilt nun:

$$\lim_{n\to\infty} \sum_{k\,:\,k>n\alpha,\ H(k/n)\leq H(c)+\varepsilon} b(k; n, c) \geq \lim_{n\to\infty} \sum_{k\,:\,|k-nc|\leq nH^{-1}(\varepsilon)} b(k; n, c) = 1.$$

Also haben wir für alle $c > \alpha$ insgesamt gezeigt:

$$\liminf_{n\to\infty} \frac{1}{n} \log P\Big(\bar{S}_n > \alpha\Big) \geq -H(c|p),$$

denn $H(c) = H(c|p)$. Daraus folgt aber mit der Stetigkeit von $H(\cdot|p)$ die Behauptung:

$$\liminf_{n\to\infty} \frac{1}{n} \log P\Big(\bar{S}_n > \alpha\Big) \geq -H(\alpha|p) = - \inf_{x\in(\alpha,\infty)} H(x|p).$$

□

Bemerkung. Die Konvergenzaussage zeigt nun, daß wir die 2. Hoeffding-Ungleichung nicht verschärfen können. Das Prinzip großer Abweichungen wurde in der angegebenen Form von *Harald Cramér* (1893-1985) 1938 bewiesen. Tatsächlich hat die Analyse der Konvergenzgeschwindigkeit beim schwachen Gesetz der großen Zahlen ihren Ursprung schon in der Physik des späten letzten Jahrhunderts. *Ludwig Boltzmann* (1844-1906) betrachtete in einer Arbeit im Jahre 1877 ein einfaches Modell eines sogenannten idealen Gases. Er teilte gedanklich einen Gasbehälter mit einem festen Volumen in zwei Teilbehälter und dachte sich die Teilchen rein zufällig und unabhängig auf die Teilbehälter verteilt. Bei der Betrachtung der „relativen Häufigkeit“ $\bar{S}_n$ der Teilchen in einem der beiden Teilbehälter kam er zu einem Resultat der Form

$$P(\bar{S}_n \approx p') \approx \exp(-nH(p'|p)),$$

wenn n die Gesamtzahl der Teilchen bezeichnet. Boltzmann erkannte weiter, daß die Abbildung $H(\cdot|p)$ einer Größe entspricht, die in der Thermodynamik und später in der durch ihn mitbegründeten statistischen Mechanik den Namen *Entropie* bekam. Formal entspricht das Prinzip großer Abweichungen dem berühmten Boltzmannschen Gesetz $S = k \log W$, wobei S die Entropie und W die Wahrscheinlichkeit bezeichnet. k ist die bekannte Boltzmannkonstante.

Runs:
Satz (4.4) ermöglicht eine genauere Untersuchung von *Runs* in einer Bernoulli-Kette. In einer Folge aus $\{0,1\}^n$ nennen wir jede maximale Teilfolge von einander benachbarten gleichen Symbolen einen *Run*. Bekannt ist das folgende Lehrexperiment. Man teile eine Schulklasse (oder auch die Gruppe der Hörerinnen und H "orer dieser Vorlesung) in zwei gleichgroße Gruppen. In der einen Gruppe wird jeder gebeten, eine Münze 200 mal zu werfen und die 0-1-Sequenz aufzuschreiben. In der anderen Gruppe wird jeder gebeten, anstelle des Münzwurfs eine *zufällige* 0-1-Folge der Länge 200 zu notieren, ohne eine Münze zu werfen. Die Zettel werden eingesammelt und die Aufgabe besteht nun darin, anhand der Resultate die Zettel wieder der jeweiligen Gruppe zuzuordnen. Die Behauptung ist, daß dies überwiegend richtig geschehen kann. Der Spielleiter hat dabei eine wichtige Information als Auswahlkriterium an der Hand: In Münzwurfserien der Länge 200 kommen 1-Runs der Länge 7 vor, während die Schüler, die von Hand eine zufällige Serie notieren, es in der Regel scheuen, mehr als 4 mal hintereinander das gleiche Symbol zu verwenden. Sind 1-Runs einer Länge größer 5 vorhanden, ordnet der Spielleiter den Zettel der Gruppe, die die Münze warfen, zu. Er liegt fast immer richtig. Um dies mathematisch begründen zu können, müssen wir also die Länge des *längsten Runs* von 1'en in einer Bernoulli-Kette der Länge n untersuchen. Ein heuristisches Argument geht so: ein 1-Run der Länge m tritt mit Wahrscheinlichkeit p^m auf. Es gibt angenähert n mögliche Positionen für einen Run, also ist $E(\text{Anzahl der 1-Runs der Länge}\, m) \approx n\, p^m$. Ist nun der längste 1-Run eindeutig, so sollte seine Länge R_n die Gleichung $1 = n\, p^{R_n}$ erfüllen. Die Lösung ist $R_n = \frac{\log n}{\log(1/p)}$. Im Fall einer fairen Münze ist $\frac{\log 200}{\log 2} \approx 7.64$.

Es bezeichne R_n die Länge des längsten Runs von 1'en in einer Bernoulli-Kette der Länge n (mit Erfolgswahrscheinlichkeit p), genauer

$$R_n := \max\left\{ l - k \,:\, 0 \leq k < l \leq n\,,\, \frac{S_l - S_k}{l - k} = 1 \right\}.$$

Es gilt:

(4.5) Satz *(Erdös-Rényi-Gesetz für Runs).* In einem Bernoulli-Experiment der Länge n mit Erfolgswahrscheinlichkeit p gilt für jedes $\varepsilon > 0$

$$\lim_{n\to\infty} P\left(\left|\frac{R_n}{\log n} - \frac{1}{\log(1/p)}\right| > \varepsilon\right) = 0.$$

Wir bemerken, daß das typische Verhalten von $\frac{R_n}{\log n}$ durch die relative Entropie $H(1|p) = \log(1/p)$ beschrieben wird. Tatsächlich können wir in einer Bernoulli-Kette der Länge n Segmente maximaler Länge betrachten, deren Mittelwert in einem gegebenen Intervall $I = (a, b)$ (oder $I = [a, b]$) mit $0 \le a < b \le 1$ liegt: sei

$$R_n^I := \max\left\{l - k \,:\, 0 \le k < l \le n\,,\ \frac{S_l - S_k}{l - k} \in I\right\}$$

und $H_I := \inf_{x\in I} H(x|p)$. Dann gilt

(4.6) Satz *(Erdös-Rényi-Gesetz).* Es sei $I = (a, b)$ oder $I = [a, b]$ mit $0 \le a < b \le 1$. In einem Bernoulli-Experiment der Länge n mit Erfolgswahrscheinlichkeit p gilt für jedes $\varepsilon > 0$

$$\lim_{n\to\infty} P\left(\left|\frac{R_n^I}{\log n} - \frac{1}{H_I}\right| > \varepsilon\right) = 0.$$

Wir führen die folgende Zufallsgröße ein:

$$T_r^I := \inf\left\{l \,:\, \frac{S_l - S_k}{l - k} \in I \text{ für ein } 0 \le k \le l - r\right\}$$

Es gilt dann $\{R_n^I \ge r\}$ genau dann wenn $\{T_r^I \le n\}$. Im Fall $I = \{1\}$ bedeutet $R_n \ge r$, daß beim n-fachen Münzwurf mindestens ein 1-Run der Länge $\ge r$ vorkommt. Somit liegt die *Ersteintrittszeit* für einen 1-Run der Länge r (oder länger) unterhalb n (oder ist gleich n). Die Argumentation für die umgekehrte Inklusion geht analog.

Beweis. Es gilt

$$(*)\qquad \left\{\left|\frac{R_n^I}{\log n} - \frac{1}{H_I}\right| > \varepsilon\right\} \subset \left\{\frac{R_n^I}{\log n} - \frac{1}{H_I} \ge \varepsilon\right\} \cup \left\{\frac{R_n^I}{\log n} - \frac{1}{H_I} < -\varepsilon\right\}.$$

Nach Definition von T_r^I gilt die folgende Inklusion:

$$\{T_r^I \le m\} \subset \bigcup_{k=0}^{m-r} \bigcup_{l=k+r}^{m} C_{k,l} \subset \bigcup_{k=0}^{m-1} \bigcup_{l=k+r}^{\infty} C_{k,l}$$

mit $C_{k,l} := \left\{\frac{S_l - S_k}{l-k} \in I\right\}$. Nun gibt es in der ersten Vereinigung auf der rechten Seite m mögliche Werte für k und es ist $P(C_{k,l}) = P\left(\frac{S_{l-k}}{l-k} \in I\right)$ für $l - k \ge r$. Also gilt

$$P(T_r^I \le m) \le m \sum_{n=r}^{\infty} P\left(\frac{S_n}{n} \in I\right).$$

Satz (4.4) liefert zu jedem $\varepsilon > 0$ eine Konstante $c = c(\varepsilon)$ mit $P\big(\frac{S_n}{n} \in I\big) \leq c\, e^{-n(H_I - \varepsilon/2)}$.
Es gibt für jedes $\varepsilon > 0$ ein hinreichend großes $n = n(\varepsilon)$ mit $\{R_n^I \geq \varepsilon \log n + \frac{\log n}{H_I}\} \subset \{R_n^I \geq \lfloor \frac{\log n}{H_I - \varepsilon} + 1 \rfloor\}$. Wir wählen $r = \lfloor (\log n)(H_I - \varepsilon)^{-1} + 1 \rfloor$. Dann ist $n \leq \lfloor e^{r(H_I - \varepsilon)} \rfloor =: m$ und wir erhalten für n hinreichend groß

$$P\left(R_n^I \geq \varepsilon \log n + \frac{\log n}{H_I}\right) \leq P\big(T_r^I \leq m\big) \leq c'\, e^{r(H_I-\varepsilon)} e^{-r(H_I - \varepsilon/2)} = c' e^{-r\varepsilon/2}.$$

Hierbei ist c' eine von ε abhängige Konstante. Setzen wir r ein, so erhalten wir

$$\lim_{n\to\infty} P\left(R_n^I \geq \varepsilon \log n + \frac{\log n}{H_I}\right) = 0.$$

Um das zweite Ereignis in (*) zu untersuchen, betrachte $B_l := \big\{\frac{1}{r}(S_{lr} - S_{(l-1)r}) \in I\big\}$. Die Ereignisse $\{B_l\}_{l\geq 1}$ sind unabhängig und $P(B_l) = P\big(\frac{S_r}{r} \in I\big)$. Die Inklusion

$$\bigcup_{l=1}^{\lfloor m/r \rfloor} B_l \subset \{T_r^I \leq m\}$$

liefert daher die Abschätzung

$$(**) \qquad P(T_r^I > m) \leq \big(1 - P(B_1)\big)^{\lfloor m/r \rfloor} \leq e^{-\lfloor m/r \rfloor P(B_1)}.$$

Es gibt für jedes $\varepsilon > 0$ ein hinreichend großes $n = n(\varepsilon)$ mit $\{R_n^I < -\varepsilon \log n + \frac{\log n}{H_I}\} \subset \{R_n^I < \lfloor \frac{\log n}{H_I + \varepsilon} - 1 \rfloor\}$. Wir setzen nun $r = \lfloor (\log n)(H_I + \varepsilon)^{-1} - 1 \rfloor$. Dann ist $n \geq \lfloor e^{r(H_I+\varepsilon)} \rfloor =: m$ und wir erhalten für n hinreichend groß

$$P\left(R_n^I < -\varepsilon \log n + \frac{\log n}{H_I}\right) \leq P\big(T_r^I > m\big) \leq \exp\left(-\frac{c}{r}\, e^{r(H_I+\varepsilon)} e^{-r(H_I + \varepsilon/2)}\right),$$

denn Satz (4.4) liefert zu jedem $\varepsilon > 0$ eine Konstante $c = c(\varepsilon)$ mit $P\big(\frac{S_n}{n} \in I\big) \geq c\, e^{-n(H_I + \varepsilon/2)}$. Setzen wir r ein, so erhalten wir

$$\lim_{n\to\infty} P\left(R_n^I < -\varepsilon \log n + \frac{\log n}{H_I}\right) = 0.$$

Somit ist der Satz bewiesen. $\square$

Bemerkung. Paul Erdös (1914-1996) und *Alfréd Rényi* (1921-1970) haben diese Resultate im Jahre 1970 in einer Arbeit mit dem Titel *On a new law of large numbers* hergeleitet. Tatsächlich haben sie in dieser Arbeit ein starkes Gesetz hergeleitet:

$$P\Big(\lim_{n\to\infty} \frac{R_n^I}{\log n} \text{ existiert und ist } = \frac{1}{H_I}\Big) = 1.$$

§5 Die eindimensionale Irrfahrt

Für $i, j \in \mathbb{Z}$, $i < j$, nennen wir eine Folge

$$((i, s_i), (i+1, s_{i+1}), \ldots, (j, s_j)) \quad \text{mit} \quad \begin{array}{lll} s_k \in \mathbb{Z} & \text{für} & i \le k \le j \quad \text{und} \\ |s_{k+1} - s_k| = 1 & \text{für} & i \le k \le j-1 \end{array}$$

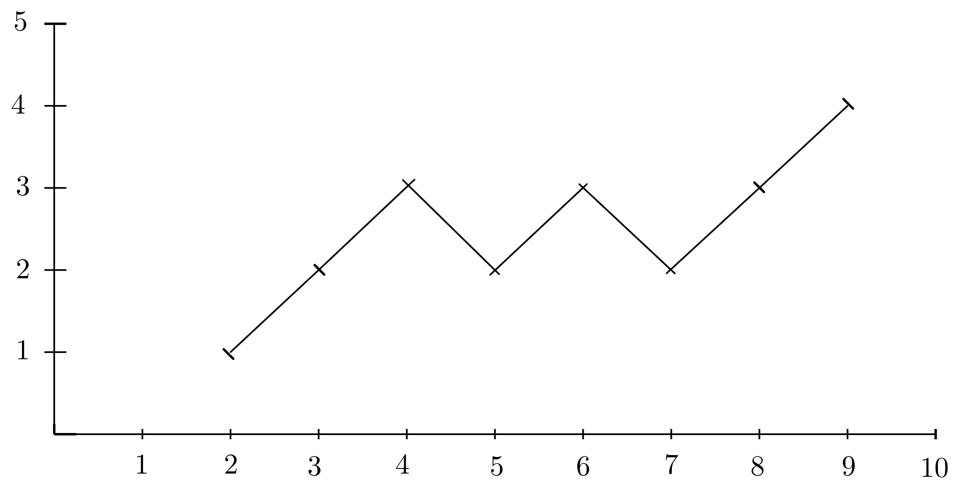
einen *Pfad (path)* von (i, s_i) nach (j, s_j). Ist ω ein derartiger Pfad, so lassen wir in Zukunft die erste Koordinate oft weg und schreiben einfach

$$(s_i, s_{i+1}, \ldots, s_j)$$

für einen derartigen Pfad. $j - i$ nennen wir die *Länge* des Pfades.

Verbinden wir die Punkte (k, s_k), $(k+1, s_{k+1})$ durch gerade Linien, so erhalten wir einen Streckenzug in der Ebene. Diese Darstellung hat keine mathematische Bedeutung, sondern dient nur der Anschauung.

Der Pfad $((2,1), (3,2), (4,3), (5,2), (6,3), (7,2), (8,3), (9,4))$ ergibt zum Beispiel den folgenden Streckenzug:



Meist werden bei uns die Pfade in $(0,0)$ beginnen; es erweist sich jedoch als günstig, allgemeinere Situationen zuzulassen.

Man beachte, daß die Parität von $s_j - s_i$ stets gleich der Parität von $j - i$ ist, das heißt $s_j - s_i$ ist genau dann gerade, wenn dies auch für $j - i$ gilt.

Wir werden zwei verschiedene Zufallsexperimente betrachten:

(I) *Der Endpunkt liegt fest:* Ist $n \in \mathbb{N}$ und hat s dieselbe Parität wie n, so bezeichne $\Omega_{(n,s)}$ die Menge der Pfade von $(0,0)$ nach (n,s). Auf dieser Menge betrachten wir die Gleichverteilung. Wir müssen zunächst die Anzahl der Pfade zählen: Hat ein Pfad $\omega \in \Omega_{(n,s)}$ p ansteigende Verbindungen und q absteigende (d. h. $p := |\{\, i \in \{0, \ldots, n-1\} : s_{i+1} = s_i + 1 \,\}|$), so gelten $p + q = n$, $p - q = s$, das heißt $p = (n+s)/2$, $q = (n-s)/2$. p und q sind also durch n und s vollständig festgelegt.

$|\Omega_{(n,s)}|$ ist die Anzahl der Möglichkeiten, die p aufsteigenden Verbindungen in der Gesamtzahl von n Schritten zu plazieren, das heißt, es gilt

$$|\Omega_{(n,s)}| = \binom{n}{(n+s)/2} = \binom{p+q}{p}. \tag{5.1}$$

(II) *Freier Endpunkt:* Ω_n bezeichne die Menge aller Pfade der Länge n mit Startpunkt $(0,0)$. $|\Omega_n|$ ist hier offenbar 2^n.

Wir betrachten zunächst den Fall (I), das heißt das Zufallsexperiment, das durch die Gleichverteilung auf $\Omega_{(n,s)} = \Omega_{(p+q,p-q)}$ beschrieben wird.

Wir können uns etwa vorstellen, daß eine Wahl zwischen zwei Kandidaten K_1, K_2 stattgefunden hat, wobei nun p Stimmen für K_1 und q Stimmen für K_2 in einer Wahlurne liegen. Diese Stimmen werden nun eine um die andere ausgezählt. Wir wollen zunächst das folgende Ereignis betrachten. Sei $p > q$ (d. h. K_1 hat gewonnen). Mit welcher Wahrscheinlichkeit liegt er stets vorn bei der Auszählung? Diese Wahrscheinlichkeit ist gleich $|A|/|\Omega_{(p+q,p-q)}| = |A|/\binom{p+q}{p}$, wobei

$$A = \{\, \omega = (0, s_1, \ldots, s_{p+q}) \in \Omega_{(p+q,p-q)} : s_k > 0 \text{ für } 1 \le k \le p+q \,\}$$

ist. Zum Abzählen der Pfade in A verwenden wir einen eleganten Trick mit einer teilweisen Spiegelung von Pfaden 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 $i \le k \le j$ existiert, für das $s_k = 0$ ist.

(5.2) 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).*

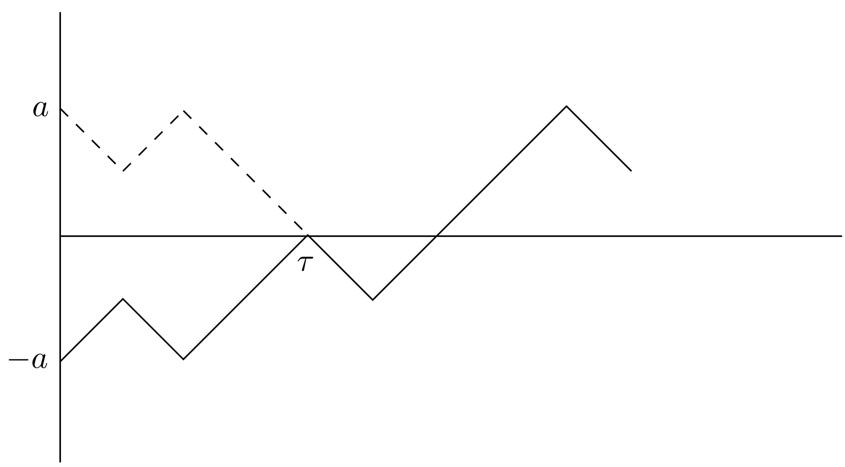
Beweis. Wir geben eine bijektive Abbildung an, die die Menge der Pfade von $(i,-a)$ nach (j,b) auf die Menge der Pfade von (i,a) nach (j,b), welche die x-Achse berühren, abbildet. Sei

$$(s_i = -a, s_{i+1}, \ldots, s_{j-1},\ s_j = b)$$

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. $\square$



Wir können mit diesem Lemma $|A|$ nun einfach bestimmen: Für $\omega = (0, s_1, \ldots, s_n) \in A$ gilt notwendigerweise $s_1 = 1$. $|A|$ ist somit die Anzahl der Pfade von $(1, 1)$ nach $(p+q, p-q)$, die die x-Achse nicht berühren. Dies ist gleich der Anzahl aller Pfade von $(1, 1)$ nach $(p+q, p-q)$, minus der Anzahl derjenigen, die die x-Achse berühren. Letztere ist nach Lemma (5.2) gleich der Anzahl aller Pfade von $(1, -1)$ nach $(p+q, p-q)$. Wenden wir (5.1) an, so ergibt sich also

$$(5.3) \qquad |A| = \binom{p+q-1}{p-1} - \binom{p+q-1}{p} = \frac{p-q}{p+q}\binom{p+q}{p}.$$

(Wir haben hier natürlich $p > q$ vorausgesetzt.) Die Anzahl aller Elemente in $\Omega_{(p+q,p-q)}$ ist nach (5.1) $\binom{p+q}{p}$. Somit ergibt sich das folgende Resultat:

(5.4) Satz *(Ballot-Theorem, von ballot (engl.) = geheime Abstimmung)***.**
Die Wahrscheinlichkeit dafür, daß der Kandidat mit der größeren Anzahl p der Stimmen während des gesamten Verlaufs der Auszählung führt, ist $(p-q)/(p+q)$, wobei q die Anzahl der Stimmen des Unterlegenen bezeichnet.

Eine kleine Modifikation des obigen Arguments gestattet auch die Diskussion des Falles $p = q$. Natürlich kann dann keiner der Kandidaten dauernd führen, da nach der Auszählung Gleichstand herrscht. Wir können aber die beiden folgenden Ereignisse betrachten:

(1) Kandidat K_1 führt während der gesamten Auszählung, erst am Schluß tritt Gleichstand ein.
(2) Kandidat K_2 führt nie.

Da der zugrunde liegende W.-Raum $\binom{2p}{p}$ Elementarereignisse hat, die alle die gleiche Wahrscheinlichkeit haben, ergeben sich aus dem folgenden Satz die Wahrscheinlichkeiten für diese beiden Ereignisse:

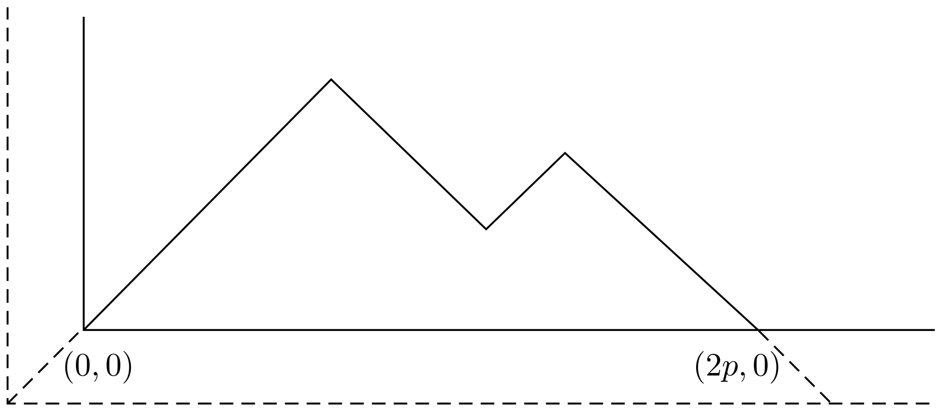
(5.5) Satz.

(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) Natürlich ist notwendigerweise $s_{2p-1} = 1$. Wir suchen somit nach der Anzahl der Pfade von $(0, 0)$ nach $(2p-1, 1)$ mit $s_1 > 0$, $s_2 > 0, \ldots, s_{2p-1} = 1$. Nach der Formel (5.3) mit $q = p-1$ ist dies gleich

$$\frac{1}{2p-1}\binom{2p-1}{p} = \frac{1}{p}\binom{2p-2}{p-1}.$$

(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). $\square$

Aus (2) des obigen Satzes folgt, daß bei Gleichstand der Stimmen mit Wahrscheinlichkeit $1/(p+1)$ der Kandidat K_2 zu keinem Zeitpunkt der Auszählung führt.

Wir wenden uns nun der Situation (II) zu, das heißt dem Zufallsexperiment, das durch die Gleichverteilung auf Ω_n beschrieben wird. Dies ist nichts anderes als eine Umformulierung unseres alten Münzwurfexperiments. Statt K, Z nehmen wir die Zahlen -1, 1. Einem Element $(a_1, \ldots, a_n) \in \{-1,1\}^n$ können wir einen Pfad $(s_0 = 0, s_1, \ldots, s_n) \in \Omega_n$ durch $s_k = \sum_{j=1}^k a_j$, $1 \le k \le n$, zuordnen. Dies definiert eine bijektive Abbildung $\{-1,1\}^n \to \Omega_n$. Der Gleichverteilung auf $\{-1,1\}^n$ entspricht dann via dieser bijektiven Abbildung die Gleichverteilung auf Ω_n.

Es ist üblich, die Positionen der Pfade als Zufallsgrößen zu beschreiben:

$$\begin{gathered} S_k \colon \Omega_n \to \mathbb{Z} \subset \mathbb{R}, \quad 0 \le k \le n, \\ \omega = (s_0 = 0, s_1, \ldots, s_n) \mapsto S_k(\omega) := s_k. \end{gathered}$$

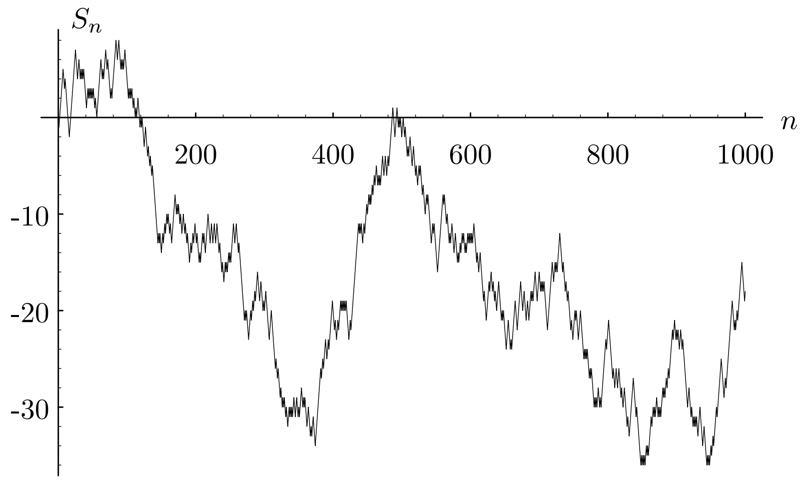
Die S_k lassen sich als Summen von unabhängigen Zufallsgrößen darstellen:

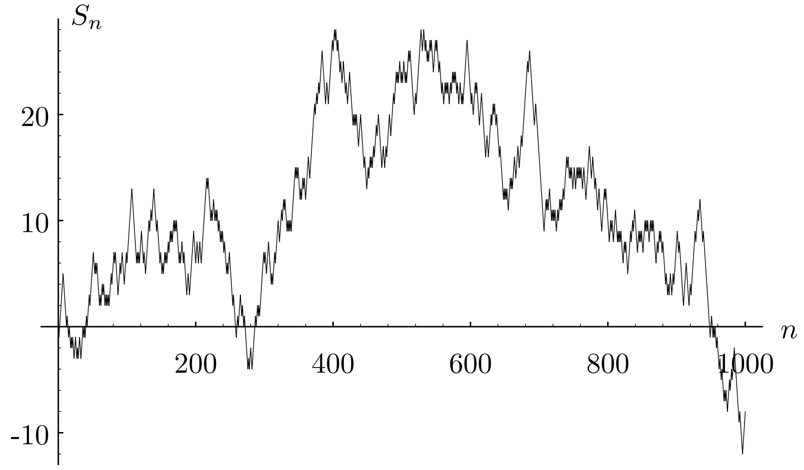
$$S_k = \sum_{j=1}^{k} X_j,$$

mit $X_j = S_j - S_{j-1}$. Die X_j sind dann unabhängige Zufallsgrößen mit $P(X_j = 1) = P(X_j = -1) = 1/2$. Aufgefasst als Abbildungen $\{-1,1\}^n \to \mathbb{R}$ sind es einfach die Projektionen auf die einzelnen Faktoren. 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

$$A = \{S_1 \ne 20, S_2 \ne 20, \ldots, S_{99} \ne 20, S_{100} = 20\}.$$

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.





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}, $2k \leq n$. Um die Anzahl der Pfade der Länge n zu bestimmen, die zu A_{2k} gehören, multiplizieren wir die Anzahl der Pfade der Länge $2k$ von $(0,0)$ nach $(2k,0)$ mit der Anzahl der Pfade der Länge $n-2k$, die in $(2k,0)$ starten (bei freiem Ende). Somit ist

$$|A_{2k}| = \binom{2k}{k} 2^{n-2k}.$$

Ω_n enthält 2^n Elemente. Also 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 Wahrscheinlichkeit, daß nach der $2k$-ten Stimmabgabe Gleichstand besteht, wird nicht davon abhängen, wieviele Stimmen nachträglich noch abgegeben werden.

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.B. Durchschnitte 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.

Man sieht zunächst nicht, von welcher Größenordnung $u_{2k} = P(A_{2k})$ für große k ist. Da

$$u_{2k} = \frac{(2k)!}{(k!)^2} 2^{-2k}$$

ist, benötigen wir eine genauere Kenntnis des Verhaltens der Fakultätsfunktion für große Argumente. Diese erhält man über die sogenannte *Stirling-Approximation*, die von *James Stirling* (1692-1770) bewiesen wurde:

(5.6) 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.

Notation. Für zwei reelle Zahlenfolgen $(a_n)_{n\in\mathbb{N}}$, $(b_n)_{n\in\mathbb{N}}$, mit $a_n, b_n > 0$ schreiben wir $a_n \sim b_n$, sofern

$$\lim_{n\to\infty} a_n/b_n = 1$$

gilt. Dies bedeutet keineswegs, daß $|a_n - b_n|$ gegen 0 konvergiert. So gilt etwa

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

(5.7) Satz.

$$u_{2k} \sim \frac{1}{\sqrt{\pi k}}.$$

Beweis. Einsetzen der Stirling-Approximation in

$$u_{2k} = \frac{(2k)!}{(k!)^2} 2^{-2k}.$$

$\square$

(5.7) ist eine recht gute Näherung für u_{2k}. Um dies genauer zu diskutieren, brauchte man gute Abschätzungen für die Differenz $n! - \sqrt{2\pi} n^{n+1/2} e^{-n}$. Wir wollen diesen Punkt jedoch nicht weiter verfolgen.

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

$$f_{2k} = P(S_1 \neq 0,\, S_2 \neq 0, \ldots, S_{2k-1} \neq 0,\, S_{2k} = 0).$$

(5.8) 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) $u_{2k} = \sum_{j=1}^{k} f_{2j} u_{2k-2j}$.

Beweis. (1) Nach (5.5 (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 (5.5 (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

$$u_{2k} = \sum_{j=1}^{k} P(B_j) = \sum_{j=1}^{k} f_{2j} u_{2k-2j}.$$

$\square$

Eine interessante Folgerung ergibt sich aus der ersten Gleichung in (2). Da nach (5.7) $\lim_{k\to\infty} u_{2k} = 0$ 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 (5.8 (1)) anwenden. Nach (5.7) 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 (5.8 (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.

(5.9) 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 (5.8 (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

$$|G_{l,m}^+| = |G_{l,m}^-| = \frac{1}{2} f_{2l} 2^{2l} P(F_m^{(n-l)}) 2^{2n-2l},$$

$$P(G_{l,m}^+) = P(G_{l,m}^-) = \frac{1}{2} f_{2l} P(F_m^{(n-l)}).$$

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

$$= \frac{1}{2}\sum_{l=1}^{j} f_{2l}\, u_{2j-2l}\, u_{2n-2j} + \frac{1}{2}\sum_{l=1}^{n-j} f_{2l}\, u_{2n-2j-2l}\, u_{2j} = u_{2j}\, u_{2n-2j} \quad \text{nach (5.8 (3)).}$$

□

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 (5.7) läßt sich eine einfach Approximation für $P(F_j^{(k)})$ für große j und $k-j$ gewinnen:

(5.10) 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 (5.10) nennt man einen lokalen 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 (5.10) für $k \to \infty$ folgern:

(5.11)
$$\gamma_k(\alpha, \beta) \sim \frac{1}{\pi} \sum_{j\,:\,\alpha \leq \frac{j}{k} \leq \beta} \frac{1}{k} \frac{1}{\sqrt{\frac{j}{k}\left(1 - \frac{j}{k}\right)}}.$$

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:

(5.12) Satz *(Arcus-Sinus-Gesetz)*.

$$\lim_{k\to\infty} \gamma_k(\alpha, \beta) = \frac{2}{\pi}(\arcsin\sqrt{\beta} - \arcsin\sqrt{\alpha}).$$

Beweis. Wir müssen (5.11) zeigen. Wir schreiben die Stirling-Approximation als $n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n F(n)$ 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

$$\sum_{\alpha \leq \frac{j}{k} \leq \beta} P(F_j^{(k)}) = \left(\frac{1}{k} \sum_{\alpha \leq \frac{j}{k} \leq \beta} \frac{1}{\pi} \frac{1}{\sqrt{(j/k)(1-(j/k))}} \right) G_{\alpha,\beta}(k).$$

Nun folgt die Behauptung gleichmäßig für $\delta \leq \alpha < \beta \leq 1-\delta$, wie auch immer $0 < \delta < 1/2$ gewählt war. Damit folgt die Behauptung. $\square$

(5.13) Bemerkung. Die Aussage von (5.12) 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, $\leq \beta$ ist — gegen $\frac{2}{\pi} \arcsin \sqrt{\beta}$ konvergiert.

Beweis. Offensichtlich gilt $\lim_{k\to\infty} \gamma_k\left(0, \frac{1}{2}\right) = 1/2$. Ist $\beta \in (0, 1/2)$, so folgt

$$\lim_{k\to\infty} \gamma_k(0,\beta) = \lim_{k\to\infty} (\gamma_k(0,1/2) - \gamma_k(\beta,1/2)) = \frac{2}{\pi} \arcsin \sqrt{\beta},$$

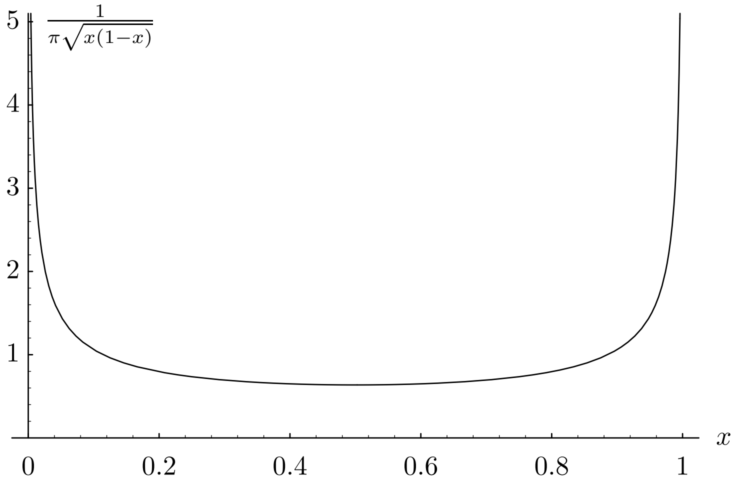
für $\beta > 1/2$

$$\lim_{k\to\infty} \gamma_k(0,\beta) = \lim_{k\to\infty} (\gamma_k(0,1/2) + \gamma_k(1/2,\beta)) = \frac{2}{\pi} \arcsin \sqrt{\beta}.$$

Für $\gamma_k(\alpha, 1)$ führt dasselbe Argument zum Ziel. $\square$

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:



§6 Poisson- und 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.

Wir wollen diese Verteilung durch eine neue Verteilung auf $\mathbb{N}_0$ approximieren, die sogenannte *Poissonverteilung.*

Für eine reelle Zahl $\alpha > 0$ betrachte man die Wahrscheinlichkeitsverteilung, die durch

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

definiert ist. Zunächst überzeugt man sich davon, daß

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

ist. π_α ist also tatsächlich eine Wahrscheinlichkeitsverteilung.

(6.1) Definition. Eine Zufallsgröße X mit $X(\Omega) = \mathbb{N}_0$ und der Verteilung π_α heißt *Poisson-verteilt mit Parameter* $\alpha > 0$.

Der Erwartungswert dieser Verteilung ist leicht auszurechnen:

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

Eine Poisson-verteilte Zufallsgröße hat also Erwartungswert α.

Als nächstes wollen wir die Varianz ausrechnen:

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

Somit gilt

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

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

Wir zeigen nun, daß die Poissonverteilung eine Approximation der Binomialverteilung ist, wenn n groß und p klein sind. Zunächst überlegt man sich, in welcher Beziehung α zu den Parametern n, p der Binomialverteilung stehen soll. Wir wählen α so, daß die Erwartungswerte übereinstimmen, daß also $\alpha = np$ ist. $b(k;n,p)$ liegt nahe bei $\pi_\alpha(k)$ für $\alpha = np$. Um das zu präzisieren, leiten wir eine konkrete Schranke für

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

her. Wir zeigen ein allgemeineres Resultat:

(6.3) Satz. Es seien $X_1, \dots, X_n$ 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 $i = 1, \dots, n$. Sei $X = X_1 + \cdots + X_n$ und $\lambda = p_1 + \cdots + p_n$, dann gilt:

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

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

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

Bevor wir Satz (6.3) beweisen, einige Kommentare: Die Schranke ist nur für kleine p_i interessant. Man kann daraus Grenzwertaussagen ableiten. Wir lassen dabei (im Falle $p = p_i$) p von n abhängen ($p := p_n$) und n nach unendlich streben. Falls $\lim_{n\to\infty} np_n^2 = 0$ gilt, so folgt aus (6.4), daß $\lim_{n\to\infty} \Delta(n, p_n) = 0$ gilt. Insbesondere folgt der sogenannte *Poissonsche Grenzwertsatz*, der von *Siméon Denis Poisson* (1781-1840) im Jahre 1832 entdeckt wurde:

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

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

(6.5) folgt sofort aus (6.4): Aus $np_n \to \alpha$ folgt $p_n \to 0$ für $n \to \infty$ und $np_n^2 \to 0$. Ferner ist $|b(k; n, p) - \pi_{np}(k)| \le \Delta(n, p)$ für jedes $k \in \mathbb{N}_0$. Demzufolge gilt

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

Wegen $\pi_{np_n}(k) \to \pi_\alpha(k)$ folgt (6.5).

Bemerkung. Die Aussage von (6.4) ist auch im Fall, wo $np_n^2 \to 0$, $np_n \to \infty$ gelten, von Interesse (z. B. $p_n = 1/n^{2/3}$).

Der wichtigste Vorzug von (6.3) und (6.4) im Vergleich zu (6.5) ist jedoch, daß eine ganz konkrete Approximationsschranke vorliegt. Satz (6.3) ist schwieriger zu beweisen als (6.5). Obwohl – wie oben gesehen – letzterer eine unmittelbare Folge von (6.3) ist, skizzieren wir den üblichen "Standardbeweis", der sehr kurz ist:

Beweis von (6.5). Setze $\alpha_n = np_n$. Nach Voraussetzung gilt $\alpha_n \to \alpha$.

$$\begin{aligned}
b(k; n, p_n) &= b\left(k; n, \frac{\alpha_n}{n}\right) = \binom{n}{k} \left(\frac{\alpha_n}{n}\right)^k \left(1 - \frac{\alpha_n}{n}\right)^{n-k} \\
&= \frac{n(n-1)\cdots(n-k+1)}{k!} \frac{\alpha_n^k}{n^k} \frac{(1-\alpha_n/n)^n}{(1-\alpha_n/n)^k} \\
&= \frac{\alpha_n^k}{k!} \frac{1}{(1-\alpha_n/n)^k} \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right) \times \cdots \times \left(1 - \frac{k-1}{n}\right)\left(1 - \frac{\alpha_n}{n}\right)^n.
\end{aligned}$$

Wegen $\alpha_n \to \alpha$ folgt $\alpha_n/n \to 0$. Weiter hat die Funktion $\log x$ an der Stelle 1 den Ableitungswert 1, also gilt $\log(1-h) = -h + h\Delta(h)$ mit $\Delta(h) \to 0$ für $h \to 0$. Es folgt

$$\left(1 - \frac{\alpha_n}{n}\right)^n = \exp\left(n\left(-\frac{\alpha_n}{n} + \frac{\alpha_n}{n}\Delta\left(\frac{\alpha_n}{n}\right)\right)\right) \to \exp(-\alpha).$$

Wir erhalten also insgesamt

$$\lim_{n\to\infty} b(k;n,p_n) = \frac{\alpha^k}{k!}\lim_{n\to\infty}\left(1-\frac{\alpha_n}{n}\right)^n = e^{-\alpha}\frac{\alpha^k}{k!}.$$

□

Die untenstehende Tabelle gibt einige numerisch ermittelte Anhaltspunkte für den Vergleich zwischen Binomial- und Poissonverteilung ($p := 0,1$).

k	$\pi_{0.5}(k)$	$b(k;5,0.1)$
0	0.6065	0.5905
1	0.3033	0.3281
2	0.0768	0.0729
3	0.0126	0.0081
4	0.00158	0.00045
5	0.00016	0.00001

k	$\pi_2(k)$	$b(k;20,0.1)$
0	0.1353	0.1216
1	0.2707	0.2702
2	0.2707	0.2852
3	0.1804	0.1901
5	0.0361	0.0319
10	0.000038	0.000052

Bevor wir den Beweis von Satz (6.3) geben, stellen wir einen wichtigen Aspekt der Poissonverteilung bereit:

(6.6) 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:

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

□

(6.7) 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.

Der Beweis des Satzes (6.3) verwendet eine Technik, die man *Kopplung (coupling)* nennt. Nehmen wir an, f und g seien zwei Wahrscheinlichkeitsverteilungen auf $\mathbb{N}_0$:

$f, g\colon \mathbb{N}_0 \to [0,1]$, $\sum_k f(k) = \sum_k g(k) = 1$. Wir wollen zeigen, daß $\sum_{k=0}^{\infty} |f(k) - g(k)|$ klein ist. Wir werden das tun, indem wir Zufallsgrößen X, Y auf einem gemeinsamen Wahrscheinlichkeitsraum Ω konstruieren, die die Verteilung f beziehungsweise g haben, und die „möglichst weitgehend" übereinstimmen. Es soll also die folgende Situation vorliegen:

$$f(k) = P(X = k), \quad g(k) = P(Y = k).$$

A sei ein Ereignis mit $X(\omega) = Y(\omega)$ für $\omega \in A$, das heißt $A \subset \{\, \omega \colon X(\omega) = Y(\omega)\,\}$. Man sagt, X und Y seien auf A „gekoppelt".

(6.8) Lemma. *Unter den obigen Bedingungen gilt*

$$\sum_{k=0}^{\infty} |f(k) - g(k)| \le 2P(A^c).$$

Beweis. Sei $M := \{\, k \in \mathbb{N}_0 : f(k) > g(k)\,\}$. Dann ist

$$\begin{aligned}
\sum_{k=0}^{\infty} |f(k) - g(k)| &= \sum_{k \in M} (f(k) - g(k)) - \sum_{k \notin M} (f(k) - g(k)) \\
&= 2 \sum_{k \in M} (f(k) - g(k)) - \sum_{k=0}^{\infty} (f(k) - g(k)) \\
&= 2(P(X \in M) - P(Y \in M)) - (1 - 1) \\
&= 2(P(X \in M, A) + P(X \in M, A^c) - P(Y \in M)) \\
&\le 2(P(Y \in M, A) + P(A^c) - P(Y \in M)) \\
&\le 2P(A^c).
\end{aligned}$$

□

Wir wenden nun dieses Kopplungsargument an, um Satz (6.3) zu beweisen. Der Hauptteil des Beweises besteht in einer geeigneten Wahl des zugrundeliegenden Wahrscheinlichkeitsraumes. Da wir nur die Verteilung von X berechnen müssen, ist es quasi egal, auf welchem Wahrscheinlichkeitsraum die Zufallsgrößen X_i definiert werden. Es ist für uns nur wichtig, daß die Zufallsgrößen unabhängig sind und $P(X_i = 1) = p_i$ sowie $P(X_i = 0) = 1 - p_i$ gilt. Diese Freiheit nutzen wir für eine Wahl derart, daß eine Poisson-verteilte Zufallsgröße zum Parameter λ möglichst weitgehend mit X in Verteilung übereinstimmt. Dazu 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 $k \ge 1$ sowie $P_i(-1) = 1 - P_i(0) - \sum_{k \ge 1} P_i(k) = e^{-p_i} - (1 - p_i)$. 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

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

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 (6.6), 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 $1 - e^{-x} \leq x$. Nach Lemma (6.8) folgt dann

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

Damit ist Satz (6.3) bewiesen.

Erstaunlich ist, daß viele verschiedene, natürliche oder künstlich erzeugte Zufallserscheinungen recht gut zum Poisson-Schema passen.

(6.9) Beispiel. Über einen Zeitraum von 20 Jahren wurde im alten Preußen die Zahl der Toten durch Hufschlag in 10 Kavallerieregimenten beobachtet. Insgesamt hatte man also 200 „Regimentsjahre“ beobachtet. Es ergab sich das folgende Bild:

k =Anzahl der Toten	Anzahl Regimentsjahre mit k Toten	$200\pi_\alpha(k)$ mit $\alpha = 0,61$ (gerundet)
0	109	109
1	65	66
2	22	20
3	3	4
4	1	1
≥ 5	0	0

(α wurde so bestimmt, daß sich die beste Übereinstimmung ergibt.)

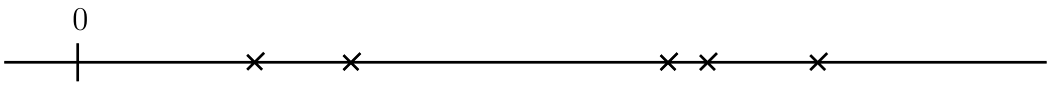
Die theoretische Begründung für die gute Übereinstimmung ist etwa die: Für den einzelnen Kavalleristen ist die Wahrscheinlichkeit p, in einem Jahr vom Pferd erschlagen zu werden, sehr klein. Hat das Regiment n Kavalleristen, so ist die Verteilung der Anzahl der Toten pro Regiment und Jahr $\approx b(k; n, p) \approx \pi_\alpha(k)$. Nach dem Gesetz der großen Zahlen ist dann bei 200 Wiederholungen des „Versuchs“ die Anzahl der Regimentsjahre mit k Toten $\approx 200\pi_\alpha(k)$. Die obige Übereinstimmung ist jedoch eher ungewöhnlich. Wegen der fast perfekten Übereinstimmung wird das Beispiel immer wieder zitiert. Die bereits 1832 endeckte Poisson-Approximation blieb lange Zeit vergessen. Erst 1898 zeigte *Ladislaus von Bortkiewicz* (1868–1931) in seinem Buch „Das Gesetz der kleinen Zahlen“ die Bedeutung dieser Approximation für die Praxis. Das obige Beispiel entstammt diesem Buch.

Als weitere Anwendung von (6.5) diskutieren wir den sogenannten Poissonschen Punktprozeß.

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. Im Rahmen dieser Vorlesung ist es leider nicht möglich, einen Wahrscheinlichkeitsraum dafür zu konstruieren.

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

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

Offensichtlich gilt $\alpha(0) = 0$, 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:

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

was wegen (P1)

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

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 $\alpha(s) = \lambda s$. $\lambda = 0$ 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 ist, daß die Wahrscheinlichkeit für zwei oder mehr Punkte in einem kleinen Intervall noch etwas kleiner ist, genauer

(P5) $$\lim_{\varepsilon \to 0} \varepsilon^{-1} P(N_\varepsilon \geq 2) = 0.$$

Man kann nachweisen, daß (P5) nur die Möglichkeit von Mehrfachpunkten ausschließt; wir wollen das jedoch nicht weiter ausführen.

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:

(6.10) 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

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

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

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

Wir stellen einige einfach zu verifizierende Eigenschaften dieser Zufallsgrößen zusammen:

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

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

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

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

(6.11) $$P(\overline{N}_s^{(k)} \geq m) \leq P(N_s \geq m).$$

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

(6.12) $\overline{N}_s^{(k)}$ ist binomialverteilt mit Parameter k, p_k.

Wir verwenden nun (P5), um nachzuweisen, daß sich für große k $\overline{N}_s^{(k)}$ nur wenig von N_s unterscheidet:

$$(6.13)\qquad P(\overline{N}_s^{(k)} \neq N_s) = P\Big(\bigcup_{i=1}^{k}\{X_i^{(k)} \geq 2\}\Big) \leq \sum_{i=1}^{k} P(X_i^{(k)} \geq 2)$$
$$= kP(N_{s/k} \geq 2) \to 0 \quad \text{für } k \to \infty.$$

Für $m \in \mathbb{N}$ und $k \in \mathbb{N}$ gilt:

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

Unter Benutzung von (6.12) und (6.13) folgt:

$$(6.14)\qquad P(N_s = m) = \lim_{k\to\infty} P(\overline{N}_s^{(k)} = m) = \lim_{k\to\infty} b(m; k, p_k)$$

und analog

$$(6.15)\qquad P(N_s \geq m) = \lim_{k\to\infty} P(\overline{N}_s^{(k)} \geq m).$$

Wir zeigen nun:

$$(6.16)\qquad \lim_{k\to\infty} kp_k = \lambda s.$$

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

$P(\overline{N}_s^{(k)} \geq l)$ ist nach (6.11) nicht größer als $P(N_s \geq l)$ und strebt nach (6.15) für $k \to \infty$ gegen diese obere Grenze. Nach einem Satz über reelle Zahlenfolgen (falls nicht bekannt oder vergessen: Übungsaufgabe!) folgt daraus

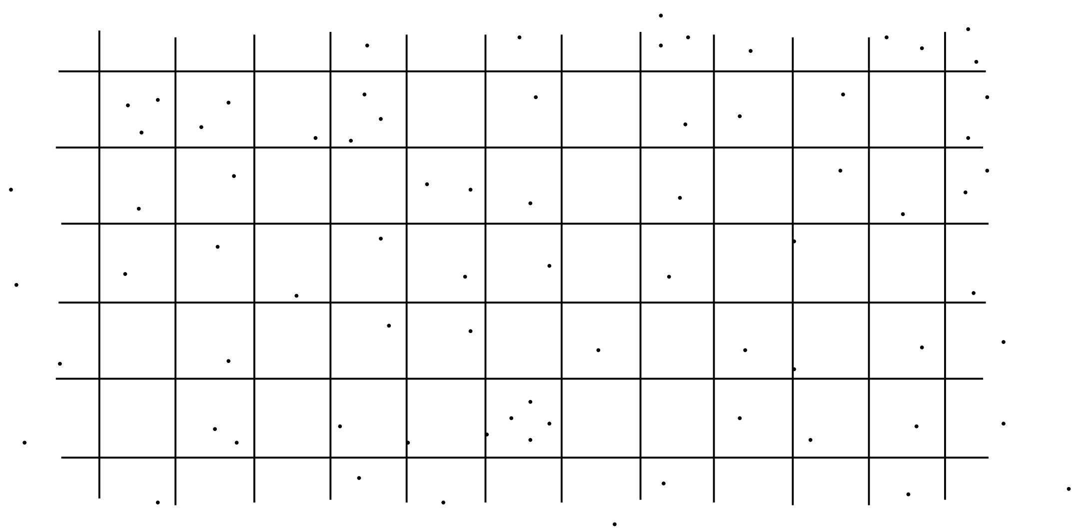
$$\lim_{k\to\infty} kp_k = \lim_{k\to\infty} \sum_{l=1}^{\infty} P(\overline{N}_s^{(k)} \geq l) = \sum_{l=1}^{\infty} P(N_s \geq l) = EN_s = \lambda s.$$

Damit ist (6.16) gezeigt. Unser Satz folgt nun aus (6.14), (6.16) und dem Satz (6.5). $\square$

Der Poissonsche Punktprozeß wird oft verwendet um etwa eintreffende Anrufe in einer Telefonzentrale, ankommende Jobs in einem Computernetzwerk etc. zu modellieren. Die Annahmen (P1)–(P5) sind natürlich nicht immer sehr realistisch oder nur näherungsweise richtig. Problematisch in Anwendungen sind oft (P1) und (P2). Es gibt viele Möglichkeiten, den Poissonschen Punktprozeß zu verallgemeinern, um dem

Rechnung zu tragen; wir wollen darauf nicht eingehen, jedoch ganz kurz auf eine Verallgemeinerung auf mehrdimensionale Räume. Die Zeitachse $(0, \infty)$ wird hier durch einen höherdimensionalen Raum $\mathbb{R}^d$ ersetzt. Man möchte also ein wahrscheinlichkeitstheoretisches Modell für zufällig im Raum $\mathbb{R}^d$ liegende Punkte konstruieren. Für beschränkte Mengen $A \subset \mathbb{R}^d$ bezeichne N_A die Anzahl der Punkte, die in A liegen. Unter Bedingungen, die analog zu (P1)–(P5) sind, läßt sich nachweisen, daß N_A Poisson-verteilt mit Parameter $\lambda|A|$ sein muß, wobei $|A|$ das d-dimensionale Volumen von A ist (wir denken uns hier einfach alle A, von denen wir bereits das Volumen kennen) und $\lambda > 0$ ein fester Parameter. Wir wollen das nicht beweisen, sondern nur ganz kurz an zwei Beispielen illustrieren.

(6.17) Beispiel. Es wird eine Probe einer Flüssigkeit auf mikroskopisch kleine Partikel untersucht, z.B. Blut auf Bakterien, Leukozyten, Trinkwasser auf Salmonellen oder ähnliches. In der Praxis wird das oft so gemacht, daß ein kleiner Teil der Probe unter dem Mikroskop untersucht und zusammen mit einem Gitternetz angeschaut wird.



Eine Zählung aller Punkte ist oft zu aufwendig und deshalb begnügt man sich damit, nur einige Gitterquadrate auszuzählen und daraus die Gesamtzahl zu schätzen. Als Basis dieser Schätzung dient meist die Modellannahme, daß die Punkte gemäß einem Poissonschen Punktprozeß verteilt sind, so daß insbesondere die Anzahl der Punkte N_A, die in einem Quadrat A des Gitters liegen, Poisson-verteilt mit Parameter $\lambda|A|$ ist. Mit Hilfe einer Auszählung von wenigen Quadraten soll dann eine Schätzung des unbekannten Parameters λ vorgenommen werden. Nehmen wir einmal an, ein Laborant zählt m Quadrate $A_1, \ldots, A_m$ aus. Aufgrund dieser Auszählung soll er den Parameter λ schätzen.

Er kann zum Beispiel alles zusammenzählen, $N_A := \sum_{j=1}^m N_{A_j}$, wobei A die Vereinigung $\bigcup_{j=1}^m A_j$ ist. N_A ist ebenfalls Poisson-verteilt mit Parameter $\lambda|A|$. Eine natürliche Schätzung von λ ist daher

$$\hat{\lambda} := N_A/|A|.$$

Wir wollen nicht weiter darauf eingehen, ob $\hat{\lambda}$ eine gute Schätzung ist, sondern kurz begründen, wieso die N_A Poisson-verteilt sein sollten, mit einem Argument, das (P1)–(P5) nicht explizit voraussetzt: Wir stellen uns vor, daß die gesamte Probe auf einer großen Fläche liegt, die in n kleine aber gleich große Quadrate aufgeteilt ist. Auf der gesamten Fläche sollen m Bakterien liegen, so daß im Mittel auf jedes der Quadrate $\mu := m/n$ kommen. Betrachten wir nun ein festes Quadrat A und nehmen an, daß jedes der Bakterien unabhängig von jedem anderen mit Wahrscheinlichkeit $1/n$ auf A fällt, so ist offenbar

$$P(N_A = k) = b(k; m, 1/n) \simeq \pi_\mu(k),$$

falls m, n sehr groß sind. Natürlich ist $\mu = \lambda|A|$, wobei λ die mittlere Anzahl der Bakterien pro Einheitsfläche ist.

Wir können auch begründen, wieso die N_A für verschiedene Quadrate (mit paarweise leerem Schnitt) im Limes unabhängig sind. Wir beschränken uns auf zwei: A_1, A_2.

Um $P(N_{A_1} = k_1,\, N_{A_2} = k_2)$ zu berechnen, stellen wir uns die Bakterien von 1 bis m durchnumeriert vor. Das Bakterium i fällt mit Wahrscheinlichkeit $1/n$ auf A_1, mit derselben Wahrscheinlichkeit auf A_2 und mit Wahrscheinlichkeit $1 - 2/n$ auf keines der beiden (A_1, A_2 sind disjunkt). Eine einfache Abzählung ergibt, daß es $\frac{m!}{k_1!\,k_2!\,(m-k_1-k_2)!}$ Bakterienfolgen gibt mit k_1 Treffern in A_1 und k_2 Treffern in A_2. Ist der Aufenthaltsort der einzelnen Bakterien unabhängig, so ergibt sich also:

$$P(N_{A_1} = k_1,\, N_{A_2} = k_2) = \frac{m!}{k_1!k_2!(m-k_1-k_2)!}\left(\frac{1}{n}\right)^{k_1}\left(\frac{1}{n}\right)^{k_2}\left(1-\frac{2}{n}\right)^{m-k_1-k_2}.$$

Dies konvergiert gegen $\pi_\mu(k_1)\pi_\mu(k_2)$ für $m, n \to \infty$ und $\frac{m}{n} \to \mu > 0$, was man sofort mit demselben Argument wie im Beweis von (6.5) sieht.

Obgleich wir in diesem Argument (P1)–(P5) nicht vorausgesetzt haben, sie sich also von selbst ergeben, so steckt die wesentliche Unabhängigkeitsannahme natürlich darin, daß die einzelnen Bakterien unabhängig von den anderen auf die einzelnen Quadrate fallen. Das ist eine oft etwas fragwürdige Annahme, wenn man nicht weiß, ob die Probe gut durchmischt ist.

Von der Binomial- zur Normalverteilung

Die Poissonapproximation der Binomialverteilung ist nur sinnvoll, wenn n groß und p klein ist. Oft möchte man jedoch $b(k; n, p)$ für einen festen Wert von p und großes n approximieren. Wir haben dies in Kapitel 5 für den Spezialfall $p = 1/2$, n gerade, $k = n/2$ schon getan und wollen hier ein allgemeineres Resultat diskutieren. Die Basis für die Approximation ist wieder die Stirlingsche Formel

$$n! \sim \left(\frac{n}{e}\right)^n \sqrt{2\pi n}.$$

Man beachte, daß aus der Approximation in Kapitel 5 $\lim_{n\to\infty} b(n; 2n, 1/2) = 0$ folgt. Dabei ist $b(n; 2n, 1/2)$ noch am größten unter den $b(k; 2n, 1/2)$, wie man aus dem Pascalschen Dreieck für die Binomialkoeffizienten ersehen kann. Allgemeiner gilt für ein beliebiges $p \in (0, 1)$

$$\lim_{n\to\infty} \max_{0\le k\le n} b(k; n, p) = 0. \tag{6.18}$$

Wir werden das später beweisen.

Es bezeichne S_n die Anzahl der Erfolge in einem Bernoulli-Experiment der Länge n und Erfolgswahrscheinlichkeit p. Dann gilt nach (6.18) für jedes $a > 0$:

$$\lim_{n\to\infty} P(|S_n - np| \leq a) = 0,$$

obschon nach dem Gesetz der großen Zahlen für jedes $\varepsilon > 0$

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - p\right| \leq \varepsilon\right) = \lim_{n\to\infty} P(|S_n - np| \leq \varepsilon n) = 1$$

gilt.

Um eine hinreichende Anzahl von möglichen Werten für S_n zu „erwischen", die dann zu einer von 0 verschiedenen Grenzwahrscheinlichkeit führen, muß der zugelassene Bereich mit n unendlich groß werden. Da wir (zumindest für $p = 1/2$) gesehen haben, daß die größten $b(k; n, p)$ von der Größenordnung $1/\sqrt{n}$ sind, ist es plausibel, wenn sich für die benötigte Anzahl eine Größenordnung von $\sqrt{n}$ ergibt und mithin die Größenordnung n aus dem Gesetz der großen Zahlen zu großzügig bemessen ist.

Zunächst müssen wir die $b(k; n, p)$ einzeln genauer unter die Lupe nehmen: Es zeigt sich, daß $b(k; n, p)$ durch eine schöne Funktion approximiert werden kann, wenn k in einer Umgebung der Größenordnung $\sqrt{n}$ um den Erwartungswert np liegt. Es ist deshalb günstig, eine Variablentransformation durchzuführen, indem k durch $(k - np)/\sqrt{n}$ ersetzt wird. Aus technischen Gründen ist es besser, noch durch $\sqrt{p(1-p)}$ zu dividieren. Wir setzen also

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

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.

Der untenstehende Satz besagt, daß $b(k; n, p)$ für große n

$$\sim \frac{1}{\sqrt{2\pi npq}} e^{-x_k^2/2}$$

ist. Was soll das aber genau heißen?

Zu schreiben

$$\lim_{n\to\infty} \sqrt{2\pi npq}\, b(k; n, p) = e^{-x_k^2/2},$$

ergibt keinen großen Sinn, denn mit $n \to \infty$ und festem k geht $x_k \to -\infty$. Was wir wollen, ist eine Approximation, wenn k in einer Umgebung der Größenordnung $\sqrt{n}$ um np liegt, das heißt x_k in einer Umgebung der Größenordnung 1 um 0. Es zeigt sich, daß in einem solchen Bereich die Konvergenz sogar gleichmäßig gilt. Wir können den Bereich sogar in n größer werden lassen:

(6.19) 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

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

(6.20) Bemerkungen.

(1) Ist $a_n = A$ eine beliebige, aber feste positive Konstante, so folgt 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 und dies auch 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 (6.19) 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.

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

Daraus folgt die Aussage sofort. □

Beweis von Satz (6.19). Es gilt

(i) $$k = np + \sqrt{npq}\, x_k,\;\; n-k = nq - \sqrt{npq}\, x_k,$$

also

(ii) $$k \sim np,\;\; n-k \sim nq.$$

Mit Hilfe der Stirlingschen Formel folgt:

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

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

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

(wir kennen diese Funktion schon recht gut aus Kapitel 4). 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 wir mit ψ das Restglied in der Taylorentwicklung bezeichnen (man denke sich zum Beispiel die Lagrangesche Form), wir also in der Umgebung von p eine Abschätzung

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

mit einer geeigneten Konstante c zur Verfügung haben. Wir erhalten somit

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

Man beachte nun, daß aus der Definition der x_k 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 die rechte Seite der Ungleichung gegen 0 für $n \to \infty$. Dies folgt unmittelbar aus der Bedingung an die Folge $(a_n)_{n\in\mathbb{N}}$. Da nun aber

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

erhalten wir

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

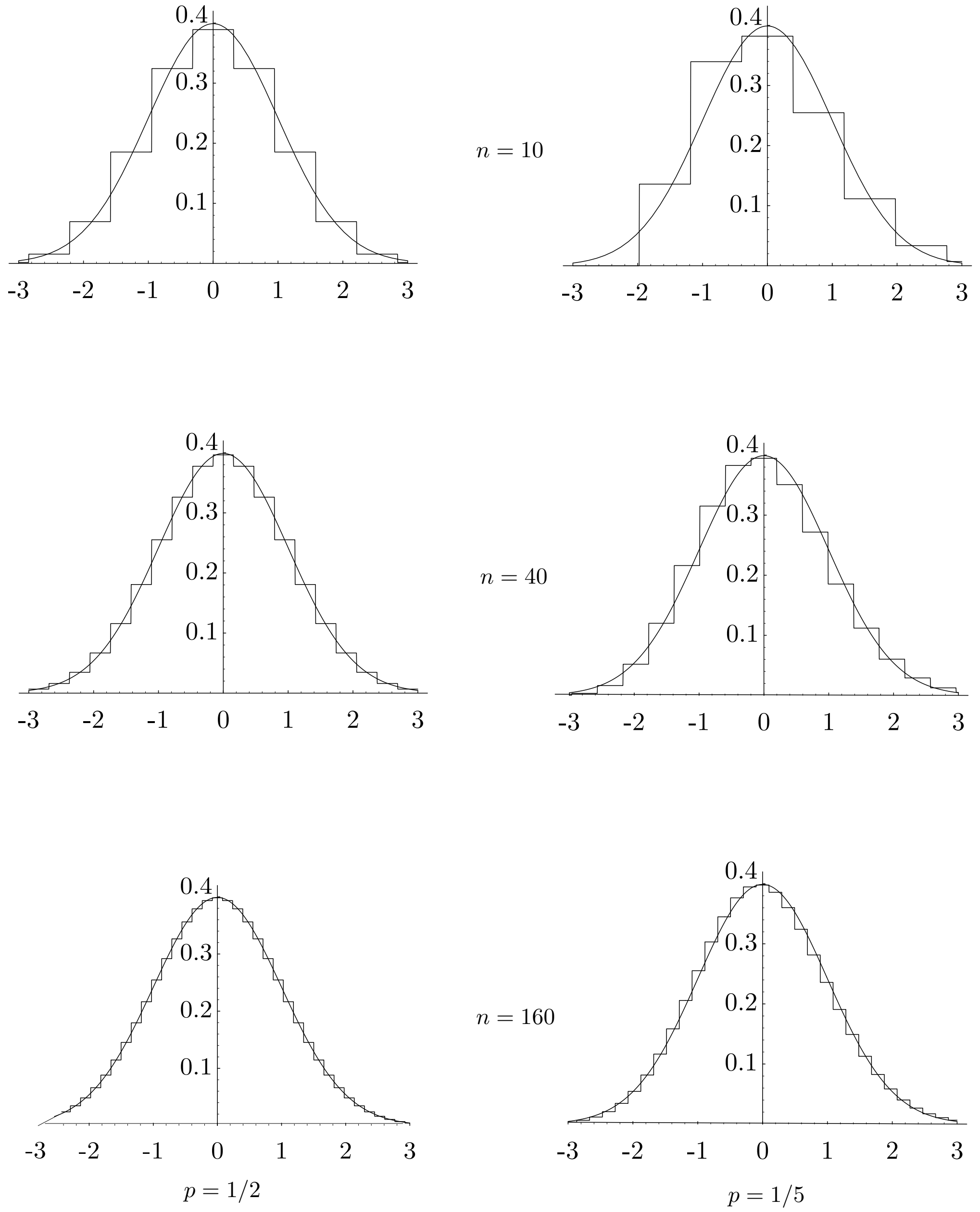
Nach (iii) ist damit der Satz gezeigt. $\square$

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 $k = 200, 250$, $n = 1200$, $p = 1/6$.

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

Nachfolgend ist eine numerische Illustration von (6.19) 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

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

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 $p = 1/5$.

Wir können noch den versprochenen Beweis von (6.18) nachholen:

Beweis. Da die Standardabweichung von S_n gleich $\sqrt{npq}$ ist, so folgt aus der Tschebyscheff-Ungleichung, daß für $a > 0$ $P(|S_n - np| \geq a\sqrt{npq}) \leq 1/a^2$ gilt. Da die

Summe nicht negativer Zahlen mindestens so groß wie jedes ihrer Glieder ist, folgt $b(k;n,p) \leq 1/a^2$ für $|x_k| \geq a$. Andererseits ergibt sich aus (6.19):

$$\lim_{n\to\infty} \sup_{k\,:\,|x_k|\leq a} b(k;n,p) = 0.$$

Da a beliebig ist, folgt (6.18). $\square$

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.

Da die Wahrscheinlichkeiten $b(k;n,p)$ für große n somit alle klein sind, ist man meist eher daran interessiert, Summen dieser Wahrscheinlichkeiten zu approximieren.

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

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

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

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

Wir setzen nun für jeden Summanden auf der rechten Seite seinen in Satz (6.19) 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:

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

Unter Verwendung der Gleichmäßigkeit der Konvergenz in $\{k\colon |x_k| \leq A\}$ für jedes A, sieht man sofort, daß der Quotient von $P\left(a \leq \frac{S_n-np}{\sqrt{npq}} \leq b\right)$ und dem obenstehenden Ausdruck gegen 1 konvergiert, das heißt, es existiert eine Nullfolge $(\varepsilon_n)_{n\in\mathbb{N}}$, $\varepsilon_n > 0$ mit

$$(*) \qquad R_n(1 - \varepsilon_n) \leq P\left(a \leq \frac{S_n - np}{\sqrt{npq}} \leq b\right) \leq R_n(1 + \varepsilon_n).$$

Die Beziehung zwischen k und x_k ist umkehrbar eindeutig, 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 $a \leq x_k \leq b$ gleich j bzw. l ist, dann ist

$$x_{j-1} < a \leq x_j < x_{j+1} < \cdots < x_{l-1} < x_l \leq b < x_{l+1}$$

und die obige Summe läßt sich schreiben als

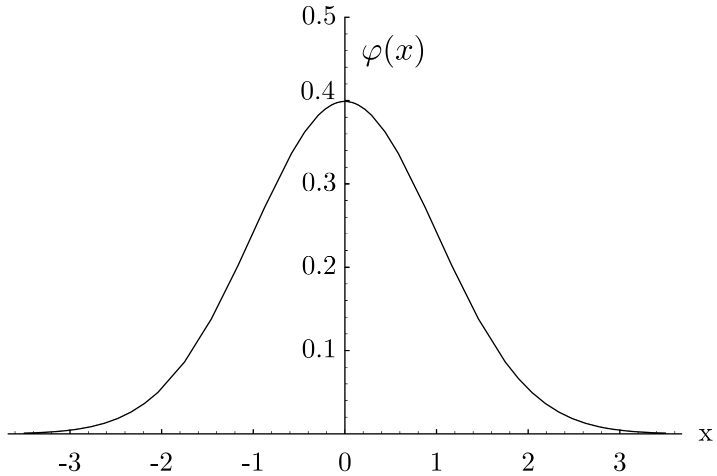
$$\sum_{k=j}^{l} \varphi(x_k)(x_{k+1} - x_k),$$

wobei $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ ist. Das ist eine Riemann'sche Summe für das bestimmte Integral $\int_a^b \varphi(x)dx$, obwohl üblicherweise in der Lehrbuchliteratur beim Riemann-Integral die Endpunkte a und b mit zu den Teilpunkten gehören.

Doch macht das für $n \to \infty$ und eine dadurch immer feiner werdende Unterteilung keinen Unterschied. Somit konvergiert R_n gegen das Integral in der Behauptung des Satzes. Dieser folgt nun sofort mit (*). □

Abraham de Moivre (1667–1754) veröffentlichte dieses Ergebnis in seiner „*Doctrine of Chances*“ 1714. Offensichtlich gebührt ihm die Priorität gegenüber James Stirling hinsichtlich der nach letzterem benannten Formel. *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 *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.



Carl Friedrich Gauss (1777–1855) hat sich intensiv mit dem zentralen Grenzwertsatz und der sogenannten Normalverteilung beschäftigt.

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. (Es ist leicht ersichtlich, daß das uneigentliche Integral konvergiert.) Wie nicht anders zu erwarten ist, gilt

$$\int_{-\infty}^{\infty} \varphi(x)dx = 1. \tag{6.22}$$

Wir geben einen Beweis: Dazu verwenden wir (6.21) und setzen $S_n^* := \frac{S_n - np}{\sqrt{npq}}$. (Für das Argument hier spielt p keine Rolle; wir könnten $p = 1/2$ 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 (6.21) 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 (6.22) bewiesen ist.

(6.23) Bemerkung.

(a) Wegen $\lim_{n\to\infty} \sup_k P(S_n = k) = 0$ ist es natürlich gleichgültig, ob in der Aussage von (6.21) $\le$ oder $<$ steht.

(b) Es gilt für $a \in \mathbb{R}$:

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

Beweis von (b). Wir beweisen die erste Gleichung; die zweite folgt analog. Wegen der Symmetrie von φ und (6.22) 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 (6.21)

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

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

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	1	2	3	4	5	6	7	8	9
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	0.8686	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	0.9756	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	0.9989	0.9989	0.9990

Der Satz (6.21) ist eine Präzisierung des Gesetzes der großen Zahlen, welches besagt, daß für jedes $\varepsilon > 0$ $\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - p\right| \geq \varepsilon\right) = 0$ ist. Letzteres können wir sofort auch aus (6.21) 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. (6.21) ist jedoch sehr viel informativer als das Gesetz der großen Zahlen.

Ein Anwendungsbeispiel
Eine Fabrik stellt ein Werkstück her mit einer Ausschußrate von 10%. Mit welcher Wahrscheinlichkeit sind unter 400 produzierten mehr als 50 defekt?

$n = 400$, $p = 0.1$, $np = 40$, $\sqrt{np(1-p)} = 6$

$$P(S_n > 50) = P\left(\frac{S_n - np}{\sqrt{np(1-p)}} > \frac{5}{3}\right) \cong 1 - \Phi\left(\frac{5}{3}\right) = \Phi\left(-\frac{5}{3}\right) \cong 0,05.$$

Mit welcher Wahrscheinlichkeit sind zwischen 35 und 45 defekt?

$$\begin{aligned} P(35 \leq S_n \leq 45) &= P\left(-\frac{5}{6} \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq \frac{5}{6}\right) \\ &\cong \Phi\left(\frac{5}{6}\right) - \Phi\left(-\frac{5}{6}\right) = 1 - 2\Phi\left(-\frac{5}{6}\right) \cong 0,6. \end{aligned}$$

Da wir keine Fehlerabschätzungen hergeleitet haben, wissen wir natürlich nicht, wie genau solche Näherungen sind. Die Genauigkeit ist etwas besser, wenn man die Mitte der möglichen Grenzpunkte nimmt; das heißt, im obigen Beispiel schreibt man besser:

$$\begin{aligned} P(S_n > 50) &= P(S_n \geq 50,5) \cong 1 - \Phi\left(\frac{21}{12}\right) \\ P(35 \leq S_n \leq 45) &= P(34,5 \leq S_n \leq 45,5) \cong 1 - 2\Phi\left(\frac{11}{12}\right). \end{aligned}$$

Für $n \to \infty$ ist die Korrektur natürlich belanglos.

(6.21) läßt sich weitgehend verallgemeinern. Wir zitieren den folgenden Satz ohne Beweis. Ein allgemeineres Ergebnis wird in einer Vorlesung „Wahrscheinlichkeitstheorie“ bewiesen.

(6.24) Satz *(Zentraler Grenzwertsatz).* Für jedes $n \in \mathbb{N}$ seien auf einem W.-Raum (Ω_n, p_n) n unabhängige Zufallsgrößen $X_1^{(n)}, \ldots, X_n^{(n)}$ definiert, die für alle i und n dieselbe Verteilung haben und deren Erwartungswert E und Varianz V existieren (notwendigerweise für alle $X_i^{(n)}$ dieselben). Dann gilt für jedes $a \in \mathbb{R}$

$$\lim_{n\to\infty} P\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i^{(n)} - E}{\sqrt{V}} \leq a\right) = \Phi(a).$$

§7 Allgemeine Wahrscheinlichkeitsräume und Zufallsgrössen mit Dichten

Im letzten Kapitel 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,

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

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 6 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

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

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. Wir werden dies nur kurz diskutieren, weil es für uns im folgenden keine große Rolle spielen wird.

Zu einer beliebigen Menge Ω möchte man - wie im diskreten Fall - allen Teilmengen A von Ω eine Wahrscheinlichkeit $P(A)$ zuordnen, und dabei natürlich die gewohnten Eigenschaften, die wir im ersten Kapitel kennengelernt haben, erhalten. Dabei trifft man aber auf mathematische Hindernisse. Wir müssen an dieser Stelle einfach akzeptieren, daß es sich bewährt hat, für die folgenden Familien von Teilmengen von Ω Wahrscheinlichkeiten zu betrachten.

(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.

(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 $a \leq b$ (d.h. $a_i \leq b_i$ für alle i) ein nach links halboffenes Intervall definiert durch

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

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) $P(A) \geq 0$ 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

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

$(\Omega, \mathcal{F}, P)$ 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 $\{a < X \leq b\}$ 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 $\mathcal{F}_X := \{A \subset \mathbb{R} \colon X^{-1}(A) \in \mathcal{F}\}$ 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". Unsere Definition ist bequemer.

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.)

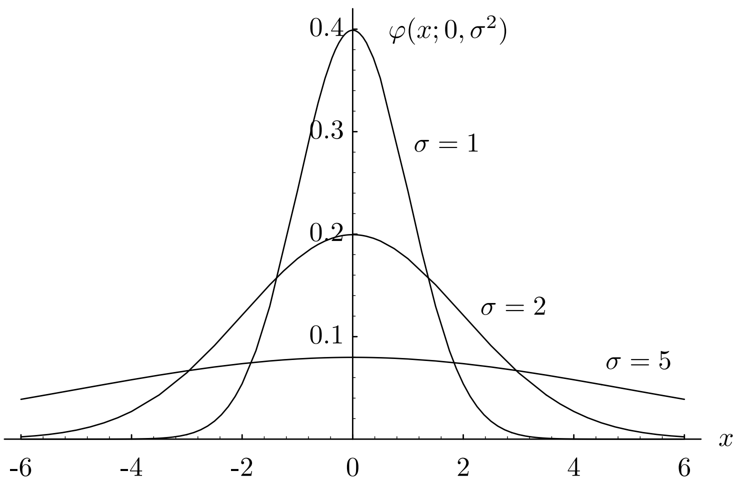
Falls das Lebesgue-Integral nicht bekannt ist, so setze man voraus, daß f Riemann-integrierbar ist und das uneigentliche Riemann-Integral $\int_{-\infty}^{\infty} f(x)\, dx$ existiert und gleich 1 ist.

(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

$$f(x) = \begin{cases} 1/(b-a) & \text{für } x \in [a, b], \\ 0 & \text{für } x \in \mathbb{R} \setminus [a, b]. \end{cases}$$

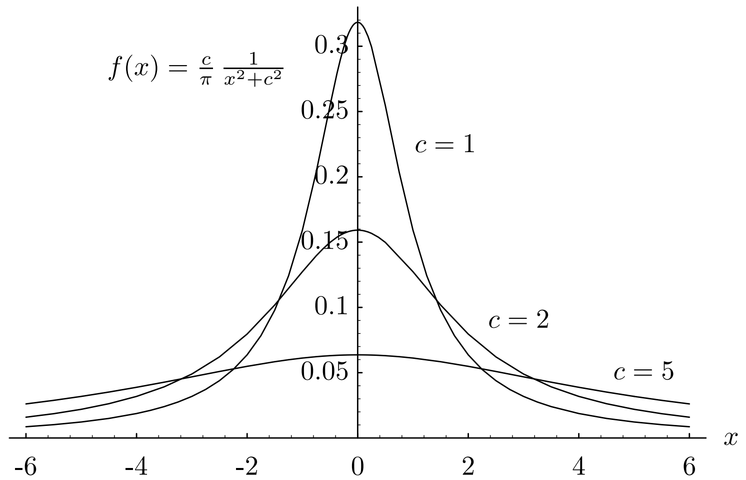
(4) Die Dichte der *Exponentialverteilung (exponential distribution) zum Parameter* $\lambda > 0$ ist definiert durch

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{für } x \geq 0, \\ 0 & \text{für } x < 0. \end{cases}$$

(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. (a) 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 $s \leq t$ gilt $F(s) \leq F(t)$.

(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 $t_n \geq t$ 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 $\lim_{t\to-\infty} F(t) = 0$.

(b) Es sei f eine Dichte. Eine Verteilungsfunktion F heißt *absolutstetig (absolutely continuous)* mit Dichte f, wenn

$$F(t) = \int_{-\infty}^{t} f(s)\, ds$$

für jedes $t \in \mathbb{R}$ gilt.

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.

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 ändern an den Integralen nichts.

Eine absolutstetige Verteilungsfunktion 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 $a < b$

$$P(a < X \le b) = P(X \le b) - P(X \le a) = \int_a^b f(x)\, dx.$$

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

$$P(a < X \le b) = P(a \le X \le b) = P(a \le X < b) = P(a < X < b).$$

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 $g\colon \mathbb{R} \to \mathbb{R}$ 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

$$V(X) = \int_{-\infty}^{\infty} (x - E(X))^2 f(x)\,dx.$$

Wer die Konstruktion des Lebesgue-Integrals kennt, wird 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

$$\int_{-\infty}^{\infty} |x| \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx = \frac{2}{\sqrt{2\pi}} \int_0^{\infty} x e^{-x^2/2}\,dx = \frac{2}{\sqrt{2\pi}} \big(-e^{-x^2/2}\big)\Big|_0^{\infty} = \sqrt{\frac{2}{\pi}} < \infty,$$

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

$$V(X) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2}\,dx = \lim_{N\to\infty} \frac{1}{\sqrt{2\pi}} \int_{-N}^{N} x(xe^{-x^2/2})\,dx,$$

und mittels partieller Integration folgt

$$V(X) = \lim_{N\to\infty} \frac{1}{\sqrt{2\pi}} (-xe^{-x^2/2})\Big|_{-N}^{N} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2}\,dx = 0 + 1 = 1.$$

(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)

$$\begin{aligned}\int_{-\infty}^{\infty} |x| \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}\, dx &= \int_{-\infty}^{\infty} |\mu + \sigma y| \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy \\ &\le |\mu| + \sigma \int_{-\infty}^{\infty} |y| \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy < \infty,\end{aligned}$$

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 $P(X \le t) = P(aX + b \le at + b)$ 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

$$E(X) = \int_0^{\infty} \lambda x e^{-\lambda x}\, dx = -xe^{-\lambda x}\Big|_0^{\infty} + \int_0^{\infty} e^{-\lambda x} dx = 0 + \Big(-\frac{1}{\lambda} e^{-\lambda x}\Big)\Big|_0^{\infty} = \frac{1}{\lambda},$$

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 $(x_1, \dots, x_n)$ aus dem $\mathbb{R}^n$ bezeichnet.

(b) Die Funktion f sei eine n-dimensionale Dichte, und $X_1, \dots, X_n$ seien n Zufallsgrößen. Man sagt, daß sie die *gemeinsame Dichte (joint density)* f haben, wenn

$$P(X_1 \le a_1,\ X_2 \le a_2, \dots, X_n \le a_n) = \int_{(-\infty,a_1]\times\cdots\times(-\infty,a_n]} f(x)\,dx$$

für alle $a_1, \dots, a_n \in \mathbb{R}$ gilt.

(7.16) Definition. $X_1, \dots, X_n$ seien n Zufallsgrößen, definiert auf einem gemeinsamen W.-Raum $(\Omega, \mathcal{F}, P)$. Sie heißen *unabhängig*, wenn für alle $a_1, \dots, a_n \in \mathbb{R}$ gilt:

$$P(X_1 \le a_1, \dots, X_n \le a_n) = P(X_1 \le a_1) \cdots P(X_n \le a_n).$$

Bemerkung. Man prüft leicht nach, daß diese Definition für diskrete Zufallsgrößen äquivalent zu der in Kapitel 3 gegebenen ist.

(7.17) Satz. $X_1, \dots, X_n$ 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 $X_1, \dots, X_n$ genau dann unabhängig, wenn eine gemeinsame Dichte für $X_1, \dots, X_n$ durch $\mathbb{R}^n \ni (x_1, x_2, \dots, x_n) \mapsto f_1(x_1) f_2(x_2) \dots f_n(x_n)$ gegeben ist.

Beweis. Ist $\mathbb{R}^n \ni (x_1, x_2, \dots, x_n) \mapsto f_1(x_1) f_2(x_2) \dots f_n(x_n)$ eine gemeinsame Dichte, so ergibt sich für alle $a_1, \dots, a_n \in \mathbb{R}$

$$\begin{aligned} P(X_1 \le a_1, \dots, X_n \le a_n) &= \int_{-\infty}^{a_1} \cdots \int_{-\infty}^{a_n} f_1(x_1) \dots f_n(x_n)\,dx_n \dots dx_1 \\ &= \prod_{j=1}^n \int_{-\infty}^{a_j} f_j(x_j)\,dx_j = \prod_{j=1}^n P(X_j \le a_j). \end{aligned}$$

Somit sind $X_1, \dots, X_n$ unabhängig. Gilt umgekehrt letzteres, so folgt

$$\begin{aligned} P(X_1 \le a_1, \dots, X_n \le a_n) &= \prod_{j=1}^n P(X_j \le a_j) \\ &= \prod_{j=1}^n \int_{-\infty}^{a_j} f_j(x_j)\,dx_j \\ &= \int_{-\infty}^{a_1} \cdots \int_{-\infty}^{a_n} f_1(x_1) \dots f_n(x_n)\,dx_n \dots dx_1, \end{aligned}$$

und somit ist $\mathbb{R}^n \ni (x_1, \dots, x_n) \mapsto f_1(x_1) \dots f_n(x_n)$ eine gemeinsame Dichte. $\square$

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 $P(X + Y \leq a)$ für alle $a \in \mathbb{R}$ bestimmen. Mit $C_a := \{(x, y) \in \mathbb{R}^2 : x + y \leq a\}$ 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):

$$\begin{aligned} P(X + Y \leq a) &= \int_{C_a} f(x) g(y)\, dx\, dy \\ &= \int_{-\infty}^{a} \int_{-\infty}^{\infty} f(u - v) g(v)\, dv\, du. \end{aligned}$$

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

$$(*) \qquad h(x) = \int_{-\infty}^{\infty} f(x - y) g(y)\, dy, \quad x \in \mathbb{R}.$$

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, $1 \leq i \leq n$, 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 $\sigma_1^2 + \sigma_2^2$. $\square$

§8 Rekurrente Ereignisse, Erneuerungstheorie

Wir wollen zunächst ein wichtiges Hilfsmittel für das Studium von Verteilungen auf den nichtnegativen ganzen Zahlen bereitstellen. Eine Arbeit aus dem Jahre 1782 von *Pierre Simon Marquis de Laplace* (1749–1827) ist eine wichtige Grundlage für die Theorie der sogenannten erzeugenden Funktionen:

(8.1) Definition. Es sei $a_0, a_1, a_2, \dots$ eine Folge reeller Zahlen. Die Funktion $A(s) := \sum_{j=0}^{\infty} a_j s^j$ bezeichnet man als die *erzeugende Funktion* der Folge (a_j). Der Definitionsbereich von A ist die Menge der Punkte $s \in \mathbb{R}$, für die die Reihe konvergiert. Es sei X eine diskrete Zufallsgröße mit Werten in $\mathbb{N}_0$. Dann heißt $G(s) := \sum_{j=0}^{\infty} P(X=j)s^j$ die *erzeugende Funktion (generating function)* von X.

(8.2) Bemerkung. Offensichtlich ist $G(s) = E(s^X)$. Da die Koeffizienten nichtnegativ sind und ihre Summe 1 ist, konvergiert die Reihe mindestens für alle s mit $|s| \leq 1$ (Abelscher Grenzwertsatz). Ist $G^{(n)}(s)$ die n-te Ableitung von G an der Stelle s und $G^{(0)}(s) = G(s)$, so ist nach der gewöhnlichen Potenzreihenentwicklung

$$P(X=n) = \frac{G^{(n)}(0)}{n!}.$$

Die Beziehung zwischen der erzeugenden Funktion von X und der Verteilung von X ist also umkehrbar eindeutig.

(8.3) Beispiele.

(a) Für die Binomialverteilung mit den Parametern $n \in \mathbb{N}$ und $p \in [0,1]$ hat die erzeugende Funktion $\mathbb{R}$ als Definitionsbereich, und es gilt

$$G(s) = \sum_{j=0}^{n} \binom{n}{j} p^j (1-p)^{n-j} s^j = (1-p+sp)^n, \quad s \in \mathbb{R}.$$

(b) Für die Poissonverteilung mit Parameter $\lambda > 0$ existiert die erzeugende Funktion auf ganz $\mathbb{R}$, und es gilt

$$G(s) = \sum_{j=0}^{\infty} \frac{(\lambda s)^j}{j!} e^{-\lambda} = e^{\lambda(s-1)}.$$

(8.4) Satz. Es seien $X_1, X_2, \dots, X_n$ unabhängige diskrete Zufallsgrößen mit Werten in $\mathbb{N}_0$. Sind $G_1, \dots, G_n$ die erzeugenden Funktionen von $X_1, \dots, X_n$ und ist G die erzeugende Funktion der Summe $X_1 + \dots + X_n$, so gilt $G(s) = G_1(s)G_2(s)\cdots G_n(s)$ für diejenigen $s \in \mathbb{R}$, für die die Reihen $G_1(s), \dots, G_n(s)$ absolut konvergieren, mindestens also für alle s mit $|s| \leq 1$.

Beweis. Nach Satz (3.25) sind $s^{X_1}, \dots, s^{X_n}$ unabhängige Zufallsgrößen. Somit folgt mit Bemerkung (3.27) $E\big(s^{X_1+\dots+X_n}\big) = E\big(s^{X_1}\big)\cdots E\big(s^{X_n}\big)$. Dabei verwenden wir aus der Theorie der Potenzreihen das folgende Resultat: Konvergieren die Reihen $\sum_{k=0}^{\infty} a_k s^k$ und $\sum_{k=0}^{\infty} b_k s^k$ für ein $s \in \mathbb{R}$ absolut, so ist

$$\left(\sum_{k=0}^{\infty} a_k s^k\right)\left(\sum_{k=0}^{\infty} b_k s^k\right) = \sum_{k=0}^{\infty} c_k s^k \quad \text{mit} \quad c_k := \sum_{j=0}^{k} a_j b_{k-j},$$

und die Reihe $\sum_{k=0}^{\infty} c_k s^k$ konvergiert ebenfalls absolut. $\square$

(8.5) Satz. Sei X eine Zufallsgröße mit Werten in $\mathbb{N}_0$ und erzeugender Funktion G. $G'(1-) := \lim_{s\uparrow 1} G'(s)$ existiert und ist kleiner als unendlich genau dann, wenn $E(X)$ existiert, und es gilt dann $G'(1-) = E(X)$.

Beweis. Sei $p_k := P(X = k)$ für $k \in \mathbb{N}_0$. Für $|s| < 1$ existiert $G'(s)$, und es gilt

$$G'(s) = \sum_{k=0}^{\infty} k p_k s^{k-1} \uparrow \sum_{k=0}^{\infty} k p_k \quad \text{für } s \uparrow 1.$$

$\square$

Erneuerungstheorie (renewal theory)
Auf einer diskreten Zeitachse $\mathbb{N}_0$ betrachten wir zufällige Punkte, die die Zeitpunkte des Eintretens eines zufälligen Geschehens sein sollen. Formal können wir diese zufälligen Punkte durch eine Folge von Zufallsgrößen Z_i, $i \in \mathbb{N}_0$, beschreiben, die nur die Werte 0 oder 1 annehmen. Dabei soll $Z_i = 1$ bedeuten, daß i durch einen zufälligen Punkt besetzt ist. Wir nehmen an, daß 0 stets besetzt ist, also daß $Z_0 = 1$ gilt.

Die Z_i, $i \in \mathbb{N}_0$, werden nicht als unabhängig vorausgesetzt, wohl aber die Zwischenzeiten. Ausgangspunkt der formalen Konstruktion sind nicht die Z_i sondern eine Folge $T_1, T_2, \ldots$ von unabhängigen Zufallsgrößen mit Werten in $\mathbb{N}$, die alle die gleiche Verteilung haben sollen. Für jedes $i \in \mathbb{N}$ sei $f_i := P(T_k = i)$. Mit Hilfe der Folge $T_1, T_2, \ldots$ definieren wir für jedes $n \in \mathbb{N}_0$ den Zeitpunkt τ_n des n-ten zufälligen Punktes durch

$$\tau_n = \sum_{k=1}^{n} T_k$$

und setzen $\tau_0 = 0$. Ferner definieren wir

$$Z_i = \begin{cases} 1, & \text{falls ein } n \in \mathbb{N}_0 \text{ existiert mit } \tau_n = i, \\ 0 & \text{sonst.} \end{cases}$$

Man kann dies wie folgt veranschaulichen: T_1 sei die Lebensdauer einer Glühbirne. In dem Moment, wo sie durchbrennt, wird sie durch eine zweite Glühbirne mit Lebensdauer T_2 ersetzt, u.s.w. Die n-te Glühbirne muß dann zum Zeitpunkt τ_n „erneuert“ werden.

Wir wollen unser Modell etwas verallgemeinern und die Möglichkeit zulassen, daß eine Glühbirne nie ausbrennt, also eine unendliche Lebensdauer hat. Dies bedeutet einfach, daß sich die f_i nicht zu 1 aufsummieren müssen. Wir stellen also an die Folge $(f_i)_{i\in\mathbb{N}}$ die Bedingungen

$$\sum_{i=1}^{\infty} f_i \leq 1$$

und $f_i \geq 0$ für alle $i \in \mathbb{N}$. (Wir könnten formal das Symbol „∞“ als möglichen Wert der Zufallsgrößen T_k, $k \in \mathbb{N}$, zulassen und die zugehörige Wahrscheinlichkeit durch $P(T_k = \infty) = 1 - \sum_{i=1}^{\infty} f_i$ definieren.) Eine Folge $(f_i)_{i\in\mathbb{N}}$ mit $f_i \geq 0$ und $\sum_{i\in\mathbb{N}} f_i \leq 1$ nennen wir eine *Sub-Wahrscheinlichkeitsverteilung*.

Es sei nun $u_n := P(Z_n = 1)$ für jedes $n \in \mathbb{N}_0$. Insbesondere gilt $u_0 = 1$. Man beachte auch, daß jedes f_n mit Hilfe der Zufallsgrößen $Z_1, Z_2, \ldots, Z_n$ mittels $f_n = P(Z_1 = 0, \ldots, Z_{n-1} = 0,\ Z_n = 1)$ definiert werden kann.

(8.6) Satz. Es gilt die *Erneuerungsgleichung (renewal equality)*

$$u_n = \sum_{k=1}^{n} f_k u_{n-k} \quad \text{für alle } n \in \mathbb{N}.$$

Wir nennen $(u_n)_{n\in\mathbb{N}_0}$ die zu $(f_n)_{n\in\mathbb{N}}$ gehörende *Erneuerungsfolge (renewal sequence).*

Beweis. Für jedes $n \in \mathbb{N}$ folgt durch Aufspalten nach dem Zeitpunkt der ersten Erneuerung

$$u_n = P(Z_n = 1) = \sum_{k=1}^{n-1} P(T_1 = k,\, Z_n = 1) + P(T_1 = n).$$

Für $1 \leq k \leq n-1$ gilt

$$\begin{aligned} P(T_1 = k,\, Z_n = 1) &= P\Big(T_1 = k,\, \exists\, l \in \mathbb{N} \text{ mit } \sum_{j=1}^{l} T_{1+j} = n-k\Big) \\ &= P(T_1 = k) P\Big(\exists\, l \in \mathbb{N} \text{ mit } \sum_{j=1}^{l} T_j = n-k\Big) \\ &= f_k u_{n-k}, \end{aligned}$$

wobei sich die zweite Gleichung aus der Voraussetzung ergibt, daß die Zwischenzeiten $T_1, T_2, \dots$ unabhängig sind und alle die gleiche Verteilung haben. $\square$

(8.7) Bemerkung. Um den Index der beiden Folgen nicht über verschiedene Bereiche laufen lassen zu müssen, setzt man meist $f_0 = 0$. Wir schreiben im folgenden einfach (f_n) bzw. (u_n). Ein *Beispiel* für eine Erneuerungsfolge haben wir schon in Kapitel 5 kennengelernt: Für die eindimensionale Irrfahrt $(S_n)_{n\in\mathbb{N}_0}$ war $u_n = P(S_n = 0)$ und $f_n = P(\, S_1 \neq 0,\, S_2 \neq 0, \dots,\, S_{n-1} \neq 0,\, S_n = 0\,)$. Die Erneuerungsgleichung ist einfach eine algebraische Beziehung zwischen den beiden Folgen (u_n) und (f_n) (stets mit $u_0 = 1$ und $f_0 = 0$). Mit Hilfe dieser Gleichung läßt sich (u_n) rekursiv aus (f_n) berechnen und umgekehrt. Aus $f_n \geq 0$ für alle $n \in \mathbb{N}_0$ und $\sum_{n=1}^{\infty} f_n \leq 1$ folgt mit Induktion sofort $0 \leq u_n \leq 1$ für alle n.

Die im Rest dieses Kapitels durchgeführten Überlegungen sind rein analytischer Art, die man völlig losgelöst von der wahrscheinlichkeitstheoretischen Interpretation durchführen kann: Gegeben sei eine Sub-Wahrscheinlichkeitsverteilung (f_n). Wir wollen Eigenschaften der zugehörigen Erneuerungsfolge (u_n) herleiten, wobei (u_n) über die Erneuerungsgleichung durch (f_n) bestimmt ist. Später kommen wir auf Beispiele aus dem Bereich der Wahrscheinlichkeitsrechnung zurück.

(8.8) Definition. Die Folge (f_n) und dann auch die zugehörige Erneuerungsfolge (u_n) heißen *rekurrent (recurrent)*, wenn $\sum_{n\in\mathbb{N}} f_n = 1$ gilt, andernfalls *transient (transient)*. Die Zahl $d := \mathrm{ggT}\{\, n \in \mathbb{N}_0 \mid f_n \neq 0 \,\}$ heißt die *Periode (period)* von (f_n). Die Folge (f_n) heißt *aperiodisch (aperiodic)*, wenn $d = 1$ ist.

Ist (f_n) ein Wahrscheinlichkeitsmaß auf $\mathbb{N}$, so ist (f_n) rekurrent. In der obigen Interpretation brennt dann jede Glühbirne sicher irgendwann einmal durch. Der größte gemeinsame Teiler der möglichen Lebensdauern einer Glühbirne ist die Periode. Da mit der Erneuerungsgleichung für alle n mit $f_n > 0$ auch $u_n > 0$ gilt, ist $d^* := \mathrm{ggT}\{\, n \in \mathbb{N}_0 \mid u_n \neq 0 \,\}$ Teiler von d. Andererseits sind die Lebensdauern aller Birnen Vielfache von d. Eine Erneuerung gibt es also nur zu Zeitpunkten, die Vielfache von d sind, das heißt d ist Teiler von d^*, womit $d = d^*$ folgt. Diese Gleichheit gilt für jede Sub-Wahrscheinlichkeitsverteilung (f_n) und die zugehörige Erneuerungsfolge (u_n):

(8.9) Lemma. *Sei d die Periode von (f_n). Dann gilt $d = \mathrm{ggT}\{\, n \in \mathbb{N}_0 \mid u_n \neq 0 \,\}$.*

Beweis. Wegen $u_n \geq f_n$ für alle $n \in \mathbb{N}$ folgt für alle n mit $f_n > 0$ auch $u_n > 0$, also ist $d \geq d^* := \mathrm{ggT}\{\, n \in \mathbb{N} \mid u_n \neq 0 \,\}$. Wir zeigen nun $d^* \geq d$. Per Definition ist $f_k = 0$ für $k \notin \{\, md \mid m \in \mathbb{N} \,\}$. Somit gilt

$$u_n = \sum_{m \,:\, md \leq n} f_{md} u_{n-md}$$

für jedes $n \in \mathbb{N}$. Daraus folgt mit vollständiger Induktion nach n sofort $u_n = 0$ für $n \notin \{\, md \mid m \in \mathbb{N} \,\}$ und somit $d^* \geq d$. $\square$

(8.10) Satz. Es seien $U(s) = \sum_{n=0}^{\infty} u_n s^n$ und $F(s) = \sum_{n=0}^{\infty} f_n s^n$ die erzeugenden Funktionen von (u_n) und (f_n). Dann gilt $U(s)(1 - F(s)) = 1$ für alle s, für die $U(s)$ und $F(s)$ existieren. (Bemerkung: $F(s)$ existiert mindestens für alle $|s| < 1$.)

Beweis. Nach der Erneuerungsgleichung und der Definition des Cauchy-Produkts gilt

$$F(s)\, U(s) = \sum_{n=1}^{\infty} u_n s^n = U(s) - 1.$$

Damit folgt aber sofort die Behauptung. $\square$

(8.11) Satz. Die Folge (f_n) ist genau dann rekurrent, wenn

$$u := \sum_{n=0}^{\infty} u_n = \infty$$

ist. Ist $u < \infty$, so gilt

$$f := \sum_{n=1}^{\infty} f_n = \frac{u-1}{u}.$$

Beweis. Die erzeugende Funktion $U(\cdot)$ ist monoton steigend auf $[0,1)$. Für jedes $N \in \mathbb{N}_0$ gilt

$$\sum_{n=0}^{N} u_n = \lim_{s \uparrow 1} \sum_{n=0}^{N} u_n s^n \leq \lim_{s \uparrow 1} U(s) \leq \sum_{n=0}^{\infty} u_n.$$

Somit ist

$$u = \lim_{s \uparrow 1} U(s).$$

Mit Satz (8.10) folgt

$$u < \infty \Leftrightarrow \lim_{s\uparrow 1} U(s) < \infty \Leftrightarrow \lim_{s\uparrow 1} F(s) < 1 \Leftrightarrow f < 1.$$

Ist $u < \infty$, so folgt $f = (u-1)/u$ ebenfalls aus Satz (8.10). □

Wir werden ein Beispiel kennenlernen (die ein- und zweidimensionalen Irrfahrt), wo im Fall der Rekurrenz $\lim_{n\to\infty} u_n = 0$ gilt. Die Beziehung von $\lim_{n\to\infty} u_n$ zum Verhalten der Folge (f_n) soll schon hier weiter untersucht werden.

(8.12) Definition. Die Folge (f_n) sei rekurrent.

(a) Die *mittlere Rückkehrzeit (mean of the recurrence time)* μ ist definiert durch $\mu = \sum_{n=1}^{\infty} n f_n$, falls die Reihe konvergiert, und durch $\mu = \infty$, falls die Reihe divergiert.

(b) Die Folge (f_n) heißt *positiv rekurrent,* wenn $\mu < \infty$ ist, andernfalls *nullrekurrent.*

(8.13) Satz *(Erneuerungssatz, renewal theorem)*. Die Folge (f_n) sei rekurrent und aperiodisch. Dann gilt

$$\lim_{n\to\infty} u_n = \frac{1}{\mu},$$

wobei $1/\infty = 0$ gesetzt wird.

Dieser Satz wurde in einer gemeinsamen Arbeit von *Paul Erdös* (1914-1996), *William Feller* (1906-1970) und *H. Pollard* im Jahre 1949 bewiesen. Die Interpretation des Erneuerungssatzes für unser Glühbirnenmodell ist durchaus einleuchtend. Hier ist μ die mittlere Lebensdauer einer Glühbirne. Somit werden n Glühbirnen etwa bis zur Zeit $n\mu$ reichen, also n Erneuerungen im Zeitraum $n\mu$ stattfinden. Im Mittel gibt es in jedem Zeitpunkt $1/\mu$ Erneuerungen, wenn wir Rekurrenz annehmen. Dies gilt im aperiodischen Fall.

Als Vorbereitung für den Beweis des Erneuerungssatzes benötigen wir eine Aussage über elementare Arithmetik.

(8.14) Lemma. *Sei Λ eine beliebige, nichtleere Teilmenge von $\mathbb{N}$ und $d := \mathrm{ggT}(\Lambda)$. Dann existieren $m, n_0 \in \mathbb{N}$ und $\lambda_1, \ldots, \lambda_m \in \Lambda$, so daß für jedes $n \geq n_0$ Zahlen $k_1, \ldots, k_m \in \mathbb{N}_0$ existieren mit*

$$nd = \sum_{j=1}^{m} k_j \lambda_j.$$

Beweis. Offensichtlich ist

$$G := \left\{ \sum_{j=1}^{m} r_j \lambda_j \;\middle|\; m \in \mathbb{N};\ r_1, \ldots, r_m \in \mathbb{Z};\ \lambda_1, \ldots, \lambda_m \in \Lambda \right\}$$

die kleinste additive Untergruppe von $\mathbb{Z}$, die Λ enthält. Sei d' das kleinste positive Element von G. Da d alle Elemente von Λ teilt, teilt d auch alle Elemente von G, insbesondere also auch d'. Dies zeigt, daß $d \leq d'$ ist. Zu jedem $n \in G$ existieren $r \in \mathbb{Z}$

und $s \in \{0, 1, \ldots, d'-1\}$ mit $n = rd' + s$. Da G eine Gruppe ist, muß $s = n - rd'$ in G sein. Aus der Definition von d' folgt $s = 0$. Diese Argumentation zeigt, daß d' alle Elemente von G teilt, insbesondere also diejenigen aus Λ. Somit folgt $d = d'$ und $d \in G$. Also existieren $m \in \mathbb{N}$ sowie $r_1, \ldots, r_m \in \mathbb{Z}$ und $\lambda_1, \ldots, \lambda_m \in \Lambda$ mit

$$d = \sum_{j=1}^{m} r_j \lambda_j.$$

Im allgemeinen können $r_1, \ldots, r_m$ jedoch negativ sein. Da d der größte gemeinsame Teiler der Zahlen in Λ ist, können wir für jedes $j \in \{1, \ldots, m\}$ die Zahl λ_j in der Form $b_j d$ mit $b_j \in \mathbb{N}$ schreiben. Sei $b := \min\{b_1, \ldots, b_m\}$ und $n_0 := \sum_{j=1}^{m} b|r_j|b_j$. Für jedes $n \geq n_0$ finden wir eine Darstellung der Form

$$n = s_1 b_1 + \cdots + s_m b_m + s$$

mit $s \in \{0, 1, \ldots, b-1\}$ und $s_j \geq b|r_j|$ für alle $j \in \{1, \ldots, m\}$. Dann gilt

$$nd = \sum_{j=1}^{m} (s_j + s r_j) \lambda_j,$$

wobei nun die Koeffizienten $k_j := s_j + s r_j$ alle in $\mathbb{N}_0$ sind. $\square$

Wir werden Lemma (8.14) im Beweis des Erneuerungssatzes auf die Menge $\Lambda_f := \{\, n \in \mathbb{N} \mid f_n \neq 0 \,\}$ anwenden. Vorher wollen wir eine Folgerung aus Lemma (8.14) notieren, die wir für den Beweis von Satz (9.19) benötigen werden.

(8.15) Lemma. *Es sei d die Periode von (f_n). Dann existiert für die zugehörige Erneuerungsfolge (u_n) ein $n_0 \in \mathbb{N}$, so daß $u_{nd} > 0$ für alle $n \geq n_0$ gilt.*

Beweis. Wir zeigen zunächst durch vollständige Induktion

(8.16) $$u_{n+m} \geq u_n u_m \quad \text{für alle } n, m \in \mathbb{N}_0.$$

Wegen $u_0 = 1$ gilt (8.16) für alle $n \in \mathbb{N}_0$ und $m = 0$. Für $m \geq 1$ folgt mit der Induktionsvoraussetzung

$$u_{n+m} = \sum_{k=1}^{n+m} f_k u_{n+m-k} \geq \sum_{k=1}^{m} f_k u_{n+m-k} \geq \sum_{k=1}^{m} f_k u_n u_{m-k} = u_n u_m.$$

Sei $\Lambda_u := \{\, n \in \mathbb{N} \mid u_n \neq 0 \,\}$. Gemäß Lemma (8.9) gilt $d = \mathrm{ggT}(\Lambda_u)$. Sind $m \in \mathbb{N}$ und $\lambda_1, \ldots, \lambda_m \in \Lambda_u$ sowie $k_1, \ldots, k_m \in \mathbb{N}_0$, so folgt durch mehrfache Anwendung von (8.16), daß $\sum_{j=1}^{m} k_j \lambda_j$ in Λ_u ist. Die Aussage des Lemmas folgt nun aus Lemma (8.14). $\square$

Beweis des Erneuerungssatzes.

1. Schritt: Für $n \in \mathbb{N}_0$ sei $r_n := \sum_{j=n+1}^{\infty} f_j$. Dann ist $r_0 = 1$ (wegen der Rekurrenz) und $\mu = \sum_{n=0}^{\infty} r_n$. Nach der Erneuerungsgleichung gilt für jedes $n \in \mathbb{N}$

$$u_n = \sum_{j=1}^{n} f_j u_{n-j} = \sum_{j=1}^{n} (r_{j-1} - r_j) u_{n-j} = \sum_{j=1}^{n} r_{j-1} u_{n-j} - \sum_{j=1}^{n} r_j u_{n-j}.$$

Da $r_0 = 1$ ist, ergibt sich durch Addition von $\sum_{j=1}^{n} r_j u_{n-j}$ auf beiden Seiten und einer Indexverschiebung

$$\sum_{j=0}^{n} r_j u_{n-j} = \sum_{j=0}^{n-1} r_j u_{n-1-j}.$$

Wegen $r_0 u_0 = 1$ ergibt sich aus dieser Gleichung mittels Induktion nach n, daß

$$\sum_{j=0}^{n} r_j u_{n-j} = 1$$

für alle $n \in \mathbb{N}_0$ gilt.

2. Schritt: Sei $\bar{\lambda} := \limsup_{n\to\infty} u_n$. Es gibt dann eine Teilfolge $(n_\nu)_{\nu\in\mathbb{N}_0}$ von $\mathbb{N}_0$ mit $u(n_\nu) \to \bar{\lambda}$ für $\nu \to \infty$, wobei wir $u(k) = u_k$ für alle $k \in \mathbb{N}_0$ setzen. Sei $j \in \mathbb{N}$ mit $f_j > 0$. Wir werden zeigen, daß $\lim_{\nu\to\infty} u(n_\nu - j) = \bar{\lambda}$ gilt.

Zu $\varepsilon > 0$ existiert $N \in \mathbb{N}_0$ mit $r_m < \varepsilon$ für alle $m \geq N$. Wir können natürlich $N \geq j$ wählen. Dann folgt aus der Erneuerungsgleichung

$$u(n_\nu) \leq \sum_{k=0}^{N} f_k u(n_\nu - k) + \varepsilon = \sum_{\substack{k=0\\k\neq j}}^{N} f_k u(n_\nu - k) + f_j u(n_\nu - j) + \varepsilon$$

und damit

$$\begin{aligned}\bar{\lambda} = \lim_{\nu\to\infty} u(n_\nu) &\leq \sum_{\substack{k=0\\k\neq j}}^{N} f_k \bar{\lambda} + f_j \liminf_{\nu\to\infty} u(n_\nu - j) + \varepsilon \\ &\leq \bar{\lambda} + f_j(\liminf_{\nu\to\infty} u(n_\nu - j) - \bar{\lambda}) + \varepsilon.\end{aligned}$$

Da $\varepsilon > 0$ beliebig war und $f_j > 0$ ist, ergibt sich $\bar{\lambda} \leq \liminf_{\nu\to\infty} u(n_\nu - j)$. Somit ist $\bar{\lambda} = \lim_{\nu\to\infty} u(n_\nu - j)$.

3. Schritt: Sind $j_1, \ldots, j_k \in \mathbb{N}$ mit $f_{j_1}, \ldots, f_{j_k} > 0$ und $m_1, \ldots, m_k \in \mathbb{N}_0$, so ergibt sich durch mehrfache Anwendung des zweiten Schrittes

$$\lim_{\nu\to\infty} u\left(n_\nu - \sum_{i=1}^{k} m_i j_i\right) = \bar{\lambda}$$

Da die Folge (f_n) aperiodisch ist, ergibt sich aus Lemma (8.14), daß ein $M \in \mathbb{N}_0$ existiert mit $\lim_{\nu\to\infty} u(n_\nu - j) = \bar{\lambda}$ für alle $j \geq M$.

Nach dem ersten Schritt folgt für alle $N \in \mathbb{N}_0$ und $n_\nu \geq M + N$

$$1 = \sum_{j=0}^{n_\nu - M} r_j u(n_\nu - M - j) \geq \sum_{j=0}^{N} r_j u(n_\nu - M - j).$$

Somit gilt $1 \geq \sum_{j=0}^{N} r_j \bar{\lambda}$ für alle $N \in \mathbb{N}_0$, das heißt $1 \geq \sum_{j=0}^{\infty} r_j \bar{\lambda} = \mu\bar{\lambda}$. Also ist $\bar{\lambda} \leq 1/\mu$.

4. Schritt: Wir müssen noch zeigen, daß

$$\underline{\lambda} := \liminf_{n\to\infty} u_n \geq \frac{1}{\mu}$$

ist. Der Fall $\mu = \infty$ ist trivial; wir können also $\mu < \infty$ voraussetzen. Es existiert eine Teilfolge (n'_ν) mit $\lim_{\nu\to\infty} u(n'_\nu) = \underline{\lambda}$. Dasselbe Argument wie oben zeigt, daß ein M' existiert mit $\lim_{\nu\to\infty} u(n'_\nu - j) = \underline{\lambda}$ für alle $j \geq M'$. Zu $\varepsilon > 0$ gibt es ein N_0, mit $\sum_{j=N_0+1}^{\infty} r_j \leq \varepsilon$ (wegen $\sum_{j=0}^{\infty} r_j = \mu < \infty$). Somit ist für $N \geq N_0$ und $n_\nu \geq M' + N$

$$1 = \sum_{j=0}^{n_\nu - M'} r_j u(n_\nu - M' - j) \leq \sum_{j=0}^{N} r_j u(n_\nu - M' - j) + \varepsilon.$$

Also folgt

$$1 \leq \sum_{j=0}^{N} r_j \underline{\lambda} + \varepsilon \leq \sum_{j=0}^{\infty} r_j \underline{\lambda} + \varepsilon = \mu\underline{\lambda} + \varepsilon.$$

Da $\varepsilon > 0$ beliebig war, folgt $\underline{\lambda} \geq 1/\mu$. $\square$

(8.17) Satz. Die Folge (f_n) sei rekurrent und habe die Periode d und die mittlere Rückkehrzeit μ. Dann gilt $\lim_{n\to\infty} u_{nd} = d/\mu$.

Beweis. Wir definieren $f'_k = f_{dk}$ und $u'_k = u_{dk}$ für alle $k \in \mathbb{N}_0$. Dann gilt die Erneuerungsgleichung für (f'_k) und (u'_k). Die Folge (f'_k) ist offensichtlich aperiodisch, und für ihre mittlere Rückkehrzeit μ' gilt

$$\mu' = \sum_{k=1}^{\infty} k f'_k = \frac{1}{d} \sum_{k=1}^{\infty} k f_k = \frac{\mu}{d}.$$

Die Aussage folgt nun durch Anwendung des Erneuerungssatzes (Satz 8.13) auf die Folgen (f'_k) und (u'_k). $\square$

(8.18) Beispiele. (a) *Rückkehr der eindimensionalen Irrfahrt zum Ursprung*
Wir betrachten die *verallgemeinerte* eindimensionale Irrfahrt $(S_n)_{n\in\mathbb{N}_0}$: $S_n = \sum_{j=1}^{n} X_j$, wobei die X_j unabhängige Zufallsgrößen sind mit $P(X_j = 1) = p$ und $P(X_j = -1) = 1 - p =: q$, $0 < p < 1$; $S_0 := 0$. Es seien $u_n = P(S_n = 0)$ und $f_n = P(S_1 \neq 0, S_2 \neq 0, \ldots, S_{n-1} \neq 0, S_n = 0)$. Es sei nun an die Definition des Binomialkoeffizienten für eine reelle Zahl x und $n \in \mathbb{N}$ erinnert:

$$\binom{x}{n} = \frac{x(x-1)\cdots(x-n+1)}{n!}.$$

Eine einfache Rechnung (siehe Übung) liefert dann

$$u_{2n} = \binom{2n}{n} p^n q^n = \binom{-\frac{1}{2}}{n} (-4pq)^n.$$

Für $\alpha \in \mathbb{R}$ und für $|x| < 1$ gilt $(1+x)^\alpha = \sum_{n=0}^{\infty} \binom{\alpha}{n} x^n$ (Binomische Reihe). Somit berechnet sich die erzeugende Funktion von (u_n) zu

$$U(s) = \sum_{n=0}^{\infty} \binom{-\frac{1}{2}}{n} (-4pq)^n \, s^{2n} = \frac{1}{\sqrt{1-4pqs^2}}.$$

Für den Fall $p = q = 1/2$ hatten wir die Erneuerungsgleichung zwischen (u_n) und (f_n) in Satz (5.8)(3) hergeleitet. Mit genau den gleichen Argumenten erhalten wir die Erneuerungsgleichung auch im Fall $p \neq q$: Für $1 \leq k \leq n$ sei $B_k = \{S_1 \neq 0,\, S_2 \neq 0, \ldots, S_{2k-1} \neq 0,\, S_{2k} = 0,\, S_{2n} = 0\}$. Diese Ereignisse sind paarweise disjunkt, und ihre Vereinigung ist $\{S_{2n} = 0\}$. $|B_k|$ ist offenbar gleich der Anzahl der Pfade von $(0,0)$ nach $(2k, 0)$, die die x-Achse dazwischen nicht berühren, multipliziert mit der Anzahl aller Pfade von $(2k, 0)$ nach $(2n, 0)$. Somit gilt $P(B_k) = f_{2k} u_{2n-2k}$, das heißt

$$u_{2n} = \sum_{k=1}^{n} P(B_k) = \sum_{k=1}^{n} f_{2k} u_{2n-2k}.$$

Wir erhalten also nach Satz (8.10)

$$F(s) = 1 - \sqrt{1-4pqs^2}.$$

Somit ist $f = F(1) = 1 - |p-q|$, da die Wurzel für $s = 1$ den Wert $|p-q|$ liefert. Im Fall der in Kapitel 5 diskutierten eindimensionalen, symmetrischen Irrfahrt ist die Folge (f_n) also *rekurrent*. Dies hatten wir bereits auf Seite 54 diskutiert. Im Fall $p \neq q$ ist die Folge (f_n) *transient*. Die Periode von (f_n) (und damit auch von (u_n)) ist natürlich 2. Im Fall $p = q$ ist (f_n) *nullrekurrent*. Dies hatten wir in Kapitel 5 nachgerechnet. Satz (8.17) bestätigt hier nur $\lim_{n\to\infty} u_{2n} = 0$.

(b) *Rückkehr der eindimensionalen Irrfahrt zum Ursprung entlang negativer Werte*
Wir betrachten erneut die verallgemeinerte eindimensionamle Irrfahrt, nun aber die Folge (f_n^-) mit $f_{2n+1}^- = 0$ und

$$f_{2n}^- = P(\, S_1 < 0,\, S_2 < 0, \ldots,\, S_{2n-1} < 0,\, S_{2n} = 0\,).$$

Dann ist $f_{2n}^- = \frac{1}{2} f_{2n}$ trotz fehlender Symmetrie, denn eine Realisierung $(X_1, \ldots, X_{2n})$ muß n-mal den Wert 1 und n-mal den Wert -1 enthalten, um zu obigem Ereignis zu zählen, und ist somit gleich in Wahrscheinlichkeit zur Realisierung $(-X_1, \ldots, -X_{2n})$. Damit folgt für die erzeugende Funktion

$$F^-(s) = \frac{1}{2} - \frac{1}{2}\sqrt{1-4pqs^2}.$$

Somit ist die Folge (f_n^-) immer *transient*, und die Wahrscheinlichkeit, daß das Ereignis $\{\, S_1 < 0,\, S_2 < 0, \ldots,\, S_{2n-1} < 0,\, S_{2n} = 0\,\}$ irgendwann einmal eintritt, berechnet sich zu $F^-(1) = 1/2 - 1/2|p-q|$.

(c) *Anwendung auf Überschreitungswahrscheinlichkeiten (first passage times)*
Wir fassen die verallgemeinerte eindimensionale Irrfahrt als Spiel auf: Der Spieler gewinnt eine Einheit, wenn „Kopf“ kommt, andernfalls verliert er eine Einheit. Er will aufhören, sobald er zum erstenmal einen Nettogewinn von x Einheiten gewonnen

hat. Sei T_x dieser Zeitpunkt, sofern dieser überhaupt existiert. Man interessiert sich für $P(T_x = n)$. Offenbar ist $\sum_{n=1}^{\infty} P(T_x = n)$ die Wahrscheinlichkeit dafür, daß der Spieler sein Ziel jemals erreicht. In der Notation von Beispiel (a) folgt $\{T_x = n\} = \{S_1 < x,\, S_2 < x, \ldots, S_{n-1} < x,\, S_n = x\}$ für alle $n \in \mathbb{N}$.

Wir diskutieren hier nur den Fall $x = 1$. Es muß nun n ungerade sein, wir ersetzen n also durch $2n+1$. Somit ist $\{T_1 = 2n+1\} = \{S_1 < 0, S_2 \leq 0, \ldots, S_{2n-1} \leq 0, S_{2n} = 0,\, S_{2n+1} = 1\}$, also $X_{2n+1} = +1$. Also gilt $P(T_1 = 2n+1) = p\, u_{2n}^-$, wobei (u_n^-) die zu (f_n^-) gehörige Erneuerungsfolge bezeichnet. Nach Satz (8.10) ist

$$U^-(s) = \frac{1 - \sqrt{1 - 4pqs^2}}{2pqs^2},$$

also berechnet sich die erzeugende Funktion $\Lambda(\cdot)$ von $\lambda_n := P(T_1 = n)$ zu

$$\Lambda(s) = ps\, U^-(s) = \frac{1 - \sqrt{1 - 4pqs^2}}{2qs} \quad \text{für} \quad |s| \leq \frac{1}{2\sqrt{pq}}. \tag{8.19}$$

$\Lambda(1)$ ist die Wahrscheinlichkeit, jemals in die „Gewinnzone 1“ zu kommen. Es ist $+\sqrt{1 - 4p(1-p)} = |2p - 1|$ und somit

$$\Lambda(1) = \frac{1 - |2p - 1|}{2(1-p)} = \begin{cases} p/(1-p), & \text{falls } p < 1/2, \\ 1, & \text{falls } p \geq 1/2. \end{cases}$$

Die Wahrscheinlichkeit dafür, jemals in die „Gewinnzone 1“ zu kommen, ist also gleich 1, wenn $p \geq 1/2$ ist. Die Folge (λ_n) ist in diesem Fall eine Wahrscheinlichkeitsverteilung auf $\mathbb{N}_0$, nämlich die Verteilung der Zufallsgröße T_1. Sie ist also für $p \geq 1/2$ *rekurrent.*

Die rechte Seite in (8.19) kann man in eine Potenzreihe entwickeln und erhält daraus die Folge (λ_n). Eine explizite Rechnung, die wir hier nicht durchführen, liefert

$$\lambda_{2n-1} = \frac{(-1)^{n-1}}{2(1-p)} \binom{\frac{1}{2}}{n} (4p(1-p))^n \quad \text{und} \quad \lambda_{2n} = 0.$$

Wir wollen abschließend die mittlere Rückkehrzeit μ im rekurrenten Fall bestimmen. Dazu muß nach (8.5) $\Lambda'(1-)$ berechnet werden. Für alle $|s| < 1$ gilt

$$\Lambda'(s) = \frac{2p}{\sqrt{1 - 4p(1-p)s^2}} - \frac{1 - \sqrt{1 - 4p(1-p)s^2}}{2(1-p)s^2}.$$

Im Fall $p = 1/2$ gilt $\Lambda'(s) \uparrow \infty$ für $s \uparrow 1$, also ist $\mu = \infty$ (Nullrekurrenz). Im Fall $p > 1/2$ folgt

$$\lim_{s \uparrow 1} \Lambda'(s) = \Lambda'(1-) = \frac{2p}{2p-1} - 1 = \frac{1}{2p-1},$$

also liegt positive Rekurrenz vor. Für beliebige Nettogewinne x gilt das gleiche Resultat. Dies betrachten wir in einer Übungsaufgabe.

§9 Markov-Ketten

Bisher haben wir uns hauptsächlich mit unabhängigen Ereignissen und unabhängigen Zufallsgrößen beschäftigt. *Andrej Andrejewitsch Markov* (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 Markov-Ketten. Wir werden sehen, daß sehr viele Modelle Markov-Ketten sind. Tatsächlich sind sie uns auch schon begegnet. 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 Markov-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) *Markov-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 Markov-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 Markov-Kette mit Startverteilung ν.

(a) Für alle $n \in \mathbb{N}_0$ und $i_0, i_1, \ldots, i_n \in I$ gilt

$$P(X_0 = i_0,\, X_1 = i_1, \ldots, X_n = i_n) = \nu(i_0) p_{i_0 i_1} p_{i_1 i_2} \cdots p_{i_{n-1} i_n}.$$

(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

$$\begin{aligned} &P(\, (X_{n+1}, \ldots, X_m) \in B \mid (X_0, \ldots, X_{n-1}) \in A,\, X_n = i_n \,) \\ &\qquad = P(\, (X_{n+1}, \ldots, X_m) \in B \mid X_n = i_n \,). \end{aligned}$$

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. $\square$

(9.4) Bemerkung. Die Aussage von (b) heißt *Markov-Eigenschaft (Markov property)*. Sie spiegelt genau die eingangs erwähnte Eigenschaft wieder, daß in einer Markov-Kette die Wahrscheinlichkeit, zur Zeit $n+1$ in einen beliebigen Zustand zugelangen, 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 Markov-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 Markov-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 Markov-Kette mit Startverteilung ν und Übergangsmatrix $\mathbb{P}$ ist.

Beweis. Es sei $\Omega := I^{\{0,\dots,N\}}$ und $p(i_0,\dots,i_N) := \nu(i_0)p_{i_0 i_1}\dots p_{i_{N-1}i_N}$ sowie $X_n(i_0,\dots,i_N) = i_n$ für alle $n \in \{0,1,\dots,N\}$ und $(i_0,\dots,i_N) \in \Omega$. Da die Summe der Komponenten der stochastischen Matrix $\mathbb{P}$ in jeder Zeile gleich eins ist, gilt für alle $n \in \{0,1,\dots,N\}$ und $(i_0,\dots,i_n) \in I^{\{0,\dots,n\}}$

$$\begin{aligned} P(X_0 = i_0,\dots,X_n = i_n) &= \sum_{(i_{n+1},\dots,i_N)\in I^{\{n+1,\dots,N\}}} P(X_0 = i_0,\dots,X_N = i_N) \\ &= \sum_{(i_{n+1},\dots,i_N)\in I^{\{n+1,\dots,N\}}} \nu(i_0)p_{i_0 i_1}\dots p_{i_{N-1}i_N} \\ &= \nu(i_0)p_{i_0 i_1}\dots p_{i_{n-1}i_n}. \end{aligned}$$

Dieses Produkt ist größer als null genau dann, wenn jeder Faktor größer als null ist. Ist dies der Fall, so ist offenbar

$$P(\, X_{n+1} = i_{n+1} \mid X_0 = i_0,\dots,X_n = i_n \,) = p_{i_n i_{n+1}}.$$

□

Bemerkung. Nachfolgend soll stets von einer unendlich langen Markov-Kette ausgegangen werden, dies jedoch nur wegen einer bequemeren Notation. Alle nachfolgenden Überlegungen benötigen die Konstruktion einer unendlichen Markov-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

$$P(X_0 = i_0,\ X_1 = i_1,\dots,X_n = i_n) = \nu(i_0)q_{i_1}\dots q_{i_n}.$$

Man sieht leicht, daß $q_j = P(X_m = j)$ für $m \geq 1$ ist. Somit gilt

$$P(X_0 = i_0,\dots,X_n = i_n) = P(X_0 = i_0)P(X_1 = i_1)\dots P(X_n = i_n),$$

d. h., die $X_0, X_1,\dots,X_n$ sind unabhängig. Satz (9.5) liefert also als Spezialfall die Konstruktion von unabhängigen, I-wertigen Zufallsgrößen.

(b) *Irrfahrt auf $\mathbb{Z}$:* Es sei $Y_1, Y_2,\dots$ 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 $X_0, X_1,\dots$ eine Markov-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 $p_{(i_1,\dots,i_d),(j_1,\dots,j_d)} = 1/(2d)$, falls $i_k = j_k$ für alle bis auf genau ein $k \in \{1,2,\dots,d\}$, 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. Eine der n Kugeln wird 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

$$\begin{aligned} p_{r,r-1} &= r/n, \quad r \in \{1, 2, \ldots, n\}, \\ p_{r,r+1} &= 1 - r/n, \quad r \in \{0, 1, \ldots, n-1\}. \end{aligned}$$

(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 Markov-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 Markov-Ketten könnte man vermuten, daß Markov selbst aus angewandten Fragestellungen heraus die Ketten analysiert hat. Markov 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. Die Vermutung, daß diese Folge der Markov-Bedingung genügt, kann man grob damit begründen, daß viele Sprachen schon verhältnismäßig kurze Folgen, die nur aus Vokalen oder nur aus Konsonanten besteht, ausschließen.

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 Markov-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

$$\begin{aligned} &P(\, X_{m+n} = j \mid X_m = i\,) \\ &\quad = \sum_{i_{m+1}, \ldots, i_{m+n-1} \in I} P(\, X_{m+1} = i_{m+1}, \ldots, \\ &\qquad\qquad X_{m+n-1} = i_{m+n-1},\ X_{m+n} = j \mid X_m = i\,) \end{aligned}$$

und mit der Definition (9.2) folgt

$$\begin{aligned} &P(\, X_{m+1} = i_{m+1}, \ldots, X_{m+n-1} = i_{m+n-1},\, X_{m+n} = j \mid X_m = i\,) \\ &\quad = P(\, X_{m+n} = j \mid X_m = i,\ X_{m+1} = i_{m+1}, \ldots, X_{m+n-1} = i_{m+n-1}\,) \\ &\qquad \times \prod_{k=1}^{n-1} P(X_{m+k} = i_{m+k} \mid X_m = i,\, X_{m+1} = i_{m+1}, \ldots, X_{m+k-1} = i_{m+k-1}) \\ &\quad = p_{i i_{m+1}} p_{i_{m+1} i_{m+2}} \cdots p_{i_{m+n-1} j}. \end{aligned}$$

Somit gilt

$$P(\, X_{m+n} = j \mid X_m = i\,) = \sum_{i_{m+1}, \ldots, i_{m+n-1} \in I} p_{i i_{m+1}} \cdots p_{i_{m+n-1} j} = p_{ij}^{(n)}.$$

□

(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 Markov-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 Markov-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 Markov-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 Markov-Kette mit Übergangsmatrix $\mathbb{P} = (p_{ij})_{i,j \in I}$ und Startverteilung ν. Die wichtigste Frage, die uns für den Rest des Kapitels beschäftigen wird, ist die Diskussion der Verteilung von X_n für große n, also

$$P_\nu(X_n = j) = \sum_{i \in I} \nu(i) p_{ij}^{(n)}, \quad j \in I.$$

Die n-stufigen Übergangswahrscheinlichkeiten $p_{ij}^{(n)}$ lassen sich fast nie explizit berechnen. Es ist daher wichtig, Approximationen für sie zu finden.

Für die nachfolgenden Überlegungen sei die Startverteilung ν in $i \in I$ konzentriert. Um dies zu betonen, schreiben wir P_i statt P_ν. Für $n \in \mathbb{N}$ sei

$$f_{ii}^{(n)} := P_i(X_1 \neq i, \dots, X_{n-1} \neq i, X_n = i).$$

(9.14) Lemma. *Es gilt die Erneuerungsgleichung*

$$p_{ii}^{(n)} = \sum_{k=1}^{n} f_{ii}^{(k)} p_{ii}^{(n-k)}, \quad n \in \mathbb{N}.$$

Beweis. Gemäß Lemma (9.10) gilt $p_{ii}^{(n)} = P_i(X_n = i)$. Aufspalten nach dem ersten Zeitpunkt, an dem die Markov-Kette wieder i erreicht, ergibt

$$\begin{aligned} p_{ii}^{(n)} &= \sum_{k=1}^{n} P_i(X_1 \neq i, \dots, X_{k-1} \neq i, X_k = i, X_n = i) \\ &= \sum_{k=1}^{n} P_i(X_n = i \mid X_1 \neq i, \dots, X_{k-1} \neq i, X_k = i) f_{ii}^{(k)}. \end{aligned}$$

Anwendung der Markov-Eigenschaft (Satz 9.3 (b)) und der Bemerkung (9.8) liefert

$$p_{ii}^{(n)} = \sum_{k=1}^{n} P_i(X_n = i \mid X_k = i) f_{ii}^{(k)} = \sum_{k=1}^{n} f_{ii}^{(k)} p_{ii}^{(n-k)}.$$

□

(9.15) Definition. $i \in I$ heißt *transient, rekurrent, nullrekurrent* bzw. *positiv rekurrent*, je nach den entsprechenden Eigenschaften der Folge $(f_{ii}^{(n)})_{n\in\mathbb{N}}$. Ferner heißt $d_i := \mathrm{ggT}\{\, n \in \mathbb{N} \mid f_{ii}^{(n)} \neq 0 \,\}$ *die Periode von i*. Ist $d_i = 1$, heißt i *aperiodisch*. Ist i aperiodisch und positiv rekurrent, so heißt i auch *ergodisch*.

Rekurrent ist $i \in I$ also genau dann, wenn

$$1 = \sum_{n=1}^{\infty} f_{ii}^{(n)} = \lim_{N\to\infty} \sum_{n=1}^{N} f_{ii}^{(n)} = \lim_{N\to\infty} P_i(\tau_i \leq N)$$

gilt, wobei $\tau_i := \inf\{\, n \in \mathbb{N} \mid X_n = i \,\}$ ist. Die rechte Seite der obigen Gleichung wird man als $P_i(\tau_i < \infty)$ interpretieren.

(9.16) Satz.

(a) i ist genau dann transient, wenn $\sum_{n=0}^{\infty} p_{ii}^{(n)} < \infty$ gilt.
(b) Ist i nullrekurrent, so gilt $\lim_{n\to\infty} p_{ii}^{(n)} = 0$.
(c) Ist i positiv rekurrent mit Periode d_i, so gilt

$$\lim_{n\to\infty} p_{ii}^{(nd_i)} = \frac{d_i}{\mu_i} \quad \text{mit} \quad \mu_i = \sum_{n=1}^{\infty} n f_{ii}^{(n)}.$$

Beweis. Teil (a) folgt aus Satz (8.11). Teil (b) und Teil (c) folgen aus dem Erneuerungssatz (8.13) und aus Satz (8.17). □

(9.17) Beispiel *(Die symmetrische Irrfahrt auf dem d-dimensionalen Gitter $\mathbb{Z}^d$).*
Sei $Y_n = (Y_{n1}, Y_{n2}, \ldots, Y_{nd})$ und alle Y_{ni} seien unabhängig mit $P(Y_{ni} = 1) = P(Y_{ni} = -1) = 1/2$. Sei weiter $X_n = Y_1 + \cdots + Y_n$. Von $i = (i_1, \ldots, i_d) \in \mathbb{Z}^d$ geht man mit Wahrscheinlichkeit 2^{-d} zu jedem der Punkte $j = (j_1, \ldots, j_d)$ mit $|i_k - j_k| = 1$ für $k = 1, \ldots, d$. Von $(0, \ldots, 0)$ kehrt man genau dann zum Zeitpunkt n nach $(0, \ldots, 0)$ zurück, wenn jede der eindimensionalen Irrfahrten $X_{ni} = Y_{1i} + \cdots + Y_{ni}$ zum Zeitpunkt n nach 0 zurückkehrt. Da diese unabhängig sind, ist

$$p_{(0,\ldots,0),(0,\ldots,0)}^{(n)} = \Big(\binom{2n}{n} 2^{-2n}\Big)^d \sim \Big(\frac{1}{\sqrt{\pi n}}\Big)^d.$$

Für d=2 ist diese Kette also *rekurrent*. Für $d \geq 3$ ist sie wegen $\sum 1/n^{3/2} < \infty$ *transient*. Den Fall $d = 1$ haben wir in Beispiel (8.18)(a) (auch im nichtsymmetrischen Fall) diskutiert. Eigentlich betrachtet man als symmetrische Irrfahrt auf $\mathbb{Z}^d$ die im

Beispiel (9.6)(c) definierte Kette, bei der man von jedem Gitterpunkt i mit der gleichen Wahrscheinlichkeit $1/2d$ zu den $2d$ Nachbarn von i geht. Tatsächlich liegt hier ebenfalls für $d = 1, 2$ Rekurrenz und für $d \geq 3$ Transienz vor. Die Bestimmung der Rückkehrwahrscheinlichkeiten ist aber schwieriger. Dieses Resultat wurde von *Georg Pólya* (1887-1985) im Jahre 1921 bewiesen.

Unser nächstes Ziel ist es, $\lim_{n\to\infty} p_{ij}^{(n)}$ auch für $i \neq j$ zu untersuchen. Dafür sei

$$f_{ij}^{(n)} := P_i(X_1 \neq j,\, X_2 \neq j, \ldots, X_{n-1} \neq j,\, X_n = j)$$

für alle $n \in \mathbb{N}$ und

$$f_{ij} := \sum_{n=1}^{\infty} f_{ij}^{(n)} \leq 1$$

für alle $i, j \in I$. Wir können f_{ij} wieder als $P_i(\tau_j < \infty)$ interpretieren. Es gilt das folgende, zu Satz (9.16) analoge Resultat:

(9.18) Satz. Sei $j \in I$.

(a) Ist j transient, so gilt für alle $i \in I$

$$\sum_{n=1}^{\infty} p_{ij}^{(n)} = f_{ij} \sum_{n=0}^{\infty} p_{jj}^{(n)} < \infty.$$

(b) Ist j nullrekurrent, so gilt für alle $i \in I$

$$\lim_{n\to\infty} p_{ij}^{(n)} = 0.$$

(c) Ist j aperiodisch und positiv rekurrent, so gilt für alle $i \in I$

$$\lim_{n\to\infty} p_{ij}^{(n)} = \frac{f_{ij}}{\mu_j}.$$

Beweis. Für alle $i \in I$ und $n \in \mathbb{N}$ gilt $p_{ij}^{(n)} = P_i(X_n = j)$ gemäß Lemma (9.10), und analog zum Beweis von Lemma (9.14) folgt

$$\begin{aligned}
p_{ij}^{(n)} &= \sum_{k=1}^{n} P_i(X_1 \neq j, \ldots, X_{k-1} \neq j,\, X_k = j,\, X_n = j) \\
&= \sum_{k=1}^{n} P_i(X_n = j \mid X_1 \neq j, \ldots, X_{k-1} \neq j,\, X_k = j) f_{ij}^{(k)} \\
&= \sum_{k=1}^{n} P_i(X_n = j \mid X_k = j) f_{ij}^{(k)} = \sum_{k=1}^{n} f_{ij}^{(k)} p_{jj}^{(n-k)}.
\end{aligned}$$

Diese zur Erneuerungsgleichung in Lemma (9.14) analoge Darstellung wird für die folgenden Beweisteile benötigt.

(a) Aus der obigen Rechnung ergibt sich

$$\sum_{n=1}^{\infty} p_{ij}^{(n)} = \sum_{n=1}^{\infty} \sum_{k=1}^{n} f_{ij}^{(k)} p_{jj}^{(n-k)} = \sum_{k=1}^{\infty} f_{ij}^{(k)} \sum_{n=k}^{\infty} p_{jj}^{(n-k)} = f_{ij} \sum_{n=0}^{\infty} p_{jj}^{(n)},$$

und die letzte Reihe konvergiert gemäß Satz (9.16 (a)).

(b) Zu $\varepsilon > 0$ existiert ein $N \in \mathbb{N}_0$ mit $\sum_{k=N+1}^{\infty} f_{ij}^{(k)} \leq \varepsilon$. Aus der obigen Rechnung folgt für alle $n \geq N$

$$p_{ij}^{(n)} \leq \sum_{k=1}^{N} f_{ij}^{(k)} p_{jj}^{(n-k)} + \sum_{k=N+1}^{n} f_{ij}^{(k)} \leq \sum_{k=1}^{N} f_{ij}^{(k)} p_{jj}^{(n-k)} + \varepsilon.$$

Gemäß Satz (9.16 (b)) konvergiert $p_{jj}^{(n-k)}$ gegen null für $n \to \infty$. Da $\varepsilon > 0$ beliebig war, folgt Teil (b).

(c) Zu $\varepsilon > 0$ existiert ein $N \in \mathbb{N}_0$ mit $\sum_{k=N+1}^{\infty} f_{ij}^{(k)} \leq \varepsilon$. Aus der obigen Rechnung folgt für alle $n \geq N$

$$\begin{aligned} \left| p_{ij}^{(n)} - \frac{f_{ij}}{\mu_j} \right| &\leq \sum_{k=1}^{N} f_{ij}^{(k)} \left| p_{jj}^{(n-k)} - \frac{1}{\mu_j} \right| + \sum_{k=N+1}^{n} f_{ij}^{(k)} + \frac{1}{\mu_j} \sum_{k=N+1}^{\infty} f_{ij}^{(k)} \\ &\leq \sum_{k=1}^{N} f_{ij}^{(k)} \left| p_{jj}^{(n-k)} - \frac{1}{\mu_j} \right| + \varepsilon + \frac{\varepsilon}{\mu_j}. \end{aligned}$$

Gemäß Satz (9.16 (c)) konvergiert $p_{jj}^{(n-k)}$ gegen $1/\mu_j$ für $n \to \infty$. Da $\varepsilon > 0$ beliebig war, folgt Teil (c). □

Viele der in Anwendungen wichtigen Markov-Ketten sind irreduzibel. Eine äußerst nützliche Tatsache ist, daß in diesem Fall Rekurrenz-, Transienz- und Periodizitätseigenschaften nicht vom einzelnen Element in I abhängen. Dies ergibt sich als Spezialfall des folgenden Satzes:

(9.19) Satz. Es seien $i, j \in I$ mit $i \sim j$. Dann gilt:

(a) i ist genau dann transient, wenn j transient ist.
(b) i ist genau dann nullrekurrent, wenn j nullrekurrent ist.
(c) i ist genau dann positiv rekurrent, wenn j positiv rekurrent ist.
(d) Die Perioden d_i und d_j von i beziehungsweise j sind gleich.

Beweis. Da $i \sim j$ ist, existiert ein $M \in \mathbb{N}_0$ mit $p_{ij}^{(M)} > 0$ und ein $N \in \mathbb{N}_0$ mit $p_{ji}^{(N)} > 0$. Dann ist gemäß Lemma (9.9)

$$p_{ii}^{(M+n+N)} \geq p_{ij}^{(M)} p_{jj}^{(n)} p_{ji}^{(N)} = \alpha \, p_{jj}^{(n)} \quad \text{mit} \quad \alpha := p_{ij}^{(M)} p_{ji}^{(N)} > 0.$$

Analog folgt $p_{jj}^{(M+n+N)} \geq \alpha \, p_{ii}^{(n)}$. Somit gilt

$$\sum_{n=0}^{\infty} p_{ii}^{(n)} < \infty \Leftrightarrow \sum_{n=0}^{\infty} p_{jj}^{(n)} < \infty,$$

und Teil (a) folgt aus Satz (9.16 (a)). Ferner gilt

$$\lim_{n\to\infty} p_{ii}^{(n)} = 0 \Leftrightarrow \lim_{n\to\infty} p_{jj}^{(n)} = 0,$$

woraus Teil (b) mittels Satz (9.16 (b)) folgt. Also muß auch Teil (c) gültig sein.

Nach Lemma (8.15) existiert ein $n \in \mathbb{N}$ mit $p_{ii}^{(nd_i)} > 0$ und $p_{ii}^{((n+1)d_i)} > 0$. Somit ist

$$p_{jj}^{(M+nd_i+N)} > 0 \quad \text{und} \quad p_{jj}^{(M+(n+1)d_i+N)} > 0.$$

Nach Lemma (8.9) teilt d_j sowohl $M + nd_i + N$ also auch $M + (n+1)d_i + N$. Also teilt d_j auch die Differenz d_i. Analog zeigt man, daß d_i die Periode d_j teilt. Also gilt $d_i = d_j$ und Teil (d) ist bewiesen. $\square$

Rekurrenzeigenschaften und die Periode sind also Klasseneigenschaften, das heißt, alle Elemente in derselben Äquivalenzklasse haben in dieser Hinsicht dieselben Eigenschaften. Man spricht daher von positiv rekurrenten, nullrekurrenten und transienten Klassen und von der Periode einer Klasse.

Ist eine Markov-Kette irreduzibel, das heißt gilt $i \sim j$ für alle $i, j \in I$, so spricht man einfach von einer transienten, nullrekurrenten oder positiv rekurrenten Markov-Kette und von *der* Periode dieser Markov-Kette.

(9.20) Beispiel. Es sei $I = \{0, 1, 2, 3, 4\}$ und die stochastische Matrix gegeben durch

$$\mathbb{P} = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 1/2 & 1/2 & 0 \\ 1/4 & 1/4 & 0 & 0 & 1/2 \end{pmatrix}.$$

Dann gibt es drei Äquivalenzklassen: $\{0, 1\}, \{2, 3\}$ und $\{4\}$. Die ersten beiden sind rekurrent, die dritte Klasse ist transient.

Rekurrente Klassen sind abgeschlossen, wie die folgenden Überlegungen zeigen.

(9.21) Lemma. *Es seien $i, j \in I$ mit $i \rightsquigarrow j$. Ist i rekurrent, so gilt auch $j \rightsquigarrow i$, und j ist dann ebenfalls rekurrent.*

Beweis. Wir können $i \neq j$ annehmen. Sei $N \in \mathbb{N}$ die kleinste Zahl mit $p_{ij}^{(N)} > 0$. Wir wollen einen Widerspruchsbeweis führen und nehmen dafür an, daß i nicht von j aus erreichbar ist, also daß $p_{ji}^{(n)} = 0$ für alle $n \in \mathbb{N}_0$ gilt. Für alle $n > N$ gilt dann $P_i(X_N = j, X_n = i) = p_{ij}^{(N)} p_{ji}^{(n-N)} = 0$. Für alle $n \in \{1, 2, \ldots, N\}$ gilt $P_i(X_N = j, X_n = i) = p_{ii}^{(n)} p_{ij}^{(N-n)} = 0$, da N definitionsgemäß die kleinste Zahl mit $p_{ij}^{(N)} > 0$ ist.

Für $M \in \mathbb{N}$ sei A_M das Ereignis, daß die Markov-Kette i im Zeitraum von 1 bis M besucht, also $A_M = \bigcup_{n=1}^M \{X_1 \neq i, \ldots, X_{n-1} \neq i, X_n = i\}$. Aus dem oben gezeigten folgt

$$P_i(A_M, X_N = j) \leq \sum_{n=1}^M P_i(X_n = i, X_N = j) = 0.$$

Also gilt

$$\begin{aligned}\sum_{n=1}^{M} f_{ii}^{(n)} &= P_i(A_M) = P_i(A_M,\, X_N = j) + P_i(A_M,\, X_N \neq j)\\ &= P_i(A_M,\, X_N \neq j) \leq P_i(X_N \neq j) = 1 - P_i(X_N = j) = 1 - p_{ij}^{(N)}.\end{aligned}$$

Da diese Abschätzung für jedes $M \in \mathbb{N}$ gilt, folgt

$$\sum_{n=1}^{\infty} f_{ii}^{(n)} \leq 1 - p_{ij}^{(N)} < 1,$$

im Widerspruch zur Annahme der Rekurrenz von i. $\square$

(9.22) Korollar. Rekurrente Klassen sind abgeschlossen.

Transiente Klassen können abgeschlossen sein, brauchen es aber nicht. Wir behandeln dies hier nicht. Aus Korollar (9.22) und der Bemerkung nach Definition (9.12) folgt, daß die Einschränkung einer stochastischen Matrix $\mathbb{P}$ auf eine rekurrente Klasse wieder eine stochastische Matrix ist, die dann natürlich irreduzibel ist. Die einzelnen rekurrenten Klassen lassen sich daher getrennt diskutieren.

(9.23) Lemma. *Sind i und j in derselben rekurrenten Klasse, so gilt $f_{ij} = f_{ji} = 1$.*

Beweis. Wir müssen nur $i \neq j$ diskutieren. Für $M \in \mathbb{N}$ sei A_M definiert wie im Beweis von Lemma (9.21). Sei $N \in \mathbb{N}_0$ die kleinste Zahl mit $p_{ji}^{(N)} > 0$. Für $M > N$ gilt

$$\begin{aligned}P_j(A_M,\, X_N = i) &= \sum_{n=1}^{N-1} P_j(X_1 \neq j, \ldots, X_{n-1} \neq j,\, X_n = j,\, X_N = i)\\ &\qquad + \sum_{n=N+1}^{M} P_j(X_1 \neq j, \ldots, X_{n-1} \neq j,\, X_n = j,\, X_N = i)\\ &= \sum_{n=1}^{N-1} f_{jj}^{(n)} p_{ji}^{(N-n)}\\ &\qquad + \sum_{n=N+1}^{M} P_j(X_1 \neq j, \ldots, X_{N-1} \neq j,\, X_N = i) f_{ij}^{(n-N)}.\end{aligned}$$

Für jedes $n \in \{1, \ldots, N-1\}$ ist $p_{ji}^{(N-n)} = 0$, und es gilt

$$P_j(X_1 \neq j, \ldots, X_{N-1} \neq j,\, X_N = i) \leq p_{ji}^{(N)}.$$

Demzufolge ist

$$P_j(A_M,\, X_N = i) \leq p_{ji}^{(N)} \sum_{n=N+1}^{M} f_{ij}^{(n-N)},$$

und wegen $\lim_{M\to\infty} P_j(A_M) = \lim_{M\to\infty} \sum_{k=1}^M f_{jj}^{(k)} = 1$ folgt

$$p_{ji}^{(N)} = \lim_{M\to\infty} P_j(A_M,\, X_N = i) \le p_{ji}^{(N)} f_{ij}.$$

Wegen $f_{ij} \le 1$ und $p_{ji}^{(N)} > 0$ ergibt sich $f_{ij} = 1$. $\square$

Als Korollar aus Lemma (9.23) und Satz (9.18)(c) folgt

(9.24) Satz. Gehören i und j zu derselben aperiodischen, positiv rekurrenten Klasse, so gilt

$$\lim_{n\to\infty} p_{ij}^{(n)} = \frac{1}{\mu_j}.$$

Bemerkung. Die Diskussion des periodischen Falls ist nicht schwierig, aber etwas lästig. Seien i und j aus derselben positiv rekurrenten Klasse, die die Periode d hat. Sei $N := \min\{\, k \in \mathbb{N}_0 \mid p_{ij}^{(k)} \neq 0\,\}$. Man überzeuge sich davon, daß $p_{ij}^{(k)} = 0$ ist, wenn d die Differenz $k - N$ nicht teilt, und daß $\lim_{n\to\infty} p_{ij}^{(N+nd)} = d/\mu_j$ gilt.

Ist $i \in I$ rekurrent, so ist die Berechnung der mittleren Rückkehrzeit μ_i mit Hilfe der Darstellung $\mu_i = \sum_{n=1}^\infty n f_{ii}^{(n)}$ in der Regel nicht möglich, da man die $f_{ii}^{(n)}$ üblicherweise nicht explizit kennt. Glücklicherweise genügen die Kehrwerte $1/\mu_i$ jedoch einem linearen Gleichungssystem, mit dessen Hilfe die μ_i in vielen Beispielen bestimmt werden können.

(9.25) Definition. Eine Wahrscheinlichkeitsverteilung π auf I heißt *stationär (stationary)* bezüglich der stochastischen Matrix $\mathbb{P} = (p_{ij})_{i,j\in I}$, wenn $\pi(j) = \sum_{i\in I} \pi(i) p_{ij}$ für alle $j \in I$ gilt.

(9.26) Bemerkung. (a) Betrachten wir eine stationäre Verteilung $\pi = (\pi(i))_{i\in I}$ als Vektor im $\mathbb{R}^I$, so erfüllt π (als Zeilenvektor aufgefaßt) die Gleichung $\pi\mathbb{P} = \pi$. Das heißt, π ist ein Linkseigenvektor von $\mathbb{P}$ zum Eigenwert 1. In der aus der Linearen Algebra üblichen Notation ist π^T also ein Eigenvektor von $\mathbb{P}^T$ zum Eigenwert 1. Man beachte, daß $\mathbb{P}$ in jedem Fall den Eigenwert 1 hat, denn es gilt

$$\mathbb{P}\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix} = \begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}.$$

Zumindest wenn I endlich ist, folgt daraus, daß auch $\mathbb{P}^T$ den Eigenwert 1 hat. Es ist jedoch im Moment noch nicht klar, ob sich ein Eigenvektor mit nichtnegativen Komponenten finden läßt.

(b) Ist π stationär, so gilt

$$\pi(j) = \sum_{i\in I} \pi(i) p_{ij}^{(n)},$$

also $P_\pi(X_n = j) = \pi(j)$ für alle $j \in I$ und $n \in \mathbb{N}_0$. Das heißt, hat die Markov-Kette die Startverteilung π, so ist die Verteilung von X_n gleich π für alle $n \in \mathbb{N}_0$.

(9.27) Satz. Ist eine Markov-Kette aperiodisch und irreduzibel, so gelten:

(a) Eine stationäre Verteilung π existiert genau dann, wenn die Kette positiv rekurrent ist. In dem Fall ist sie eindeutig gegeben durch $\pi(i) = 1/\mu_i$ für $i \in I$.

(b) Ist die Kette positiv rekurrent, so gilt für jede Startverteilung ν und alle $i \in I$

$$\lim_{n\to\infty} P_\nu(X_n = i) = 1/\mu_i.$$

Beweis. Sei ν eine beliebige Verteilung auf I. Gemäß (9.18) existiert $\lim_{n\to\infty} p_{ji}^{(n)}$ für alle $i, j \in I$. Ein Satz über majorisierte Konvergenz für Reihen liefert

$$\lim_{n\to\infty} P_\nu(X_n = i) = \lim_{n\to\infty} \sum_{j\in I} \nu(j) p_{ji}^{(n)} = \sum_{j\in I} \nu(j) \lim_{n\to\infty} p_{ji}^{(n)}.$$

Mit Satz (9.18) folgt daraus Teil (b) und die Tatsache, daß im transienten und nullrekurrenten Fall keine stationäre Verteilung existiert, da dann $\lim_{n\to\infty} p_{ji}^{(n)} = 0$ ist.

Ferner folgt, daß, wenn es überhaupt eine stationäre Verteilung π gibt, diese durch $\pi(i) := 1/\mu_i$ für alle $i \in I$ gegeben ist. Es bleibt zu zeigen, daß dies für positiv rekurrente Markov-Ketten tatsächlich eine stationäre Verteilung definiert. Für alle $j \in I$ gilt

$$\sum_{i\in I} \frac{1}{\mu_i} p_{ij} = \sum_{i\in I} (\lim_{n\to\infty} p_{ji}^{(n)}) p_{ij} \leq \lim_{n\to\infty} \sum_{i\in I} p_{ji}^{(n)} p_{ij} = \lim_{n\to\infty} p_{jj}^{(n+1)} = \frac{1}{\mu_j}.$$

(Beweis von „$\leq$“ als Übungsaufgabe)

Andererseits gilt

$$\sum_{j\in I} \sum_{i\in I} \frac{1}{\mu_i} p_{ij} = \sum_{i\in I} \frac{1}{\mu_i} = \sum_{j\in I} \frac{1}{\mu_j}.$$

Somit folgt

$$\sum_{i\in I} \frac{1}{\mu_i} p_{ij} = \frac{1}{\mu_j},$$

für alle $j \in I$, denn es gilt

$$\sum_{i\in I} \frac{1}{\mu_i} = \sum_{i\in I} \lim_{n\to\infty} p_{ji}^{(n)} \leq \lim_{n\to\infty} \sum_{i\in I} p_{ji}^{(n)} = 1.$$

Daraus ergibt sich, daß die Verteilung

$$\pi(i) = \frac{1}{\mu_i} \left(\sum_{j\in I} \frac{1}{\mu_j} \right)^{-1}, \quad i \in I,$$

stationär ist. Nach dem oben schon Bewiesenen muß also $\pi(i) = 1/\mu_i$ sein. $\square$

(9.28) Bemerkung. Satz (9.27 (a)) ist auch richtig im periodischen Fall, und Teil (b) muß leicht modifiziert werden: Für positiv rekurrente, irreduzible Ketten mit Periode d gilt:

$$\lim_{n\to\infty} \frac{1}{d} \sum_{k=0}^{d-1} P_\nu(X_{n+k} = i) = \frac{1}{\mu_i}$$

für jede Startverteilung ν und jedes $i \in I$. Der Beweis geht wie oben, nur muß in den Argumenten $p_{ij}^{(n)}$ durch $\sum_{k=0}^{d-1} p_{ij}^{(n+k)}/d$ ersetzt werden. Die Details seien dem Leser überlassen.

Die Frage, ob eine vorgegebene irreduzible Kette rekurrent oder transient ist, ist nicht immer einfach. Nach dem folgenden Satz ist eine irreduzible Kette mit endlichem I stets positiv rekurrent.

(9.29) Satz. Ist I endlich, so existiert stets ein positiv rekurrentes Element von I, und keines ist nullrekurrent.

Beweis. Wären alle Elemente von I transient oder nullrekurrent, so folgte nach Satz (9.18) (a), (b) für ein beliebiges $i \in I$

$$(*) \qquad 1 = \lim_{n\to\infty} \sum_{j\in I} p_{ij}^{(n)} = \sum_{j\in I} \lim_{n\to\infty} p_{ij}^{(n)} = 0.$$

Somit existiert also mindestens ein positiv rekurrentes Element. Wäre dieses nullrekurrent, so wäre auch die zugehörige Äquivalenzklasse $A \subset I$ nullrekurrent. Nach Korollar (9.22) ist A abgeschlossen, das heißt, es gilt $\sum_{j\in A} p_{ij}^{(n)} = 1$ für alle $i \in A$. Das Argument in (*) mit A anstelle von I führt dann ebenfalls auf einen Widerspruch. $\square$

(9.30) Beispiele.

(a) In Beispiel (9.20) folgt nun unmittelbar aus der Transienz der Klasse $\{4\}$, daß die anderen beiden Klassen positiv rekurrent sind.

(b) Es sei $I = \{0, 1, 2, 3\}$ und die stochastische Matrix gegeben durch

$$\mathbb{P} = \begin{pmatrix} 0 & 0 & 1/2 & 1/2 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}.$$

Die Kette ist irreduzibel, also positiv rekurrent.

(c) Endliche Ketten können natürlich transiente Elemente enthalten:

$$\mathbb{P} = \begin{pmatrix} 1/2 & 1/2 \\ 0 & 1 \end{pmatrix}.$$

Die Kette hat zwei Äquivalenzklassen, nämlich $\{1\}$ und $\{2\}$. 1 ist transient, 2 ist positiv rekurrent.

(d) *Ehrenfests Urnenmodell:* Diese Markov-Kette hat die Periode 2. Die invariante Verteilung kann man erraten: nach einem langen Zeitraum dürfte jede Kugel unabhängig von den anderen mit Wahrscheinlichkeit 1/2 in der linken Schachtel liegen. Man kann leicht nachrechnen, daß die Binomialverteilung

$$\pi_k = \binom{n}{k} 2^{-n}, \quad k \in I,$$

stationär für $\mathbb{P}$ ist. Wegen der offensichtlichen Irreduzibilität ist das also *die* stationäre Verteilung. Ist $|i-j|$ gerade, gilt (entsprechend der Bemerkung nach Satz (9.24)) $p_{ij}^{(2n+1)} = 0$ und $\lim_{n\to\infty} p_{ij}^{(2n)} = 2\pi_j$. Bei ungeradem $|i-j|$ gilt $p_{ij}^{(2n)} = 0$ und $\lim_{n\to\infty} p_{ij}^{(2n+1)} = 2\pi_j$.

(e) *Wettervorhersage:* Diese Kette ist aperiodisch und irreduzibel, also positiv rekurrent. Die Limesverteilung berechnet sich mittels der Gleichungen $\pi(0) = \alpha\pi(0) + \beta\pi(1)$, $\pi(1) = (1-\alpha)\pi(0) + (1-\beta)\pi(1)$ und $\pi(0) + \pi(1) = 1$. Die Lösung ist

$$\pi(0) = \frac{\beta}{1+\beta-\alpha}, \quad \pi(1) = \frac{1-\alpha}{1+\beta-\alpha}.$$

§10 Schätzen und Testen

Im letzten Kapitel wollen wir einen kurzen Einblick in die mathematische Statistik geben. Man unterscheidet zwischen der *deskriptiven Statistik* und der *schließenden Statistik*. Die deskriptive Statistik faßt Datensätze zusammen und macht deren Besonderheiten mit Hilfe von Kennzahlen und Grafiken sichtbar. Wir wollen uns damit hier nicht befassen. Die schließende Statistik betrachtet Beobachtungen als Realisierungen von Zufallsgrößen und zieht Rückschlüsse auf die zugrunde liegende Verteilung. Dabei ist das Ziel, Zufallsfehler und „echte Effekte" unter der Angabe der statistischen Unsicherheit zu beschreiben. Die bisher behandelte Wahrscheinlichkeitstheorie ist deduktiver Natur, die Statistik induktiver Natur.

Die mathematische Statistik unterscheidet sich von den bisher betrachteten Modellen darin, daß man mit Klassen von möglichen Verteilungen arbeitet. In den hier betrachteten Fällen enthalten diese einen *strukturellen Parameter*, der meist reellwertig ist und direkt mit der ursprünglichen Fragestellung zusammenhängt. 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 Parameterwerte mit den Daten verträglich sind (*statistische Tests*), und man möchte Schranken berechnen, die einen unbekannten Parameter mit vorgegebener Wahrscheinlichkeit einfangen (*Konfidenzintervalle*). Den dritten Punkt betrachten wir hier nicht.

Schätzprobleme

Wir starten direkt mit der Konstruktion von Schätzern. Gegeben sei eine Familie $\{P_\theta : \theta \in \Theta\}$ von Wahrscheinlichkeitsmaßen auf einem beliebigen Stichprobenraum $\mathcal{X}$ (so bezeichnet man häufig den Raum in der Statistik), versehen mit einer σ-Algebra $\mathcal{F}$. Man möchte aus vorliegenden Beobachtungen (Realisierungen von Zufallsgrößen), die nach P_θ verteilt sind, den tatsächlich zugrunde liegenden Parameter θ schätzen. Es sei eine zu schätzende Funktion $g \colon \Theta \to \mathbb{R}$ gegeben. Ist der Parameter selbst zu schätzen, so ist $g(\theta) = \theta$. Aber es gibt einfache Fälle, in denen g etwas komplizierter aussieht. 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 $g(\theta) = \mu$.

(10.1) Definition.

(a) Ist $\mathcal{X}$ eine endliche oder abzählbare Menge, so heißt die Funktion $\theta \mapsto L_x(\theta) = P_\theta(x)$ mit $x \in \mathcal{X}$ *Likelihood-Funktion*. Es seien X eine Zufallsvariable, definiert auf einem allgemeinen W.-Raum, mit Werten in $\mathcal{X} = \mathbb{R}^n$ 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 $g(\theta)$.

(10.2) Bemerkung. $L_x(\theta)$ gibt an, wie wahrscheinlich die gemachte Beobachtung x ist, wenn die zugrunde liegende Verteilung P_θ ist. Wenn man nun θ nicht kennt, ist es plausibel anzunehmen, daß man einen typischen Wert beobachtet hat. Typisch soll hier also der $P_\theta(x)$ bzw. $f(x|\theta)$ maximierende Wert für θ sein. In vielen Fällen ist Θ ein Intervall in $\mathbb{R}$, und ein Maximum-Likelihood-Schätzer kann durch Differentiation gefunden werden. Es ist häufig zweckmäßig, statt L_x die Funktion $\log L_x$ zu betrachten. Sie hat wegen der Monotonie des Logarithmus das Maximum an der gleichen Stelle.

Eine binomialverteilte Zufallsgröße X stand bisher im Mittelpunkt des Interesses. Sie stellt eine typische Klasse von Zufallsexperimenten dar. Die Normalverteilung ist die weitaus wichtigste Verteilung. Für viele statistische Anwendungen wird vorausgesetzt, daß die diskutierten Größen normalverteilt sind (z. B. Meßfehler bei physikalischen Beobachtungen, Intelligenzquotienten in einer Population etc.). Viele Größen, die oft und unter identischen Bedingungen gemessen werden können, sind tatsächlich wenigstens genähert normalverteilt. Eine gewisse theoretische Rechtfertigung gibt der Satz von de Moivre-Laplace und der zentrale Grenzwertsatz aus Kapitel 6. Man stellt sich etwa vor, daß Meßfehler zustande kommen, indem sich kleine Fehler unabhängig überlagern. Ist dies der Fall, so ist nach dem zentralen Grenzwertsatz der gesamte Meßfehler genähert normalverteilt. Wir bestimmen daher für die genannten Verteilungen die Maximum-Likelihood-Schätzer:

(10.3) 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 $L_x(p) = b(x; n, p)$, und $(\log L_x(p))' = \frac{x}{p} - \frac{n-x}{1-p}$. Die Nullstelle findet man zu $\hat{p}(x) = \frac{x}{n}$, und es ist leicht zu sehen, daß es sich um ein Maximum von $\log L_x(p)$ handelt. $\frac{x}{n}$ ist also der Maximum-Likelihood-Schätzer für p und entspricht der naiven Mittelwertbildung.
(b) *Normalverteilung:*
Seien $X_1, X_2, \ldots, X_n$ unabhängig und normalverteilt zu den Parametern μ und σ^2 (wir schreiben im folgenden $N(\mu, \sigma^2)$-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.
(2) *(Erwartungswert bekannt, Schätzung der Varianz)*
Sei $\mu = \mu_0$ bekannt und $\sigma^2 > 0$ unbekannt. Hier ist $\Theta = \{(\mu, \sigma^2) \colon \mu = \mu_0, \sigma^2 > 0\}$. 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.

Wir haben bisher nur Maximum-Likelihood Schätzer kennengelernt. Es bleibt einen ganz wesentlichen Aspekt zu untersuchen: wie gut sind diese Schätzer? Dazu müssen wir uns zunächst sinnvolle Kriterien zur *Beurteilung* von Schätzern verschaffen. Eine allgemeine theoretische Abhandlung dazu sprengt den Rahmen einer einführenden Vorlesung. Wir begnügen uns daher an dieser Stelle damit, zwei plausible Minimalforderungen an einen guten Schätzer zu stellen. Ein guter Schätzer sollte im Mittel wenig von der zu schätzenden Größe abweichen.

(10.4) Definition. Es sei $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ eine Beobachtung, also n Zufallsgrößen, definiert auf einem gemeinsamen W.-Raum, mit Verteilungen $\{P_\theta, \theta \in \Theta\}$. Ein *Schätzer (estimator)* einer zu schätzenden Größe $g(\theta)$ mit $g\colon \Theta \to \mathbb{R}$ ist eine Zufallsgröße $S\colon \mathbb{R}^n \to \mathbb{R}$. S heißt *erwartungstreu (unbiased)* für $g(\theta)$, wenn für alle $\theta \in \Theta$ die Gleichung

$$E_\theta S = g(\theta)$$

gilt. S heißt *konsistent (consistent)* für $g(\theta)$, wenn für alle $\theta \in \Theta$ und alle $\delta > 0$

$$\lim_{n\to\infty} P_\theta\Big(|S(X_1, \ldots, X_n) - g(\theta)| > \delta\Big) = 0.$$

Wir haben nur die Klasse der Maximum-Likelihood-Schätzer betrachtet und hatten Glück bei unseren Beispielen, daß diese auch existieren (im Allgemeinen braucht natürlich ein Maximum nicht zu existieren). Nun wollen wir untersuchen, ob diese Schätzer erwartungstreu und konsistent sind. Im Sinne dieser Kriterien wären sie dann „gut“.

(10.5) Beispiele. (a) *Bernoulli-Experiment:*
Ist X binomialverteilt mit Parametern n und p, so ist $E(X/n) = p$, also ist der Schätzer $S = \frac{X}{n}$ erwartungstreu für $g(p) = p$. Das schwache Gesetz der großen Zahlen (Satz (3.31)) liefert die Konsistenz dieses Schätzers für p.
(b) *Normalverteilung:*
(1) *(Varianz bekannt, Schätzung des Erwartungswertes)*
Gegeben sei die Situation von Beispiel (10.3 (b)). 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 Markov-Ungleichung (Satz (3.29)) 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.
(2) *(Erwartungswert bekannt, Schätzung der Varianz)*
Den Maximum-Likelihood-Schätzer $\tilde\sigma^2$ können wir schreiben:

$$S_2 = \frac{\sigma^2}{n}\sum_{i=1}^n \Big(\frac{X_i - \mu_0}{\sigma}\Big)^2.$$

Nach Beispiel (7.14)(2) sind die Zufallsgrößen $X_i^* := (X_i - \mu_0)/\sigma$ standardnormalverteilt. Nach Übung 43 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) $E(S_2) = \sigma^2$ und $V(S_2) = \frac{2\sigma^4}{n}$. Damit ist S_2 erwartungstreu für σ^2, und wir erhalten entsprechend der Diskussion im Fall (1) die Konsistenz von S_2 für σ^2.

Wir verschieben die Diskussion der Güte der Schätzer $\hat\mu$ und $\hat\sigma^2$ bei unbekanntem Erwartungswert und unbekannter Varianz.

Statistische Tests
Wir motivieren die Fragestellungen der Testtheorie durch ein einfaches Problem: Ein neues Medikament soll mit einem bisher verwendeten Medikament verglichen werden. Man möchte natürlich wissen, welches besser wirkt. Bei 10 von 20 Testpersonen habe das neue Medikament eine bessere Heilung bewirkt. Wie können wir nun erfassen, ob dieses Testergebnis reiner Zufall ist oder nicht? Der Schluß, das neue Medikament auf Grund einer Testreihe auf den Markt zu bringen (oder auch nicht), kann ein fataler Fehler sein. Man kann sich stark geirrt haben. Nun kann die statistische Testtheorie Irrtümer nicht vermeiden. Sie soll uns aber Kriterien liefern, nach denen sich die sogenannten Irrtumswahrscheinlichkeiten kontrollieren lassen.

Wieder sei $\{P_\theta : \theta \in \Theta\}$ die Menge der in Frage kommenden Verteilungen, zunächst gegeben auf einer diskreten Menge $\mathcal{X}$. Es sei weiter eine Teilmenge $H \subset \Theta$ durch zusätzliche Bedingungen ausgezeichnet. Ein *statistischer Test* ist eine Entscheidungsregel $\varphi : \mathcal{X} \to \{0, 1\}$, die für jeden möglichen Wert $x \in \mathcal{X}$ festlegt, ob man sich für die *Hypothese* $\theta \in H$, beschrieben durch $\varphi(x) = 0$, oder für die *Alternative* $\theta \in \Theta \setminus H$, beschrieben durch $\varphi(x) = 1$, entscheiden soll. Die Entscheidung für die Hypothese nennt man *Annahme der Hypothese (accept the hypothesis)*, die Entscheidung für die Alternative nennt man *Verwerfen der Hypothese (reject the hypothesis)*. Es sind zwei Arten von Fehlern möglich. Ist $\theta \in H$ und wird die Hypothese verworfen, so spricht man von einem *Fehler erster Art (type I error)*, ist $\theta \in K := \Theta \setminus H$ 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)*.

(10.6) Definition. Zu einer Entscheidungsregel φ betrachten wir

$$E_\theta(\varphi) = P_\theta(\varphi = 1),$$

also die Wahrscheinlichkeit, die Hypothese zu verwerfen zu $\theta \in \Theta$. Wir sagen, daß φ ein Test zum *Niveau (level of significance)* $\alpha > 0$ ist, wenn

$$\sup_{\theta \in H} E_\theta(\varphi) \le \alpha$$

gilt. (Die Wahrscheinlichkeit eines Fehlers erster Art ist dann maximal α.) Für $\theta \in K$ heißt $E_\theta(\varphi)$ die *Macht (power)* des Tests in θ. (Ist die Macht nahe bei 1, so ist die Wahrscheinlichkeit $1 - E_\theta(\varphi)$ eines Fehlers zweiter Art klein.)

(10.7) Beispiel. Wir führen 20 unabhängige 0-1 Experimente mit unbekanntem Erfolgsparameter p durch. Also ist $\Theta = [0, 1]$. Die Anzahl der Erfolge X ist $b(k; 20, p)$-verteilt. Sei die Hypothese $H = [0, 1/2]$ (obiges Medikament führt zu einer Häufigkeit der Krankheit, die kleiner oder gleich 1/2 ist; dies möge einer Verbesserung gegenüber dem herkömmlichen Medikament entsprechen). Wir suchen den Verwerfungsbereich $R = \{c, c + 1, \ldots, n = 20\}$ in Abhängigkeit vom Niveau α. Es gilt

$$E_p(\varphi) = \sum_{k=c}^{20} \binom{20}{k} p^k (1 - p)^{20-k}.$$

Nun ist dies in p monoton wachsend, das Supremum über alle $\theta \in H$ liegt also bei $p = 1/2$. 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 $\alpha \in [0.021, 0.058] \Rightarrow c = 15$. Wir können noch die Macht diskutieren. Es gilt:

p	0.6	0.7	0.8	0.9
$E_p(\varphi)$	0.126	0.416	0.804	0.989

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 Vergrößerung der Macht, ohne dabei das Niveau des Tests zu vergrößern, ist also hier allein durch eine größer gewählte Stichprobe möglich.

Eine Hypothese oder Alternative heißt *einfach (simple)*, wenn sie aus einer einzigen Verteilung besteht. Wir setzen $\Theta = \{\theta_0, \theta_1\}, H = \{\theta_0\}$ und $P_H = P_{\theta_0}$ sowie $P_K = P_{\theta_1}$. Wir fragen uns nun in dieser recht einfachen Situation, ob unter allen möglichen Tests φ zu einem festen Niveau α ein Test mit maximaler Macht existiert. Natürlich wollen wir ihn auch konstruieren können.

(10.8) Definition. Ein Test φ^* heißt *Neyman-Pearson-Test*, wenn eine Konstante c^* mit $0 \leq c^* \leq \infty$ existiert mit

$$\varphi^*(x) = \begin{cases} 1, & \text{falls} \quad P_K(x) > c^* P_H(x), \\ 0, & \text{falls} \quad P_K(x) < c^* P_H(x). \end{cases}$$

Ein Test φ_1 heißt *schärfer (more powerful)* als φ_2, wenn $E_K(\varphi_1) > E_K(\varphi_2)$ gilt. Auf der Menge der x mit $P_K(x) = c^* P_H(x)$ darf die Entscheidungsregel φ^* beliebige Werte $\gamma(x)$ mit $0 \leq \gamma(x) \leq 1$ annehmen.

Es gibt eine Heuristik zu der so definierten Klasse von Tests: Da man die Macht maximieren möchte unter der Nebenbedingung, das Niveau unter α zu halten, sind diejenigen Beobachtungen x möglichst im Verwerfungsbereich, für die P_K groß und P_H klein ist. Der Quotient P_K/P_H soll also groß werden.

Wir kommen nun zu einem zentralen Ergebnis dieses Kapitels. *Egon Sharpe Pearson* (1895–1980) und *Jerzy Neyman* (1894–1981) haben im Jahre 1933 den folgenden Satz bewiesen. Er ist ein wichtiges Fundament der Testtheorie.

(10.9) Satz *(Neyman-Pearson-Lemma)*. Für das Testen einer einfachen Hypothese gegen eine einfache Alternative gilt: Ist φ^* Neyman-Pearson-Test, so ist φ^* mindestens so scharf wie alle anderen Tests φ mit $E_H(\varphi) \leq E_H(\varphi^*)$. Zu $0 \leq \alpha \leq 1$ existiert ein Neyman-Pearson-Test φ^* mit $E_H(\varphi^*) = \alpha$.

Beweis. Auf der Menge $S^+ := \{x : \varphi^*(x) > \varphi(x)\}$ ist $\varphi^*(x) > 0$ und damit $P_K(x) \geq c^* P_H(x)$. Entsprechend gilt auf $S^- := \{x : \varphi^*(x) < \varphi(x)\}$: $\varphi^*(x) < 1$ und damit $P_K(x) \leq c^* P_H(x)$. Damit gilt

$$\begin{aligned} E_K(\varphi^*) - E_K(\varphi) &= \sum_{x \in S^+} (\varphi^*(x) - \varphi(x)) P_K(x) + \sum_{x \in S^-} (\varphi^*(x) - \varphi(x)) P_K(x) \\ &\geq c^* \sum_{x \in \mathcal{X}} (\varphi^*(x) - \varphi(x)) P_H(x) \\ &= c^* (E_H(\varphi^*) - E_H(\varphi)) \geq 0. \end{aligned}$$

Damit ist der erste Teil gezeigt.
Für den Fall $\alpha = 0$ setze $c^* = \infty$. Es folgt unmittelbar $E_H(\varphi^*) = 0$. Für $\alpha > 0$ setze mit $c \geq 0$

$$\alpha(c) := P_H\big(P_K(x)/P_H(x) > c\big) \quad \text{und} \quad \alpha(c-0) := P_H\big(P_K(x)/P_H(x) \geq c\big).$$

Wir können annehmen, daß für jedes x $P_H(x) + P_K(x) > 0$ ist. Andere Ereignisse haben keinen Einfluß auf die Irrtumswahrscheinlichkeiten. Nun ist $\alpha(0-0) = 1$, und da $P_K(x)/P_H(x) < \infty$ für alle x mit $P_H(x) > 0$, ist $\alpha(c)$ fallend mit $\alpha(c) \to 0$ für $c \to \infty$. Weiter ist für eine Folge c_n, die strikt gegen c fällt, jedes Element der Menge $\{x\colon P_K(x)/P_H(x) > c\}$ für hinreichend großes n auch ein Element der Menge $A_{c_n} := \{x\colon P_K(x)/P_H(x) > c_n\}$. Damit folgt $\alpha(c_n) \to \alpha(c)$, die Abbildung $c \mapsto \alpha(c)$ ist also rechtsseitig stetig. Da die Mengen A_{c_n} eine fallende Folge bilden und der Durchschnitt über alle n die Menge $\{x\colon P_K(x)/P_H(x) \geq c\}$ ist, folgt insbesondere $\lim_{n\to\infty} \alpha(c_n) = \alpha(c-0)$.

Wir wählen nun $c^* = \inf\{c\colon \alpha(c) \leq \alpha\}$. Dann ist $\alpha(c^*-0) \geq \alpha \geq \alpha(c^*)$. Im Fall $\alpha(c^*-0) = \alpha(c^*)$ sei $\gamma^* = 0$. Im Fall $\alpha(c^*-0) > \alpha(c^*)$ sei

$$\gamma^* = \frac{\alpha - \alpha(c^*)}{\alpha(c^*-0) - \alpha(c^*)}.$$

Nun ist der gesuchte Test wie folgt gewählt: Auf $\{x\colon P_K(x) = c^* P_H(x)\}$ sei $\varphi^*(x) = \gamma^*$, auf dem Komplement sei φ^* durch die Definition des Neyman-Pearson-Tests gegeben. Dann ist

$$\begin{aligned} E_H(\varphi^*) &= P_H\big(P_K(x)/P_H(x) > c^*\big) + \gamma^* P_H\big(P_K(x) = c^* P_H(x)\big) \\ &= \alpha(c^*) + \gamma^*\big(\alpha(c^*-0) - \alpha(c^*)\big) = \alpha. \end{aligned}$$

Damit ist der Satz bewiesen. $\square$

Eine für das weitere Vorgehen wichtige Bemerkung ist, daß dieser Satz einfach auf die Situation absolutstetig verteilter Zufallsgrößen übertragbar ist. Man ersetzt dabei P_{θ_0} durch eine Dichte $f(\cdot|\theta_0)$ und P_{θ_1} durch eine andere Dichte $f(\cdot|\theta_1)$. Dann kann man den Beweis Schritt für Schritt übernehmen. Wir müssen allerdings schon in der Definition eines Tests die Abbildung φ als meßbar bezüglich der σ-Algebra $\mathcal{F}$ annehmen.

(10.10) Beispiele. (a) *Bernoulli-Experiment:*
Sei X binomialverteilt mit Parameter n und p. Die Hypothese sei $p_H = 1/2$ und die Alternative $p_K > 1/2$ (p_K fest gewählt). Nun ist

$$q(x) = \frac{\binom{n}{x} p_K^x (1-p_K)^{n-x}}{\binom{n}{x} p_H^x (1-p_H)^{n-x}}$$

strikt wachsend in x. Damit ist der Verwerfungsbereich $\{x\colon q(x) > a\}$ für jedes a ein Intervall $\{c, c+1, \ldots, n\}$. Für einen optimalen Test φ im Sinne des Satzes von Neyman-Pearson existiert also eine Zahl c mit $\varphi(x) = 1$ für $x \geq c$ und $\varphi(x) = 0$ für $x < c-1$. Im Fall $q(c-1) = a$ ist $\varphi(c-1)$ eine beliebige Zahl γ. Im Fall

$q(c-1) < a$ muß $\varphi(c-1) = 0$ sein. Das Niveau des Tests berechnet sich nun zu $E_H(\varphi) = P_H(X \geq c) + \gamma P_H(X = c-1)$. Zu vorgegebenem $\alpha > 0$ bestimmt man nun c und γ aus der Identität $E_H(\varphi) = \alpha$. Dies liefert den schärfsten Test zum Niveau α. Wir stellen bei diesem einfachen Beispiel noch eine kleine Überraschung fest: In der Gleichung $E_H(\varphi) = \alpha$ kommt der fest gewählte Parameter p_K nicht vor. Wir erhalten hier also automatisch für alle $p_K > 1/2$ einen schärfsten Test zum Niveau α. Dies bedeutet, daß wir den schärfsten Test für die sogenannte *zusammengesetzte* Alternative $K = \{p\colon p > 1/2\}$ gefunden haben. Dies liefert der Satz von Neyman-Pearson im allgemeinen nicht.

(b) *Normalverteilung:*

Es seien $X_1, X_2, \ldots, X_n$ unabhängig und $N(\mu, 1)$-verteilt. Wir testen die Hypothese $H = \{(0,1)\}$ gegen die Alternative $K = \{(1,1)\}$. Nun ist

$$f(x|(0,1)) = \Big(\frac{1}{\sqrt{2\pi}}\Big)^n \exp\Big(-\frac{1}{2}\sum_{i=1}^n x_i^2\Big)$$

und

$$f(x|(1,1)) = \Big(\frac{1}{\sqrt{2\pi}}\Big)^n \exp\Big(-\frac{1}{2}\sum_{i=1}^n (x_i-1)^2\Big).$$

Somit ist

$$q(x) = \frac{f(x|(1,1))}{f(x|(0,1))} = \exp\Big(-\frac{1}{2}\Big(\sum_{i=1}^n (x_i-1)^2 - \sum_{i=1}^n x_i^2\Big)\Big).$$

Mit $\sum_{i=1}^n (x_i-1)^2 - \sum_{i=1}^n x_i^2 = -2n(1/n \sum_{i=1}^n x_i) + n$ folgt für den Neyman-Pearson Verwerfungsbereich:

$$\{x : q(x) > k\} = \Big\{x : \frac{1}{n}\sum_{i=1}^n x_i > \frac{1}{n}\log k + \frac{1}{2}\Big\}.$$

Da $\overline{X} = 1/n \sum_{i=1}^n X_i$ unter der Hypothese $N(0, 1/n)$-verteilt ist, bestimmt sich die Schranke $c = \frac{1}{n}\log k + \frac{1}{2}$ (und daraus dann die eigentliche Schranke k) mittels der Identität $P(\overline{X} > c) = \alpha$. In der Praxis ermittelt man zu vorgegebenem α aus der Tabelle einer standardnormalverteilten Zufallsgröße Y den Wert c' mittels $\Phi(c') = P(Y < c') = 1 - \alpha$. Dann ist der gesuchte Wert $c = \frac{1}{\sqrt{n}}c'$, und daraus bestimmt man k.

Wir kommen abschließend zu einem sehr wichtigen Testverfahren, bekannt unter dem Namen *t-Test*. Es seien erneut $X_1, X_2, \ldots, X_n$ unabhängig und $N(\mu, \sigma^2)$-verteilt, hier nun mit unbekanntem μ und σ^2. Dies ist sicherlich die in der Praxis am häufigsten vorkommende Situation. Für ein gegebenes festes μ_0 wollen wir die Hypothese $\mu = \mu_0$ gegen die Alternative $\mu \neq \mu_0$ testen. Es sei also

$$H = \{(\mu, \sigma^2)\colon \mu = \mu_0, \sigma^2 > 0\} \tag{10.11}$$

und

$$K = \{(\mu, \sigma^2)\colon \mu \neq \mu_0, \sigma^2 > 0\}. \tag{10.12}$$

Sei $\Theta = H \cup K$. Wir beschäftigen uns also mit nicht einfachen Hypothesen und Alternativen. Man spricht von sogenannten *zusammengesetzten Hypothesen (composite Hyphothesis)* und Alternativen. Hier ist nun das Resultat von Neyman und Pearson nicht ohne weiteres anwendbar. Dort war der Verwerfungsbereich konstruiert worden aus den Beobachtungen x, für die $P_K(x)$ groß und $P_H(x)$ klein ist. Kommen nun mehrere Parameter aus der Menge Θ in Betracht, mag $P_\theta(x)$ für manche Werte von $\theta \in \Theta$ groß, für andere hingegen klein sein. Das obige Verfahren ist nicht mehr klar. Wir wollen nun an dieser Stelle nicht allgemeine (gute) Tests für zusammengesetzte Alternativen entwickeln. Wir müssen uns mit der Bemerkung begnügen, daß eine naheliegende Verallgemeinerung der Neyman-Pearson Methode zu tatsächlich guten Tests führt. Wir betrachten dies speziell für den Fall der normalverteilten Zufallsgrößen. Gegeben sei der *Likelihood-Quotient*

$$q(x) = \frac{\sup\{L_x(\theta) : \theta \in K\}}{\sup\{L_x(\theta) : \theta \in H\}}.$$

Eine Entscheidungsregel φ heißt nun *Likelihood-Quotienten-Test*, wenn für ein geeignetes $c \in \mathbb{R}$ im Fall $q(x) > c$ die Hypothese verworfen wird und sie im Fall $q(x) < c$ angenommen wird. Im Falle einfacher Hypothesen und Alternativen ist dies natürlich der Neyman-Pearson-Test.

Für die in Beispiel (10.3)(b) gegebenen Dichten $f(x|\theta)$ von $X = (X_1, \ldots, X_n)$ und die in (10.11) und (10.12) gegebenen Mengen gilt nun $\sup\{f(x|\theta) : \theta \in K\} = \sup\{f(x|\theta) : \theta \in \Theta\}$, denn K liegt dicht in Θ. Das Supremum über alle $\theta \in \Theta$ wird nach Beispiel (10.3)(b)(3) für eine feste Beobachtung x in $(\hat{\mu}, \hat{\sigma}^2)$ angenommen. Ist $\mu = \mu_0$ fest, so folgt nach Beispiel (10.3)(b)(2)

$$\sup\{f(x|\theta) : \theta \in H\} = f(x|(\mu_0, \tilde{\sigma}^2)).$$

Nun ist

$$f(x|(\hat{\mu}, \hat{\sigma}^2)) = \Big(\frac{1}{\sqrt{2\pi}\hat{\sigma}}\Big)^n \exp(-n/2) \quad \text{und} \quad f(x|(\mu_0, \tilde{\sigma}^2)) = \Big(\frac{1}{\sqrt{2\pi}\tilde{\sigma}}\Big)^n \exp(-n/2).$$

Daraus folgt

$$q(x) = \Big(\frac{\tilde{\sigma}}{\hat{\sigma}}\Big)^n. \tag{10.13}$$

Ist nun φ ein Likelihood-Quotienten-Test, so gilt für ein geeignetes c

$$\{x : q(x) > c\} \subset \{x : \varphi(x) = 1\} \quad \text{und} \quad \{x : q(x) < c\} \subset \{x : \varphi(x) = 0\}.$$

Für $q(\cdot)$ wie in (10.13) ist somit für $c' = c^{2/n}$ $\varphi(x) = 1$ auf der Menge $\{x : \tilde{\sigma}^2 > c'\hat{\sigma}^2\}$ und $\varphi(x) = 0$ auf $\{x : \tilde{\sigma}^2 < c'\hat{\sigma}^2\}$. Eine kleine Zwischenrechnung liefert nun

$$\frac{\tilde{\sigma}^2}{\hat{\sigma}^2} = 1 + \frac{(\hat{\mu} - \mu_0)^2}{\hat{\sigma}^2}. \tag{10.14}$$

Wir definieren nun eine neue Zufallsgröße:

$$T(x) = \frac{\sqrt{n}(\hat{\mu} - \mu_0)}{s(x)} \quad \text{mit} \quad s(x) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2}.$$

Damit ist der zweite Summand in (10.14) ein Vielfaches von $|T(x)|^2$ und somit ist für ein geeignetes $k \in \mathbb{R}$ die Bedingung $q(x) > c$ äquivalent zu $|T(x)| > k$. Um nun erneut zu einem vorgegebenen Niveau $\alpha > 0$ die Zahl k bestimmen zu können, für die φ gerade Niveau α hat, muß man abschließend die Verteilung von $T(X)$ unter der Hypothese kennen. Dies ist recht aufwendig. Die Umformulierung des Tests φ mit Hilfe der Zufallsgröße $T(X)$ wird erst verständlich, wenn man erkannt hat, daß man die Verteilung dieser Größe gut im Griff hat. Das man an der Verteilung von $T(X)$ allgemeines Interesse hat, zeigt uns aber auch die folgende einfache Beobachtung. Mit $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ ist

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

$N(0,1)$-verteilt. Ersetzen wir nun die unbekannte Standardabweichung σ durch den erwartungstreuen Schätzer $s(X)$ (wir werden noch sehen, daß dies ein erwartungstreuer Schätzer ist), kommen wir unmittelbar auf $T(X)$. Konnten wir bei bekannter Varianz also bisher die Tabelle der $N(0,1)$-Verteilung nutzen, müssen wir nun zunächst die Verteilung von $T(X)$ bestimmen. Dies geschieht in mehreren Schritten.

1. Schritt: Setzen wir $Y_i = (X_i - \mu_0)/\sigma$, so ist Y_i unter der Hypothese unabhängig und $N(0,1)$-verteilt. Ist $Y = (Y_1, \dots, Y_n)$, so gilt $s(X) = \sigma s(Y)$ und somit ist

$$T(X) = \frac{\sqrt{n}\,\bar{Y}}{s(Y)}, \tag{10.15}$$

wobei $\bar{Y} := \frac{1}{n}\sum_{i=1}^{n} Y_i$. Damit sehen wir, daß die Verteilung von $T(X)$ unabhängig von μ_0 und σ^2 ist. Dies ist sehr wichtig, denn wir kennen diese Parameter nicht.

2. Schritt: Wir zeigen hier, daß $\bar{Y}$ und $s^2(Y)$ unabhängig sind.

(10.16) Lemma. *Es seien A eine orthogonale $n \times n$-Matrix und $Z = (Z_1, \dots, Z_n)$ der Zufallsvektor $A(Y)$. Dann sind $Z_1, \dots, Z_n$ unabhängig und $N(0,1)$-verteilt.*

Beweis. Es bezeichne $g(y_1, \dots, y_n)$ die Dichte von Y. Für jedes n-dimensionale Rechteck $[a,b[$ gilt nach der Transformationsformel für orthogonale Transformationen:

$$\begin{aligned} P\big(A(Y) \in [a,b[\big) &= P\big(Y \in A^{-1}([a,b[\big) = \int_{A^{-1}([a,b[)} g(y_1, \dots, y_n)dy_1 \cdots dy_n \\ &= \int_{[a,b[} g(y_1, \dots, y_n)dy_1 \cdots dy_n = P\big(Y \in [a,b[\big). \end{aligned}$$

□

Wir wenden dieses Lemma auf die spezielle orthogonale Matrix A an, die in der ersten Zeile den Vektor $(1/\sqrt{n}, \dots, 1/\sqrt{n})$ als Eintrag hat. Diese Vorgabe kann nach dem Gram-Schmidtschen Orthonormalisierungs-Verfahren zu einer orthogonalen Matrix aufgefüllt werden. Hier ist dann

$$Z_1 = \frac{1}{\sqrt{n}}(Y_1 + \cdots + Y_n) = \sqrt{n}\,\bar{Y}.$$

Es bezeichne $\langle\cdot,\cdot\rangle$ das gewöhnliche Skalarprodukt in $\mathbb{R}^n$. Dann gilt wegen der Orthogonalität von A

$$\begin{aligned} Z_2^2 + \cdots + Z_n^2 &= \langle Z, Z\rangle - n\bar{Y}^2 = \langle Y, Y\rangle - n\bar{Y}^2 \\ &= \sum_{i=1}^n (Y_i - \bar{Y})^2 = (n-1)s^2(Y). \end{aligned}$$

Da die Z_i unabhängig sind, ist Z_1 von $Z_2^2 + \cdots + Z_n^2$ unabhängig, und somit $\bar{Y}$ von $s^2(Y)$.

3. Schritt: Wir haben in einer Übungsaufgabe gesehen:

(10.17) Lemma. *Die Verteilung der Summe der Quadrate von n unabhängigen $N(0,1)$-verteilten Zufallsgrößen nennt man eine χ_n^2-Verteilung (χ^2-Verteilung mit n Freiheitsgraden, χ^2-distribution with n degrees of freedom). Ihre Dichte ist gegeben durch*

$$g_n(x) = \frac{1}{2^{n/2}\Gamma(n/2)}\, x^{(n/2)-1}\, e^{-x/2}, \quad x > 0.$$

Der Erwartungswert einer χ_n^2-verteilten Zufallsgröße ist n, die Varianz $2n$.

Für $n = 2$ ist $g_2(x) = (1/2)\exp(-x/2)$. Dies ist die Dichte der Exponentialverteilung zum Parameter $1/2$, siehe Beispiel (7.9)(4).

Aus diesem Lemma und Schritt 2 folgt: $(n-1)s^2(Y)$ ist χ_{n-1}^2-verteilt.

4. Schritt: Wir nutzen nun aus, daß Zähler und Nenner in (10.15) nach Schritt 2 unabhängig sind. Wir erhalten die Verteilung von $T(X)$ als Spezialfall des folgenden Lemmas:

(10.18) Lemma. *Sind W und U_n unabhängige Zufallsvariable, und ist W $N(0,1)$-verteilt und U_n χ_n^2-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_n^2-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 $\Gamma(1/2) = \sqrt{\pi}$:

$$P(T_n < \lambda) = \int_0^\infty \int_{-\infty}^{\lambda} \frac{1}{\sqrt{n}2^{\frac{n+1}{2}}\Gamma(n/2)\Gamma(1/2)} \exp\left(-\frac{1}{2}\Big(y + \frac{y+t^2}{n}\Big)\right) y^{\frac{n+1}{2}-1}\, dt\, dy.$$

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 zusammen:

(10.19) Satz. Sind $X_1, \ldots, X_n$ unabhängige $N(\mu_0, \sigma^2)$-verteilte Zufallsgrößen, dann ist $T(X)$ t_{n-1}-verteilt.

(10.20) Beispiel. Wir haben ein praktisches Testverfahren gewonnen: Wir wollen auf der Grundlage von n Beobachtungen $\mu = \mu_0$ gegen $\mu \neq \mu_0$ testen. Die Verwerfungswahrscheinlichkeit unter der Hypothese ist $P(|T(X)| > k)$. 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

$$|\hat{\mu} - \mu_0| > t_{n-1,1-\alpha/2} \frac{s(x)}{\sqrt{n}}.$$

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 $s(x)$ 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.

Wir schließen diese Vorlesung ab mit der Diskussion, ob die in Beispiel (10.3(b))(3) gewonnenen Maximum-Likelihood-Schätzer erwartungstreu und konsitent sind. $\hat{\mu}$ ist erwartungstreu und konsistent für μ. Dies hatten wir bereits diskutiert. Es gilt

$$S_3 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2 = \frac{n-1}{n}s^2(X).$$

Somit ist mit $s(X) = \sigma s(Y)$ und nach Lemma (10.17)

$$E(S_3) = \frac{\sigma^2}{n}E((n-1)s^2(Y)) = \sigma^2\frac{n-1}{n}.$$

Also ist S_3 *nicht* erwartungstreu für σ^2, aber $s^2(X)$ ist ein erwartungstreuer Schätzer der Varianz. Weiter ist $(n-1)s^2(Y) = (n-1)/\sigma^2 s^2(X)$ χ^2_{n-1}-verteilt, hat also Varianz $2(n-1)$. Daraus folgt $V(s^2(X)) = \frac{2\sigma^4}{n-1}$, $s^2(X)$ ist also auch konsistent für σ^2.

§11 Historischer Anhang

Diesem historischen Anhang liegen das Buch *Die Entwicklung der Wahrscheinlichkeitstheorie von den Anfängen bis 1933* von Ivo Schneider, die Sammlung *Paradoxa* von Gábor Székely sowie die Ausarbeitungen von Ulrich Krengel zur *Wahrscheinlichkeitstheorie* und zur *Mathematischen Statistik* von Hermann Witting, erschienen im Band *Ein Jahrhundert Mathematik 1890–1990, Festschrift zum Jubiläum der DMV (Deutsche Mathematiker Vereinigung)*, zugrunde.

Es wird die Entwicklung der Glücksspielrechnung bis ins 17. Jahrhundert, ein kleiner Einblick in eine philosophische Diskussion des Wahrscheinlichkeitsbegriffs, die erste Phase der Wahrscheinlichkeitsrechnung, eingeleitet durch *Jakob Bernoulli (1654–1705)*, die Geschichte des schwachen Gesetzes der großen Zahlen, die Anstöße aus der Physik und die Axiomatisierung der Wahrscheinlichkeitsrechnung durch *Andrey Nikolayevich Kolmogoroff (1903–1987)* 1933 betrachtet. Weiter wird ein kurzer Einblick der Entwicklung der Wahrscheinlichkeitstheorie bis 1945 sowie die Grundlegung der Mathematischen Statistik vorgestellt.

Die Glücksspielrechnung bis zum 17. Jahrhundert

Die Glücksspielrechnung vor Jakob Bernoulli, also bis hinein in das 17. Jahrhundert, stellt sich in groben Zügen wie folgt dar:

Man versuchte lange Zeit, Aufgaben aus dem Bereich der Glücksspiele mit mathematischen Mitteln zu lösen, ohne den Begriff der Wahrscheinlichkeit dabei entwickelt zu haben. Quellen aus der Mitte des 13 Jahrhunderts deuten an, daß man sich in der islamischen Welt des Mittelalters bereits mit Glücksspielen beschäftigt hat, doch das strikte Glücksspielverbot des Korans wird die Ursache sein, daß man keine schriftlichen Quellen finden kann. Bis zum Ende des Mittelalters war das Würfeln das populärste Glücksspiel. Der Ausdruck *Hasard* für Glücksspiel bezieht sich auf den Würfel: das Wort kommt aus dem arabischen „az-zahr", was „Würfel zum Spielen" bedeutet. Die Kartenspiele verbreiten sich in Europa erst im 14. Jahrhundert. Griechischer Überlieferungen nach war es *Palamedeo*, der den Würfel erfand, um während der langwierigen Belagerung von Troja die sich langweilenden griechischen Soldaten zu unterhalten. Eine häufig genannte Quelle des 13. Jahrhunderts ist die *Pseudo-Ovidius, De Vetula*, entstanden zwischen 1222 und 1268, in der man eine Antwort auf die Frage gibt, mit welchen Erfolgsaussichten auf Augenzahlen zwischen 3 und 18 beim Wurf mit drei Würfeln gesetzt werden kann. Diese Quelle enthält einige arabische Fachtermini. Dies ist wohl in der zeitlichen Nähe dieser Quelle zu den Kreuzzügen begründet, die das christliche Abendland und den Islam auch in einer friedlichen Art und Weise zueinander geführt haben.

In einer häufig genannten weiteren Quelle, um 1400 entstanden, deren Herkunft nicht geklärt ist, beschäftigt man sich mit einem Spezialfall des sogenannten Teilungsproblems (man nennt es in der Literatur auch häufig das Aufteilungsproblem). Dort heißt es gleich zu Beginn:

Zwei Männer spielen Schach und setzen (je) einen Dukaten ein auf drei Gewinnspiele. Es trifft sich, daß der erste zwei Spiele gegen den zweiten gewinnt. Ich frage, welchen Anspruch auf Gewinn von diesem Dukaten der erste gegenüber dem zweiten haben wird, wenn sie nicht weiterspielen. (Handschrift Magl. Cl. XI, 120, Nationalbibliothek Florenz)

Tatsächlich wurde in dieser Schrift das Problem gelöst. Man vermutet heute, daß dieses Problem arabischen Ursprungs ist oder zumindest durch arabische Vermittlung nach Italien gelangte. In einem Buch von *Frac Luca Paccioli (1445-1517)* mit dem Titel *Summa de Arithmetica, Geometrica, Proportioni et Proportinalita* (1494) erscheint das Teilungsproblem wie folgt:

Eine Gesellschaft spielt Ball auf 60, 10 Punkte für das Einzelspiel (dies bedeutet: die Partei gewinnt, die zuerst 6 Einzelspiel für sich entscheidet) Weiter heißt es: *Sie setzen 10 Dukaten ein. Aufgrund gewisser Umstände können sie nicht zu Ende spielen; dabei hat eine Partei 50 und die andere 30. Man fragt, welcher Anteil des Einsatzes jeder Partei zusteht. Für dieses Problem habe ich verschiedene Lösungsvorschläge, in die eine oder andere Richtung, vorgefunden; alle kommen mir ungereimt vor in Bezug auf einige Argumente. Aber die Wahrheit ist das, was ich sagen werde, zusammen mit dem richtigen Weg.*

So kann man sich täuschen: tatsächlich war die Lösung von Paccioli falsch. In unserer Sprache kann man das Teilungsproblem so formulieren: zwei Personen spielen ein gerechtes Spiel (beide haben dieselbe Gewinnchance), und sie besprechen, daß derjenige von ihnen, der zuerst sechs Runden gewinnt, den ganzen Gewinn davonträgt. Wie soll der ganze Gewinn auf gerechte Weise aufgeteilt werden, wenn das Spiel abgebrochen wird zu einem Zeitpunkt, zu dem noch keiner die Bedingung erfüllt hat? Wir betrachten, wie Paccioli, das Beispiel, daß der erste fünf und der zweite nur drei Runden gewonnen hat. Paccioli schlug vor, im Verhältnis der bei Abbruch gewonnenen Spiele aufzuteilen, also 5:3. *Nicolo Fontana Tartaglia (1499–1557)* schlug 2:1 vor. Man interpretiert seine Ausführungen so: der erste Spieler hat zwei Runden mehr gewonnen, also ein Drittel der notwendigen sechs Runden. Dafür soll er ein Drittel des Gewinns erhalten. Vom Rest soll jeder der beiden die Hälfte bekommen. Diese Lösung ist auch falsch, obwohl Tartaglia ein begnadeter Mathematiker war. Tatsächlich haben erst 1654 unabhängig voneinander *Blaise Pascal (1623–1662)* und *Pierre de Fermat (1601–1665)* die richtige Antwort auf die Frage gegeben. Wir diskutieren dies und die Lösung etwas weiter unten.

Interessant erscheint, daß Tartaglia das Teilungsproblem ohne den Begriff des Glücksspiels diskutierte. In der zweiten Hälfte des 16. Jahrhunderts schienen sich die Zeiten dann aber gelockert zu haben. Um 1564 verfasste *Girolamo Cardano (1501–1576)* ein Manusskript über Glücksspiele, welches im Jahre 1663 unter dem Titel *De ludo aleae* erschien. In diesem Buch werden viele Näherungslösungen konkreter Glücksspiele aufgeführt. Da aber auch das Prinzip des fairen Spiels diskutiert wurde, glaubt man an ein intuitives Erfassen des schwachen Gesetzes der großen Zahlen. Die eigentlichen Wegbereiter der Glücksspielrechnung und Wegbereiter der Bernoullischen Wahrscheinlichkeitsrechnung, es sind dies Pascal, Fermat und *Christiaan Huygens (1629–1695)*, haben die Leistungen der genannten Italiener niemals erwähnt, obwohl die Veröffentlichungen der italienischen Schule bis in die zweite Hälfte des 17. Jh. reichen. Einen Hauptgrund für dieses Verhalten sieht man heute darin, daß in jener Zeit durch die Wissenschaftler *François Viète (1540–1603)* und *René Descartes (1596–1650)* die absolute Richtigkeit und Beweisbarkeit von mathematischen Aussagen in den Vordergrund rückte. So sah man die vielen Vorschläge zur Lösung des Teilungsproblems wohl eher nur als mathematische Folklore an. Der *Chevalier de Méré, Antoine Gombaud (1610–1685)* gilt daher als Symbol dieser Folklore, mit der sich Pascal und Fermat in ihrem berühmten Briefwechsel im Jahre 1654 befaßten. Das zentrale Thema dieses Briefwechsels ist das Teilungsproblem. Mit kombinatorischen

Methoden bestimmt Pascal, wie sich der Anspruch eines Spielers auf den Einsatz mit jedem gewonnenen Spiel ändert. Man empfindet heute seine Darstellung der Lösung des Teilungsproblems als sehr elegant. Pascal notierte die Lösung in seinem Werk *Triangle arithmétique* im Jahr 1654. Er legte dieses Werk dem mathematisch überlegenen Fermat zur Beurteilung vor. Dieser war auf anderem Weg zu denselben Resultaten gelangt.
Wir diskutieren die Lösung zunächst für das oben gegebene Zahlenbeispiel. Zunächst bestimmt man die Gewinnchancen der beiden Spieler, also die Chance, daß der erste Spieler nur eine einzige Runde gewinnen muß, der zweite jedoch drei Runden. Führt man nun in Gedanken - und dies ist der Lösungsweg von Fermat - noch drei weitere Runden aus, so gibt es 8 mögliche Resultate, die gleichwahrscheinlich sind. Von diesen erhält der zweite Spieler nur in genau einem Fall den Gewinn (wenn er alle drei Runden gewinnt). In allen anderen Fällen ist der erste Spieler der Gewinner. Somit ist das gerechte Aufteilungsverhältnis 7:1. Wir erwähnen noch eine Lösungsformel für den allgemeinen Fall, daß für den Erhalt des Gewinns der erste Spieler noch n weitere Runden gewinnen muß und der zweite m. Die Formel stammt auch von Pascal und Fermat. Die Chance, daß der erste Spieler den Gewinn erhält, ist gleich

$$\frac{1}{2^{m+n-1}} \sum_{j=n}^{n+m-1} \binom{n+m-1}{j},$$

wobei die Anzahl der fiktiven Runden $n+m-1$ ist und alle 2^{n+m-1} Ergebnisse gleichwahrscheinlich sind. Wir verstehen diese Formel nach einem Kurs zur Stochastik heute unmittelbar. Das bedeutet aber nicht, daß dieses Problem etwa sehr einfach ist.
In Paris sprach man im Jahre 1654 vermutlich von der Entdeckung der Wahrscheinlichkeitsrechnung. In diesem Jahr besuchte Huygens Paris, traf Pascal und Fermat allerdings nicht an. Er hörte aber von ihren wichtigsten Ergebnissen. Mit Hilfe des Prinzips der fairen Wette (welches auch Cardano schon kannte) fand er nun einen Zugang zum Begriff eines zufälligen Ereignisses mit Hilfe von Erwartungswerten. Im Jahre 1656 legte er Pascal und Fermat ein Manuskript vor. Die beiden billigten es und schickten ihm drei Probleme. Eines von ihnen ist das berühmte Problem des Ruins von Spielern, auf das die *Irrfahrten* von heute zurückgehen. Huygens schrieb dann ein Buch über Wahrscheinlichkeitsrechnung. Dieses Werk wurde im Jahre 1657 unter dem Titel *De Ratiociniis in Aleae Ludo* als ein Teil (genauer als fünftes Buch) von Schootens *Exerciationes Mathematicarum* publiziert. Es enthält die Lösung des Teilungsproblems für den Fall von drei Spielern. Die Wirkung dieses Traktats übertrifft die der *Triangle arithmétique* von Pascal oder des Briefwechsels zwischen Pascal und Fermat. Ein Grund dafür ist auch, daß es fünf offene Probleme enthält, die Huygens seinen Lesern zur Lösung überließ. Im folgenden halben Jahrhundert stützen sich alle bekannten Mathematiker, die sich der Probleme der Wahrscheinlichkeitsrechnung annahmen, auf Huygens Werk. Besonders auffällig ist dies bei Jakob Bernoulli, doch dazu kommen wir später. Huygens 4.Problem lautet beispielsweise:
Zwei Spieler A und B nehmen sich 12 Steine, 4 weiße und 8 schwarze; A wettet mit B darauf, daß er mit verbundenen Augen 7 Steine davon wegnehmen wird, unter denen 3 weiße sein werden. Gesucht ist das Verhältnis der Erwartung des A zu der Erwartung des B. Huygens löste das Problem erst 1665: das Verhältnis ist 35:64.

Wir wollen abschließend ein paar Zitate aus dem Briefwechsel von Pascal und Fermat betrachten und dabei neben dem Teilungsproblem auch an das Würfelproblem von de Méré (siehe Beispiel (1.5)(4)) erinnern. Dabei wollen wir die Verwirrung von de Méré aufklären.
Vermutlich war es *Gottfried Wilhelm von Leibniz (1646–1716)*, der zu berichten wußte, daß der Chevalier de Méré auf dem Wege zu seinem Landsitz bei Pitou Pascal traf und an ihn zwei Fragen über Glücksspiele richtete. Über diese Fragen, die das Teilungsproblem und Beispiel (1.5)(4) betreffen, führte Pascal dann den berühmten Briefwechsel mit Fermat.
Auf einen Brief von Fermat an Pascal erwiderte Pascal am Mittwoch, den 29. Juli 1654 unter anderem:
Ich bewundere die Lösungsmethode beim Spielabbruch mehr als die für die Würfel; ich habe einige Personen die für die Würfel finden sehen, wie Herrn Chevalier de Méré, der mir übrigens diese Probleme vorgelegt hat, ...; aber Herr de Méré hat niemals den richtigen Wert beim Spielabbruch und auch keinen Ansatz, um dahin zu gelangen, finden können, so daß ich mich für den einzigen hielt, dem dieses Verhältnis bekannt war.... Denn ich möchte von nun an, wenn möglich, mein Herz öffnen, so sehr freut es mich, uns in Übereinstimmung zu sehen. Ich sehe wohl, daß die Wahrheit in Toulouse und in Paris dieselbe ist.
Heute weiß man, daß die de Méré zugesprochenen Paradoxon (siehe das Buch von Székely) schon seit langem allgemein bekannt waren. Bezüglich des Teilungsproblems hatten wir dies weiter oben bereits ausgeführt.
De Méré war nicht zufrieden, daß Pascal das Problem der Würfel nur löste und so die Richtigkeit der Antwort bestätigte; er konnte aus ihr nicht ersehen, wie der von ihm gesehene Widerspruch gelöst wurde. Pascal schrieb weiter in seinem Brief vom 29.7.1654:
Ich habe nicht die Zeit, Ihnen eine Schwierigkeit zu erläutern, die M... (de Méré) sehr befremdet, denn er ist ein sehr tüchtiger Kopf, aber er ist kein Mathematiker (das ist, wie Sie wissen, ein großer Mangel), und er begreift nicht einmal, daß eine mathematische Linie bis ins unendliche teilbar ist, und ist zutiefst davon überzeugt, daß sie sich aus einer endlichen Zahl von Punkten zusammensetzt; ich habe ihn niemals davon abbringen können. Wenn Sie das zustande brächten, würden Sie ihn vollkommen machen.
Was waren die Schwierigkeiten von de Méré? Zur Erinnerung: wir hatten gesehen, daß der gesuchte Wert der Wahrscheinlichkeit für „mindestens einmal eine Sechs bei vier Würfen" größer als 1/2, für „mindestens einen Doppelsechser bei 24 Würfe" hingegen kleiner als 1/2 ist. Im zweiten Fall ist die Wahrscheinlichkeit nur bei 25 Würfen (mindestens einen Doppelsechser zu würfeln) größer als 1/2. Einmal liegt der „kritische Wert" bei 4, im zweiten Fall bei 25. Dies schien für de Méré ein Widerspruch zur bis dahin von vielen als gültig angenommenen *Proportionalitätsregel der kritischen Werte* zu sein. Danach müßte zu einem Sechstel der Wahrscheinlichkeit der sechsfache kritische Wert gehören. Erst *Abraham de Moivre (1667–1754)* führte in seinem Buch *Doctrine of Chances*, 1718 erschienen, genau aus, daß diese Proportionalitätsregel nicht weit entfernt ist von der Wahrheit. Man kann natürlich den kritischen Wert k durch Auflösung der Gleichung

$$(1-p)^x = \frac{1}{2}$$

bestimmen, wenn $p \in (0,1)$ die Wahrscheinlichkeit eines Ereignisses bezeichnet. k ist die kleinste ganze Zahl, die größer als x ist. Man erhält

$$x = -\frac{\log 2}{\log(1-p)} = \frac{\log 2}{p + p^2/2 + \cdots},$$

wobei log den natürlichen Logarithmus bezeichne. Ist nun p^2 vernachlässigbar klein, wächst der kritische Wert in demselben Maße wie p klein wird, so wie de Méré es auch glaubte. Die Verwunderung bei dem gewählten Zahlenbeispiel in Beispiel (1.5)(4) ist nun klar: für $p = 1/6$ ist $p^2/2$ nicht klein genug, um vernachlässigt werden zu können. Je größer p, desto ungenauer wird die „Proportionalitätsregel". Sicher scheint aber auch, daß man spätestens am Ende einer Stochastikvorlesung das Beispiel rechnen kann, die richtigen Resultate herausbekommt und sich darüber nicht so sehr verwundern wird, da man die oben genannte Regel erst gar nicht aufstellen würde.

Der Begriff der Wahrscheinlichkeit

Mit der raschen Einführung der Axiomatik nach Kolmogoroff sind wir in der Vorlesung einer Frage ausgewichen, die die Geister seit Jahrhunderten beschäftigt. Was bedeutet Wahrscheinlichkeit genau? Kann man den Begriff der Wahrscheinlichkeit gut beschreiben und eingrenzen? Es sollen einige Ausführungen nahmhafter Wissenschaftler aus den vergangenen Jahrhunderten einen Eindruck vermitteln, welcher Art Gedanken und Einordnungsversuchen man hier begegnet. *Carl Friedrich von Weizsäcker (1912)* führte in einem Aufsatz *Probability and Quantum Mechanics* 1973 zum Thema Wahrscheinlichkeit und Erfahrung aus:

Die Wahrscheinlichkeitstheorie hatte ihren Ursprung in einer empirischen Frage: dem Würfelspielproblem des Chevalier de Méré. Ebenso findet auch der heutige Physiker keine Schwierigkeit darin, empirisch eine theoretisch vorhergesagte Wahrscheinlichkeit zu überprüfen, indem er die relative Häufigkeit des Eintretens eines gewissen Ereignisses mißt. Andererseits ist die erkenntnistheoretische Diskussion über den Sinn der Anwendung des sogenannten mathematischen Wahrscheinlichkeitsbegriffs auf die empirische Wirklichkeit keineswegs zu Ende. Noch immer tobt die Schlacht zwischen „objektivistischen", „subjektivistischen" und noch anderen Deutungen des Wahrscheinlichkeitsbegriffs. Der Wahrscheinlichkeitsbegriff ist eines der auffallendsten Beispiele für das erkenntnistheoretische Paradoxon, daß wir unsere Grundbegriffe erfolgreich anwenden können, ohne sie wirklich zu verstehen.

Dies mag am Ende einer einführenden Vorlesung ernüchternd klingen. Doch verfolgen wir die Ausführung von Weizsäckers einfach weiter. Wenig später schreibt er über einen speziellen Weg zur Einführung des Wahrscheinlichkeitsbegriffs:

Zur Definition einer Wahrscheinlichkeit brauchen wir eine experimentelle Situation, in der verschiedene „Ereignisse" $E_1, E_2, \ldots$ die verschiedenen möglichen Ergebnisse eines und desselben Experiments sind. Wir müssen ferner sinnvoll sagen können, eine gleichartige experimentelle Situation ... liege in verschiedenen Fällen vor ..., und, gegeben diese Situation, werde in jedem Falle ein gleichartiges Experiment ausgeführt. Das Experiment sei in N Fällen ausgeführt worden, und das Ereignis E_k möge dabei n_k-mal eingetreten sein. In dieser Versuchsreihe wollen wir den Bruch

$$f_k = \frac{n_k}{N}$$

die relative Häufigkeit nennen, mit der E_k in der Serie vorgekommen ist. Nun denken wir an eine zukünftige Serie von Ausführungen desselben Versuchs. Nehmen wir an, unsere (theoretische und empirische) Kenntniss befähige uns, eine Wahrscheinlichkeit p_k für das Ereignis E_k in dem Versuch anzugeben. Dann wollen wir als den Sinn dieser Zahl p_k annehmen, sie sei eine Vorhersage der relativen Häufigkeit f_k für die zukünftige Versuchsserie (Diese Formulierung hat M. Drieschner 1970 vorgeschlagen). Man wird diese Vorhersage p_k empirisch überprüfen, indem man sie mit den Werten von f_k vergleichen wird, die sich in dieser und weiteren Serien des betrachteten Versuchs ergeben werden.

Wir haben in der Vorlesung seit der Formulierung des schwachen Gesetzes der großen Zahlen das Gefühl, daß diese Einführung korrekt ist. Weizsäcker schreibt weiter:

Dies ist die vereinfachende Denkweise des normalen Experimentators. Ich halte sie im wesentlichen für korrekt; sie muß nur gegen die Einwände der Erkenntnistheoretiker verteidigt werden. Natürlich hoffen wir, sie, indem wir sie verteidigen, besser zu verstehen. Ein einfaches Beispiel möge dazu dienen, den Haupteinwand zu formulieren. Unser Versuch bestehe im einmaligen Werfen eines Würfels. Es gibt 6 mögliche Ereignisse. Wählen wir das Ereignis, daß eine fünf erscheint, als dasjenige Ereignis, für das wir uns speziell interessieren. Seine Wahrscheinlichkeit p_5 wird den Wert 1/6 haben, wenn der Würfel „gut" ist. Nun wollen wir den Würfel N-mal werfen. Selbst wenn N durch 6 teilbar ist, wird der Bruch f_5 nur in seltenen Fällen genau gleich 1/6 sein; und, was noch wichtiger ist, die Wahrscheinlichkeitstheorie erwartet gar nicht, daß f_5 gleich 1/6 sein soll. Die Theorie sagt eine Verteilung des gemessenen Werts von f_5 um die theoretische Wahrscheinlichkeit p_5 herum voraus, wenn mehrere Serien von Würfen gemacht werden. Die Wahrscheinlichkeit ist nur der Erwartungswert der relativen Häufigkeit. Aber der Begriff Erwartungswert wird üblicherweise so definiert, daß dabei der Begriff Wahrscheinlichkeit schon benützt wird. Also sieht es so aus, als könne man die Wahrscheinlichkeit selbst grundsätzlich nicht durch Bezugnahme auf meßbare relative Häufigkeiten definieren, da diese Definition bei strenger Formulierung den Begriff der Wahrscheinlichkeit selbst schon benützen müßte; es entstünde - so scheint es - eine zirkelhafte Definition.

Der Ursprung der Schwierigkeit liegt nicht in dem speziellen Begriff der Wahrscheinlichkeit, sondern allgemein im Gedanken der empirischen Überprüfung irgendeiner theoretischen Vorhersage. Betrachten wir das Beispiel der Messung einer Ortskoordinate x eines Planeten zu einem bestimmten Zeitpunkt. Für sie sei von der Theorie der Wert ξ vorhergesagt. Eine einzelne Messung wird einen Wert ξ_1 ergeben, der von ξ verschieden ist. Die einzelne Messung wird vermutlich nicht genügen, uns zu überzeugen, ob man dieses Meßresultat als Bestätigung oder Widerlegung der theoretischen Vorhersage ansehen soll. Also werden wir die Messung N-mal wiederholen und die Fehlertheorie anwenden. Sei $\bar{\xi}$ der Mittelwert der gemessenen Werte. Vergleichen wir nun die Distanz $|\xi - \bar{\xi}|$ mit der mittleren Streuung der gemessenen Werte, so können wir formal eine Wahrscheinlichkeit dafür ausrechnen, daß der vorhergesagte Wert von ξ sich von dem wirklichen Wert ξ_r („ξ real") um die Größe $d = |\xi - \bar{\xi}|$ unterscheidet. Diese Wahrscheinlichkeit gibt selbst eine Vorhersage der relativen Häufigkeit, mit welcher die gemessene Distanz $|\xi - \bar{\xi}|$ den Wert d annehmen wird, wenn wir die Versuchsreihe oft wiederholen. Diese Struktur der empirischen Überprüfung einer theoretischen Vorhersage ist etwas kompliziert, aber wohlbekannt. Wir können sie in die abgekürzte Behauptung zusammenpressen: „Die empirische

Bestätigung oder Widerlegung einer theoretischen Vorhersage ist nie mit Gewißheit möglich, sondern nur mit einem höheren oder geringeren Grad von Wahrscheinlichkeit. Dies ist ein Grundzug aller Erfahrungen. Und weiter sagt er:

...folgt, daß unsere „abgekürzte Behauptung" auch für den Wahrscheinlichkeitsbegriff selbst gilt: Die empirische Überprüfung einer theoretisch gewonnenen Wahrscheinlichkeit ist nur mit einem gewissen Grad von Wahrscheinlichkeit möglich. Das Auftreten des probabilistischen Begriffs des Erwartungswerts in der Definition von Wahrscheinlichkeit ist daher kein Paradoxon, sondern eine notwendige Konsequenz aus der empirischen Bedeutung des Wahrscheinlichkeitsbegriffs; oder es ist ein „Paradoxon", das dem Begriff der Erfahrung selbst anhaftet.

Ich empfehle jedem die Lektüre der Arbeit von von Weizsäcker, erschienen zum Beispiel auch in seinem Buch *Aufbau der Physik* (dtv 10899).

Ein altbekanntes „Paradoxon" wird in einem weiteren, unveröffentlichten, Aufsatz von Carl Friedrich von Weizsäcker (1971) diskutiert: *Der Lehrer sagt den Schülern: „In der kommenden Woche werde ich eine Klassenarbeit schreiben lassen, aber ihr werdet nicht vorher wissen, an welchem Tag." Präzisierungsfrage: „ Werden wir es auch am Morgen des betreffenden Tages nicht wissen?" Antwort: „Auch am Morgen nicht." Das Paradox besteht darin, daß diese Aussage 1. einen Widerspruch impliziert, 2. empirisch leicht bestätigt werden kann.*

1. Der Widerspruch: Am Samstag kann er die Arbeit nicht schreiben lassen. Denn wenn auch der Freitag vorbeigegangen ist, ohne daß sie geschrieben wurde, so wissen die Schüler am Morgen des Samstag, daß sie heute geschrieben wird. Also muß sie an einem der fünf Tage Montag bis Freitag geschrieben werden. Am Freitag kann sie folglich auch nicht geschrieben werden, mit demselben Argument wie soeben. Also an einem der vier Tage Montag bis Donnerstag, also auch nicht am Donnerstag usf. Also gar nicht. Zur Analyse des Widerspruchs zerlege man die Behauptung des Lehrers in ihre zwei Bestandteile: A. In der kommenden Woche werde ich eine Arbeit schreiben lassen. B. Am Morgen keines Tages, ehe sie geschrieben ist, werdet ihr wissen, ob sie heute geschrieben wird. Der Widerspruch wäre direkt, wenn die Woche nur einen Arbeitstag hätte. Dann reduziert sich A und B auf : A'. Am Tag X werde ich eine Arbeit schreiben lassen. B'. Am Morgen des Tages X werdet ihr nicht wissen, ob ich heute eine Arbeit schreiben lasse. A und B, also auch A' und B' sind vom Lehrer gemeint als Prognose, die die Schüler von nun an als wahr glauben sollen. So interpretiert, implizieren A' und B': A''. Am Morgen des Tages X werdet ihr wissen, daß die Arbeit geschrieben wird. B''. Am Morgen des Tages X werdet ihr nicht wissen, ob die Arbeit geschrieben wird.

Hat die Woche mehrere Arbeitstage, so treffen A'' und B'' auf den letzten Tag zu, falls vorher nicht geschrieben wurde. Die obige Überlegung ist eine vollständige Induktion nach dem Schema: Wenn A'' und B'' auf den n-ten Tag zutreffen, so auch auf den $(n-1)$-ten.

2. Die empirische Bestätigung: Wenn der Lehrer z.B. am Mittwoch schreiben läßt, so waren A und B richtig: Er hat in dieser Woche schreiben lassen, und die Schüler konnten nicht vorher wissen, daß es gerade am Mittwoch sein würde.

Zur Auflösung dieses Paradoxons zitieren wir erneut Carl Friedrich von Weizsäcker: *Keine Prognose ist an sich wahr oder falsch. Sie hat nur eine gewisse Wahrscheinlichkeit, die man praktisch häufig mit Sicherheit gleichsetzen kann. Das Paradox entsteht, wenn man diese praktische Gleichsetzung prinzipiell versteht. Mit subjektiven*

Wahrscheinlichkeiten und Bayesscher Theorie läßt sich das Paradox beispielsweise wie folgt auflösen: Wir akzeptieren A als gewiß. Statt B werde im Laplaceschen Sinn behauptet: C: Alle noch verbleibenden Tage haben dieselbe Wahrscheinlichkeit p, der Tag der Arbeit zu sein. Am Montagmorgen ist $p = 1/6$. *Wird am Montag geschrieben, so ist A empirisch bestätigt. Auch B war empirisch wahr, denn B besagt nur* $p \neq 1$. *Wurde am Montag nicht geschrieben, so ändert sich für die restlichen Tage der Wert von p; es wird* $p = 1/5$. *Und so fort. Wurde bis und mit Freitag nicht geschrieben, so wird nunmehr für den Samstag* $p = 1$. *In diesem Falle wird B falsch. Aus A und C würde man also folgern, daß am Montagmorgen B die Wahrscheinlichkeit 5/6 hat. Wurde bis Freitag nicht geschrieben, so bekommt B die Wahrscheinlichkeit Null. Das Paradoxon reduziert sich nun auf die Behauptung: Die Prognose „A und B" kann höchstens die Wahrscheinlichkeit 5/6 haben.*

Man kann auch A nur die Wahrscheinlichkeit $1 - q$ *geben. Dann ist am Montagmorgen* $p = (1 - q)/6$, *am Samstagmorgen* $p = 1 - q$. *In diesem Fall erhält B nie die Wahrscheinlichkeit Null.*

Es mag durchaus sein, daß diese Art Begriffsdiskussionen verwirren. Die Zitate aus den Aufsätzen der Jahre 1971 und 1973 sollen verdeutlichen, daß man bis in jüngster Zeit den durchaus schwierig zu erfassenen Begriff der Wahrscheinlichkeit diskutiert. Wir wollen nun noch ein paar Äußerungen aus vergangenen Jahrhunderten betrachten.

Leibniz reflektierte über verschiedene Aspekte des Wahrscheinlichkeitsbegriffs. Eine Zusammenfassung seiner Gedanken über das Wahrscheinliche findet man in dem 1704 vollendeten Werk *Nouveaux essais sur l'entendement humain* (Neue Abhandlungen über den menschlichen Verstand). Darin heißt es in Kapitel 2 („von den Graden der Erkenntnis"):

Die Meinung, die auf das Wahrscheinliche gegründet ist, verdient vielleicht auch den Namen der Erkenntnis. Andernfalls würden fast alle historischen und viele andere Erkenntnisse hinfällig. Ohne aber über den Namen zu streiten, behaupte ich, daß die Untersuchung des Wahrscheinlichkeitsgrades sehr wichtig ist und uns noch fehlt, was ein großer Mangel unserer Logik ist. Denn wenn man eine Frage nicht absolut entscheiden kann, so könnte man doch immer den Wahrscheinlichkeitsgrad aus den Beobachtungsdaten bestimmen und folglich auf vernünftige Weise urteilen, welche Seite den größten Anschein für sich hat.

Leibniz nahm direkten Einfluß auf die Wahrscheinlichkeitsrechnung, unter anderem durch einen Briefwechsel mit Jakob Bernoulli in den Jahren 1703 bis 1705. Er erhob Einwände gegen Bernoullis Überzeugung, daß man mit endlich vielen, aber großen Anzahlen von Beobachtungen unbekannte Wahrscheinlichkeiten beliebig gut schätzen kann. Die Einwände von Leibniz fanden eine gewisse Reaktion in der *Ars conjectandi* von Bernoulli, auf die wir später noch eingehen werden. In einem Brief vom 3.12.1703 schrieb Leibniz an Bernoulli:

Die Schätzung von Wahrscheinlichkeiten ist für die Praxis von außerordentlicher Bedeutung, obwohl bei juristischen und politischen Fällen häufig eher eine lückenlose Aufzählung aller Umstände als eine genaue Rechnung erforderlich ist. Ich erinnere mich, daß ich über diesen von Dir behandelten Gegenstand zunächst nicht von Deinem Bruder (Johann I), *sondern anderswoher erfahren habe. Du fragst an, wenn man Wahrscheinlichkeiten empirisch schätzt aufgrund von Versuchen für das erfolgreiche Eintreten, ob man auf diese Weise schließlich eine vollkommen richtige Schätzung* *erhalten kann. Du schreibst auch, daß Du die Lösung gefunden hast. Das Hauptproblem scheint mir darin zu bestehen, daß zufällige Ereignisse bzw. das, was von unendlich vielen Umständen abhängt, nicht durch endlich viele Versuche bestimmt werden kann. Zwar hat die Natur ihre Gewohnheiten, die aus der Wiederkehr der Ursachen erwachsen, aber nur im Regelfall. Wer sagt deshalb, ob nicht der nächste Versuch gerade wegen der Veränderlichkeit der Dinge beträchtlich von der Regel aller vorhergehenden abweicht? Es treten immer wieder neue Krankheiten bei den Menschen auf; wenn man also auch beliebig viele Untersuchungen über die Sterblichkeit gemacht hat, hat man die Grenzen für die natürlichen Dinge keineswegs so festgelegt, daß sich in Zukunft nichts ändern könnte. ...*

In der *Ars conjectandi* von Jakob Bernoulli finden wir in Kapitel 1:

Die Sicherheit des Eintretens irgendeiner Sache hat einmal einen objektiven, für sich bestehenden Aspekt und bedeutet nichts anderes als die unverbrüchliche Tatsache der gegenwärtigen oder zukünftigen Existenz dieser Sache. Sie hat zum anderen einen subjektiven, auf uns bezogenen Aspekt und besteht nach Maßgabe unserer Kenntnis über diese Tatsache. In Kapitel II findet man:

Wir sprechen davon, das, was sicher und unzweifelhaft ist, zu wissen oder zu verstehen, alles übrige nur zu vermuten oder zu meinen. Irgendeine Sache zu vermuten bedeutet, ihre Wahrscheinlichkeit zu messen. Deshalb definiere ich als die Technik des Vermutens oder als Stochastik die Technik der möglichst genauen Bestimmung von Wahrscheinlichkeiten mit dem Ziel, uns bei unseren Bestimmungen und Handlungen immer für das entscheiden zu können, was uns besser, befriedigender, sicherer oder ratsamer erscheinen wird; darin allein besteht die ganze Weisheit der Philosophen und die ganze Klugheit der Politiker.

Soweit der kleine Eindruck zu Diskussionen um den Begriff der Wahrscheinlichkeit.

Jakob Bernoulli und das schwache Gesetz der großen Zahlen

Von Jakob Bernoulli an kann man in der Entwicklung der Wahrscheinlichkeitsrechnung eine Tendenz feststellen, bewußt durch Verallgemeinerung von Problemen zu Vorstufen einer Theorie zu gelangen. Dies wird in der *Ars conjectandi* besonders deutlich. Dort diskutiert Bernoulli eine Aufgabe aus Huygens Werk, deren Verallgemeinerung direkt zu der von Bernoulli gefundenen Binomialverteilung führt. Für die erste Fassung des Gesetzes der großen Zahlen war bei Bernoulli das folgende Problem ausschlaggebend: Kann man Wahrscheinlichkeiten aufgrund von wiederholten Beobachtungen zumindest näherungsweise bestimmen. Die Reaktionen von Leibniz hatten wir schon betrachtet. Bernoulli erkannte, daß es wesentlich war, nachweisen zu können, daß die Zuverlässigkeit der relativen Häufigkeit als Schätzwert für die gesuchte Wahrscheinlichkeit des vorgegebenen Ereignisses mit wachsender Anzahl der Beobachtungen ansteigt. Der Nachweis erfolgte in einem „Hauptsatz", in der Ars conjectandi auf Seite 236 und 239.

Hauptsatz: Nun folgt endlich der eigentliche Satz, dessentwegen all dies (Bernoulli verweist hier auf fünf vorausgehende Hilfssätze) *ausgeführt wurde, dessen Beweis nun aber erbracht wird allein durch die Anwendung der vorausgeschickten Hilfssätze auf das gegenwärtige Vorhaben. Um aber jede lästige Umschreibung zu vermeiden, bezeichne ich die Fälle, in welchen ein bestimmtes Ereignis eintreten kann, als fruchtbar, und jene Fälle, in welchen dasselbe Ereignis nicht eintreten kann, als unfruchtbar. Ebenso bezeichne ich die Versuche als fruchtbar, in welchen ersichtlich einer der fruchtbaren Fälle eintritt, und als unfruchtbar jene, in welchen man das Eintreten* *eines der fruchtbaren Fälle beobachtet. Verhalte sich also die Anzahl der fruchtbaren zur Anzahl der unfruchtbaren Fälle genau oder angenähert wie* $\frac{r}{s}$ *oder vielmehr zur Anzahl aller wie* $\frac{r}{r+s}$ *oder* $\frac{r}{t}$ (hierbei ist t als $r+s$ bereits eingeführt worden), *welches Verhältnis die Grenzen* $\frac{r+1}{t}$ *und* $\frac{r-1}{t}$ *einschließen. Zu zeigen ist: Es können so viele Versuche angestellt werden, daß es sich als beliebig vorgegeben, z.B. c-mal wahrscheinlicher herausstellt, daß die Anzahl der fruchtbaren Beobachtungen innerhalb diese Grenzen als außerhalb fallen wird, d.h., daß das Verhältnis der Anzahl der fruchtbaren zur Anzahl aller Beobachtungen nicht größer als* $\frac{r+1}{t}$ *und nicht kleiner als* $\frac{r-1}{t}$ *sein wird …*

Daraus scheint sich diese bemerkenswerte Folgerung zu ergeben, daß, wenn die Beobachtungen aller Ereignisse in alle Ewigkeit fortgesetzt würden, wobei schließlich die Wahrscheinlichkeit in vollkommene Sicherheit übergehen würde, alles in der Welt als sich in bestimmten Verhältnissen und nach einer festen Gesetzmäßigkeit des Wandels ereignend erkannt würde, so daß wir sogar gehalten wären, selbst bei den zufälligsten Dingen gleichsam eine geweisse Notwendigkeit und sozusagen eine Bestimmung anzuerkennen.

Tatsächlich hat Bernoulli mit den obigen Gedanken bewiesen, daß die relative Häufigkeit des Eintretens eines (beliebig oft unabhängig von seiner Vorgeschichte reproduzierbaren) Ereignisses nach Wahrscheinlichkeit mit wachsender Beobachtungszahl immer weniger von der vorgegebenen Wahrscheinlichkeit dieses Ereignisses abweicht. Zunächst zeigte er 1689 in den *Meditationes* aufgrund einer sehr eindrucksvollen heuristischen Überlegung, daß für den einfachen Fall einer Wahrscheinlichkeit von $1/2$ für das Ereignis das Verhältnis der Wahrscheinlichkeit, daß die relative Häufigkeit in dem Intervall $[1/3, 2/3]$ liegt, zu der Wahrscheinlichkeit, daß die relative Häufigkeit außerhalb dieses Intervalls liegt, abhängig von der Anzahl n der Beobachtungen beliebig groß gemacht werden kann. Kurz darauf stellt der in den *Meditationes* für beliebige Wahrscheinlichkeiten fest:

Es ist möglich, so viele Beobachtungen anzustellen, daß es mit einer vorgegebenen beliebigen Wahrscheinlichkeit wahrscheinlicher ist, daß (das Verhältnis der) Anzahlen der von beiden Seiten gewonnenen Spiele innerhalb vorgegebener, beliebig enger Grenzen zu liegen kommt als außerhalb.

Bernoulli führte einen strengen Beweis. Er setzt voraus, daß die Wahrscheinlichkeit dafür, daß die relative Häufigkeit in das Intervall $[p-\varepsilon, p+\varepsilon]$ fällt, wenn p die bekannte Wahrscheinlichkeit des Ereignisses ist, die Summe all der Glieder in der Entwicklung $(p+q)^n = \sum_{\nu=0}^{n} \binom{n}{\nu} p^\nu q^{n-\nu}$ entspricht, für die gilt, daß $\frac{\nu}{n}$ in $[p-\varepsilon, p+\varepsilon]$ liegt. Er symbolisierte diese Voraussetzung in anderer Form und zeigte:

Also folgt, da das Maximum innerhalb der Grenzen das Maximum außerhalb unendlich übertrifft, wie auch das zweit (-größte Glied innerhalb) *das zweit,* (größte außerhalb) *das dritte das dritte, und da auch das Verhältnis der übrigen* (davon) *nicht abweicht, daß alle* (Glieder) *innerhalb der Grenzen eine gleiche Anzahl der größten* (Glieder) *außerhalb der Grenzen ebenfalls unendlichfach übertreffen.*

Und so weiter. Diese Originalquelle ist aus heutiger Sicht teils schwer verständlich. Bevor wir zur weiteren Geschichte des schwachen Gesetzes der großen Zahlen kommen, wollen wir an dieser Stelle ein paar Aspekte des Lebens von Jakob Bernoulli schildern.

Jakob Bernoulli und sein Bruder *Johann Bernoulli (1667–1748)* haben vor allem durch ihren Ausbau des von Leibniz entdeckten Infinitesimalkalküls historische Bedeutung erlangt. Die Kenntnisse der Leibnizschen Analysis des Unendlichen, welche sich die beiden eroberten, gab ihnen schon zu Lebzeiten einen mit einem Nimbus umgebenen Namen, den noch zwei spätere Generationen in Ehre halten sollten. Die berühmte Mathematikerdynastie der Bernoulli, welche mit der Musikerdynastie der Bach verglichen wird, hat ihren Ursprung in dem „Familiengeheimnis" des Leibnizschen Kalküls. Die Bernoullis verschafften der Stadt Basel einen unvergänglichen Ruhm. In der ersten Hälfte des achzehnten Jahrhunderts spielte der Kreis der Basler Mathematiker, zu denen auch *Leonhard Euler (1707–1783)* gehörte, eine vergleichbare Rolle wie zum Beispiel der Florentiner Kreis der Mathematiker um *Galileo Galilei (1564–1642)* oder der Pariser Kreis der kartesianischen Akademiker im siebzehnten Jahrhundert. Die Entwicklung der Analysis durch diese Mathematiker erscheint als ein Vorspiel zu den großen Leistungen der Bernoullis und Eulers. Der Genius Euler ist ein Schüler von Johann gewesen. Der Keim vieler Arbeiten Eulers liegt in den Bernoullischen Arbeiten. Als Leibniz 1684 in der von ihm mitbegründeten ersten deutschen wissenschaftlichen Zeitschrift, den Leipziger *Acta Eruditorum* unter dem Titel *Nova methodus pro maximis et minimis* seinen Differentialkalkül veröffentlichte, galt er als schwerverständlich. Jakob Bernoulli gelang es durch jahrelange Studien, den Sinn zu erfassen. Auf den Wunsch des Vaters hatte Jakob zunächst Theologie studiert, sich im geheimen aber dem Studium der Mathematik gewidmet. Dies war in Basel zunächst nur das Erlernen der Elementarmathematik. Bei einer Reise nach Holland und England lernte er die moderne kartesische Geometrie und die Infinitesimalrechnung des siebzehnten Jahrhunderts kennen. Nach der Rückkehr faßte er den Entschuß, sich ausschließlich der Mathematik zu widmen und lehnte eine ihm angebotene Predigerstelle in Straßburg ab. Er hielt sogleich eine Vorlesung über Experimentalphysik und studierte die moderne Mathematik, welche seit Descartes eine besondere Entwicklung genommen hatte. Neben einigen physikalischen Arbeiten schrieb Bernoulli bis 1686 einige Studien zur Logik. Während bei Leibniz die mathematischen Probleme nur Spezialfälle einer „Universallogik" darstellten, sind in seinen Schriften kleine mathematische Einzelprobleme versteckt. Vor der „Entdeckung" des Kalküls formuliert er Abhandlungen zu unendlichen Reihen. Die erste Abhandlung enthält zum Beispiel den Divergenzbeweis für die harmonische Reihe, den übrigens, wie Jakob mitteilt, sein Bruder Johann zuerst gefunden hat. Ferner steht hier die berühmte Bernoullische Ungleichung

$$(1+a)^n \geq 1 + n\,a \qquad n = 1, 2, 3, \ldots; a \geq 0,$$

welche Jakob beim Vergleich von geometrischen und arithmetischen Reihen findet. Diese Abhandlungen sind durchaus fehlerhaft, da sie Rechnungen mit und an divergenten Reihen durchführen. Doch hat der mathematische Instinkt Jakobs doch das Richtige gefunden. 1687 trat Jakob eine mathematische Proffesur in Basel an. Er wandte sich an Leibniz, um nähere Auskunft über dessen schwerverständliche Abhandlung von 1684 zu erhalten. Leibniz war zu jener Zeit verreist und antwortete erst am 24. September 1690. Daher war in der Zwischenzeit Jakob genötigt, selbstständig den Leibniz Kalkül zu ergründen. Er löste mit Hilfe seiner Erkenntnisse schnell das Problem der Isochrone (wir lassen dies weg). So konnte Leibniz in seinem Schreiben ihm bereits bescheinigen, daß er keiner Hilfe von außen mehr bedürfe, da er den Sinn der neuen Methode vollkommen erfaßt habe. Jakob hatte zu dieser Zeit seinen Bruder Johann in die eroberten Geheimnisse der neuen Methode eingeweiht. Die beiden gelten durch ihre detaillierte Ausarbeitung des Kalküls als Wegbereiter. Leibniz, selbst durch den Dreißigjährigen Krieg an weiteren Ausarbeitungen gehindert, kannte das Verdienst der Bernoullis um den neuen Kalkül neidlos an. Die Brüder reagierten keineswegs ohne Eifersucht auf die Leistungen des jeweils anderen. Sie waren schnell in der Öffentlichkeit für ihre Streitereien bekannt. So stellte Jakob am Schluß einer Abhandlung die Frage nach der Gestalt einer sogenannten Kettenlinie, die Galileo Galilei für eine Parabel hielt. Johann löste diese Aufgabe. Dieser betont später, daß das Problem für Jakob zu schwer gewesen sei. Die gegenseitige Eifersucht wurde auch spürbar, als Johann in Paris als ein Repräsentant des neuen Leibnizschen Kalküls empfangen wurde. In Basel machte es großen Eindruck, daß sich der illustre *Guillaume Francois Antoine Marquis de L'Hôpital (1661–2.2.1704)*, damals wohl der begabteste Mathematiker Frankreichs, von Johann in die Infinitesimalrechnung einführen ließ. Das erste Lehrbuch der Differentialrechnung, die *Analyse des infiniment petits* (1696) mit welchem der Marquis seinen Namen verewigte, ist aus diesen Vorlesungen Johanns und einem Briefwechsel entstanden.

Die *Ars conjectandi* erschien erst acht Jahre nach Jakobs Tod als separates Buch mit einem Vorwort seines Neffen *Nikolaus I Bernoulli (1687–1759)* versehen, 1713 in Basel. Man kann aus den Manuskripten Jakobs ersehen, daß die wesentlichen Lehrsätze dieses Buches aber schon vor 1690 gefunden wurden. Diese Zeit gilt ohnehin als die logische und kombinatorische Phase seiner mathematischen Entwicklung. Jakob beginnt sein Werk mit den klassischen Problemen des Glücksspiels. Der erste der vier Teile des Buches bringt die Abhandlungen von Huygens *De ratiociniis in ludo alea* mit Anmerkungen Jakobs. Die an speziellen Zahlenbeispielen entwickelten Formeln von Huygens treten hier bereits in allgemeiner Form auf. Im zweiten Teil werden allgemeine Formeln der Kombinatorik abgeleitet, um sie im dritten Teil auf allgemeine Probleme der kombinatorischen Wahrscheinlichkeitsrechnung anzuwenden. Der Übergang zur Statistik wird im vierten Teil vollzogen. Der frühzeitige Tod Jakobs scheint verhindert zu haben, daß Schlußfolgerungen für eine Lehre der Statistik erfolgten. Das Werk gipfelt und schließt mit dem oben diskutierten Hauptsatz, dem schwachen Gesetz der großen Zahlen.

Vor der *Ars conjectandi* erschien noch ein weiteres Werk. *Bernard le Bouyer Fontenelle (1657–1757)* hielt nach Jakobs Tod eine Laudatio, die ein Jahr später veröffentlicht wurde. Man findet darin eine Zusammenfassung eines Maunskripts von Jakob Bernoulli. In dem Glauben, dieses Manuskript würde nicht veröffentlicht werden, beschloß *Pierre Rémond de Montmort (1678–1719)*, ein Buch nach den Angaben der Zusammenfassung herauszugeben. Es erschien 1708 unter dem Titel *Essai d'Analyse sur les jeux de hasard*. Zahlreiche Würfel- und Kartenspiele sowie die fünf Probleme von Huygens werden untersucht. In der zweiten Auflage aus dem Jahr 1713 wird ein großer Teil der Korrespondenz, die Jakob über verschiedene Probleme mit Nikolaus Bernoulli führte, aufgeführt.

In einem der Briefe aus dem Jahr 1927 stellte Nikolaus das *Petersburger Problem*, welches von *Daniel Bernoulli (1700–1784)* in vielen Abhandlungen untersucht wurde. Eine der Abhandlungen wurde 1730 an der Akademie von St.Petersburg veröffentlicht. Im Jahre 1937 zeigte *William Feller (1906–1970)*, als er den Begriff des gerechten Spieles im Zusammenhang mit dem Gesetz der großen Zahlen untersuchte, daß bei unendlich großem Erwartungswert ein verallgemeinertes Gesetz der großen Zahlen zur Anwendung kommt. Wir haben dies im Detail in Kapitel 3 diskutiert.

Schauen wir noch etwas auf die weitere Entwicklung des schwachen Gesetzes der großen Zahlen. Abraham de Moivre griff im Jahre 1733 den Hauptsatz von Jakob Bernoulli auf. Er zeigte: wenn in einem Intervall $[p-\varepsilon, p+\varepsilon]$ für ε ein Wert proportional zu $n^{-1/2}$ bei sehr großem n gewählt wird, dann fällt die Hauptmasse der Wahrscheinlichkeit in dieses Intervall. Er konnte diese einfache Version eines zentralen Grenzwertsatzes mit den Methoden der Newtonschen Analysis für den Fall $p = 1/2$ beweisen. Man findet diesen Beweis in der berühmten Abhandlung *The doctrine of chances*. Dieses 1738 entstandene Werk hat einen Vorläufer, einen siebenseitigen Sonderdruck in lateinischer Sprache, 1733 unter dem Titel *Approximatio ad summam terminorum binomii* $(a+b)^n$ *in seriem expansi* erschienen. De Moivre hatte ein schweres Leben. Er floh mit 18 Jahren vor den Hugenottenverfolgungen nach England. Es gelang ihm nie, seinen Traum von der Karriere als Professor zu verwirklichen. Er lebte in Armut und ernährte sich von Privatstunden.

Thomas Bayes (1702–1761) hinterfragte 1763 erneut die Brauchbarkeit der relativen Häufigkeit als Schätzwert für die unbekannte Wahrscheinlichkeit. Dabei setzte er in seiner Arbeit *An essay towards solving a problem in the doctrine of chances* in Gedanken voraus, daß die unbekannte Wahrscheinlichkeit a priori in dem Intervall $(0,1)$ gleichverteilt sein soll. Er bestimmte unter dieser Voraussetzung die Wahrscheinlichkeit dafür, daß bei vorgegebener relativer Häufigkeit des Eintretens eines Ereignisses die unbekannte Wahrscheinlichkeit dieses Ereignisses in einem vorgegebenen Intervall liegt. *Pierre-Simon Laplace (1749–1827)* befaßte sich 1820 mit der Frage, inwieweit bei sehr großen, aber endlichen Beobachtungszahlen die Normalverteilung die Binomialverteilung annähert, d.h., welches Korrekturglied erforderlich ist, wenn die Binomialverteilung durch die Normalverteilung ersetzt wird. In seinem Buch *Théorie analytique des probabilités*, genauer im Buch II, Kapitel III mit der Überschrift „Wahrscheinlichkeitsgesetze, die sich aus der unbeschränkten Vervielfältigung der Ereignisse ergeben", findet man diesen zentralen Grenzwertsatz für beliebiges p. Die Tendenz, durch Abschwächung der Voraussetzungen zu allgemeineren Aussagen zu kommen, setzt mit *Siméon Denis Poisson (1781–1840)* ein. Er bewies 1837 die Gültigkeit des schwachen Gesetzes für den Fall, daß die Wahrscheinlichkeiten eines Ereignisses von Versuch zu Versuch verschieden sind, wenn nur der Mittelwert dieser Wahrscheinlichkeiten stabil bleibt. Poisson ist der Namensgeber dieses Gesetzes. In seinem *Lehrbuch der Wahrscheinlichkeitsrechnung und deren wichtigste Anwendungen* findet man einleitend:

Die Erscheinungen jeglicher Art sind einem allgemeinen Gesetze unterworfen, welches man das Gesetz der großen Zahlen nennen kann. Es besteht darin, daß, wenn man sehr große Anzahlen von Erscheinungen derselben Art beobachtet, welche von constanten und von unregelmäßigen veränderlichen Ursachen abhängen, die aber nicht progressiv veränderlich sind, sondern bald in dem einen bald in dem anderen Sinne; man zwischen diesen Zahlen Verhältnisse findet, welche fast unveränderlich sind.

Wir betrachten nun die sogenannte *Bienaymé-Tschebyscheff-Ungleichung*. Wir waren in Kapitel 3 der Historie von Satz 3.29 nicht weiter nachgegangen. Tatsächlich hat *Irénée-Jules Bienaymé (1796–1878)* 1853 diese Ungleichung in der heute noch üblichen Form abgeleitet, um damit „das Verhalten des Wertes (der Fehlerwahrscheinlichkeit), den sie mit wachsender Anzahl von Beobachtungen annimmt, zu berechnen". *Pafnuty*

Lvovich Tschebyscheff (1821–1894) hatte eine äquivalente Form der Ungleichung 1867 in russischer und in französischer Sprache veröffentlicht. Seine Herleitung war elementar und war von ihm als Verallgemeinerung des Gesetzes der großen Zahlen in der Form von Poisson verstanden worden. Der französischen Version seiner Arbeit war der Wiederabdruck der vorher erwähnten Bienayméschen Arbeit von 1853 unmittelbar vorausgegangen. Dies veranlaßte ihn 1874, die Beweismethoden von Bienaymé zu analysieren. Er erkannte als Kernstück die als gegeben vorausgesetzten Momente verschiedener Ordnung einer Zufallsgröße. Es sei bemerkt, daß zu jener Zeit der Begriff der Zufallsgröße noch nicht eingeführt war. Tschebyscheff benutzte diese Methode der Momente für einen Beweisversuch des zentralen Grenzwertsatzes für Summen von unabhängigen Zufallsgrößen. In seiner Arbeit *Über zwei Wahrscheinlichkeiten betreffende Sätze* aus dem Jahr 1887 findet man:
...Dieser Satz kann so formuliert werden: Wenn die Erwartungswerte der Größen

$$u_1, u_2, u_3, \ldots$$

alle gleich Null sind, und wenn die Erwartungswerte aller ihrer Potenzen eine beliebige endliche Grenze nicht überschreiten, wird die Wahrscheinlichkeit, daß die Summe

$$u_1 + u_2 + \cdots + u_n$$

einer Anzahl n dieser Größen, geteilt durch die Quadratwurzel der doppelten Summe der Erwartungswerte ihrer Quadrate, zwischen zwei beliebigen Grenzen t und t' enthalten ist, für n gegen Unendlich gleich

$$\frac{1}{\sqrt{\pi}} \int_t^{t'} e^{-x^2}\, dx.$$

Tschebyscheffs Schüler *Andrei Andreyevich Markoff (1856–1922)* bewies 1898 einen zentralen Grenzwertsatz unter wesentlich schwächeren Voraussetzungen mit der „Methode der Momente", im Jahre 1908 auch für anhängige Zufallsgrößen. *Aleksandr Mikhailovich Lyapunov (1857–1918)* konnte mit einer Methode, die die Dirichlet-Sprungfunktion verwendet, den zentralen Grenzwertsatz unter den sehr schwachen Voraussetzungen der Existenz von Erwartungswert und Varianz sowie der Erfüllung der sogenannten Ljapunow-Bedingung beweisen. Eine weitere Verallgemeinerung gelang erst 1922, als *Jarl Waldemar Lindeberg (1876–1932)* die Güligkeit des zentralen Grenzwertsatzes für unabhängige Zufallsvariablen unter Abschwächung der Ljapunoff-Bedingung zur Lindeberg-Bedingung bewies. 1935 konnte Feller auch die Notwendigkeit der von Lindeberg angegebenen Bedingung für die Gültigkeit des zentralen Grenzwertsatzes nachweisen, womit diese Entwicklung zu einem Abschluß kam.

Anstöße aus der Physik
Die Anfänge der Anwendung der Wahrscheinlichkeitsrechnung in der Physik betreffen die kinetische Gastheorie und die sogenannte Statistische Mechanik. Um 1900 hatten die Anwendungen einen Umfang erreicht, der *David Hilbert (23.1.1862–14.2.1943)* dazu veranlaßte, die Wahrscheinlichkeitsrechnung als physikalische Disziplin zu bezeichnen. Erst die Überzeugung von der Gültigkeit des Energieerhaltungssatzes und die Erfindung der Dampfmaschine in der Mitte des 19. Jahrhunderts machten deutlich, daß eine mechanische Wärmetheorie eine richtige Theorie

sein müsse. Es kam zur Entwicklung der kinetischen Gastheorie. Gegen Ende des 19. Jahrhunderts kam es zu Vorbehalten gegen die Hypothese einer atomaren Struktur der Materie. Diese Stagnation konnte im wesentlichen erst durch Arbeiten von *Albert Einstein (1879–1955)* überwunden werden. Doch betrachten wir genauer die Begründung der Statistischen Mechanik durch *James Clerk Maxwell (1831–1879)* und *Ludwig Boltzmann (1844–1906)*. 1859 Maxwell verfeinerte das kinetische Gasmodell von *Rudolf Julius Emmanuel Clausius (2.1.1822–24.8.1888)*, indem er statt der von Clausius verwendeten mittleren, für alle Gasmoleküle gleichen Geschwindigkeiten eine Geschwindigkeitsverteilung ableitete. Dabei führten die speziellen Annahmen Maxwells für das gaskinetische Modell zu einer Funktionalgleichung, die zu der von Gauss bei der ersten Begründung der Methode der kleinsten Quadrate verwendeten äquivalent ist. Aus der Annahme der Rotationsinvarianz der Verteilung und der Unabhängigkeit der drei Geschwindigkeitskomponenten folgerte er, daß diese normalverteilt sein müssen, und daß die freie Energie eine χ^2-Verteilung mit drei Freiheitsgraden hat. 1866 gab er eine weitere Ableitung an und zeigte, daß die von ihm angegebene Verteilung die Eigenschaft hat, daß sie durch zufällige Zusammenstöße der Teilchen nicht verändert wird. Boltzmann verallgemeinerte 1868 diesen Ansatz. Er ließ mehratomige Moleküle zu, die zusätzlichen Kräften ausgesetzt sein dürfen. 1872 entstand seine berühmte Arbeit zur wahrscheinlichkeitstheoretischen Deutung der Entropie. In seinem Modell besteht ein Gas aus N Atomen. Ist $Ng(u,t)$ die Zahl der Atome im Volumenelement $[u, u+du]$ des 6-dimensionalen Phasenraums (drei Orts- und drei Geschwindigkeitskoordinaten) eines Atoms zur Zeit t, so genügt g der sogenannten Boltzmann-Gleichung. Es ist erst 1988 gelungen, für sie einen brauchbaren Existenzsatz zu beweisen. Man kann zeigen, daß

$$H(t) = \int g(u,t) \, \log g(u,t) \, du$$

nicht zunimmt. $H(t)$ nimmt ab, bis g eine Maxwell-Verteilung geworden ist. 1877 führte Boltzmann die Einteilung des Phasenraums in r gleich große Zellen ein. Wir wissen bereits, daß

$$W = \frac{N!}{n_1! n_2! \cdots n_r!}$$

die Anzahl der Möglichkeiten ist, die N Atome so auf die Zellen zu verteilen, daß n_i Atome in der i-ten Zelle sind. Vernachlässigen wir die Normierung, so ist W die Wahrscheinlichkeit der Besetzung $(n_1, \dots, n_r)$. Wenn τ das Volumen einer Zelle bezeichnet, findet man

$$H \approx \tau \sum n_i \log n_i .$$

Mit Hilfe der Stirling-Formel erhält man $H \approx -\tau \log W$. Wir haben diese Ergebnisse von Boltzmann im Kapitel 4 kennengelernt und es in Form eines Prinzips großer Abweichungen formuliert. Boltzmann zeigte weiter, daß unter der Nebenbedingung $\sum n_i = N$ und $\sum \varepsilon_i n_i = U$ (wobei ε_i das Energieniveau der i-ten Zelle und U die Gesamtenergie bezeichne), die wahrscheinlichste Verteilung von der Form

$$n_i = N\lambda \exp(\beta\varepsilon_i)$$

ist. Ist T die Temperatur und k die Boltzmann Konstante, so ist $\beta = \frac{1}{kT}$. Der Aussage, daß H nur abnehmen kann, entspricht die Aussage, daß die Entropie S

nur zunehmen kann. Boltzmanns Arbeiten riefen lebhafte Diskussionen hervor, da sie mathematische Fehler aufwiesen. So gab *Loschmidt* 1876 den Einwand, daß eine mechanische Deutung der Wärmelehre unmöglich sein müßte, da mechanische Systeme reversibel sind. *Ernst Zermelo (1871–1953)* gab den Hinweis, daß nach dem Wiederkehrsatz von *Jules Henri Poincaré (1854–1912)* jedes begrenzte System immer wieder beliebig nahe an seinen Ausgangspunkt zurückkehren müßte. Eine wichtige Rolle bei der Klärung dieser Fragen spielte ein Artikel von *Paul Ehrenfest (1880–1933)* aus dem Jahr 1911. In dieser Arbeit führt er das Modell ein, welches wir in Kapitel 9 als ein Beispiel für eine Markoff-Kette diskutiert haben. Die heutige Sicht der Dinge ist, daß die Wahrscheinlichkeit eines Zustands des Gesamtsystems mit kleiner Entropie um so viele Größenordnungen kleiner ist als die des Gleichgewichtszustands, daß der Erwartungswert der Rückkehrdauer unvorstellbar groß ist. Als *Max Karl Ernst Ludwig Planck (1858–1947)* seine Strahlungsformel im Jahr 1900 ableitete, stützte er sich auf die Boltzmannsche Wahrscheinlichkeitsdeutung der Entropie. Man kann also sagen, daß die Arbeiten Boltzmanns den Weg zur Quantentheorie ebneten. Eine zentrale Rolle spielte weiter die Ergodenhypothese. *Plancherel* und *Rosenthal* zeigten unabhängig voneinander, daß diese für die betrachteten Systeme im schärfsten Sinn („auf einer Energiefläche geht jede Bahn durch jeden Punkt") nicht gelten kann. Was in der Statistischen Mechanik statt der Ergodenhypothese benötigt wurde, war nur die Gleichheit der Raummittel und Zeitmittel auf Energieniveau. Das Wort *Ergode* verwendete Boltzmann bereits 1884. Stationäre Systeme in einem Gebiet, das durch Gleichungen begrenzt ist, nennt er Monoden, und Monoden, welche nur durch „die Gleichung der lebendigen Kraft" begrenzt sind, heißen Ergoden. Man würde heute sagen, daß die Energie konstant gehalten wird.
John von Neumann (1903–1957) konnte 1931 zeigen, daß der Grenzwert der Zeitmittelwerte existiert. Er zeigte, daß für eine sogenannte maßtreue Transformation τ und eine 2-fach integrierbare Abbildung f die Mittel

$$\frac{f + f \circ \tau + \cdots + f \circ \tau^{n-1}}{n}$$

im L_2-Sinn konvergieren. *George David Birkhoff (1884–1944)* bewies 1931 die fast sichere Konvergenz. Diese Arbeiten führten zur Entwicklung der Ergodentheorie mit zahlreichen Bezügen zu den verschiedensten mathematischen Gebieten. Die Verbindung zur Wahrscheinlichkeitstheorie stelle *Aleksandr Yakovlevich Khinchin (1894–1959)* 1934 durch Einführung des Begriffs des stationären stochastischen Prozesses her.
Albert Einsteins berühmte Arbeit aus dem Jahr 1905 (*annus mirabilis*) in den Annalen der Physik, in der er die *Brownsche Bewegung* untersuchte, führte zur Entwicklung der stochastischen Prozesse. 1827 hatte der britische Botaniker *Robert Brown* beobachtet, daß anorganische winzige Partikel, die in einer Flüssigkeit schwebten, unter dem Mikroskop eine unregelmäßige tanzende Bewegung zeigen. Einstein nahm an, daß die Teilchen durch die Flüssigkeit diffundieren und Gleichgewicht zwischen osmotischem Druck und der durch die Zähigkeit der Flüssigkeit verursachten Kraft herrschen müsse. Er leitete eine Formel für die Diffusionskonstante her. Unter der Annahme, daß die Partikel sich unabhängig voneinander bewegen und in disjunkten Zeitintervallen unabhängige Bewegungen vollführen, konnte er zeigen, daß die Konzentration zum Zeitpunkt t normalverteilt ist mit einer zu t proportionalen Var-

ianz. Mit Einsteins Arbeit wurde der positivistische Widerstand gegen den Atomismus einer statistischen Mechanik endgültig überwunden. In dieser Arbeit liegt die konkrete Betrachtung eines stochastischen Prozesses vor. In wichtigen Arbeiten von *Norbert Wiener (1894–1964)* und *Kolmogoroff* wurden Einsteins Grundlagen ausgeweitet.
Krengel diskutiert in seinem Artikel zur Wahrscheinlichkeitstheorie 2 Fragen grundsätzlicher Bedeutung im Zusammmenhang von Physik und Wahrscheinlichkeitstheorie. Sie lauten: Welche physikalische Bedeutung hat eine Wahrscheinlichkeitsaussage? Ist eine Wahrscheinlichkeit, die durch die Theorie geliefert wird, auch experimentell meßbar? Die anhaltende Diskussion zu dieser Frage kann kaum überblickt werden. Man ist etwa der Ansicht, daß Wahrscheinlichkeitsaussagen durch experimentell gewonnene relative Häufigkeiten bestätigt oder widerlegt werden. Dies entspricht einer statistischen Interpretation der Quantentheorie, bei der man konsequent Wahrscheinlichkeit nur als relative Häufigkeit interpretiert. *Max Born (1882–1970)* vertrat zuerst diesen Ansatz.
In Bezug auf die Bedeutung der Wahrscheinlichkeitstheorie in der Quantentheorie begnügen wir uns mit einer einzigen Ausführung aus dem Buch „Aufbau der Physik" von Carl Friedrich von Weizsäcker:
Der Versuch, mit den Wahrscheinlichkeitsaussagen der Quantentheorie zu beginnen, scheint in der Tat auf ein fundamentales Hindernis zu stoßen: es ist zweifelhaft, ob Kolmogoroffs Axiome überhaupt in der Quantentheorie gelten. Das Grundphänomen ist die „Interferenz der Wahrscheinlichkeiten"; die Grundgesetze beziehen sich nicht direkt auf die meßbaren Wahrscheinlichkeiten, sondern auf die „Wahrscheinlichkeitsamplituden". Im System von Kolmogoroff bedeutet dies, daß an die Stelle des ersten Axioms, das besagt, die möglichen Ereignisse bildeten einen Booleschen Verband, (es genügt die Kenntnis des Begriffs σ-Algebra, so wie er in Kapitel 7 eingeführt wurde) *ein anderes Axiom tritt, nach dem der Ereignisverband durch die Teilräume eines Hilbertraums gebildet wird.*

Die Axiomatik Kolmogoroffs

In Band 7 des Jahresberichts der Deutschen Mathematiker Vereinigung erschien 1898 ein fast 300 Seiten langer Übersichtsartikel von *Emanuel Czuber (1851–1925)* mit dem Titel „Die Entwicklung der Wahrscheinlichkeitstheorie und ihrer Anwendungen". Eine Fülle von Aufgaben zu Urnenmodellen, Münzwurfprobleme und viele trickreiche Ansätze zur Bestimmung von Wahrscheinlichkeiten sind in dieser Arbeit enthalten. Aber die Grundlagen dieser Disziplin waren nicht klar, und so stellte sich zu Beginn dieses Jahrhunderts die Frage, ob die Wahrscheinlichkeitsrechnung eine mathematische Disziplin werden kann, oder eher der Physik oder der Philosophie zugerechnet werden sollte.
David Hilbert stellte in seinem berühmten Pariser Vortrag vom 8. August 1900 auf dem 2. Internationalen Mathematikerkongreß als sechstes Problem die Aufgabe der „mathematischen Behandlung der Axiome der Physik":
Durch die Untersuchungen über die Grundlagen der Geometrie wird uns die Aufgabe nahegelegt, nach diesem Vorbilde diejenigen physikalischen Disziplinen axiomatisch zu behandeln, in denen schon heute die Mathematik eine hervorragende Rolle spielt; dies sind in erster Linie die Wahrscheinlichkeitsrechnung und die Mechanik. Was die Axiome der Wahrscheinlichkeitsrechnung angeht, so scheint es mir wünschenswert, daß mit der logischen Untersuchung derselben zugleich eine strenge und befriedigende

Entwicklung der Methode der mittleren Werte in der mathematischen Physik, speziell in der kinetischen Gastheorie in Hand gehe.

Hilbert verwies auf vier Vorlesungen von *Georg Bohlmann (1869–1928)*, die dieser im Rahmen eines Ferienkurses für Lehrer Ostern 1900 in Göttingen hielt. Bohlmann hatte in der zweiten Vorlesung fünf Axiome oder Hypothesen der Wahrscheinlichkeitsrechnung angekündigt, die er dann in einem Artikel über Lebensversicherung in der Encyclopädie der Mathematischen Wissenschaften, Band I, Teil 2, aufführte. Dort ist die Wahrscheinlichkeit eines Ereignisse E eine nichtnegative Zahl $p(E)$, für die gilt:

Ist E das sichere Ereignis, so ist $p(E) = 1$;

Sind E_1, E_2 zwei Ereignisse, die nur mit Wahrscheinlichkeit 0 gemeinsam eintreten, so ist die Wahrscheinlichkeit des Ereignisses, daß E_1 oder E_2 eintritt, $p(E_1) + p(E_2)$.

Weiter definierte Bohlmann die Unabhängigkeit zweier Ereignisse, wobei der Begriff des Ereignisses noch nicht richtig geklärt war. 1908 trug er auf dem Internationalen Mathematikkongreß in Rom über eine modifizierte Fassung seines Ansatzes vor. Dabei definierte er die Unabhängigkeit von mehreren Ereignissen so, wie wir es auch heute in den Lehrbüchern finden. Er zeigte den Unterschied zu paarweiser Unabhängigkeit auf.

Bohlmanns Ansatz forderte also nur die endliche Additivität. 1907 schrieb *Ugo Broggi (1880–1965)* bei Hilbert eine Dissertation mit dem Titel „Die Axiome der Wahrscheinlichkeitsrechnung". Ereignisse erscheinen bei ihm als Teilmengen einer gegebenen Menge, jedoch immer noch etwas unklar. Broggi stellt weiter dem Axiom der endlichen Additivität das Axiom der σ-Additivität entgegen und behauptet, daß diese Axiome äquivalent sind. Man schließt heute daraus, daß Hilbert die Dissertation allenfalls überflogen hat. Da Broggi aber auch bemerkt: *es sind die meßbaren Mengen die Einzigen, die wir in Betracht ziehen wollen*, sieht man seine Arbeit trotz ihrer Mängel als einen wichtigen Vorläufer der Axiomatik von Kolmogoroff an. Es ist wichtig zu bemerken, daß zu jener Zeit der Begriff eines Maßes zur Beschreibung von Wahrscheinlichkeiten noch nicht zur Verfügung stand. So schrieb zum Beispiel *Felix Émile Borel (1871–1956)* 1905:

Wenn man die Konvention benutzt, daß die Wahrscheinlichkeit einer Menge proportional zu deren Länge (oder Fläche, oder Volumen) ist, sollte es explizit ausgesprochen werden, daß dies eine Konvention ist und nicht die eigentliche Bedeutung von Wahrscheinlichkeit.

Die Maßtheorie, entwichelt vor allem von *Maurice René Fréchet (1878–1973)* und *Constantin Carathéodory (1873–1950)*, war nun ganz entscheidend für die Formulierung des Axiomensystems durch Kolmogoroff, der 1933 in seinem berühmten Buch *Grundbegriffe der Wahrscheinlichkeitsrechnung* die heute akzeptierte axiomatische Formulierung der Begriffe Wahrscheinlichkeitsraum, Wahrscheinlichkeit und Ereignis gab. Wir geben an dieser Stelle nur den Hinweis, daß die Entwicklung einer Maßtheorie, die sich von den geometrischen Elementen befreit hatte, entscheidend war. Im Vorwort schreibt Kolmogoroff:

Der diesen allgemeinen Gesichtspunkten entsprechende Aufbau der Wahrscheinlichkeitsrechnung war in den betreffenden Kreisen seit einiger Zeit geläufig; es fehlte jedoch eine vollständige und von überflüssigen Komplikationen freie Darstellung des ganzen Systems. Im zweiten Kapitel („Unendliche Wahrscheinlichkeitsfelder") definiert Kolmogoroff den Begriff des Wahrscheinlichkeitsfeldes folgendermaßen:

Es seien E eine beliebige Menge, $\mathcal{F}$ ein Körper der Untermengen von E, welcher E enthält, und $P(A)$ eine nichtnegative auf $\mathcal{F}$ definierte vollständig additive Mengenfunktion; der Mengenkörper $\mathcal{F}$ zusammen mit der Mengenfunktion $P(A)$ bildet dann ein Wahrscheinlichkeitsfeld.

Kolmogoroff hat in seiner Arbeit über die schon von anderen benutzte Analogie zwischen dem Maß einer Menge und der Wahrscheinlichkeit eines Ereignisses, dem Integral einer Funktion und dem Erwartungswert einer zufälligen Größe hinaus die Theorie der unendlichdimensionalen Wahrscheinlichkeitsräume und die Theorie der bedingten Wahrscheinlichkeiten und Erwartungen hinzugefügt. Kolmororoff legte sein Werk relativ bald nach Aufenthalten in Göttingen vor, wo er nach 1930 mehrfach vortrug.

Erwähnt sei noch die Kontroverse um die Axiomatik von *Richard von Mises (1883–1953)*, der 1919 die Wahrscheinlichkeitsrechnung auf axiomatische Weise aus der Häufigkeitsinterpretation heraus entwickelte. Seiner Meinung nach ist der Begriff der Wahrscheinlichkeit nur sinnvoll als Grenzwert von relativen Häufigkeiten. Grundlegend war bei von Mises der Begriff des Kollektivs. Sei $e_1, e_2, \ldots$ eine unendliche Folge (e_i) gedachter Dinge, die Elemente genannt werden (er führt als Beispiel die Folge der Würfe eines Würfels auf). Jedem e_i sei als Merkmal eine reelle Zahl x_i (oder ein Vektor) zugeordnet, zum Beispiel jedem Wurf das bei diesem Wurf beobachtete Ergebnis. Die Folge (e_i) heißt *Kollektiv*, wenn die Zuordnung der Merkmale die folgenden Axiome erfüllt:

(M1) (Existenz der Grenzwerte) Für jede Teilmenge A des Merkmalraums konvergiert die relative Häufigkeit, mit der die ersten n der x_i zu A gehören, also $h_n(A) =$ Anzahl der $i \leq n$ mit $x_i \in A/n$ gegen einen Grenzwert W_A. Dieser heißt die Wahrscheinlichkeit von A innerhalb des Kollektivs.

(M2) (Regellosigkeit der Zuordnung) Für beliebige disjunkte Teilmengen A, B des Merkmalraums, mit $W_A + W_B > 0$, gelte die folgende Aussage: Streicht man aus (e_i) alle e_i, für die das zugehörige x_i nicht zu $A \cup B$ gehört, und wählt man aus dieser Teilfolge eine unendliche Teilfolge (e_i') daraus aus, daß über die Indices der ausgewählten Elemente ohne Benutzung ihrer Merkmalunterschiede verfügt wird, so sollen innerhalb von (e_i') die zu A und B gehörigen Grenzwerte W_A', W_B' der relativen Häufigkeiten existieren, und es soll $W_A' : W_B' = W_A : W_B$ gelten.

Felix Hausdorff (8.11.1869–26.1.1942) wies von Mises darauf hin, daß die relativen Häufigkeiten im allgemeinen nicht für alle Teilmengen, nicht einmal für alle meßbaren konvergieren. Weiter bemerke er, daß die von von Mises ohne Beweis unterstellte σ-Additivität im Allgemeien nicht gilt. Weiter wies man daraufhin, daß von Mises so allgemeine Auswahlen zugelassen hatte, daß es keine Folge gibt, die der Forderung genügt. Die Diskussion um die Inkonsistenz des Kollektivbegriffs wurde teils polemisch geführt. Ein Durchbruch für die Rechtfertigung des Kollektivbegriffs gelang *Abraham Wald (1902–1950)* 1937. Er konnte zeigen: Schränkt man (M1) auf die Mengen A einer abzählbaren Algebra von Teilmengen ein, und die Regellosigkeitsforderung in (M2) nur für ein vorgegebenes abzählbares System von Stellenauswahlen, so gibt es kontinuierlich viele Kollektivs. Die Wahrscheinlichkeit W_A ist dann eine auf der Algebra definierte additive, normierte, nichtnegative Mengenfunktion. Wald stellt seine Arbeit während einer Konferenz an der Universität Genf 1937 vor, die sich mit den Grundlagen der Wahrscheinlichkeitstheorie beschäftigte. Feller zeigte 1939, daß sich mit Hilfe des starken Gesetzes der großen Zahlen leicht zeigen läßt, daß fast

jede Realisierung einer Folge von unabhängigen identisch verteilten Zufallsvariablen ein Kollektiv ist. Doch der durchschlagende Erfolg des Kolmogoroffschen Ansatzes führte dazu, daß man kein Interesse an der Theorie von von Mises mehr hatte. Es scheint bis heute kein Beispiel einer Anwendung dieser Theorie zu existieren, welches nicht auch mit dem Begriff des Wahrscheinlichkeitsraums nach Kolmogoroff erfolgreich diskutiert werden kann.
Historisch hat die Schaffung des Kollektivbegriffs aber durchaus Bedeutung. So waren mit Hilfe dieses Begriffs erste Schritte in Richtung der Klärung der Frage, welche Folgen als zufällig angesehen werden können, möglich (es geht grob um die Frage, warum die Folge 1001100100 eher den Eindruck macht, zufällig entstanden zu sein, als etwa 10101010). Kolmogoroff nahm 1957 Bezug auf von Mises und führte einen Komplexitätsbegriff ein, der die Regelmäßigkeit einer endlichen Folge beschreibt. Es entwickelte sich eine tiefliegende Theorie der zufälligen Folgen, die sich auf die Theorie der rekursiven Funktionen und der Turing-Maschinen stützt.

Weitere Beiträge vor 1945

Wir stellen ein paar nichtaxiomatische Beiträge vor 1945 zusammen. *Ladislaus von Bortkiewicz (1868–1931)* hat im Jahr 1889 mit seinem Büchlein *Das Gesetz der kleinen Zahlen* auf die Bedeutung der Poissonverteilung aufmerksam gemacht. Die Poissonapproximation war in Vergessenheit geraten. Er erinnerte an den klassischen Poissonschen Grenzwertsatz und gab Schätzer für den Parameter α der Poissonverteilung an sowie für die Standardabweichung des Schätzfehlers. Weiterhin leitete er die Normalapproximation der Poissonverteilung für große α her. Das Beispiel der Hufschläge hatten wir in Kapitel 6 diskutiert. Es erregte damals besondere Aufmerksamkeit. Im Jahr 1913 legte er ein neues Büchlein *Die radiaktive Strahlung als Gegenstand wahrscheinlichkeitstheoretischer Untersuchungen* vor. Er bewies, daß die Wahrscheinlichkeit, in einem Zeitintervall der Länge t genau k Scintillationen beobachten zu können, sich zu $e^{-\alpha t}(\alpha t)^k(k!)$ berechnet. Ferner zeigte er, daß die Wartezeit zur ersten Scintillation exponentialverteilt ist. Es scheint, als seien seine Untersuchungen die erste mathematische Analyse eines Poisson-Prozesses. 1917 betrachtet Bortkiewicz Iterationen in Bernoulli-Experimenten. Man spricht von einer Iteration oder einem *run* der Länge m, wenn in einer Folge das gleiche Ergebnis m mal hintereinander auftritt. Er bestimmte mit einfachen Methoden den Erwartungswert und die Varianz der Zahl der Iterationen gegebener Länge. *Jean Le Rond D'Alembert (1717–1783)* hatte 1761 die Vermutung ausgesprochen, daß lange Folgen von aufeinanderfolgenden gleichen Ereignissen in einer Folge eines Bernoulli-Experiments zwar logisch möglich seien, aber viel unwahrscheinlicher als dies nach den Regeln der Wahrscheinlichkeitstheorie zu erwarten sei. Dies wurde durch die Resultate von Bortkiewicz nicht bestätigt (aus heutiger Sicht ist die These von D'Alembert etwas unklar formuliert). So ergibt sich zum Beispiel bei einem fairen Bernoulli-Experiment eine mittlere Wartezeit von 1 Minute, um einen 5-run zu beobachten, wobei man annehme, man werfe die Münze einmal pro Sekunde. Für einen 15-run ergibt sich der Erwartungswert zu ca. 18 Stunden. Ist p die Erfolgswahrscheinlichkeit des Experiments und r die vorgegebene Mindestlänge der Erfolgs-runs, so erhält man für die mittlere Wartezeit auf den ersten Erfolgs-r-run:

$$\frac{(1-p^r)}{(p^r(1-p))}.$$

Borel gab in seiner berühmten Arbeit aus dem Jahr 1909 einen Beweis des starken Gesetzes der großen Zahlen. *Marc Kac* sagte einmal über diese Arbeit: „All of its theorems are true, but almost all of the proofs are false". Obwohl man auch heute Borel dieses Gesetz zuschreibt, war sein Beweis fehlerhaft. Der deutlichste Einwand war, daß Borel mit der Normalapproximation der Binomialverteilung einfach so rechnete, als sei sie die exakte Verteilung. Den ersten vollständig korrekten Beweis von Borels starkem Gesetz findet man in Hausdorffs 1914 erschienenem Buch über Mengenlehre. Dort ersetzte er die Anwendung der Normalapproximation durch eine Abschätzung der vierten Momente. Dieser Beweis enthielt bereits das Argument, daß auch ohne Unabhängigkeit für eine beliebige Folge von Ereignissen A_n mit $\sum_{n\geq 1} P(A_n) < \infty$ die Wahrscheinlichkeit des Ereignisses A^*, daß unendlich viele A_n eintreten, 0 ist. Dies ist die von *Francesco Paolo Cantelli (1875–1966)* bewiesene Hälfte des Borel-Cantelli-Lemmas. Darüberhinaus lieferte Hausdorffs Beweis auch eine erste Aussage über die Geschwindigkeit der Konvergenz: die Folge

$$(S_n - n/2)/n^{1/2+\varepsilon}$$

ist für jedes $\varepsilon > 0$ fast sicher beschränkt. Dies war ein erster Schritt zu Khinchines berühmten Gesetz des iterierten Logarithnus aus dem Jahre 1924, welches wir in der Vorlesung aber nicht behandelt haben.

Eine wichtige Rolle spielten die von *Oskar Perron (1880–1975)* und *Ferdinand Georg Frobenius (1849–1917)* in den Jahren 1907 bis 1912 erschienenen Arbeiten über Matrizen mit nichtnegativen Elementen. 1907 fragte Markoff bekanntlich, ob Unabhängigkeit eine notwendige Bedingung für ein schwaches Gesetz der großen Zahlen sei. Das Gegenbeispiel führte dann zu den Markoff-Ketten. Es dauerte bis 1931, als von Mises in seinem Buch den Zusammenhang zu der Matrizentheroie erkannte und eine Reihe interessanter Resultate über Markoffsche Ketten aus den Resultaten von Perron und Frobenius herleitete.

Georg Pólya (1887–1985) entwickelte in einer Arbeit über den zentralen Grenzwertsatz 1920 die Momentenmethode als allgemeine Methode zum Beweis von Verteilungskonvergenz. Wir hatten bereits bemerkt, daß Tschebyscheff, Markoff und andere die Konvergenz von Momenten zum Beweis der Konvergenz von Verteilungen standardisierter Summen eingesetzt hatten. Pólya zeigte: ist t_m das m-te Moment der Verteilung F und ist

$$\limsup_{m\to\infty} \sqrt[2m]{t_{2m}}/m < \infty,$$

und konvergieren für jedes m die m-ten Momente einer Folge F_k von Verteilungen gegen t_m, so konvergiert F_k in Verteilung gegen F. Laut Feller wird in der Arbeit von Pólya zum erstenmal der Begriff „Zentraler Grenzwertsatz" verwendet. Berühmt ist Pólyas Arbeit über Irrfahrten im Straßennetz, im $\mathbb{Z}^d$, aus dem Jahr 1921. Er zeigt, daß die symmetrische Irrfahrt auf dem Gitter, bei der man in jedem Zeitpunkt mit gleicher Wahrscheinlichkeit $1/2\,d$ von dem gerade besuchten Zustand $i \in \mathbb{Z}^d$ zu den $2d$ Nachbarn geht, im Fall der Dimension $d \leq 2$ rekurrent ist und im Fall $d \geq 3$ transient. Wir hatten dieses Beispiel nur in einer stark vereinfachten Variante diskutiert. Seine Methoden führten zu analogen Rekurrenz-Untersuchungen bei Markoffschen Ketten. Schließlich führte Pólya 1928 in einer gemeinsamen Arbeit mit Eggenberger ein Modell für Prozesse mit Ansteckung ein, was ein Urnenmodell ist. Hier wird die Zahl der weißen und schwarzen Kugeln, die in den einzelnen Stufen eines Experiments

in der Urne sind, entsprechend der Zahl der schon gezogenen weißen und schwarzen Kugeln verändert. 1931 wurde dieses sogenannte Pólyasche Urnenschema gründlich analysiert.
Ein Blick in die existierende Literatur zeigt, daß wir die Auflistung interessanter Arbeiten zur Wahrscheinlichkeitstheorie noch enorm ausdehnen könnten. Wir begnügen uns einfach mit den genannten Resultaten, die in engem Bezug zum Vorlesungsstoff stehen.

Grundlegung der Mathematischen Statistik

Die Statistik ist etwa um 1890 eine eigenständige Disziplin geworden. Sie hat, wie der Name andeutet, ihren Ursprung im Staatswesen, insbesondere in der Erhebung von Daten in einer Bevölkerung. Der Göttinger Staatswissenschaftler *Achenwall (1719–1772)* hielt eine Vorlesung mit dem Titel *Notitia politica vulgo statistica* (ital.: statista = Staatsmann). Diesen Titel sieht man heute als namensgebend an.
Zunächst stand zum Ende des letzten Jahrhunderts in der englischen Schule die *beschreibende Statistik* im Vordergrund. Die Begriffe der Regression und Korrelation wurden 1885 bzw. 1888 durch *Sir Francis Galton (1822–1911)* eingeführt. Recht bald darauf war das Hauptinteresse die Untersuchung von Aussagen über die zugrundeliegende Verteilung bzw. die sie kennzeichnenden Parameter. Zu erwähnen ist hier insbesondere die Einführung des χ^2-Test durch *Karl Pearson (1857–1936)* in einer Arbeit des Jahres 1900. Pearson führte diesen Test zur Überprüfung der Anpassungsgüte eines Modells an vorliegende Beobachtungen ein. Man sieht seine Arbeit heute als Beginn der *schließenden Statsitik* an. Anfang dieses Jahrhunderts erkannte man insbesondere in England als Aufgabe der *Mathematischen Statistik* das Ziehen von Rückschlüssen aus den Beobachtungen auf die tatsächlich vorliegenden Parameterwerte aufgrund wahrscheinlichkeitstheoretischer Überlegungen.
Entscheidende Fortschritte wurden in der englischen Schule erzielt. Zunächst leitete man Verfahren her, die speziell auf durch endlich-dimensionale Parameter charakterisierte Klassen von Verteilungen zugeschnitten waren. In diesen Bereich der *parametrischen Statistik* fallen die bekannten t und F-Tests oder die Verfahren der Varianz- und Regressionsanalyse, die auf der Normalverteilungsannahme beruhen und vornehmlich mit den Namen *William Sealy Gosset (1876–1937)* (bekannter unter dem Pseudonym „Student", da er als Angestellter der Guiness-Brauerei nicht publizieren durfte) und *Sir Ronald Aylmer Fisher (1890–1962)* verknüpft sind. Student berechnete bereits 1908 für Normalverteilungsmodelle die exakten Verteilungen der Stichproben-Standardabweichung und der Prüfgröße des t-Tests, Fisher dann 1915 diejenige des Stichproben-Korrelationskoeffizienten. Student erkannte bereits damals die Bedeutung der Gegenhypothese.
Ab 1930 kam in die Statistik der optimierungstheoretische Gesichtspunkt hinzu. Für praktische Anwendungen interessiert nicht irgendein, sondern ein „möglichst gutes" Verfahren. Fisher wies in Arbeiten um 1922 auf die Notwendigkeit einer optimalen Festlegung hin. Es dauerte noch weitere 10 Jahre, bis die Testtheorie eine entsprechende Klarheit gewonnen hatte. Verbunden damit sind die Namen *Jerzy Neyman (1894–1981)* und *Egon Sharpe Pearson (1895–1980)*. Die bereits von Gosset erwähnte Erkenntnis, daß zur Angabe eines opitmalen Tests zur Prüfung einer Nullhypothese auch gesagt werden muß, wogegen diese Prüfung erfolgen soll, führte zu dem heute auch in der Praxis üblichen *Signifikanztest*. Er beruht darauf, daß man unter Einhaltung einer Schranke für die Wahrscheinlichkeit von Fehlern 1. Art (Ablehnung

einer richtigen Nullhypothese) die Wahrscheinlichkeit für Fehler 2.Art (Annahme einer falschen Nullhypothese) minimiert. Wir hatten gesehen, daß die Lösung dieses Optimierungsproblems im Spezialfall von einfacher Null- und Gegenhypothese explizit angegeben werden kann. Man gewinnt einen gleichmäßig besten Test.
Wir belassen es bei diesem kurzen Einblick in die Geschichte der Mathematischen Statistik, da dieser Teil in der Vorlesung nur kurz behandelt wird.