Zufallsvariable und Wahrscheinlichkeiten

Eine elementare Einführung in die Stochastik

Sommersemester 06

Prof. Dr. Götz Kersting

16. Juni 2006

Inhaltsverzeichnis

Kapitel 1

Elementare Ansätze

Die Wahrscheinlichkeitstheorie und die Statistik sind alte Wissenschaften, die man heutzutage zu einer Disziplin, der Stochastik zusammenfaßt. Kurz gesagt ist die Stochastik also die mathematische Lehre vom Zufall - diese Charakterisierung ist aber noch recht grob, denn ‚Zufall' kann verschiedenes heißen. Die Stochastik hat Situationen vor Augen, in denen eine Anzahl von Alternativen bestehen, von denen dann eine oder mehrere ‚zufällig' realisiert werden[1]. Prototypisch ist der Münzwurf, der eine Zufallswahl zwischen 0 und 1 (Kopf und Zahl) erlaubt. Die Chancen brauchen dabei nicht ausgeglichen, die Münze nicht fair zu sein. Um unterschiedliche Chancen quantitativ zu bewerten, benutzt man Wahrscheinlichkeiten.

Die Stochastik bedient sich gern Beispiele aus der Welt des Glückspiels, sie ist deswegen aber noch lange keine „Würfelbudenmathematik". Ihr geht es darum, die Vorstellung einer Zufallsentscheidung so allgemein zu fassen, daß sie auch in ganz anderen Bereichen - von der Genetik bis zur Börse - zum Tragen kommen kann. Dazu hat man in der Stochastik den Begriff der **Zufallsvariablen** geprägt, er ist von fundamentaler Bedeutung. Formal gehören zu einer Zufallsvariablen X eine Menge S, ihr **Wertebereich**, sowie **Ereignisse** $\{X \in B\}$, wobei B geeignete Teilmengen von S durchläuft. Man stellt sich vor, daß X einen zufälligen Wert in S annimmt, $\{X \in B\}$ steht dann für das zufällige Ereignis, daß dieser Wert B angehört. Analog steht $\{X = x\}$ für das Ereignis, dass X einen vorgegebenen Wert $x \in S$ annimmt. Die Chance, daß das Ereignis $\{X \in B\}$ eintrifft, wird durch seine **Wahrscheinlichkeit** quantifiziert, einer Zahl $\mathbf{Ws}\{X \in B\}$ zwischen 0 und 1. Die Gesamtheit der Wahrscheinlichkeiten

$$\mu(B) \;:=\; \mathbf{Ws}\{X \in B\}\;, \quad B \subset S\;,$$

[1] Aristoteles begriff das Zufällige als dasjenige, was weder unmöglich noch notwendig ist und darum auch nicht oder auch anders sein könnte.

nennt man die **Verteilung** μ von X.

Dies sind geläufige Sprech- und Schreibweisen der Stochastik. Wir benutzen sie von Anfang an, auch wenn sie bei aller Suggestivität noch nicht die Ansprüche erfüllen, die man in der Mathematik an Begriffsbildungen stellt. Auf die mathematischen Grundlagen gehen wir im nächsten Kapitel ein, zunächst wollen wir uns anhand verschiedener Beispiele mit Zufallsvariablen vertraut machen, dabei einige wichtige Verteilungen kennenlernen und ein paar Wahrscheinlichkeiten berechnen, teils exakt, teils approximativ. Ein wesentliches Hilfsmittel ist die Stirlingsche Approximationsformel für Fakultäten. Wie man Phänomene der realen Welt mit Zufallsvariablen und Wahrscheinlichkeiten modellhaft erfaßt, können wir in diesem Kapitel nur ansatzweise ansprechen.

1.1 Uniforme Verteilungen

Definition. *Sei S eine endliche Menge. Eine Zufallsvariable X mit Werten in S heißt* ***uniform (gleichförmig) in S verteilt****, falls für alle $B \subset S$*

$$\mathbf{Ws}\{X \in B\} \;=\; \frac{\text{card } B}{\text{card } S}$$

gilt (mit card $B :=$ *Anzahl der Elemente von B).*

Bei einer gleichförmigen Verteilung wird kein Element von S bevorzugt, man spricht daher auch von einer **rein zufälligen Wahl** eines Elements aus S.

Beispiel. Um einen schnellen Zugriff auf Daten zu haben, kann man sie in Listen aufteilen. Nur bei kurzen Listen sind auch die Suchzeiten kurz, daher stellt sich die Frage, mit welcher Wahrscheinlichkeit es zu „Kollisionen“ kommt, zu Listen, die mehr als einen Eintrag enthalten. Wir berechnen diese Wahrscheinlichkeit für n Listen und k Daten unter der Annahme, daß alle möglichen Belegungen der Listen mit den Daten gleich wahrscheinlich sind. Wir werden sehen, daß mit Kollisionen schon dann zu rechnen ist, wenn k von der Größenordnung $\sqrt{n}$ ist.

Diese Fragestellung ist in der Stochastik unter dem Namen **Geburtstagsproblem** bekannt. Gefragt ist nach der Wahrscheinlichkeit, daß in einer Klasse mit k Schülern alle verschiedene Geburtstage haben. Wir lassen uns von der Vorstellung leiten, daß das Tupel $X = (X_1, \ldots, X_k)$ der k Geburtstage ein rein zufälliges Element aus

$$S \;:=\; \{(x_1, \ldots, x_k) \;:\; x_i \in \{1, \ldots, n\}\}$$

ist, mit $n = 365$. Gesucht ist die Wahrscheinlichkeit, daß X zu

$$B \;:=\; \{(x_1, \ldots, x_k) \in S \;:\; x_i \neq x_j \text{ für alle } i \neq j\}$$

gehört. Es gilt card $B = n(n-1)\cdots(n-k+1)$. Nehmen wir also an, daß es sich um eine rein zufällige Wahl der Geburtstage aus S handelt, so ist die gesuchte Wahrscheinlichkeit

$$\mathbf{Ws}\{X \in B\} \;=\; \frac{\text{card } B}{\text{card } S} \;=\; \frac{n(n-1)\cdots(n-k+1)}{n^k} \;=\; \prod_{i=1}^{k-1}\Big(1 - \frac{i}{n}\Big)\,.$$

Diese Formel ist noch nicht befriedigend, denn sie vermittelt keine Vorstellung, wie groß die Wahrscheinlichkeit ist. Dafür ist die Abschätzung

$$\prod_{i=1}^{k-1}\Big(1 - \frac{i}{n}\Big) \;\le\; \exp\Big(-\sum_{i=1}^{k-1}\frac{i}{n}\Big) \;=\; \exp\Big(-\frac{(k-1)k}{2n}\Big)\,,$$

nützlich, die auf der Ungleichung $1 - t \le e^{-t}$ beruht. Es folgt

$$\mathbf{Ws}\{X \in B\} \;\le\; \exp\Big(-\frac{(k-1)k}{2n}\Big)\,. \tag{1.1}$$

Unklar bleibt, wann diese Abschätzung brauchbare Näherungen ergibt. Wir wollen deswegen eine Approximationsformel ableiten, die immer gute Näherungswerte liefert. Sie beruht auf der **Stirling-Approximation** für Fakultäten $n! = 1 \cdot 2 \cdots n$,

$$n! \;\approx\; \sqrt{2\pi n}\; n^n e^{-n}\,.$$

Da

$$\mathbf{Ws}\{X \in B\} \;=\; \frac{n!}{n^k(n-k)!}\,,$$

erhalten wir die Approximation

$$\mathbf{Ws}\{X \in B\} \;\approx\; \Big(\frac{n}{n-k}\Big)^{n-k+\frac{1}{2}} e^{-k}\,, \tag{1.2}$$

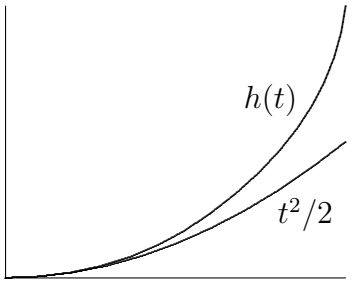
die immer sehr gute Näherungen liefert (und nur im Fall $n = k$ versagt). Die Güte der Formel beruht darauf, daß die Stirlingschen Formeln schon für kleines n gute Approximationswerte liefern (für $n = 1$ lauten sie $1 \approx 0,92$). Ein numerisches Beispiel illustriert dies: Für $k = 25$ und $n = 365$ ist $\mathbf{Ws}\{X \in B\} = 0,431300$, die beiden Näherungswerte sind 0,440 und 0,431308.

Ignorieren wir in (1.2) den Faktor $\sqrt{n/(n-k)}$, der typischerweise nahe bei 1 liegt, so lässt sich die letzte Approximation kompakt als

$$\mathbf{Ws}\{X \in B\} \;\approx\; \exp\big(-n \cdot h\big(\tfrac{k}{n}\big)\big) \tag{1.3}$$

schreiben, mit

$$h(t) \;:=\; t + (1-t)\ln(1-t)\;, \quad 0 \le t \le 1\;.$$



Wegen $h(0) = h'(0) = 0$ und $h''(0) = 1$ gilt $h(t) \approx t^2/2$ für kleines t, wir können daher (1.1) als Taylor-Näherung für (1.3) verstehen. □

Wahrscheinlichkeiten von der Gestalt

$$\frac{\mathrm{card}B}{\mathrm{card}S} \;=\; \frac{\text{Anzahl der ‚günstigen Fälle‘}}{\text{Anzahl der ‚möglichen Fälle‘}}$$

nennt man **Laplace-Wahrscheinlichkeiten** (nach dem Mathematiker Laplace (1749-1827)). Ein **Laplace-Mechanismus** ist ein Mechanismus, der eine rein zufällige Wahl eines Elements aus S realisiert. Die Frage, ob es solche Mechanismen „in Wirklichkeit“ gibt, hat sich nicht recht klären lassen und überhaupt als wenig fruchtbar erwiesen. Für uns ist das nicht besonders wichtig. In der Stochastik dienen uniform verteilte Zufallsvariable und Laplace-Mechanismen als Gedankenmodelle, die man an die Wirklichkeit heranträgt, und deren Angemessenheit sich von Fall zu Fall erweisen muß. So kann man natürlich den oben für das Geburtstagsproblem gewählten Ansatz, die Verteilung der Geburtstage durch eine rein zufällige Wahl zu beschreiben, in Frage stellen, etwa durch den Verweis auf Schaltjahre, Zwillingsgeburten etc. Deswegen wird man aber diesen Ansatz nicht von vornherein verwerfen.

Bei der Berechnung von Laplace-Wahrscheinlichkeiten stellt sich die Aufgabe, Mächtigkeiten endlicher Mengen zu bestimmen. Dies ist ein Thema der Kombinatorik. Wichtige kombinatorische Größen sind Fakultät und Binomialkoeffizient (bzw. Multinomialkoeffizient). Die **Fakultät**

$$n! \;:=\; 1 \cdot 2 \cdots n$$

der natürlichen Zahl n gibt an ($0! := 1$), in wieviel verschiedenen Weisen sich n Dinge nebeneinander aufreihen lassen (wieviel Permutationen der Länge n existieren). Der **Binomialkoeffizient**

$$\binom{n}{x} := \frac{n(n-1)\cdots(n-x+1)}{1\cdot 2\cdots x} = \frac{n!}{x!(n-x)!} ,$$

mit ganzen Zahlen $0 \le x \le n$ bestimmt, wieviele Teilmengen H der Mächtigkeit x in einer Menge G der Mächtigkeit n enthalten sind. Es gibt nämlich $n(n-1)\cdots(n-x+1)$ Möglichkeiten, der Reihe nach x verschiedene Elemente aus G auszuwählen. Dabei entsteht jede x-elementige Teilmenge H auf $x!$ Weisen, weil ihre Elemente in verschiedenen Reihenfolgen gezogen werden können.

Beispiele. Das Gerät, mit dem die Lotto-Zahlen gezogen werden, kann man als Laplace-Mechanismus zur Wahl einer 6-elementigen Menge aus $G = \{1, \ldots, 49\}$ ansehen. Die Wahrscheinlichkeit für einen Hauptgewinn ist folglich

$$1 \Big/ \binom{49}{6} = 1 \,/\, 13.983.816 \simeq 7,15 \cdot 10^{-8} ,$$

die Wahrscheinlichkeit für 4 Treffer

$$\binom{6}{4}\binom{43}{2} \Big/ \binom{49}{6} \simeq 0,0010 .$$

Die Laplace-Wahrscheinlichkeit, in einem Skatblatt (10 Karten aus 32) alle 4 Asse zu finden, ist

$$\binom{28}{6} \Big/ \binom{32}{10} \simeq 0,0058 .$$

□

Die folgende Fragestellung führt zur Verallgemeinerung des Binomialkoeffizienten. n Dinge sollen so auf k Fächer verteilt werden, daß das i-te Fach genau x_i Objekte enthält (mit $x_1 + \cdots + x_k = n$). Wir stellen uns vor, daß wir die Gegenstände in einer willkürlichen Reihenfolge verteilen, erst x_1 Stück ins erste Fach, die nächsten x_2 ins zweite Fach und so fort. Es gibt $n!$ verschiedene Reihenfolgen, davon enthält in jeweils $x_1!$ Fällen das erste Fach dieselben Objekte, in jeweils $x_2!$ das zweite Fach dieselben Objekte und so weiter. Die Anzahl der Möglichkeiten ist also

$$\binom{n}{x_1, x_2, \ldots, x_k} := \frac{n!}{x_1! \cdot x_2! \cdots x_k!} .$$

Dieser Ausdruck heißt **Multinomialkoeffizient** ($n = x_1 + \cdots + x_k$).

Beispiel. Die Laplace-Wahrscheinlichkeit, daß beim Bridge jeder der 4 Spieler unter seinen 13 Karten genau ein As hat, ist

$$\binom{48}{12,\ldots,12}\binom{4}{1,\ldots,1} \Big/ \binom{52}{13,\ldots,13} \simeq 0,11 \ .$$

(Natürlich ist die Frage berechtigt, ob es angemessen ist, bei einem Kartenspiel alle Blätter als gleichwahrscheinlich anzusehen.) □

Im nächsten Beispiel geht es um das Testen einer Hypothese, eine Fragestellung, die systematisch in der Statistik behandelt wird.

Beispiel. Im Hörsaal sitzen in der ersten Reihe x Studenten. Es fällt auf, daß jeder für sich allein sitzt. Darf man vermuten, daß sie diese Platzwahl bewußt getroffen haben, oder könnte hier auch der Zufall eine Rolle gespielt haben? - Wir berechnen dazu die Wahrscheinlichkeit, daß sich eine solche Sitzverteilung rein zufällig ergibt. Sei n die Zahl der Plätze in der ersten Reihe. Insgesamt gibt es dann $\binom{n}{x}$ verschiedene Möglichkeiten, die Hörer auf die Sitze zu verteilen. Sitzen sie voneinander getrennt, so kann man gedanklich zwischen je zwei Personen einen Sitzplatz, insgesamt also $x-1$ Plätze entfernen. Es gibt also genauso viele Möglichkeiten zum getrennten Sitzen, wie man x Personen auf $n-(x-1)$ Sitze verteilen kann. Die gesuchte Laplace-Wahrscheinlichkeit ist folglich

$$p_{n,x} := \binom{n-x+1}{x} \Big/ \binom{n}{x} = \frac{(n-x)(n-x-1)\cdots(n-2x+2)}{n(n-1)\cdots(n-x+2)} \ .$$

Für $n = 25, x = 8$ erhält man $p_{n,x} = 0,04$. Bei einem solch kleinen Wert darf man wohl bezweifeln, daß die Platzwahl rein zufällig getroffen wurde. Orientiert man sich an den Vorstellungen der Statistik, so würde man sich vorneweg einen Maximalwert für $p_{n,x}$ vorgeben, bei dessen Überschreitung man die Annahme einer rein zufälligen Platzwahl nicht mehr in Frage stellt – gängig ist der Wert 0,05 (und auch 0,01). Im Jargon der Statistiker könnte man dann feststellen: Für $p_{x,n} = 0,04$ wird die (Null-)Hypothese reiner Zufälligkeit auf dem Signifikanzniveau von 0,05 verworfen. □

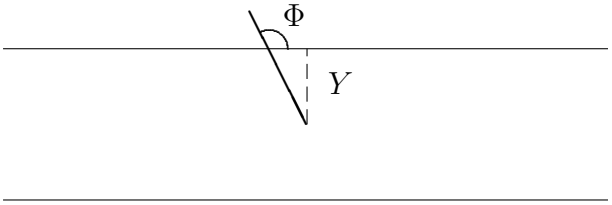
Uniforme Verteilungen betrachtet man nicht nur auf endlichen Mengen.

Definition. *Sei S eine Teilmenge des $\mathbb{R}^d$, $d \geq 1$, von endlichem Inhalt $|S|$. Dann heißt eine S-wertige Zufallsvariable X* ***uniform (gleichförmig) verteilt in S****, falls für alle $B \subset S$ mit wohldefiniertem Inhalt $|B|$ gilt*

$$\mathbf{Ws}\{X \in B\} = \frac{|B|}{|S|} \ .$$

Wieder ist die Vorstellung die, daß kein Element aus S bevorzugt ausgewählt wird. In der Maßtheorie lernt man, für welche Teilmengen des $\mathbb{R}^d$ sich ein Inhalt (Volumen, Fläche oder Länge) definieren läßt; zur Behandlung von Beispielen benötigt man diese Kenntnisse meist nicht. Die Modellannahme einer uniformen Verteilung hat sich vielfach bewährt.

Beispiel. Buffons Nadelproblem. Eine Nadel der Länge ℓ wird zufällig auf ein liniertes Blatt geworfen. Wie groß ist die Wahrscheinlichkeit, daß sie eine Linie schneidet? Wir beschränken uns auf den Fall, daß der Abstand d zwischen den Linien größer als ℓ ist. Die Lage der Nadel beschreiben wir durch den Winkel Φ zwischen 0 und π, den sie mit den Linien bildet, und dem Abstand Y, den das untere Nadelende von der nächsthöher gelegenen Linie hat:



Zum Schnitt kommt es, falls $Y \le \ell \cdot \sin\Phi$ gilt. Wir machen nun den Ansatz, daß $X = (\Phi, Y)$ eine uniform in $S = [0,\pi) \times [0,d)$ verteilte Zufallsvariable ist. Unter Beachtung von $d > \ell$ bestimmt sich der Flächeninhalt von $B = \{(\phi, y) \in S : y \le \ell \cdot \sin\phi\}$ als

$$|B| \;=\; \int_0^\pi \ell \cdot \sin\phi \; d\phi \;=\; 2\ell \;,$$

die gesuchte Wahrscheinlichkeit ist also

$$\mathbf{Ws}\{X \in B\} \;=\; \frac{|B|}{|S|} \;=\; \frac{2\ell}{\pi d} \;.$$

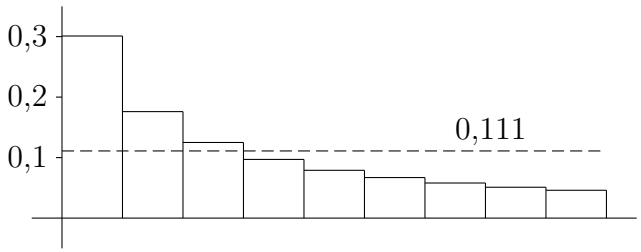
Damit kann man $\pi = 3,14\ldots$ durch wiederholten Wurf einer Nadel statistisch schätzen. (Das Lehrbuch *Elementare Wahrscheinlichkeitsrechnung* von Pfanzagl enthält dazu Datenmaterial.) □

Beispiel. Benfords Gesetz. Der Physiker Benford hat 1936 (wie schon vor ihm 1881 der Astronom Newcomb) für verschiedene Datensätze von positiven Zahlen eine merkwürdige Beobachtung gemacht: Wählt man aus dem

Datensatz zufällig eine Zahl aus und betrachtet die erste signifikante Ziffer in ihrer Dezimaldarstellung, so ist diese Ziffer bevorzugt eine 1, viel seltener dagegen eine 9. Benford stützte sich auf ganz unterschiedliche Datensätze, wie die Fläche von Flüssen, Konstanten der Physik, aus Zeitungen ausgewählte Zahlen und andere Daten. Für das Phänomen hat er sogar eine Gesetzmäßigkeit angegeben: Die Wahrscheinlichkeit, daß es sich bei der Anfangsziffer um k handelt, ist durch den Ausdruck

$$p_k \;=\; \log_{10}\left(1+\tfrac{1}{k}\right)$$

gegeben, $k = 1, 2, \ldots, 9$. Diese Verteilung, die er empirisch aus seinem Datenmaterial gewann, weicht merklich von der uniformen Verteilung auf $\{1, \ldots, 9\}$ ab.



Benfords Verteilung

Das Phänomen ist auf den ersten Blick unglaublich, warum sollten die Ziffern 1,...,9 nicht gleichberechtigt sein? Daß es Unterschiede gibt, erkennt man, wenn man sich fragt, in welchem Maße man eine Zahl vergrößern muß, damit sich ihre Anfangsziffer ändert: Die Zahl 1 (oder 10) muß man dafür mindestens verdoppeln, bei der Zahl 9 (oder 90) langt schon der Faktor 10/9.

Das Gesetz von Benford läßt sich nicht „beweisen“, man kann nur versuchen, es plausibel zu machen, indem man es in einem Modell nachbildet (dessen Angemessenheit jederzeit infrage gestellt werden darf). Wir stellen ein stochastisches Modell auf, daß Benfords Gesetz aus einer uniformen Verteilung heraus erklärt. Die aus den Daten zufällig gezogene Zahl bezeichnen wir mit X und ihre zufällige Anfangsziffer mit D. Dann nimmt D genau dann den Wert k an, wenn $k \cdot 10^n \le X < (k+1) \cdot 10^n$ für ein $n \in \mathbb{Z}$ gilt, bzw.

$$\log_{10} k + n \;\le\; \log_{10} X \;<\; \log_{10}(k+1) + n \;.$$

Wegen $0 \le \log_{10} k < \log_{10}(k+1) \le 1$ für $k = 1, \ldots, 9$ ist insbesondere n die größte ganze Zahl kleiner oder gleich $\log_{10} X$, wir schreiben $n = [\log_{10} X]$.

Insgesamt können wir feststellen, daß $D = k$ genau dann gilt, wenn für die Zufallsvariable

$$U \;:=\; \log_{10} X - [\log_{10} X]$$

mit Werten in $[0, 1)$ die Bedingung $\log_{10} k \;\le\; U \;<\; \log_{10}(k+1)$ erfüllt ist. In Ereignissen ausgedrückt bedeutet dies

$$\{D = k\} \;=\; \{U \in [\log_{10} k, \log_{10}(k+1))\}\ .$$

Wir betrachten nun den *Ansatz*, daß U eine in $[0, 1)$ uniform verteilte Zufallsvariable ist. Dies erscheint insbesondere für Daten plausibel, die über einen weiten Bereich streuen. Der Ansatz impliziert das Benfordsche Gesetz, es folgt nämlich

$$\begin{aligned} \mathbf{Ws}\{D = k\} \;&=\; \mathbf{Ws}\{U \in [\log_{10} k, \log_{10}(k+1))\} \\ &=\; \log_{10}(k+1) - \log_{10} k \;=\; \log_{10}(1 + \tfrac{1}{k})\ . \end{aligned}$$

Eine wichtige Eigenschaft unseres Ansatzes ist, daß sie invariant unter einem Skalenwechsel ist: Gehen wir von X zu $X' = c \cdot X$ mit einer Konstanten $c > 0$, so folgt $\log_{10} X' = \log_{10} X + \log_{10} c$ und damit $U' \equiv U + \log_{10} c \bmod 1$. Mit U ist dann auch U' uniform verteilt, und Benfords Gesetz gilt auch für X'.

Daten, für die es keine ausgezeichnete Skala gibt (wie die Fläche von Flüssen oder physikalische Konstanten), sind damit Kandidaten für Benfords Gesetz. Dagegen kommen Daten, die an eine spezielle Skala adjustiert sind (etwa an einen Index oder, wie Preise, an eine Währung), für Benfords Gesetz weniger in Betracht. (Für Beispiele und weitere Details vgl. T. Hill, The Significant-Digit Phenomen, *American Mathematical Monthly* **102**, 1995, 322-327). □

1.2 Binomial-, Poisson- und Hypergeometrische Verteilung

Wir kommen nun zu drei wichtigen Typen von Verteilungen. Zunächst wollen wir eine Formel für die Verteilung der Zufallsvariablen

$$X \;=\; \text{Anzahl der Erfolge bei } n\text{-fachem unabhängigen Wiederholen eines Bernoulli-Experiments}$$

angeben. Unter einem **Bernoulli-Experiment** versteht man ein Experiment mit zwei möglichen Ausgängen, genannt ‚Erfolg' und ‚Mißerfolg'. Man kann

an das Werfen einer Münze oder eines Würfels denken, oder auch an das Drehen eines Glücksrades mit zwei Sektoren. p sei die Wahrscheinlichkeit für einen Erfolg und $q = 1 - p$ die Gegenwahrscheinlichkeit, die Wahrscheinlichkeit für einen Mißerfolg. Bei unabhängiger Wiederholung des Experiments multiplizieren sich diese Wahrscheinlichkeiten, die Wahrscheinlichkeit für x Erfolge und $n - x$ Mißerfolge in einer vorgegebenen Reihenfolge ist also $p^x q^{n-x}$. Da es $\binom{n}{x}$ Möglichkeiten gibt, x Erfolge in einer Versuchsserie der Länge n unterzubringen, ist die Wahrscheinlichkeit, daß X den Wert x annimmt, gerade $\binom{n}{x} p^x q^{n-x}$. Dies führt uns zu folgender Sprechweise.

Definition. *Sei $n \in \mathbb{N}$ und $p \in [0,1]$. Eine Zufallsvariable X mit Werten in $\{0, 1, \ldots, n\}$ heißt* ***binomialverteilt zum Parameter (n, p)****, kurz* ***$B(n, p)$-verteilt****, falls für $x = 0, \ldots, n$ gilt*

$$\mathbf{Ws}\{X = x\} \;=\; \binom{n}{x} p^x q^{n-x} \;.$$

$\binom{n}{x} p^x q^{n-x}$, $x = 0, \ldots, n$, heißen die **Gewichte** der Binomialverteilung. Nach dem Binomischen Lehrsatz summieren sie sich zu 1 auf,

$$\sum_{x=0}^{n} \binom{n}{x} p^x q^{n-x} \;=\; (p+q)^n \;=\; 1 \;.$$

Im einfachsten Fall $n = 1$ nimmt X nur die Werte 1 oder 0 an, und zwar mit Wahrscheinlichkeit p bzw. q. Man spricht dann von einer Zufallsvariablen mit einer **Bernoulli-Verteilung** zur Erfolgswahrscheinlichkeit p oder kurz einer **B(p)-Verteilung**.

Beispiele.

1. Eine Folge von Bernoulli-Experimenten denkt man sich gern in einem **Urnenmodell** realisiert: Aus einer Urne mit r roten und s schwarzen Kugeln, insgesamt $t = r + s$ Kugeln, wird eine Stichprobe der Länge n **mit Zurücklegen** gezogen, d.h. jede gewählte Kugel wird zurückgelegt, bevor die nächste Kugel gezogen wird. Dann sind (unter Berücksichtigung der Reihenfolge, in der die Kugeln erscheinen) t^n verschiedene Stichproben möglich. Darunter gibt es $\binom{n}{x} r^x s^{n-x}$, die genau x rote Kugeln enthalten. Die Laplace-Wahrscheinlichkeit für dieses Ereignis ist

$$\frac{\binom{n}{x} r^x s^{n-x}}{t^n} \;,$$

die Anzahl X der roten Kugeln in der Stichprobe ist also $\mathrm{B}(n,p)$-verteilt mit $p = r/t$. - Ähnlich ist die Zahl der Sechsen bei n-fachem Würfeln $\mathrm{B}\big(n, \frac{1}{6}\big)$-verteilt.

2. **Runs.** $Z_0, Z_1, \ldots, Z_n$ sei eine per Münzwurf erzeugte, rein zufällige Folge aus Nullen und Einsen. Wir betrachten die Anzahl Y von Runs, von Maximalserien aus Nullen oder aus Einsen in $Z_0, \ldots, Z_n$ (die Folge 0 1 1 0 0 1 0 beispielsweise enthält 5 Runs). Da mit dem Ereignis $Z_i \neq Z_{i-1}$ immer ein neuer Run beginnt, gilt

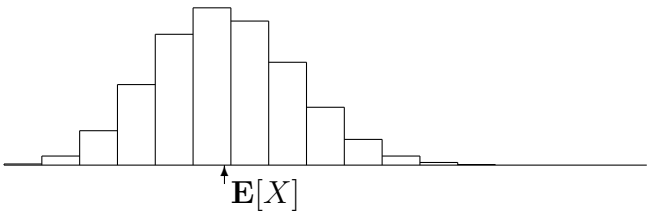
$$Y \;:=\; 1 + \mathrm{card}\{i : i = 1, \ldots, n,\; Z_i \neq Z_{i-1}\}\ .$$

$Y - 1$ ist eine binomialverteilte Zufallsvariable zum Parameter $\big(n, \frac{1}{2}\big)$. Fassen wir nämlich das Eintreten von $Z_i \neq Z_{i-1}$ als Erfolg im i-ten Versuch auf, so ist $Y-1$ die Anzahl der Erfolge in einer unabhängigen Versuchsserie der Länge n. (Dies setzt voraus, daß die Erfolgswahrscheinlichkeit $1/2$ ist. Benutzt man eine unfaire Münze, so ist die Verteilung von Y nicht so leicht zu bestimmen.) □

Das mit den Wahrscheinlichkeiten gewichtete Mittel $\mathbf{E}[X]$ aller möglichen Werte einer binomialverteilten Zufallsvariablen X wird als ihr **Erwartungswert** bezeichnet,

$$\mathbf{E}[X] \;:=\; \sum_{x=0}^{n} x \cdot \mathbf{Ws}\{X = x\}\ .$$

Häufig wird er als der „mittlere Wert“ von X interpretiert. Präziser ist die Aussage, daß $\mathbf{E}[X]$ genau der Schwerpunkt einer Massenverteilung ist, die im Punkt x die Masse $\binom{n}{x} p^x q^{n-x}$ plaziert.



Die $\mathrm{B}\big(16, \frac{1}{3}\big)$-Verteilung

Wir berechnen $\mathbf{E}[X]$ unter Beachtung von $\binom{n}{x} = \frac{n}{x}\binom{n-1}{x-1}$ mit dem binomischen Lehrsatz:

$$\sum_{x=0}^{n} x \binom{n}{x} p^x q^{n-x} \;=\; np \sum_{x=1}^{n} \binom{n-1}{x-1} p^{x-1} q^{n-x} \;=\; np\,(p+q)^{n-1}\ ,$$

also

$$\mathbf{E}[X] \;=\; np\;.$$

Man wird vermuten, daß X seine Werte bevorzugt um den Erwartungswert herum annimmt. Das folgende Beispiel gibt einen Hinweis, mit welchen Abweichungen man zu rechnen hat.

Beispiel. Wirft man n Münzen, dann ist $\binom{n}{n/2}\cdot 2^{-n}$ die Wahrscheinlichkeit, daß genau die Hälfte der Münzen Kopf zeigt (n gerade). Sie ist approximativ gleich $(\pi n/2)^{-1/2}$, denn aus der Stirling-Approximation folgt

$$\binom{n}{n/2} \;\approx\; \frac{\sqrt{2\pi n}\;n^n e^{-n}}{(\sqrt{\pi n}\;(n/2)^{n/2}e^{-n/2})^2} \;=\; \frac{2^{n+\frac{1}{2}}}{\sqrt{\pi n}}\;. \qquad \square$$

Allgemeiner gilt, daß die Gewichte einer Binomialverteilung in der Nähe ihres Erwartungswerts von der Größenordnung $n^{-1/2}$ sind. Daher wird man zwischen X und $\mathbf{E}[X]$ mit Abweichungen von der Größenordnung $n^{1/2}$ rechnen müssen. Wir werden dies später präzisieren.

Von mindestens so großer Bedeutung wie die Binomialverteilung ist die Poisson-Verteilung.

Definition. *Eine Zufallsvariable X mit Werten in $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$ heißt* ***Poisson-verteilt*** *zum Parameter $\lambda \geq 0$, kurz* ***$\boldsymbol{P(\lambda)}$-verteilt****, falls für $x = 0, 1, \ldots$ gilt*

$$\mathbf{Ws}\{X = x\} \;=\; e^{-\lambda}\;\frac{\lambda^x}{x!}\;.$$

Man beachte, daß die Gewichte sich zu 1 aufsummieren,

$$\sum_{x=0}^{\infty} e^{-\lambda}\;\frac{\lambda^x}{x!} \;=\; 1\;.$$

Der Erwartungswert $\mathbf{E}[X]$ einer $\mathrm{P}(\lambda)$-verteilten Zufallsvariablen X ist

$$\sum_{x=0}^{\infty} x\cdot\mathbf{Ws}\{X = x\} \;=\; \lambda e^{-\lambda}\sum_{x=1}^{\infty}\;\frac{\lambda^{x-1}}{(x-1)!}\;.$$

Die Summe hat den Wert e^{λ}, es gilt also

$$\mathbf{E}[X] \;=\; \lambda\;.$$

Poisson-Verteilungen sind brauchbare Approximationen für Binomialverteilungen bei einem großen Stichprobenumfang n, aber nicht allzugroßem Erwartungswert np. Dies zeigt ein Grenzübergang $n \to \infty$ für die Gewichte der Binomialverteilung, bei dem man den Erwartungswert festhält.

Satz 1.1. *Sei X_n, $n \in \mathbb{N}$, eine Folge von $B(n,p)$-verteilten Zufallsvariablen (mit variablem $p = p_n$), und sei $\lambda \geq 0$. Gilt $\mathbf{E}[X_n] \to \lambda$ für $n \to \infty$, d.h. $p \sim \lambda/n$, so folgt*

$$\lim_{n\to\infty} \mathbf{Ws}\{X_n = x\} \;=\; e^{-\lambda}\,\frac{\lambda^x}{x!}$$

für alle $x \in \mathbb{N}_0$.

Beweis. Die Behauptung folgt aus

$$\binom{n}{x} p^x q^{n-x} \;=\; \frac{1}{x!}\,\frac{n(n-1)\cdots(n-x+1)}{n^x}\,\Big(\frac{np}{q}\Big)^x\,(1-p)^n$$

unter Beachtung von $np/q \to \lambda$ und $(1-p)^n \sim (1-\frac{\lambda}{n})^n \to e^{-\lambda}$. □

Beispiele.

1. Verteilt man n Kugeln unabhängig voneinander rein zufällig auf n Schachteln, so ist die Anzahl der Kugeln in einer einzelnen Schachtel $\mathrm{B}\big(n, \frac{1}{n}\big)$-verteilt und für großes n approximativ $\mathrm{P}(1)$-verteilt.

2. Radioaktives Material besteht aus einer immensen Zahl n von Teilchen, die unabhängig voneinander mit sehr kleiner Wahrscheinlichkeit p zerfallen. Die Gesamtzahl der in einem festen Zeitintervall zerfallenen Teilchen ist daher eine Poissonsche Zufallsvariable, eine Tatsache, die empirisch gut belegt ist.

   In einem bekannten Experiment hat man insgesamt 2608-mal beobachtet, wieviele Teilchen während 7,5 sec zerfallen. Die folgende Tabelle enthält die relativen Häufigkeiten h_x der Fälle, bei denen genau x Teilchen registriert werden, sowie die Gewichte p_x einer an die Daten angepaßten Poisson-Verteilung.

x	0	1	2	3	4	5	6	7	≥ 8
h_x	0,022	0,078	0,147	0,201	0,204	0,156	0,105	0,053	0,034
p_x	0,022	0,083	0,159	0,203	0,195	0,150	0,096	0,053	0,039

   Der Wert von λ in $p_x = e^{-\lambda}\lambda^x/x!$ ist aus den Daten geschätzt, er ist in Anbetracht von $\lambda = \sum_x x \cdot p_x$ als $\sum_{x=0}^{8} x \cdot h_x = 3,84$ gewählt. Man erkennt, daß die angepaßten Gewichte gut mit den beobachteten Häufigkeiten übereinstimmen. Da sich eine perfekte Übereinstimmung bei zufälligen

Häufigkeiten h_x nicht einstellen wird, stellt sich die Frage, wie große Abweichungen zwischen (h_x) und (p_x) noch toleriert werden können. Diese Frage wird in der Statistik beantwortet, es ergibt sich, daß die vorliegenden Abweichungen zwischen h_x und p_x von einer plausiblen Größe sind. □

Andere Daten passen ähnlich gut zur Poisson-Verteilung. Bekannt sind Datensätze aus der Schweiz über die Anzahl der 100-jährigen Geburtstage pro Jahr oder die jährliche Zahl der Reitunfälle in der preußischen Kavallerie.

Die hypergeometrische Verteilung, auf die wir nun zu sprechen kommen, ist für die Stichprobentheorie von besonderer Bedeutung. Eine Urne enthalte r rote und s schwarze Kugeln, die totale Zahl der Kugeln ist also $t = r + s$. Wir ziehen aus der Urne rein zufällig eine Stichprobe vom Umfang n und betrachten die Zufallsvariable

$$X \;=\; \text{Anzahl der roten Kugeln in der Stichprobe} \quad .$$

Wird die Stichprobe mit Zurücklegen gezogen, so ist X binomialverteilt. Jetzt betrachten wir den Fall, daß die Stichprobe **ohne Zurücklegen** gezogen wird. Was ist die Wahrscheinlichkeit, daß X den Wert x annimmt? Dazu müssen x rote und $y = n - x$ schwarze Kugeln ausgewählt werden, was (ohne Berücksichtigung der Reihenfolge) auf $\binom{r}{x}$ bzw. $\binom{s}{y}$ Weisen möglich ist. Insgesamt gibt es $\binom{t}{n}$ Stichproben, die gesuchte Wahrscheinlichkeit ist daher

$$\binom{r}{x}\binom{s}{y}\Big/\binom{t}{n} \; .$$

Definition. *Eine Zufallsvariable X mit Werten in $\{0,\ldots,n\}$ heißt* ***hypergeometrisch verteilt*** *zum Parameter (n,r,t), $n,r \le t$, falls für alle $x \in \{0,\ldots,n\}$ gilt*

$$\mathbf{Ws}\{X = x\} \;=\; \binom{r}{x}\binom{t-r}{n-x}\Big/\binom{t}{n} \; .$$

Bei dieser Definition ist die Konvention $\binom{n}{x} = 0$ für $x > n$ oder $x < 0$ zu beachten. Eine hypergeometrisch verteilte Zufallsvariable nimmt also mit positiver Wahrscheinlichkeit nur Werte zwischen $\max(0, n-s)$ und $\min(n, r)$ an. Um auf die Beispiele des ersten Abschnitts zurückzukommen: Die Anzahl der Treffer beim Lotto und die Anzahl der Asse in einem Skatblatt sind hypergeometrisch verteilt, die Parameter sind (6,6,49) bzw. (10,4,32).

Die hypergeometrische Verteilung kommt zum Beispiel in der Qualitätskontrolle zur Anwendung. Will man die Güte einer Lieferung durch eine

Stichprobe überprüfen, so müssen sich die Beteiligten darauf einigen, wieviele fehlerhafte Stücke X die Stichprobe enthalten darf. Der Verkäufer wird darauf achten, daß eine Lieferung mit einem geringen Anteil von Ausschuß mit hoher Wahrscheinlichkeit akzeptiert wird. Der Käufer hat das Interesse, daß eine Lieferung von schlechter Qualität die Kontrolle mit nur geringer Wahrscheinlichkeit passiert. Diese unterschiedlichen Interessen werden sich nur dann vereinbaren lassen, wenn die Stichprobengröße groß genug gewählt ist. Die Wahrscheinlichkeiten werden unter der Annahme bestimmt, daß X hypergeometrisch verteilt ist.

Unsere Ableitung der hypergeometrischen Verteilung ergibt als Nebenresultat die kombinatorische Identität

$$\binom{t}{n} \;=\; \sum_{x=0}^{n} \binom{r}{x}\binom{t-r}{n-x}\;, \tag{1.4}$$

denn die Gewichte der hypergeometrischen Verteilung summieren sich zu 1 auf. Wir nutzen sie zur Berechnung des Erwartungswertes einer hypergeometrisch verteilten Zufallsvariablen X:

$$\begin{aligned}\sum_{x=0}^{n} x\binom{r}{x}\binom{t-r}{n-x} \;&=\; r\,\sum_{x=1}^{n}\binom{r-1}{x-1}\binom{t-r}{n-x}\\ &=\; r\,\binom{t-1}{n-1} \;=\; \frac{nr}{t}\,\binom{t}{n}\;,\end{aligned}$$

und damit

$$\mathbf{E}[X] \;=\; \sum_{x=0}^{n} x\cdot\mathbf{Ws}\{X=x\} \;=\; np \quad \text{mit}\;\; p=\frac{r}{t}\;.$$

Auf den Erwartungswert hat es also keinen Einfluß, ob man eine Stichprobe mit oder ohne Zurücklegen zieht, er ist in beiden Fällen gleich np. - Sind r und s groß im Vergleich zu n, so wird die Unterscheidung zwischen Stichproben mit und ohne Zurücklegen belanglos, und die hypergeometrische Verteilung nähert sich der Binomialverteilung an. Dann gilt ($y=n-x$)

$$\frac{\binom{r}{x}\binom{s}{y}}{\binom{t}{n}} \;\approx\; \frac{\frac{r^x}{x!}\cdot\frac{s^y}{y!}}{\frac{t^n}{n!}} \;=\; \binom{n}{x}p^x q^{n-x}\;,\qquad p=\frac{r}{t}\;.$$

Zum Abschluß behandeln wir eine statistische Anwendung der hypergeometrischen Verteilung.

Beispiel. Ein Maximum-Likelihood-Schätzer. Ein Fischteichbesitzer möchte seinen Fischbestand t schätzen. Er markiert dazu einige Fische. In einem späteren Fang findet er dann markierte wie unmarkierte Fische. Der Teichbesitzer überlegt: Der Anteil der markierten Fische im Fang wird vermutlich die Verhältnisse im Teich widerspiegeln. Ist also

r = Anzahl der markierten Fische,
n = Anzahl der Fische im Fang,
x = Zahl der markierten Fische im Fang,

so ist zu erwarten, daß r/t und x/n einen ähnlichen Wert haben. Dies macht $\frac{rn}{x}$ zu einem plausiblen Schätzer für t.

Zu demselben Resultat führt ein allgemeines statistisches Prinzip, das besagt:

> *Wähle als Schätzer von t diejenige ganze Zahl $\widehat{t}$, für die das beobachtete Ereignis maximale Wahrscheinlichkeit bekommt (Maximum-Likelihood-Prinzip).*

Wir machen die Annahme, daß die Anzahl X der markierten Fische in dem Fang eine hypergeometrisch verteilte Zufallsvariable ist, zum Parameter (n, r, t). Gesucht ist dasjenige t, das

$$\ell_x(t) \;=\; \binom{r}{x}\binom{t-r}{n-x}\Big/\binom{t}{n}$$

(die Statistiker sprechen von der *Likelihoodfunktion*) maximiert. Eine einfache Rechnung ergibt

$$\frac{\ell_x(t-1)}{\ell_x(t)} \;=\; \frac{t^2 - tr - tn + tx}{t^2 - tr - tn + nr}\;,$$

daher gilt $\ell_x(t-1) \le \ell_x(t)$ genau dann, wenn $xt \le nr$. $\ell_x(t)$ wächst also für kleine Werte und fällt für große Werte von t. Der Wechsel findet bei $[nr/x]$ statt, der größten ganzen Zahl kleiner als nr/x. Als **Maximum-Likelihood-Schätzer** von t erhalten wir

$$\widehat{t} \;=\; \Big[\frac{nr}{X}\Big]\;.$$

□

1.3 Besetzungszahlen

n Kugeln werden zufällig auf k Schachteln verteilt werden. Die zufälligen Besetzungszahlen der Schachteln bezeichnen wir mit $X_1, \ldots, X_k$. Wir betrachten zwei Szenarien:

A. Die Kugeln werden der Reihe nach rein zufällig und unabhängig voneinander in die Schachteln gelegt.

B. Man legt rein zufällig n Kugeln und $k-1$ Stäbchen nebeneinander. Die Kugeln zwischen dem $(i-1)$-ten und i-ten Stäbchen kommen in die i-te Schachtel ($i = 2, \ldots, k-1$), die Kugeln links vom ersten Stäbchen in die erste und die Kugeln rechts vom letzten Stäbchen in die letzte Schachtel. Ein Beispiel: Für $n = 6, k = 5$ führt die Reihe $\circ\ \circ\ ||\ \circ\ \circ\ \circ\ |\ \circ\ |$ zu den Besetzungszahlen $2, 0, 3, 1, 0$.

Gegenüber von Methode A, die natürlicher erscheint, liegt die besondere Bedeutung von Methode B darin, daß sie keine mögliche Wahl der Besetzungszahlen bevorzugt - jeder solchen Möglichkeit entspricht nämlich genau einer Anordnung der Kugeln und Stäbchen. Statistische Physiker kennen die Szenarien A und B unter den Namen **Maxwell-Boltzmann Verteilung** und **Bose-Einstein Verteilung**, sie dienen dort – zusammen mit der **Fermi-Dirac Verteilung** (jede Schachtel nimmt höchstens eine Kugel auf) – als Modelle für Gase. Die Begründung, daß sich die Verteilung von Gasteilchen nicht immer angemessen mit der Maxwell-Boltzmann Verteilung beschreiben läßt, liefert die Quantenmechanik (das Stichwort ist ‚ununterscheidbare Teilchen').

Wir betrachten nun die Zufallsvariable

$$X_1 \ = \ \text{Zahl der Kugeln in Schachtel 1 .}$$

Offenbar ist X_1 in Modell A binomialverteilt:

$$\mathbf{Ws}_A\{X_1 = x\} \ = \ \binom{n}{x} k^{-x}(1-k^{-1})^{n-x} \ .$$

Bei Modell B gibt es $\binom{n+k-1}{n}$ Möglichkeiten, die n Kugeln und $k-1$ Stäbchen aneinanderzureihen, die wir alle als gleichwahrscheinlich betrachten. Nimmt X_1 den Wert x an, so beginnt die Reihe mit x Kugeln und einem Stäbchen, und es können noch $k-2$ Stäbchen und $n-x$ Kugeln untereinander vertauscht werden. Also gilt

$$\mathbf{Ws}_B\{X_1 = x\} \ = \ \binom{n-x+k-2}{n-x} \Big/ \binom{n+k-1}{n} \ .$$

Durch Kürzen erhält man die Gleichung

$$\mathbf{Ws}_B\{X_1 = x\} \;=\; \frac{k-1}{n-x+k-1} \prod_{i=1}^{x} \frac{n+1-i}{n+k-i} \,. \tag{1.5}$$

Im Unterschied zu Modell A fallen diese Wahrscheinlichkeiten mit wachsendem x, kleine Besetzungszahlen werden bevorzugt. Instruktiv ist ein Grenzübergang $n, k \to \infty$, so daß $n/k \to \lambda > 0$. Für Modell A kommt Satz 1.1 zur Anwendung: X_1 ist approximativ $\mathrm{P}(\lambda)$-verteilt. Dagegen gilt

$$\mathbf{Ws}_B\{X_1 = x\} \;\to\; \frac{1}{\lambda+1} \Big(\frac{\lambda}{\lambda+1}\Big)^x \,.$$

Dies ist nicht schwer zu verstehen. Wir stellen uns vor, daß in Modell B die Kugeln und Stäbchen nacheinander von links nach rechts abgelegt werden. Durch das Ablegen einiger Kugeln und Stäbchen bleibt bei großem n und k der Restbestand praktisch unverändert. Stäbchen und Kugeln folgen daher anfangs einander wie bei einer unabhängigen Serie von Erfolgen und Mißerfolgen, wobei Erfolge mit Wahrscheinlichkeit $p = \frac{1}{1+\lambda}$ eintreten. Um $\{X_1 = x\}$ zu realisieren, benötigt man x Mißerfolge (Kugeln) vor dem ersten Erfolg (Stäbchen). (Ähnlich läßt sich Gleichung (1.5) verständlich machen.)

Definition. *Sei $p \in (0,1)$, $q = 1-p$. Eine Zufallsvariable X mit Werten in $\mathbb{N}_0 = \{0, 1, \ldots\}$ heißt* ***geometrisch verteilt*** *zum Parameter p, falls*

$$\mathbf{Ws}\{X = x\} \;=\; pq^x$$

für $x = 0, 1, \ldots$ gilt.

Nach der Formel für die geometrische Reihe summieren sich die Gewichte einer geometrischen Verteilung zu 1 auf. Beispielsweise ist die Anzahl der Mißerfolge vor dem ersten Erfolg eine geometrisch verteilte Zufallsvariable, wenn man ein Bernoulli-Experiment mit Erfolgswahrscheinlichkeit p unabhängig wiederholt. Den Erwartungswert bestimmen wir mit Hilfe der Formel

$$\sum_{y=0}^{\infty} x \cdot u^{x-1} \;=\; \frac{d}{du}\Big(\sum_{x=0}^{\infty} u^x\Big) \;=\; \frac{d}{du}\,\frac{1}{1-u} \;=\; \frac{1}{(1-u)^2} \,.$$

Es folgt

$$\sum_{x=0}^{\infty} x \cdot \mathbf{Ws}\{X = x\} \;=\; pq \sum_{x=0}^{\infty} x \cdot q^{x-1} \;=\; \frac{pq}{(1-q)^2} \,,$$

also

$$\mathbf{E}[X] \;=\; \frac{q}{p} \;=\; \frac{1}{p} - 1 \,.$$

Dieses Resultat entspricht der Anschauung: Beim Würfeln ($p = 1/6$) rechnet man im Durchschnitt mit 5 Würfen, bevor die erste ‚Sechs' fällt.

Bemerkung. Eine Zufallsvariable X mit Gewichten $\mathbf{Ws}\{X = x\} = pq^{x-1}$ für $x = 1, 2, \ldots$ wird ebenfalls als **geometrisch verteilte Zufallsvariable** bezeichnet. Sie lässt sich auffassen als die Anzahl von Versuchen bis zum ersten Erfolg, wenn man ein Bernoulli-Experiment mit Erfolgswahrscheinlichkeit p unabhängig wiederholt. Nun ist der Erwartungwert $\mathbf{E}[X] = \sum_{x=1}^{\infty} xpq^{x-1} = 1/p$. □

Zurück zu den Modellen A und B: Wir betrachten nun den gesamten Zufallsvektor $X = (X_1, \ldots, X_k)$ der Besetzungszahlen aller Schachteln. Die Besonderheit von Modell B ist, daß keine der möglichen Konstellationen bevorzugt ist, d.h. X ist uniform in

$$S \;=\; \{(x_1, \ldots, x_k) \;:\; x_i \in \mathbb{N}_0 \;,\; x_1 + \cdots + x_k = n\}$$

verteilt. Dies trifft auf Modell A nicht zu, denn dort läßt sich das Ereignis $\{X = (x_1, \ldots, x_k)\}$ auf $\binom{n}{x_1,\ldots,x_k}$ Weisen realisieren, so daß

$$\mathbf{Ws}_A\{X = (x_1, \ldots, x_k)\} \;=\; \binom{n}{x_1, \ldots, x_k} k^{-n}$$

gilt. Nimmt man für Modell A allgemeiner an, daß die Kugeln jeweils mit Wahrscheinlichkeit p_i in die i-te Schachtel gelegt werden, so ist $(X_1, \ldots, X_k)$ multinomialverteilt im Sinne der folgenden Definition.

Definition. *Sei $n \in \mathbb{N}$ und $p_1, \ldots, p_k \geq 0$, so daß $\sum_i p_i = 1$. Dann heißt ein Zufallsvariable X mit Werten in*

$$S \;=\; \{(x_1, \ldots, x_k) : x_i \in \mathbb{N}_0,\; x_1 + \cdots + x_k = n\}$$

multinomialverteilt *zum Parameter $(n, p_1, \ldots, p_k)$, falls*

$$\mathbf{Ws}\{X = (x_1, \ldots, x_k)\} \;=\; \binom{n}{x_1, \ldots, x_k} p_1^{x_1} \cdots p_k^{x_k}$$

gilt für alle $(x_1, \ldots, x_k) \in S$.

1.4 Zufallsvariable mit Dichten, Normalapproximation der Binomialverteilung

Neben den Verteilungen, die durch Gewichte gegeben sind, spielen Verteilungen mit Dichten eine besondere Rolle. Sei $S \subset \mathbb{R}$ ein (endliches oder unendliches) Intervall mit den Endpunkten $l < r$ und sei $p : S \to \mathbb{R}$ eine stetige (oder allgemeiner integrierbare) nicht-negative Funktion, so daß

$$\int_l^r p(x)\,dx \;=\; 1\ .$$

Wir nennen dann p eine Dichtefunktion.

Sei weiter X Zufallsvariable mit Werten in S und einer Verteilung von der Gestalt

$$\mathbf{Ws}\{a \le X \le b\} \;=\; \int_a^b p(x)\,dx\ ,$$

für alle $l \le a \le b \le r$. Man sagt dann, daß X die **Dichte** $p(x)\,dx$ besitzt, und schreibt kurz

$$\mathbf{Ws}\{X \in dx\} \;=\; p(x)\,dx\ .$$

Eine Möglichkeit, zu Verteilungen mit Dichten zu gelangen, ist durch Grenzübergang aus Verteilungen mit Gewichten. Ein einfaches Beispiel bieten geometrische Verteilungen. Sei $\lambda > 0$ und sei X_n geometrisch verteilt mit Erfolgswahrscheinlichkeit $p_n = \lambda/n$. Sie hat den Erwartungswert n/λ, deswegen ist es plausibel, die Verteilungen von X_n/n zu betrachten. Wir können $\{X_n/n > a\}$ als das Ereignis auffassen, daß in einer Serie von unabhängigen Bernoulliexperimenten die Anzahl der Misserfolge vor dem ersten Erfolg an übertrifft. Daher folgt für $a \ge 0$ mit $n \to \infty$

$$\mathbf{Ws}\{X_n/n > a\} \;=\; q_n^{[an]+1} \;=\; \Bigl(1 - \frac{\lambda}{n}\Bigr)^{an+O(1)} \;\to\; e^{-\lambda a}$$

und folglich $\mathbf{Ws}\bigl\{a < \frac{X_n}{n} \le b\bigr\} \;\to\; e^{-\lambda a} - e^{-\lambda b}$ für $0 \le a < b$, oder auch

$$\mathbf{Ws}\Bigl\{a \le \frac{X_n}{n} \le b\Bigr\} \;\to\; e^{-\lambda a} - e^{-\lambda b} \;=\; \int_a^b \lambda e^{-\lambda x}\,dx\ .$$

Dies ist für uns Anlass, eine besonders wichtige Verteilung einzuführen.

Definition. *Sei $\lambda > 0$. Eine Zufallsvariable X mit Werten in $S = [0, \infty)$ heißt* **exponential verteilt zum Parameter $\boldsymbol{\lambda}$**, *falls*

$$\mathbf{Ws}\{a \le X \le b\} \;=\; \int_a^b \lambda e^{-\lambda x}\, dx$$

für all $0 \le a < b$, falls sie also die Dichte

$$\mathbf{Ws}\{X \in dx\} \;=\; \lambda e^{-\lambda x}\, dx \quad , \quad x \ge 0$$

besitzt.

Also sind geometrisch verteilte Zufallsvariable - passend normiert - im Grenzwert exponential verteilt. Wir werden auf exponential verteilte Zufallsvariable zurückkommen.

Das Konzept von Verteilungen mit Dichten benutzen wir nun, um uns ein genaueres Bild der Verteilung einer $\mathrm{B}(n,p)$-verteilten Zufallsvariablen X bei wachsendem n zu verschaffen. Bleibt ihr Erwartungswert

$$\mathbf{E}[X] \;=\; np$$

dabei beschränkt, so ist die Verteilung im wesentlichen durch den Erwartungswert bestimmt. Dies ist Konsequenz der Poisson-Approximation aus Satz 1.1. Andernfalls kommt die zweite wichtige Kenngröße einer Binomialverteilung ins Spiel, ihre **Varianz**

$$\mathbf{Var}[X] \;=\; npq$$

(ausführlich behandeln wir Varianzen in Abschnitt 3.2). Ist die Varianz ausreichend groß, so ist X in erster Näherung symmetrisch um $\mathbf{E}[X]$ verteilt, und man muß mit Abweichungen rechnen, die typischerweise von der Größe $\sqrt{\mathbf{Var}[X]}$, der sogenannten **Standardabweichung** von X, sind.

Um dies zu zeigen, leiten wir im Folgenden für die Wahrscheinlichkeit

$$\mathbf{Ws}\{\alpha \le X \le \beta\} \;=\; \sum_{\alpha \le x \le \beta} \binom{n}{x} p^x q^{n-x}\ ,$$

daß eine binomialverteilte Zufallsvariable X einen Wert zwischen α und β annimmt, eine Approximationsformel ab, die sich auch für praktische Zwecke als nützlich erweist. Erneut arbeiten wir mit der **Stirlingschen Formel**, und zwar in folgender Form, wie man sie in Lehrbüchern der Analysis findet.

Satz 1.2. *Für $n \to \infty$ gilt*

$$n! \;=\; \sqrt{2\pi n}\; n^n e^{-n+o(1)} \;.$$

Dabei bezeichnet $o(1)$ wie üblich eine Folge, die mit $n \to \infty$ gegen 0 konvergiert. Mit Hilfe dieser Formel erhalten wir für die Gewichte der Binomialverteilung die für $x \to \infty$ und $n - x \to \infty$ gültige Asymptotik

$$\frac{n!}{x!(n-x)!}\; p^x q^{n-x} \;=\; \sqrt{\frac{n}{2\pi x(n-x)}}\; \Big(\frac{pn}{x}\Big)^x \Big(\frac{qn}{n-x}\Big)^{n-x} \exp(o(1)) \;.$$

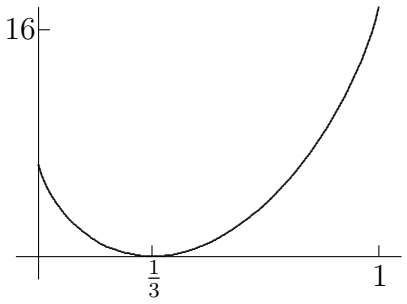
Wir formen diese Gleichung (ähnlich wie schon in anderen Fällen) um zu

$$\frac{n!}{x!(n-x)!}\; p^x q^{n-x} \;=\; \sqrt{\frac{n}{2\pi x(n-x)}}\; \exp\Big(- nh\Big(\frac{x}{n}\Big) + o(1)\Big) \;, \qquad (1.6)$$

mit

$$h(t) \;:=\; t \ln \frac{t}{p} + (1-t) \ln \frac{1-t}{q} \;, \quad 0 \le t \le 1 \;. \qquad (1.7)$$

Es gilt $h(p) = h'(p) = 0$. Außerdem ist h wegen $h''(t) = (t(1-t))^{-1} > 0$ eine strikt konvexe Funktion. Es folgt $h(t) > 0$ für alle $t \neq p$. $nh(t)$ heißt die *Entropiefunktion der Binomialverteilung*. Das folgende Bild zeigt die Entropiefunktion der $\mathrm{B}\big(16, \frac{1}{3}\big)$-Verteilung



Aus (1.6) erkennt man, daß die Binomialgewichte exponentiell klein sind, wenn die relative Häufigkeit x/n der Erfolge deutlich von p abweicht, die Hauptmasse der Binomialverteilung konzentriert sich daher um np herum. Wir werden deswegen (1.6) zu vorgegebenem $c > 0$ in dem Bereich

$$|x - np| \;\le\; c\sqrt{npq} \qquad (1.8)$$

weiter analysieren. Wir approximieren $h(t)$ um p herum nach Art einer Taylor-Näherung durch eine Parabel. Sie ist wegen $h(p) = h'(p) = 0$ und $h''(p) = (pq)^{-1}$ durch $(2pq)^{-1}(t-p)^2$ gegeben. Es folgt

$$n\, h\big(\tfrac{x}{n}\big) \;=\; \frac{n}{2pq}\, \Big(\frac{x}{n} - p\Big)^2 + r(x,n,p) \;=\; \frac{(x-np)^2}{2npq} + r(x,n,p) \qquad (1.9)$$

mit einem Approximationsfehler $r(x,n,p)$, der sich nach der Taylor-Formel als

$$r(x,n,p) \;=\; n\,\frac{h'''(\xi)}{6}\left(\frac{x}{n}-p\right)^3$$

ergibt, wobei ξ zwischen x/n und p liegt.

Diese Formel werten wir unter der Annahme $npq \geq 4c^2$ weiter aus. Nach (1.8) gilt dann $x/n \geq p - c(pq)^{1/2}n^{-1/2} \geq p - cp(npq)^{-1/2} \geq p/2$ und analog $1 - x/n \geq q/2$, und damit $\xi \geq p/2$, $1-\xi \geq q/2$. Eine kurze Rechnung zeigt $|h'''(\xi)| \leq \xi^{-2}(1-\xi)^{-2}$, und es folgt $|h'''(\xi)| \leq 16(pq)^{-2}$ und mit (1.8)

$$|r(x,n,p)| \;\leq\; \frac{16|x-np|^3}{6(npq)^2} \;\leq\; \frac{3c^3}{\sqrt{npq}}\;. \tag{1.10}$$

Unser Resultat nimmt eine besonders übersichtliche Gestalt im Grenzübergang $n \to \infty$ an. Es genügt $npq \to \infty$, deswegen darf p mit n variieren, solange nur die Varianz npq gegen ∞ strebt. Aus (1.6) erhalten wir dann unter Beachtung von (1.8) - (1.10) insgesamt die asymptotische Darstellung

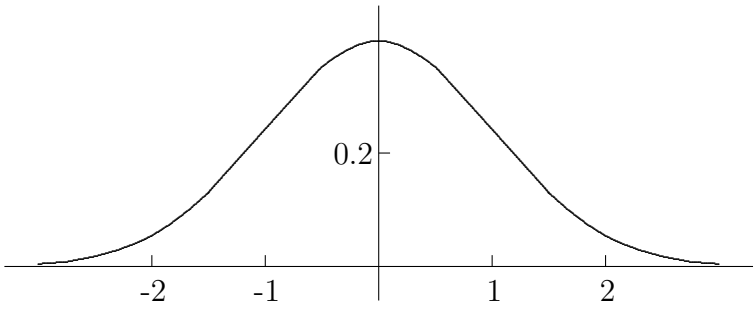
$$\binom{n}{x}p^x q^{n-x} \;=\; \sqrt{\frac{1}{2\pi npq}}\;\exp\Big(-\frac{(x-np)^2}{2npq}+o(1)\Big)\;, \tag{1.11}$$

den **lokalen Grenzwertsatz** für die Binomialverteilung. Der Approximationsfehler ist in dem Term $o(1)$ erfaßt, er geht gegen 0, und zwar, wie unsere Rechnung zeigt, gleichmäßig für alle x, die zu vorgegebenem $c > 0$ der Bedingung (1.8) genügen.

Mit dieser Formel passen wir nun die Binomialgewichte an die **Gaußsche Glockenkurve**

$$\mathsf{n}(z) := \;\frac{1}{\sqrt{2\pi}}\;e^{-z^2/2}\;,\quad z \in \mathbb{R}\;,$$

an. Ihr Graph sieht so aus:



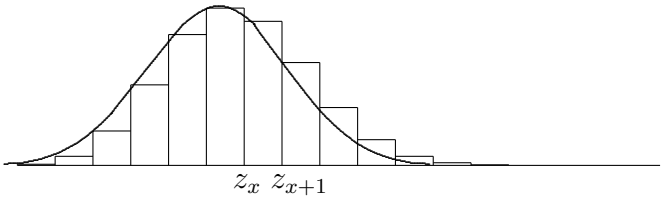
Unter Benutzung der Notation

$$z_x \;:=\; \frac{x-np}{(npq)^{1/2}}$$

erhalten wir aus (1.11) die Formel

$$\binom{n}{x} p^x q^{n-x} \;=\; \mathsf{n}(z_x)(z_{x+1}-z_x)\exp(o(1))\;. \tag{1.12}$$

Stellen wir die Gewichte der Binomialverteilung als Flächen von Rechtecken dar, so ergibt sich folgendes Bild.



Normalapproximation der $\mathrm{B}\big(16, \frac{1}{3}\big)$-Verteilung

Insgesamt gelangen wir zu einem klassischen Resultat der Stochastik, dem **Satz von de Moivre-Laplace**.

Satz 1.3. *Sei $X_1, X_2, \ldots$ eine Folge binomialverteilter Zufallsvariabler mit $\mathbf{Var}[X_n] \to \infty$ für $n \to \infty$. Dann gilt für die normierten Zufallsvariablen*

$$X_n^* \;:=\; \frac{X_n - \mathbf{E}[X_n]}{\sqrt{\mathbf{Var}[X_n]}} \;=\; \frac{X_n - np}{\sqrt{npq}}$$

und für alle reellen Zahlen $a \le b$

$$\lim_{n\to\infty} \mathbf{Ws}\{a \le X_n^* \le b\} \;=\; \int_a^b \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2}\, dz\;.$$

Beweis. Aus $a \le z_x \le b$ folgt $|x-np| \le c\sqrt{npq}$ mit $c = \max(|a|,|b|)$. Für diese x konvergieren die Fehlerterme in (1.12) gleichmäßig gegen 0, daher gilt

$$\sum_{a\le z_x\le b} \binom{n}{x} p^x q^{n-x} \;=\; \Big(\sum_{a\le z_x\le b} \mathsf{n}(z_x)(z_{x+1}-z_x)\Big)\;\exp(o(1))\;,$$

(mit einer einzigen Nullfolge $o(1)$). Die rechte Summe fassen wir als Approximation eines Integrals mit Hilfe von Treppenfunktionen auf. Da $z_{x+1} - z_x$ mit $n \to \infty$ gegen 0 geht, konvergiert der Ausdruck wie behauptet gegen $\int_a^b \mathsf{n}(z)\,dz$. □

Der Satz von de Moivre-Laplace führt uns zu folgender Sprechweise.

Definition. *Eine reellwertige Zufallsvariable Z heißt* ***standard normalverteilt****, kurz* ***N(0,1)-verteilt****, falls*

$$\mathbf{Ws}\{a \le Z \le b\} \;=\; \frac{1}{\sqrt{2\pi}} \int_a^b \exp\Big(-\frac{z^2}{2}\Big)\,dz$$

für alle $-\infty \le a \le b \le \infty$ gilt, falls sie also die Dichte

$$\mathbf{Ws}\{Z \in dz\} \;=\; \mathsf{n}(z)\,dz$$

hat.

Daß es sich bei $\mathsf{n}(z)\,dz$ um eine Dichte handelt, ist aus der Gleichung

$$\int_{-\infty}^{\infty} e^{-z^2/2}\,dz \;=\; \sqrt{2\pi}$$

ersichtlich, die in der Analysis bewiesen wird (vgl. auch den Abschnitt über Dichten in Kapitel 2). Man spricht von der **Dichte der Standardnormalverteilung**.

In Anlehnung an diese Sprechweise besagt der Satz von de Moivre-Laplace, daß X_n^* asymptotisch standard normalverteilt ist. Anders als bei der Poissonapproximation kann man X_n^* im Grenzwert nicht mehr als eine Zufallsvariable betrachten, die nur abzählbar viele Werte annimmt, die Grenzverteilung ist nicht durch Gewichte gegeben, sondern eben durch eine Dichte.

Die Normalapproximation ist auch für das explizite Berechnen von Wahrscheinlichkeiten nützlich. Zu diesem Zweck empfiehlt es sich, folgende Integralnäherung zu benutzen (die ‚Tangentenregel' der Numerischen Mathematik):

$$\begin{aligned}\binom{n}{x} p^x q^{n-x} \;&\approx\; \mathsf{n}(z_x)(z_{x+1} - z_x) \\ &=\; \mathsf{n}\Big(\frac{t_{x+1} + t_x}{2}\Big)(t_{x+1} - t_x) \;\approx\; \int_{t_x}^{t_{x+1}} \mathsf{n}(z)\,dz\;,\end{aligned}$$

mit

$$t_x \;:=\; \frac{x - np - \frac{1}{2}}{(npq)^{1/2}} \;.$$

Für $\mathrm{B}(n,p)$-verteiltes X und ganzzahliges $\alpha \le \beta$ führt dies zu der Approximationsformel

$$\mathbf{Ws}\{\alpha \le X \le \beta\} \;\approx\; \int_{t_\alpha}^{t_{\beta+1}} \mathsf{n}(z)\,dz \;.$$

Man kann mit einer brauchbaren Näherung rechnen, falls $t_{x+1} - t_x = (npq)^{-1/2}$ genügend klein ist. ($npq > 9$ ist eine Faustregel, die sich in Lehrbüchern der Stochastik findet.)

Da sich $\mathsf{n}(z)$ nicht in elementarer Weise integrieren läßt, hat man

$$\Phi(x) \;:=\; \int_{-\infty}^{x} \mathsf{n}(z)\,dz$$

tabelliert. $\Phi(x)$ heißt **Gaußsches Fehlerintegral** oder **Verteilungsfunktion der Standardnormalverteilung**. Unsere Näherung lautet damit für ganzzahliges $\alpha \le \beta$

$$\mathbf{Ws}\{\alpha \le X \le \beta\} \;\approx\; \Phi\Big(\frac{\beta - np + \frac{1}{2}}{\sqrt{npq}}\Big) - \Phi\Big(\frac{\alpha - np - \frac{1}{2}}{\sqrt{npq}}\Big) \;.$$

Einige häufig benutzte Werte von Φ sind

x	0	1	$1,28$	$1,64$	$1,96$	$2,33$
$\Phi(x)$	$0,5$	$0,84$	$0,9$	$0,95$	$0,975$	$0,99$

Für negatives x beachte man die Formel

$$\Phi(x) \;=\; 1 - \Phi(-x) \;,$$

die aus der Symmetrie von $\mathsf{n}(z)$ um Null folgt.

Bemerkung. Ist Z standard normalverteilt, so hat $X = \mu + \sigma Z$ mit reellen Zahlen μ und $\sigma \neq 0$ die Dichte

$$\mathbf{Ws}\{X \in dx\} \;=\; \mathsf{n}_{\mu,\sigma^2}(x)\,dx$$

mit

$$\mathsf{n}_{\mu,\sigma^2}(x) \;:=\; \frac{1}{\sqrt{2\pi\sigma^2}}\;\exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big) \;.$$

Man erhält nämlich im Fall $\sigma > 0$ (der andere Fall ist analog) für $a < b$ durch Substitution $z = (x-\mu)/\sigma$ (also $dz = dx/\sigma$)

$$\begin{aligned}\mathbf{Ws}\{a \le X \le b\} &= \mathbf{Ws}\Big\{\frac{a-\mu}{\sigma} \le Z \le \frac{b-\mu}{\sigma}\Big\} \\ &= \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz \;=\; \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}\, dx\ .\end{aligned}$$

X heißt dann **normalverteilt mit Erwartungswert $\boldsymbol{\mu}$ und Varianz $\boldsymbol{\sigma^2}$** bzw. $N(\mu,\sigma^2)$-verteilt. $\mathsf{n}_{\mu,\sigma^2}(x)\,dx$ heißt **Dichte der $\boldsymbol{N(\mu,\sigma^2)}$-Verteilung**.

Wir können nun den Satz von de Moivre-Laplace kurz wie folgt ausdrücken: Eine $B(n,p)$-verteilte Zufallsvariable ist bei ausreichend großer Varianz approximativ $N(np,npq)$-verteilt. □

Anwendungen der Normalapproximation

1. Jemand hat „rein zufällig“ ein 01-Folge der Länge 101 per Hand fabriziert. Die Zahl der Nullen und Einsen erscheint wohlaustariert, allerdings fällt die große Zahl der Runs auf: Die Folge enthält insgesamt 62 Runs aus Nullen bzw. Einsen. Wir wissen, daß in einer rein zufälligen Folge die um 1 verminderte Anzahl Y von Runs $\mathrm{B}(100,\frac{1}{2})$-verteilt ist. Die Wahrscheinlichkeit, eine mindestens so große Zahl von Runs vorzufinden, ist also

$$\mathbf{Ws}\{Y \ge 61\} \;\approx\; 1 - \Phi\Big(\frac{61 - np - \frac{1}{2}}{\sqrt{npq}}\Big) \;=\; 1 - \Phi(2,1) \;\simeq\; 0,018\ .$$

Dieser auffällig kleine Wert ist ein deutlicher Hinweis darauf, daß es sich nicht um eine rein zufällig generierte 01-Folge handelt. - Bemerkung: Die übliche Methode, um eine 01-Folge auf Zufälligkeit zu testen, benutzt nicht die Gesamtanzahl der Runs, sondern die Tatsache, daß die Länge eines Runs geometrisch verteilt ist zum Parameter $p = 1/2$.

2. Ein Konfidenzintervall für die Binomialverteilung. Zwischen 1871 und 1900 wurden in der Schweiz bei $n = 2.644.757$ Geburten $x = 1.285.086$ Mädchen zur Welt gebracht. Was läßt sich für die Wahrscheinlichkeit p folgern, daß (zur damaligen Zeit) ein neugeborenes Kind ein Mädchen ist? Naheliegend ist es, p als $\widehat{p}(x) := x/n = 0,4858$ zu schätzen, wie gut ist aber eine solche Schätzung?

Informativer ist es, ein **Konfidenzintervall** zu konstruieren. Im vorliegenden Fall bedeutet dies das Folgende: Man bestimme ein (zufälliges) Intervall so, daß es das zu schätzende p mit großer Wahrscheinlichkeit überdeckt, was der Wert von p auch immer sein mag. Wir nehmen dazu an, daß

die Anzahl X der Mädchengeburten eine binomialverteilte Zufallsvariable zum Parameter (n, p) ist. Die Aufgabe besteht darin, Zahlen $p_u(x) \le p_o(x)$, $x = 0, 1, \ldots, n$, zu finden, so daß bei beliebigem p

$$\mathbf{Ws}\{p_u(X) \le p \le p_o(X)\} \;\ge\; 1 - \alpha$$

gilt. α ist die **Irrtumswahrscheinlichkeit** (das Sicherheitsniveau) des Verfahrens, üblicherweise wird $\alpha = 0,05$ oder $\alpha = 0,01$ gewählt.

Diese Aufgabe läßt sich auf verschiedene Weise lösen. Wir begnügen uns mit einer approximativen Methode: Wähle $a^* > 0$ so, daß für eine $\mathrm{N}(0,1)$-verteilte Zufallsvariable Z

$$\mathbf{Ws}\{-a^* \le Z \le a^*\} \;=\; 1 - \alpha \;.$$

Nach dem Satz von de Moivre-Laplace gilt dann

$$\begin{aligned}\mathbf{Ws}\Big\{\frac{X}{n} - a^*\sqrt{\frac{pq}{n}} \le p \le \frac{X}{n} + a^*\sqrt{\frac{pq}{n}}\Big\} \\ = \;\mathbf{Ws}\Big\{-a^* \le \frac{X - np}{\sqrt{npq}} \le a^*\Big\} \;\approx\; 1 - \alpha \;.\end{aligned}$$

Berücksichtigt man noch die Ungleichung $pq \le \frac{1}{4}$, so erkennt man, daß das Konfidenzintervall

$$p_u(X) \;:=\; \widehat{p}(X) - \frac{a^*}{2\sqrt{n}} \;, \quad p_o(X) \;:=\; \widehat{p}(X) + \frac{a^*}{2\sqrt{n}} \tag{1.13}$$

für jedes p asymptotisch das Niveau α einhält. - Für die Wahrscheinlichkeit einer Mädchengeburt ergibt sich bei der Wahl $\alpha = 0,05$, also $a^* = 1,96$, das Intervall $0,4858 \;\pm\; 0,0006$.

Man bemerke: Natürlich ist nicht gewährleistet, daß das zu schätzende p in diesem Intervall liegt, wie sollte das auch in einer mit Unsicherheit behafteten Situation möglich sein. Der Statistiker kann nur garantieren, daß für dieses statistische Verfahren die Aussage „p liegt im Konfidenzintervall“ auf lange Sicht in 95% aller Anwendungsfälle korrekt ist. Die Schreibweise $\mathbf{Ws}\{0,4852 \le p \le 0,4864\} \gtrsim 0,95$ ist falsch und irreführend. Sie suggeriert, daß p zufällig ist, und nicht das Konfidenzintervall.

1.5 Kartenmischen*

Wir beenden das Kapitel mit einem aufwendigeren Beispiel einer stochastischen Modellierung, das auf kombinatorisch anspruchsvollere Fragestellungen führt. Da es sich um ein Spezialthema handelt, kann der Abschnitt überschlagen werden.

Wie lange muß man ein Blatt von b Spielkarten mischen, damit es ordentlich durchgemischt ist? Dies hängt davon ab, wie man mischt - wir denken an die professionelle Technik, bei der man das Blatt erst abhebt und anschließend die beiden Teilstapel ineinanderblättert. Wie oft sollte man diesen Vorgang wiederholen? Zur Beantwortung dieser Frage stützen wir uns auf ein Modell, das auf Gilbert, Shannon (1955) und Reeds (1981) zurückgeht. Wir werden sehen: Für ein Bridgespiel ($b = 52$) ist *siebenmaliges Mischen* ausreichend.

Mischen bedeutet, daß man die Karten untereinander vertauscht, sie mehrfach zufällig permutiert. Um den Vorgang mathematisch zu beschreiben, identifizieren wir das Blatt mit der Menge $B := \{1, 2, \ldots, b\}$; die 1 steht für die Karte oben auf dem Stapel und b für die Karte ganz unten. Einmaliges Mischen entspricht dann einer zufälligen Permutation Π von B, einem zufälligen Element der Menge

$$S \;:=\; \{\pi : B \to B \;:\; \pi \text{ ist eine Bijektion}\}\;,$$

und mehrfaches Mischen einer Hintereinanderausführung

$$X_n \;:=\; \Pi_n \circ \Pi_{n-1} \circ \cdots \circ \Pi_1$$

von mehreren zufälligen Permutationen Π_1, $\Pi_2, \ldots$ Wir nehmen an, daß sie voneinander unabhängig generiert werden und die gleiche Verteilung besitzen.

Für konkrete Rechnungen müssen wir die Verteilung der Zufallsvariablen $\Pi = \Pi_1$ festlegen. Dazu beschreiben wir den Vorgang des einmaligen Mischens im Detail. Er erfolgt in zwei Schritten. Erst wird das Blatt abgehoben, d.h. in ein oberes Päckchen $P_o = \{1, \ldots, K\}$ und ein unteres $P_u = \{K+1, \ldots, b\}$ vom Umfang K und $b - K$ geteilt. Anschließend werden die beiden Päckchen ineinander geblättert. Dies ändert die Reihenfolge des Blattes, nicht jedoch die Reihenfolge innerhalb P_o und P_u. Für die resultierende Permutation Π bedeutet dies, daß die Karten aus P_o in die Positionen $\Pi(1) < \Pi(2) < \cdots < \Pi(K)$ und die aus P_u in die Positionen $\Pi(K+1) < \Pi(K+2) < \cdots < \Pi(b)$ wandern.

Insgesamt beschreiben wir den Vorgang durch die zufällige Größe $R = (K, \Pi)$ mit Werten in

$$S_2 \;:=\; \{(k, \pi) \;:\; k = 0, 1, \ldots, b\;,\; \pi \in S \text{ wächst monoton auf } p_o \text{ und } p_u\}\;,$$

mit $p_o := \{1, \ldots, k\}$ und $p_u := \{k+1, \ldots, b\}$. Um einen kurzen Begriff zu haben, sprechen wir wie im Englischen von einem **Riffle-Shuffle**. Die Elemente $r = (k, \pi)$ von S_2 nennen wir deterministische Riffle-Shuffle.

Es bleibt, die Verteilung eines Riffle-Shuffle R festzulegen. Das Modell von Gilbert-Shannon-Reed macht dazu die folgenden Annahmen:

1. Die Wahrscheinlichkeit, daß das Päckchen P_o genau k Karten enthält, ist $\binom{b}{k}2^{-b}$. Mit anderen Worten: Die Schnittstelle K im Blatt ist binomialverteilt zum Parameter $(b, 1/2)$.

2. Hat K den Wert k, so wird aus den $\binom{b}{k}$ Möglichkeiten, P_o und P_u ineinanderzublättern, eine rein zufällig ausgewählt, jeweils mit Wahrscheinlichkeit $\binom{b}{k}^{-1}$.

Insgesamt enthält S_2

$$\sum_{k=0}^{b} \binom{b}{k} \;=\; 2^b$$

Elemente, und jedes wird von R mit derselben Wahrscheinlichkeit 2^{-b} angenommen. Mit anderen Worten: R ist uniform auf S_2 verteilt.

Die Permutation Π besteht normalerweise aus 2 aufsteigenden Sequenzen $\Pi(1) < \cdots < \Pi(K) > \Pi(K+1) < \cdots < \Pi(b)$, aus denen sich der gesamte Shuffle (K, Π) leicht rekonstruieren läßt. Jedoch ist hier folgendes zu beachten: Wir haben nicht ausgeschlossen, daß ein Päckchen leer bleibt, d.h. K den Wert 0 oder b hat, oder daß im zweiten Schritt das Päckchen P_o wieder ganz auf P_u zurückgelegt wird. In diesen Fällen ändert sich die Reihenfolge im Blatt nicht, so dass der Wert von Π die identische Permutation id ist und Π den Shuffle nicht mehr eindeutig festlegt. Dies ist nicht besonders störend: Es handelt sich um $b+1$ verschiedene deterministische Riffle-Shuffle (k, π) mit $\pi = id$, und das Ereignis $\{\Pi = id\}$ hat damit die verschwindend kleine Wahrscheinlichkeit $(b+1)2^{-b}$.

Der Vorteil des Modells von Gilbert-Shannon-Reed ist, daß es explizite Rechnungen erlaubt. Insbesondere lassen sich die n-Schritt-Übergangswahrscheinlichkeiten

$$p_\pi^n \;:=\; \mathbf{Ws}\{X_n = \pi\}\;, \quad \pi \in S$$

explizit berechnen, wobei sich nun $X_n = \Pi_n \circ \cdots \circ \Pi_1$ aus den Permutationen von n unabhängigen Riffle-Shuffle zusammensetzt. Dazu vereinbaren wir folgende Sprechweise. Wir sagen, eine Permutation π zerfällt in s *wachsende Sequenzen*, wenn es natürliche Zahlen $0 = r_0 < r_1 < \ldots < r_s = b$ gibt mit $\pi(r_{i-1}+1) < \pi(r_{i-1}+2) < \cdots < \pi(r_i)$ für alle $i = 1, \ldots, s$ und $\pi(r_i) > \pi(r_i+1)$ für alle $i = 1, \ldots, s-1$. (Beispielsweise zerfällt die Permutation 461352 in die drei wachsenden Sequenzen 46, 135 und 2.)

Behauptung. *Sei π eine Permutation, die in s wachsende Sequenzen zerfällt. Dann gilt*

$$p_\pi^n \;=\; \binom{b+2^n-s}{b} 2^{-nb} \;=\; \frac{1}{b!}\,\prod_{i=1}^{b}\Big(1+\frac{i-s}{2^n}\Big)\;. \tag{1.14}$$

(*Für $s > 2^n$ ist diese Wahrscheinlichkeit gleich 0 zu setzen.*)

Bevor wir die Formel beweisen, wollen wir einige Folgerungen ziehen. Wegen $1 \le s \le b$ folgt unmittelbar

$$\frac{1}{b!}\,\prod_{i=1}^{b-1}\Big(1-\frac{i}{2^n}\Big) \;\le\; p_\pi^n \;\le\; \frac{1}{b!}\,\prod_{i=1}^{b-1}\Big(1+\frac{i}{2^n}\Big)\;. \tag{1.15}$$

Die Produkte auf der rechten und linken Seite konvergieren mit $n \to \infty$ exponentiell schnell gegen 1, und damit p_π^n gegen $1/b!$, die Gewichte der uniformen Verteilung auf S. Auf lange Sicht ist damit das Mischen erfolgreich. Um die Abweichung von der Gleichverteilung genauer zu quantifizieren, betrachtet man die **Totalvariation** zwischen der Verteilung von X_n und der uniformen Verteilung auf S, gegeben durch den Ausdruck

$$v_n \;:=\; \frac{1}{2}\sum_{\pi\in S}\Big|p_\pi^n-\frac{1}{b!}\Big|\;.$$

Bemerkung. Sind p_x und q_x die Gewichte zweier Wahrscheinlichkeitsverteilungen μ und ν auf einer abzählbaren Menge S, so definiert man die Totalvariation zwischen μ und ν als

$$d(\mu,\nu) \;:=\; \frac{1}{2}\sum_{x\in S}|p_x-q_x|\;.$$

Der Faktor $1/2$ dient dazu, $d(\mu,\nu)$ auf Werte zwischen 0 und 1 zu normieren: $d(\mu,\nu) \le \frac{1}{2}\sum_x(p_x+q_x) = 1$. Eine alternative Formel für die Totalvariation ist (Übung)

$$d(\mu,\nu) \;=\; \sup_{B\subset S}|\mu(B)-\nu(B)|\;.$$

□

Mit Formel (1.14) folgt

$$v_n \;=\; \frac{1}{2}\sum_{s=1}^{b}\frac{a(b,s)}{b!}\Big|\prod_{i=1}^{b}\Big(1+\frac{i-s}{2^n}\Big)-1\Big|\;.$$

Dabei bezeichnet $a(b,s)$ die Anzahl der Permutationen der Länge b, die in genau s wachsende Sequenzen zerfallen. Die Zahlen $a(b,s)$ heißen *Euler-Zahlen*, man kann sie rekursiv aus den Gleichungen

$$a(b,s) \;=\; s^b - \sum_{t=1}^{s-1} \binom{b+s-t}{b} a(b,t)\;, \qquad a(b,1) \;=\; 1 \tag{1.16}$$

berechnen, wie wir später zeigen werden.

Mit diesen Gleichungen läßt sich v_n per Computer leicht berechnen. Für ein Bridgespiel mit 52 Karten erhält man

n	≤ 4	5	6	7	8	9
v_n	1,000	0,924	0,614	0,334	0,167	0,085

Die Tabelle liefert die Begründung für die Behauptung, daß 7-maliges Mischen genügt. Daß sich bis $n = 4$ kein Mischeffekt bemerkbar macht, ist nicht überraschend: Für $b = 52$ gibt es 2^{52} verschiedene deterministische Riffle-Shuffle, so daß man mit 4-maligem Mischen höchstens $2^{4\cdot 52} \simeq 4 \cdot 10^{62}$ verschiedene Permutationen erreichen kann - andererseits gibt es insgesamt $52! \simeq 8 \cdot 10^{67}$ Permutationen. Um so auffälliger ist, wie rapide ab $n = 5$ der Effekt des Mischens zur Geltung kommt.

Es bleibt der *Beweis* der Formeln (1.14) und (1.16). Dazu müssen wir die Komposition von Permutationen unabhängiger Riffle-Shuffle behandeln, deswegen betrachten wir nun allgemeiner α-Shuffle, mit $\alpha \in \mathbb{N}$. Bei einem *deterministischen α-Shuffle* ist das Blatt in α Päckchen

$$\begin{aligned} p_1 &= \{1,\dots,k_1\}\;, \\ p_2 &= \{k_1+1,\dots,k_1+k_2\}\;, \\ &\;\;\vdots \\ p_\alpha &= \{k_1+\cdots+k_{\alpha-1}+1,\dots,k_1+\cdots+k_\alpha\} \end{aligned}$$

zerlegt. Wir nehmen $k_1,\dots,k_\alpha \ge 0$, $k_1+\cdots+k_\alpha = b$ an, einzelne p_i dürfen also leer sein. Die Stapel werden dann ineinander geschoben, dabei ist einzig zu gewährleisten, daß die Anordnung in den Teilstapeln beim Mischen nicht verloren geht. Wie ein Riffle-Shuffle (also einem 2-Shuffle in neuer Terminologie) induziert damit ein α-Shuffle eine Permutation π, die auf den Abschnitten p_i monoton wächst. Erneut wird ein α-Shuffle durch seine Permutation π nicht vollständig festgelegt, aus π lassen sich im allgemeinen die Teilstapel p_i

nicht vollständig rekonstruieren. Die Menge aller deterministischen α-Shuffle $m = (k_1, \ldots, k_\alpha, \pi)$ bezeichnen wir mit S_α. Sie enthält

$$\sum_{k_1+\cdots+k_\alpha=b} \binom{b}{k_1, \ldots, k_\alpha} \;=\; \alpha^b$$

Elemente.

Ein *uniformer α-Shuffle* ist eine Zufallsvariable $M = (K_1, \ldots, K_\alpha, \Pi)$, die uniform in S_α verteilt ist,

$$\mathbf{Ws}\{M = m\} \;=\; \alpha^{-b}\;, \quad b \in S_\alpha\;.$$

Es gilt:

Behauptung. *Sei π eine Permutation mit genau s wachsenden Sequenzen und sei $\alpha \geq s$. Dann gilt für den uniformen α-Shuffle $(K_1, \ldots, K_\alpha, \Pi)$*

$$\mathbf{Ws}\{\Pi = \pi\} \;=\; \binom{b+\alpha-s}{b}\alpha^{-b}\;. \tag{1.17}$$

Beweis. Es ist zu klären, wieviele Wahlmöglichkeiten für $k_1, \ldots, k_\alpha \geq 0$ bestehen, so daß π auf allen $p_1 = \{1, \ldots, k_1\}, \ldots, p_\alpha = \{b - k_\alpha + 1, b\}$ monoton wächst. Alle p_i müssen also vollständig in einer wachsenden Sequenz von π enthalten sein. Sei $r_i \in B$ die Stelle, an der die i-te wachsende Sequenz in π endet $(1 \leq r_1 < r_2 < \ldots < r_s = b)$.

Um die Anzahl der Möglichkeiten abzuzählen, stellen wir uns vor, daß b Kugeln und $\alpha - 1$ Stöckchen von links nach rechts nebeneinander gelegt werden. Jede Anordnung dieser $b + \alpha - 1$ Gegenstände repräsentiert eine Wahlmöglichkeit, die Anzahl der Kugeln zwischen dem $(i-1)$ten und dem $i$ten Stöckchen (bzw. vor dem ersten oder nach dem letzten Stöckchen) gibt k_i an. Die Bedingung, daß alle p_i vollständig in wachsenden Sequenzen von π enthalten sind, ist äquivalent dazu, daß direkt nach der r_1-ten bis r_{s-1}-ten Kugel ein Stöckchen folgt. Diese $s - 1$ Stöckchen können wir aus der Reihe herauslegen bzw. später einfügen. Die restlichen $b + \alpha - s$ Gegenstände können wir in beliebiger Reihenfolge nebeneinander legen. Dies ist auf $\binom{b+\alpha-s}{b}$ verschiedene Weisen möglich. Da bei einem uniformen α-Shuffle jeder dieser Fälle mit Wahrscheinlichkeit α^{-b} realisiert wird, folgt die Behauptung. □

Wir kommen nun zu der entscheidenden Eigenschaft des Modells von Gilbert-Shannon-Reed: Das Hintereinanderausführen eines α-Shuffles und eines davon unabhängigen β-Shuffles entspricht einem einzigen $\alpha\beta$-Shuffle.

Seien zunächst $m = (k_1, \ldots, k_\alpha, \pi)$ und $m' = (k'_1, \ldots, k'_\beta, \pi')$ deterministische Shuffle und $p_1, \ldots, p_\alpha, p'_1, \ldots, p'_\beta$ die zugehörigen Päckchen. Um die Hintereinanderausführung der beiden Shuffle als einzelnen Shuffle zu beschreiben, setzen wir

$$p_{i,j} \;:=\; p_i \cap \Pi^{-1}(p'_j)\;, \quad k_{i,j} \;:=\; \text{card}\; p_{i,j}\;.$$

Es gilt:

- Durch $p_{1,1}, \ldots, p_{1,\beta}, \ldots, p_{\alpha,1}, \ldots, p_{\alpha,\beta}$ wird das Blatt von oben nach unten in Päckchen zerlegt: Erstens sind alle Karten aus $p_{i,j}$ oberhalb von $p_{i+1,k}$ (weil p_i oberhalb von p_{i+1} ist) und zweitens sind alle Karten aus $p_{i,j}$ oberhalb von $p_{i,j+1}$ (weil p'_j oberhalb von p'_{j+1} ist und π auf p_i monoton wächst).

- $\pi' \circ \pi$ wächst monoton auf allen $p_{i,j}$, denn π ist auf p_i monoton, und π' monoton auf p'_j.

$(k_{1,1}, \ldots, k_{\alpha,\beta}, \pi' \circ \pi)$ ist also ein deterministischer $\alpha\beta$-Shuffle. Wir bezeichnen ihn mit $m' \circ m$.

- Aus $m \circ m'$ lassen sich (gegeben α und β) m und m' zurückgewinnen: Es gilt $\pi'(p'_j) = \bigcup_i (\pi' \circ \pi)(p_{i,j})$, und diese Mengen legen m' fest (denn k'_j ist die Anzahl von $\pi'(p'_j)$ und π' ist dann durch Monotonie bestimmt). Damit erhalten wir $\pi = (\pi')^{-1} \circ (\pi' \circ \pi)$. Außerdem gilt $p_i = \bigcup_j p_{i,j}$, und wir haben auch m rekonstruiert.

Die Abbildung $(m, m') \mapsto m' \circ m$ ist also eine Injektion von $S_\alpha \times S_\beta$ nach $S_{\alpha\beta}$, und folglich eine Bijektion, denn beide Mengen enthalten $(\alpha\beta)^b$ Elemente. Daher gibt es keine $\alpha\beta$-Shuffle, die nicht wie eben aus einem α-Shuffle m und einem β-Shuffle m' zusammengesetzt werden können.

Sei nun M ein uniformer α-Shuffle und M' ein davon unabhängiger uniformer β-Shuffle. Dann können wir nach demselben Schema den zufälligen $\alpha\beta$-Shuffle $M' \circ M$ bilden. Unsere Überlegungen zeigen, daß die Gleichung $\{M' \circ M = m \circ m'\} = \{M = m, M' = m'\}$ gilt. Aus der Unabhängigkeit folgt

$$\mathbf{Ws}\{M' \circ M = m' \circ m\} \;=\; \mathbf{Ws}\{M = m\}\mathbf{Ws}\{M' = m'\} \;=\; (\alpha\beta)^{-b}\;.$$

$M' \circ M$ ist also ein uniformer $\alpha\beta$-Shuffle.

Beweis von Formel (1.14). Insbesondere ist das n-malige unabhängige Hintereinanderausführen eines Riffle-Shuffle äquivalent zu einem uniformen 2^n-Shuffle, und (1.14) erweist sich als Spezialfall von (1.17). □

Beweis von Formel (1.16). Die einzige Permutation mit einer einzigen aufsteigenden Sequenz ist die Identität, daher gilt $a(b,1) = 1$. Weiter folgt aus (1.17)

$$\mathbf{Ws}\{\Pi \text{ enthält genau } t \text{ aufsteigende Sequenzen}\} \;=\; a(b,t)\binom{b+\alpha-t}{b}\alpha^{-b}$$

und folglich

$$\sum_{t=1}^{\alpha} a(b,t)\binom{b+\alpha-t}{b}\alpha^{-b} \;=\; 1\ .$$

Diese Formel ist zu (1.16) äquivalent. □

Kapitel 2

Zufallsvariable und Wahrscheinlichkeiten

Im vorigen Kapitel haben wir uns anhand von Beispielen mit Zufallsvariablen vertraut gemacht, nun wollen wir die Zufallsvariable als Begriff der Mathematik kennenlernen. Die Frage, „was eine Zufallsvariable denn nun eigentlich ist“, drängt sich auf. Was ist es, das etwa bei einer Serie von unabhängigen Bernoulli-Experimenten den Unterschied zwischen der relativen Häufigkeit der Erfolge (einer Zufallsvariablen) und der Erfolgswahrscheinlichkeit der Einzelexperimente (einer Zahl) ausmacht?

Die Mathematik beantwortet die Frage nach dem Inhalt ihrer Begriffe, indem sie festlegt, wie man mit ihnen formal verfährt, wie man also mit ihnen ‚rechnet‘. So wollen wir auch hier verfahren: Im Abschnitt 2.1 geben wir einen Überblick über formale Eigenschaften von Zufallsvariablen und Ereignissen und ihren Zusammenhang.

Für einen strengen Aufbau wird das aber in der Mathematik noch nicht als ausreichend angesehen. Man erinnere sich an zwei Weisen, wie man die natürlichen Zahlen in der Mathematik einführen kann. Entweder gewinnt man sie durch eine mathematische Konstruktion (sozusagen durch ein mathematisches Modell). Man kann sie sich etwa aus der leeren Menge verschaffen, $0 := \emptyset$, $1 := \{\emptyset\}$, $2 := \{\emptyset, \{\emptyset\}\}$, ... gemäß der Vorschrift $n+1 := n \cup \{n\}$. Das Willkürliche des Vorgehens liegt auf der Hand, deswegen wird häufig vorgezogen, die natürlichen Zahlen nicht einzeln zu definieren, sondern axiomatisch als Elemente einer Menge mit einer Struktur, die durch die Peano-Axiome gegeben ist. Das System aller natürlichen Zahlen samt ihrer Eigenschaften tritt in den Vordergrund, und nicht die einzelne Zahl („die natürlichen Zahlen sind die Elemente der Menge der natürlichen Zahlen“).

Auch bei der Einführung von Zufallsvariablen und Ereignissen stehen beide Wege offen. Der axiomatische ist aufwendiger und wegen seiner Abstrakt-

heit vielleicht für den Anfang weniger geeignet. Deswegen verfolgen wir hier den üblichen Weg der Wahrscheinlichkeitstheorie, Ereignisse und Zufallsvariable in einem mengentheoretischen Kontext zu behandeln. Dies geschieht in Abschnitt 2.2. Man sollte sich aber bewußt sein, daß erst eine axiomatische Behandlung, fern von speziellen mengentheoretischen Konstruktionen, letzte Klarheit bieten kann, was es eigentlich mit Ereignissen und Zufallsvariablen mathematisch auf sich hat.

Im Abschnitt 2.3 gehen wir auf die Forderungen ein, die man an Wahrscheinlichkeiten stellt, definieren stochastische Unabhängigkeit von Zufallsvariablen und Ereignissen und behandeln schließlich Zufallsvariable mit Werten im Euklidischen Raum, deren Verteilungen durch Dichten gegeben sind. Dies ist dann eine ausreichende Grundlage, um im Abschnitt 2.4 elementare Eigenschaften des Poisson-Prozesses zu behandeln, ein fundamentales stochastischen Modells für zufällige Punktmengen.

Abgesehen vom Abschnitt über Dichten beschränken wir uns in diesem Kapitel auf **diskrete Zufallsvariable**, auf Zufallsvariable mit abzählbaren Wertebereichen. Wir vermeiden so die Diskussion technischer Details.

2.1 Diskrete Zufallsvariable und Ereignisse

Systeme von diskreten Zufallsvariablen und Ereignissen nennen wir **diskrete Zufallsräume**. Sie bestehen aus zwei sich komplementär ergänzenden Bestandteilen.

Einerseits ist eine Gesamtheit $\mathcal{Z}$ von Zufallsvariablen gegeben, wobei jeder Zufallsvariablen eine abzählbare Menge, ihr Wertebereich zugeordnet ist. Dabei sind folgende Eigenschaften erfüllt:

A. Aus einer S-wertigen Zufallsvariablen X und einer Abbildung $\varphi : S \to S'$ lässt sich eine neue Zufallsvariable $\varphi(X)$ mit Werten in S' bilden. Dabei gilt $(\psi \circ \varphi)(X) = \psi(\varphi(X))$.

B. Aus Zufallsvariablen $X_1, \ldots, X_n$ mit Werten in $S_1, \ldots, S_n$ lässt sich eine neue Zufallsvariable $X = (X_1, \ldots, X_n)$ mit Werten in $S = S_1 \times \cdots \times S_n$ bilden. Sie erfüllt $\pi_i(X) = X_i$, wobei π_i die Projektionsabbildung von S auf S_i bezeichnet, $\pi_i(x_1, \ldots, x_n) := x_i$.

Sind etwa X_1, X_2 Zufallsvariable mit Werten in $S := \mathbb{Z}$, so lässt sich die neue Zufallsvariable $X_1 + X_2 := \varphi(X)$ mit Werten in $\mathbb{Z}$ bilden, mit $X := (X_1, X_2)$ und $\varphi(x_1, x_2) := x_1 + x_2$, $x_1, x_2 \in \mathbb{Z}$. Es ist nicht schwer, das Assoziativitätsgesetz $(X_1 + X_2) + X_3 = X_1 + (X_2 + X_3)$ und andere bekannte Rechenregeln zu folgern.

Andererseits ist eine System $\mathcal{A}$ von Ereignissen gegeben, wir sprechen von einem **Ereignisfeld**. Es hat folgende Eigenschaften:

C. Auf $\mathcal{A}$ ist eine Halbordnung $\subset$ gegeben. $A \subset A'$ wird interpretiert als: „mit dem Ereignis A tritt sicher auch das Ereignis A' ein".

D. $\mathcal{A}$ enthält zwei Ereignisse Φ und Ω, charakterisiert durch $\Phi \subset A \subset \Omega$ für alle $A \in \mathcal{A}$. Φ heißt das **unmögliche Ereignis** und Ω das **sichere Ereignis**.

E. Zu jeder endlichen oder unendlichen Folge $A_1, A_2, \ldots$ von Ereignissen lassen sich die Ereignisse $\bigcup_n A_n$ und $\bigcap_n A_n$ bilden, die **Vereinigung** und der **Durchschnitt** der Ereignisfolge. Ihre anschauliche Bedeutung ist: „$\bigcup_n A_n$ tritt ein, wenn mindestens eines der Ereignisse A_n eintreten, und $\bigcap_n A_n$ tritt ein, wenn alle A_n eintreten". Für zwei Ereignisse A, A' schreiben wir $A \cup A'$ und $A \cap A'$.

F. Jedes Ereignis A besitzt ein **komplementäres Ereignis** A^c, charakterisiert durch die Eigenschaften $A \cup A^c = \Omega$, $A \cap A^c = \Phi$. Dies bedeutet anschaulich: „A^c tritt genau dann ein, wenn A nicht eintritt".

Wir benutzen die Notationen der Mengenlehre, auch wenn man Ereignisse zunächst einmal nicht als Mengen begreifen möchte. Gleichwohl sind die Rechenregeln diejenigen, wie sie aus der Mengenlehre bekannt sind, etwa

$$\Big(\bigcup_n A_n\Big)^c \;=\; \bigcap_n A_n^c \;.$$

Gilt $\bigcap_n A_n = \Phi$, so sprechen wir von **sich gegenseitig ausschließenden**, von **disjunkten Ereignissen** A_n. Wie in der Mengenlehre kann man auch **Differenzereignise** $A - A' := A \cap (A')^c$ und **symmetrische Differenzen** $A \Delta A' := (A - A') \cup (A' - A)$ betrachten.

Der Zusammenhang zwischen den Zufallsvariablen und den Ereignissen entsteht auf zweierlei Weise. Zum Einen lässt sich zu jeder diskreten Zufallsvariablen X und jeder Teilmenge B ihres Wertebereichs S ein Ereignis $\{X \in B\}$ bilden, das anschaulich gesprochen genau dann eintritt, wenn X seinen Wert in B annimmt. Man benutzt hier Schreibweisen wie

$$\begin{aligned} \{X = x\} \;&:=\; \{X \in \{x\}\} \\ \{X_1 \in B_1, \ldots, X_n \in B_n\} \;&:=\; \{X_1 \in B_1\} \cap \cdots \cap \{X_n \in B_n\} \;. \end{aligned}$$

Es gilt:

$$\{X \in S\} \;=\; \Omega\;, \quad \{X \in \emptyset\} \;=\; \Phi\;, \quad \{X \in B\}^c \;=\; \{X \in B^c\}\;,$$

und für $B_1, B_2, \ldots \subset S$

$$\Big\{X \in \bigcup_n B_n\Big\} \;=\; \bigcup_n \{X \in B_n\}\;, \quad \Big\{X \in \bigcap_n B_n\Big\} \;=\; \bigcap_n \{X \in B_n\}\;.$$

Der Zusammenhang zu den Prinzipien A. und B. ist gegeben durch

$$\{\phi(X) \in B\} \;=\; \{X \in \phi^{-1}(B)\}\;,$$
$$\{(X_1, \ldots, X_n) \in B_1 \times \cdots \times B_n\} \;=\; \bigcap_{i=1}^n \{X_i \in B_i\}\;.$$

Damit lassen sich dann für Zufallsvariable X, Y mit demselben Wertebereich S Ereignisse bilden wie

$$\{X \le Y\} \;:=\; \{(X,Y) \in B_{\le}\}$$

mit $B_{\le} := \{(x,y) \in S \times S : x \le y\}$, dabei bezeichne $\le$ eine Ordnungsbeziehung in S (oder irgend eine andere Relation). Ist $\{X \le Y\}$ das sichere Ereignis, so schreiben wir

$$X \le Y\;.$$

Gleichbedeutend für diskrete Zufallsvariable ist $\{X = x\} \cap \{Y = y\} = \Phi$ für alle $x \not\le y$, wie sich aus $\{X \le Y\}^c = \bigcup_{x \not\le y} \{X = x, Y = y\}$ ergibt (Übung).

Insbesondere lässt sich immer das Ereignis

$$\{X = Y\} \;:=\; \{(X,Y) \in B_{=}\}$$

mit $B_{=} := \{(x,y) \in S \times S : x = y\}$ bilden. Wir heben hervor: Zwei Zufallsvariable X, Y mit demselben Wertebereich sind genau dann gleich, wenn $\{X = Y\}$ das sichere Ereignis ist. Man kann das auch so ausdrücken, dass $X = Y$ gleichbedeutend ist mit $\{X = x\} = \{Y = x\}$ für alle $x \in S$ bzw. $\{X \in B\} = \{Y \in B\}$ für alle $B \subset S$ (Übung).

Zum Anderen gehört zu jedem Ereignis A eine Zufallsvariable I_A mit Werten in $\{0, 1\}$, so dass

$$\{I_A = 1\} \;=\; A\;, \quad \{I_A = 0\} \;=\; A^c\;.$$

I_A heißt **Indikatorvariable** von A.

Ähnlich gibt es zu jeder unendlichen Folge von disjunkten Ereignissen $A_1, A_2, \ldots$ eine Zufallsvariable X mit Werten in $\overline{\mathbb{N}} = \{1, 2, \ldots, \infty\}$ mit

$$A_n \;=\; \{X = n\}\;,\;\Big(\bigcup_n A_n\Big)^c \;=\; \{X = \infty\}\;.$$

Allgemeiner lassen sich aus einer beliebigen Folge A_n von Ereignissen disjunkte Ereignisse $A_n \cap A_{n-1}^c \cap \cdots \cap A_1^c$ bilden. Für die Zufallsvariable X mit

$$A_n \cap A_{n-1}^c \cap \cdots \cap A_1^c \;=\; \{X = n\}\;,\;\Big(\bigcup_n A_n\Big)^c \;=\; \{X = \infty\}$$

schreiben wir

$$X \;=\; \min\{n : A_n \text{ tritt ein}\}\;.$$

Angesichts dieser Beziehungen zwischen Zufallsvariablen und Ereignissen bedingen sich $\mathcal{Z}$ und $\mathcal{A}$ gegenseitig und können auseinander gewonnen werden. Dies ist in unterschiedlicher axiomatischer Weise möglich, wir gehen darauf nicht weiter ein.

2.2 Messbare Räume und Abbildungen

Wir betten die Begriffe des letzten Abschnitts nun in einen mengentheoretischen Kontext ein. Dazu verwendet man folgende mathematischen Begriffe.

Definition. *Ein Mengensystem $\mathcal{A}$ in einer nicht-leeren Grundmenge Ω heißt* ***σ-Algebra****, falls gilt:*

i) $\Omega \in \mathcal{A}$,

ii) mit $A \in \mathcal{A}$ *gilt* $A^c := \Omega - A \in \mathcal{A}$,

iii) für abzählbare viele $A_1, A_2, \ldots \in \mathcal{A}$ *gilt* $\bigcup_n A_n \in \mathcal{A}$.

Das Paar $(\Omega, \mathcal{A})$ heißt dann **messbarer Raum**.

Es folgt $\emptyset = \Omega^c \in \mathcal{A}$ und $\bigcap_n A_n = \big(\bigcup_n A_n^c\big)^c \in \mathcal{A}$.

Definition. *Sei $(\Omega, \mathcal{A})$ messbarer Raum und S abzählbar. Dann heißt eine Abbildung $X : \Omega \to S$* **messbar** *(genauer $\mathcal{A}$-messbar), falls*

$$X^{-1}(B) \;\in\; \mathcal{A} \quad \textit{für alle } B \subset S\;.$$

Die für uns wichtigen Sachverhalte sind in folgender Proposition zusammengefasst.

Proposition 2.1. *Es gilt:*

i) Seien S, S' abzählbar und $\phi : S \to S'$. Mit $X : \Omega \to S$ ist dann auch $\phi(X) = \phi \circ X : \Omega \to S'$ messbar.

ii) Seien $S_1, \ldots, S_n$ abzählbar. Mit $X_1 : \Omega \to S_1$, …, $X_n : \Omega \to S_n$ ist dann auch $(X_1, \ldots, X_n) : \Omega \to S_1 \times \cdots \times S_n$ messbar.

Beweis. Die erste Behauptung ergibt sich aus

$$\phi(X)^{-1}(B') \;=\; X^{-1}(B) \;\in\; \mathcal{A}$$

für alle $B' \subset S'$, mit $B := \phi^{-1}(B') \subset S$. Die zweite Behauptung folgt aus

$$(X_1, \ldots, X_n)^{-1}(B) \;=\; \bigcup_{(x_1,\ldots,x_n)\in B} X_1^{-1}(\{x_1\}) \cap \cdots \cap X_n^{-1}(\{x_n\}) \;\in\; \mathcal{A}$$

für alle $B \subset S_1 \times \cdots \times S_n$. □

Der Zusammenhang zu den Begriffen des letzten Abschnitts ergibt sich, indem wir Ereignisse mit den Elementen einer σ-Algebra $\mathcal{A}$ auf einer Grundmenge Ω identifizieren und Zufallsvariable mit Wertebereich S mit messbaren Abbildungen $X : \Omega \to S$. Vereinigungen, Durchschnitte und Komplemente von Ereignissen stimmen nun mit den üblichen mengentheoretischen Operationen innerhalb Ω überein, das sichere Ereignis und das unmögliche Ereignis werden durch $\mathit{\Omega} := \Omega$ und $\mathit{\Phi} := \emptyset$ repräsentiert. Aufgrund der Messbarkeit von X gehört

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

zu $\mathcal{A}$ und ist also Ereignis. Die Indikatorvariable von A ist nun die charakteristische Funktion $1_A : \Omega \to \{0, 1\}$ der Menge A, die auf A den Wert 1 und sonst den Wert 0 annimmt. Man überzeuge sich, dass alle im vorigen Abschnitt angesprochenen Eigenschaften von Ereignissen und diskreten Zufallsvariablen erfüllt sind.

Beispiel. $\Omega := \{(x, y) : x, y = 1, \ldots, 6\}$, $\mathcal{A} := \{A : A \subset \Omega\}$ und die Abbildungen $X, Y : \Omega \to S$ mit $S := \{1, \ldots, 6\}$, $X(x, y) := x$, $Y(x, y) := y$ geben ein Modell für 2-faches Würfeln. □

Bemerkung. Dieser mengentheoretische Ansatz hat Aspekte, die für die Stochastik von sekundärer oder von überhaupt keiner Bedeutung sind. Dies gilt für die Elemente ω von Ω. Ihr einziger Zweck ist, dass man messbare Abbildungen und damit Zufallsvariable bilden kann, sonst treten sie in keiner relevanten Aussage der Wahrscheinlichkeitstheorie als Gegenstand der Untersuchung auf. Die Teilmengen von Ω, die nicht zu $\mathcal{A}$ gehören, sind völlig bedeutungslos. Auch Anderes erscheint willkürlich: Das sichere Ereignis Ω kann prinzipiell durch jede nicht-leere Menge Ω repräsentiert werden, während das unmögliche Ereignis Φ immer durch die leere Menge $\emptyset$ dargestellt wird. Ginge man axiomatisch vor, gäbe es solch überflüssige und manchmal verwirrende Details nicht.

Man kann deswegen den mengentheoretischen Ansatz mit einigem Recht als mathematisches Modell für Zufallsvariable und Ereignisse auffassen. Vielleicht würde es das Verständnis fördern, wenn man in der Stochastik noch über ganz andersartige Modelle für Zufallsvariable verfügen würde, im Sinne von COXETER, der zu den Modellen der hyperbolischen Geometrie feststellt: „Wenn wir Modelle verwenden, ist es wünschbar deren zwei anstatt nur eines zu haben, um nicht einem von ihnen ungebührlichen Vorrang zu erteilen; denn unser ... Schließen sollte nur von den Axiomen abhängen“ (Unvergängliche Geometrie, 1981, S. 352). □

2.3 Wahrscheinlichkeiten und stochastische Unabhängigkeit

Wir betrachten nun einen Zufallsraum, dessen Ereignisse A mit Wahrscheinlichkeiten $\mathbf{Ws}\{A\}$ versehen sind. Eine Minimalforderung ist, daß sich die Wahrscheinlichkeiten additiv verhalten, daß also die Wahrscheinlichkeit der Vereinigung von endlich vielen, paarweise disjunkten Ereignissen gleich der Summe der Einzelwahrscheinlichkeiten ist. Wie sich herausgestellt hat, reicht das aber für eine substantielle Theorie nicht aus, man muß fordern, daß sich Wahrscheinlichkeiten auch für unendliche Folgen von disjunkten Ereignisse additiv verhalten. Die Grundannahmen für Wahrscheinlichkeiten, die **Axiome von Kolmogorov**, lauten daher

i) $0 \le \mathbf{Ws}\{A\} \le 1$ *für alle* $A \in \mathcal{A}$, $\mathbf{Ws}\{\Phi\} = 0$, $\mathbf{Ws}\{\Omega\} = 1$.

ii) ***σ-Additivität.*** *Ist* $A_1, A_2, \ldots$ *eine endliche oder unendliche Folge von paarweise disjunkten Ereignissen, so gilt*

$$\mathbf{Ws}\Big\{\bigcup_m A_m\Big\} = \sum_m \mathbf{Ws}\{A_m\} .$$

Man spricht dann von einem Wahrscheinlichkeitsmaß auf dem Ereignisfeld $\mathcal{A}$, kurz von einem **W-Maß**. Ein mit einem W-Maß versehener Zufallsraum heißt ein Wahrscheinlichkeitsraum, kurz ein **W-Raum**. Aus den Annahmen leiten sich die bekannten Eigenschaften von Wahrscheinlichkeiten ab:

iii) ***Monotonie.*** *Für $A_1 \subset A_2$ gilt* $\mathbf{Ws}\{A_2 - A_1\} = \mathbf{Ws}\{A_2\} - \mathbf{Ws}\{A_1\}$, *insbesondere* $\mathbf{Ws}\{A_1\} \le \mathbf{Ws}\{A_2\}$,
denn A_1 und $A_2 - A_1$ sind disjunkt und ergeben vereinigt A_2, so dass $\mathbf{Ws}\{A_1\} + \mathbf{Ws}\{A_2 - A_1\} = \mathbf{Ws}\{A_2\}$ gemäß Additivität. Speziell folgt für $A_1 = A, A_2 = \Omega$, also $A_2 - A_1 = A^c$:

iv) $\mathbf{Ws}\{A^c\} = 1 - \mathbf{Ws}\{A\}$.

v) ***σ-Stetigkeit.*** *Für unendlich viele Ereignisse $A, A_1, A_2, \ldots$ gilt*

$$A_n \uparrow A \quad \Rightarrow \quad \lim_n \mathbf{Ws}\{A_n\} = \mathbf{Ws}\{A\} \ ,$$
$$A_n \downarrow A \quad \Rightarrow \quad \lim_n \mathbf{Ws}\{A_n\} = \mathbf{Ws}\{A\} \ .$$

Dabei benutzen wir die Notation

$$A_n \uparrow A \quad :\Leftrightarrow \quad A_1 \subset A_2 \subset \cdots \ , \ A = \bigcup_{n=1}^{\infty} A_n \ ,$$
$$A_n \downarrow A \quad :\Leftrightarrow \quad A_1 \supset A_2 \supset \cdots \ , \ A = \bigcap_{n=1}^{\infty} A_n \ .$$

Zum Beweis der ersten Aussage bilde man die disjunkten Ereignisse $\widetilde{A}_n = A_n - A_{n-1}$, $n \ge 1$, mit $A_0 = \Phi$. Wegen $A_n = \bigcup_{m=1}^{n} \widetilde{A}_m$ und $A = \bigcup_{m=1}^{\infty} \widetilde{A}_m$ folgt

$$\lim_n \mathbf{Ws}\{A_n\} = \lim_n \sum_{m=1}^{n} \mathbf{Ws}\{\widetilde{A}_m\} = \sum_{m=1}^{\infty} \mathbf{Ws}\{\widetilde{A}_m\} = \mathbf{Ws}\{A\} \ .$$

Die zweite Aussage ergibt sich mittels *iv)* durch Übergang zu Komplementärereignissen.

vi) ***σ-Subadditivität.*** *Für abzählbar viele (nicht notwendig paarweise disjunkte) Ereignisse $A_1, A_2, \ldots$ gilt*

$$\mathbf{Ws}\Big\{\bigcup_m A_m\Big\} \le \sum_m \mathbf{Ws}\{A_m\} \ .$$

Zum Beweis langt es, zwei Ereignisse zu betrachten (der Rest folgt für endliche Ereignisfolgen per Induktion und für unendliche Ereignisfolgen mittels σ-Stetigkeit):

$$\mathbf{Ws}\{A_1 \cup A_2\} \;=\; \mathbf{Ws}\{A_1\} + \mathbf{Ws}\{A_2 - A_1\} \;\leq\; \mathbf{Ws}\{A_1\} + \mathbf{Ws}\{A_2\} \;.$$

Wir können nun den Anschluß an die in Kapitel 1 benutzten Sprechweisen herstellen. In einem W-Raum besitzt jede Zufallsvariable X eine **Verteilung** $\mu = \mu_X$, gegeben durch

$$\mu(B) \;:=\; \mathbf{Ws}\{X \in B\} \;, \quad B \subset S \;,$$

dabei steht $\mathbf{Ws}\{X \in B\}$ für $\mathbf{Ws}\{\{X \in B\}\}$. Zufallsvariable mit derselben Verteilung heißen **identisch verteilt**, man sagt dann auch, sie seien **Kopien einer Zufallsvariablen**. Die Verteilung einer Produktvariablen $(X_1, \ldots, X_n)$ nennt man die **gemeinsame Verteilung** von $X_1, \ldots, X_n$. Die Verteilung einer diskreten Zufallsvariablen ist durch die Formel

$$\mathbf{Ws}\{X \in B\} \;=\; \sum_{x \in B} p(x)$$

gegeben, mit Zahlen $p(x)$, die die Bedingungen $p(x) \geq 0$ und $\sum_x p(x) = 1$ erfüllen. $\mu = (p(x))_{x \in S}$ nennt man dann eine **W-Verteilung** mit den **Gewichten** $p(x)$.

Bemerkung. Ereignisse der Wahrscheinlichkeit 0 heißen **Nullereignisse** und ihre Komplementärereignisse **fast sichere Ereignisse**. Man geht davon aus, daß ein Nullereignis nicht eintritt, genauso, wie es nicht gelingen wird, in einer beliebig langen Serie von unabhängigen Münzwürfen immer nur Kopf zu werfen. Dies legt es nahe, die (sichere) Gleichheit zwischen Ereignissen bzw. Zufallsvariablen zu ergänzen durch einen Begriff **fast sicherer Gleichheit**. Zwei Ereignisse A und A' heißen fast sicher gleich, falls sie sich nur um ein Nullereignis unterscheiden (falls also $A \triangle A'$ ein Nullereignis ist), und zwei Zufallsvariable X und Y mit demselben Wertebereich S heißen fast sicher gleich, falls für alle $B \subset S$ die Ereignisse $\{X \in B\}$ und $\{Y \in B\}$ fast sicher gleich sind (oder äquivalent, falls $\{X = Y\}$ ein fast sicheres Ereignis ist). Es handelt sich hier um Äquivalenzrelationen.

Ein Beobachter ist nicht in der Lage, zwei fast sicher gleiche Zufallsvariable anhand ihrer Werte zu unterscheiden. Dies legt es nahe, fast sicher gleiche Zufallsvariable zu identifizieren. Man spricht dann von **fast sicher definierten Zufallsvariablen**, sie bilden, wie man sich leicht überzeugt, in kanonischer Weise einen diskreten Zufallsraum. Wir machen davon im Folgenden keinen weiteren Gebrauch. □

Unabhängigkeit

Bisher war von unabhängigen Zufallsvariablen nur in einem anschaulichen Sinne die Rede. Nun können wir Unabhängigkeit mathematisch definieren. Wir beginnen mit diskreten Zufallsvariablen.

Definition. *Zufallsvariable $X_1, \dots, X_n$ mit abzählbaren Wertebereichen $S_1, \dots, S_n$ heißen* ***(stochastisch) unabhängig****, falls*

$$\mathbf{Ws}\{X_1 \in B_1, \dots, X_n \in B_n\} \;=\; \mathbf{Ws}\{X_1 \in B_1\} \cdots \mathbf{Ws}\{X_n \in B_n\}$$

für alle $B_1 \subset S_1, \dots, B_n \subset S_n$ gilt. Eine unendliche Folge von Zufallsvariablen heißt unabhängig, wenn jede endliche Teilfolge unabhängig ist.

Es ist dann auch jede Teilfamilie $X_{i_1}, \dots, X_{i_k}$ mit $1 \le i_1 < \cdots < i_k \le n$ unabhängig. Man erkennt dies, indem man in der Gleichung $B_j = S_j$ für alle $j \neq i_1, \dots, i_k$ wählt.

Unabhängigkeit lässt sich mit folgendem einfachen Kriterium feststellen.

Proposition 2.2. *Seien $X_1, \dots, X_n$ diskrete Zufallsvariable mit Werten in $S_1, \dots, S_n$ und seien $\mu_1, \dots, \mu_n$ W-Verteilungen auf $S_1, \dots, S_n$ mit den Gewichten $p_1(x_1), \dots, p_n(x_n)$. Dann sind folgende Aussagen äquivalent:*

i) Die $X_1, \dots, X_n$ sind unabhängig, und X_m hat die Verteilung μ_m, $m = 1, \dots, n$.

ii) Es gilt

$$\mathbf{Ws}\{X_1 = x_1, \dots, X_n = x_n\} \;=\; p_1(x_1) \cdots p_n(x_n)$$

für alle $x_1 \in S_1, \dots, x_n \in S_n$.

Beweis. *i)* $\Rightarrow$ *ii)* ist offensichtlich. Zu *ii)* $\Rightarrow$ *i)*: Aus *ii)* folgt unter Beachtung der σ-Additivität von Wahrscheinlichkeiten

$$\begin{aligned}\mathbf{Ws}\{X_1 \in B_1, \dots, X_n \in B_n\} \;&=\; \sum_{x_1 \in B_1} \cdots \sum_{x_n \in B_n} \mathbf{Ws}\{X_1 = x_1, \dots, X_n = x_n\} \\ &=\; \sum_{x_1 \in B_1} p_1(x_1) \cdots \sum_{x_n \in B_n} p_n(x_n)\ .\end{aligned}$$

Insbesondere gilt $\mathbf{Ws}(X_m \in B_m\} = \sum_{x_m \in B_m} p_m(x_m)$, wie die Wahl $B_i = S_i$ für alle $i \neq m$ zeigt, und es folgt *i)*. □

Beispiel. Fehlstellen bei Permutationen. Einer Permutation $\pi = (\pi_1, \ldots, \pi_k)$ der Zahlen $1, \ldots, k$ ordnen wir für jedes $i = 2, \ldots, k$ ihre *Zahl der Fehlstellen (Inversionen)*

$$x_i \;=\; \phi_i(\pi) \;:=\; \mathrm{card}\{j : j < i, \pi_j > \pi_i\}$$

zu, sie gibt an, wieviele paarweise Vertauschungen in der Permutation nötig sind, damit vor π_i keine größeren Zahlen mehr stehen. Umgekehrt kann man zu nicht-negativen ganzen Zahlen $x_2 < 2, \ldots, x_k < k$ eindeutig eine passende Permutation π konstruieren: x_k bestimmt π_k, x_{k-1} sagt dann, welche der übrigen Zahlen π_{k-1} ist etc.

Nun sei Π eine rein zufällige Permutation von $1, \ldots, k$ und $X_i := \phi_i(\Pi)$ ihre Inversionszahlen mit den Wertebereichen $S_i = \{0, 1, \ldots, i-1\}$. Da es insgesamt $k!$ Permutationen gibt, gilt

$$\mathbf{Ws}\{X_2 = x_2, \ldots, X_k = x_k\} \;=\; \frac{1}{k!} \;=\; \frac{1}{2} \cdot \frac{1}{3} \cdots \frac{1}{k} \;.$$

Nach Proposition 2.2 folgt, daß X_i uniform in S_i verteilt ist und daß $X_2, \ldots, X_k$ unabhängig sind.

Umgekehrt kann man aus unabhängigen, uniform verteilten $X_2, \ldots, X_k$ eine rein zufällige Permutation gewinnen. Man mischt also Spielkarten perfekt auf die folgende Weise: Stecke im Blatt die i-te Spielkarte von oben an eine rein zufällige Stelle zwischen die Karten, die sich über ihr befinden, nacheinander für $i = 2, 3, \ldots, k$ und unabhängig voneinander (dabei darf die Karte auch ganz oben auf den Stapel kommen oder ihre Position beibehalten). □

Manche Verteilungen sind für das Rechnen mit unabhängigen Zufallsvariablen besonders geeignet. Dazu gehört die Poissonverteilung.

Beispiel. Poissonverteilung. Die Poissonverteilung besitzt folgende fundamentale Eigenschaft: *Sind X und Y unabhängige Poisson-verteilte Zufallsvariable zu den Parametern λ und ν, so ist $Z := X + Y$ Poisson-verteilt zum Parameter $\lambda + \nu$.* Denn aus der Additivität von Wahrscheinlichkeiten und der Unabhängigkeit ergibt sich

$$\begin{aligned} \mathbf{Ws}\{Z = z\} \;&=\; \sum_{x=0}^{z} \mathbf{Ws}\{X = x, Y = z - x\} \\ &=\; \sum_{x=0}^{z} \mathbf{Ws}\{X = x\}\mathbf{Ws}\{Y = z - x\} \;, \end{aligned}$$

und aus der Verteilungsannahme folgt

$$\begin{aligned}\mathbf{Ws}\{Z=z\} &= \sum_{x=0}^{z} \frac{e^{-\lambda}\lambda^x}{x!}\cdot\frac{e^{-\nu}\nu^{z-x}}{(z-x)!} \;=\; \frac{e^{-(\lambda+\nu)}}{z!}\sum_{x=0}^{z}\binom{z}{x}\lambda^x\nu^{z-x}\\ &= \frac{e^{-(\lambda+\nu)}}{z!}(\lambda+\nu)^z\ .\end{aligned}$$

□

Bemerkung. Seien allgemeiner X und Y unabhängige Zufallsvariable mit Werten in $\mathbb{Z}$ und seien p'_x und p''_y die Gewichte der zugehörigen Verteilungen. Dann gilt für die Gewichte p_z der Verteilung von $Z := X + Y$

$$p_z \;=\; \sum_x p'_x p''_{z-x}\ , \quad z\in\mathbb{Z}\ . \tag{2.1}$$

Für die Verteilungen μ', μ'' und μ von X, Y und Z schreibt man

$$\mu \;=\; \mu' * \mu''$$

und nennt μ die **Faltung** von μ' und μ''. □

Bemerkung. Es ist nicht schwer zu zeigen, daß Unabhängigkeit unter Transformation erhalten bleibt: Sind $X_1,\dots,X_n$ stochastisch unabhängig, so auch Zufallsvariable der Gestalt $\phi_1(X_1,\dots,X_{i_1})$, $\phi_2(X_{i_1+1},\dots,X_{i_2}),\dots$, $\phi_k(X_{i_k+1},\dots,X_n)$ mit $1\le i_1<i_2<\cdots<i_k\le n$. □

Für allgemeine Zufallsvariable definiert man Unabhängigkeit im wesentlichen wie im diskreten Fall. Wir gehen auf reellwertige Zufallsvariable ein.

Definition. *Zufallsvariable $X_1,\dots,X_n$ mit Werten in $\mathbb{R}$ heißen* ***unabhängig****, falls für alle reellen Zahlen $a_1\le b_1,\dots,a_n\le b_n$ gilt*

$$\begin{aligned}&\mathbf{Ws}\{X_1\in[a_1,b_1],\dots,X_n\in[a_n,b_n]\}\\ &\qquad = \mathbf{Ws}\{X_1\in[a_1,b_1]\}\cdots\mathbf{Ws}\{X_n\in[a_n,b_n]\}\ .\end{aligned}$$

Beispiel. Ordnungsstatistiken. Ordnet man Zufallsvariable $X_1, X_2\dots, X_n$ mit Werten in den reellen Zahlen der Größe nach an, so entstehen die sogenannten **Ordnungsstatistiken**

$$X_{(1)} \;\le\; X_{(2)} \;\le\; \cdots \;\le\; X_{(n)}\ .$$

Wir wollen ihre Verteilungen bestimmen unter der Annahme, daß $X_1,\dots,X_n$ unabhängige Kopien einer Zufallsvariablen X sind, die eine Dichte $f(x)\,dx$ besitzt. Dazu betrachten wir zu vorgegebenem $x\in\mathbb{R}$

$$Y_x \;:=\; I_{\{X_1\le x\}}+\cdots+I_{\{X_n\le x\}}\ ,$$

die Anzahl der X_i, die einen Wert kleiner oder gleich x annehmen. Y_x ist binomialverteilt zum Parameter $(n, F(x))$ mit $F(x) := \mathbf{Ws}\{X \le x)$. $F(x)$ heißt die **Verteilungsfunktion** von X. Es folgt

$$\mathbf{Ws}\{X_{(k)} \le x\} \;=\; \mathbf{Ws}\{Y_x \ge k) \;=\; \sum_{j=k}^{n} \binom{n}{j} F(x)^j (1-F(x))^{n-j} \;.$$

Die Ableitung nach x dieses Ausdrucks errechnet sich unter Beachtung von $F'(x) = f(x)$ als

$$f_{k,n}(x) \;:=\; \frac{n!}{(k-1)!(n-k)!} F(x)^{k-1}(1-F(x))^{n-k} f(x) \;,$$

damit folgt

$$\mathbf{Ws}\{X_{(k)} \le x\} \;=\; \int_{-\infty}^{x} f_{k,n}(x)\, dx$$

bzw. in Kurznotation

$$\mathbf{Ws}\{X_{(k)} \in dx\} \;=\; f_{k,n}(x)\, dx \;.$$

Aufschlußreich ist eine heuristische Begründung dieses Sachverhalts: Damit $X_{(k)}$ ihren Wert in dem infinitesimalen Intervall der Länge dx an der Stelle x annimmt, müssen $k-1$ der X_i einen Wert kleiner oder gleich x annehmen, $n-k$ einen Wert größer oder gleich x und ein X_i seinen Wert in dem infinitesimalen Intervall. Da $X_1, \ldots, X_n$ als unabhängig angenommen sind, ist die Wahrscheinlichkeit durch die Trinomialverteilung gegeben, zu den Wahrscheinlichkeiten $p_1 = F(x)$, $p_2 = 1 - F(x)$ und $p_3 = f(x)\, dx$. Also wie eben:

$$\begin{aligned}&\mathbf{Ws}\{X_{(k)} \in dx\} \\ &\quad= \binom{n}{k-1, n-k, 1} F(x)^{k-1}(1-F(x))^{n-k} f(x)\, dx \;=\; f_{k,n}(x)\, dx \;.\end{aligned}$$

Sind die X_i speziell uniform verteilt auf dem Intervall (0,1), so hat $X_{(k)}$ auf (0,1) die Dichte

$$\mathbf{Ws}\{X_{(k)} \in dx\} \;=\; \frac{n!}{(k-1)!(n-k)!}\, x^{k-1}(1-x)^{n-k}\, dx \;.$$

Die Ordnungsstatistiken sind dann Beta-verteilt im Sinne der folgenden Definition: Seien $a, b > 0$ reelle Zahlen. Dann heißt eine Zufallsvariable X mit Werten im Intervall (0,1) **Beta-verteilt zum Parameter (a, b)**, falls sie Werte in (0,1) annimmt und ihre Verteilung durch

$$\mathbf{Ws}\{X \in dx\} \;=\; c_{a,b}\, x^{a-1}(1-x)^{b-1}\, dx$$

gegeben ist, mit der Normierungskonstante $c_{a,b} := 1/\int_0^1 x^{a-1}(1-x)^{b-1}\, dx$. □

Wir kommen nun zum Begriff der stochastischen Unabhängigkeit von Ereignissen.

Definition. *Eine endliche oder unendliche Folge $A_1, A_2, \ldots$ von Ereignissen heißt* ***unabhängig****, wenn die Indikatorvariablen $I_{A_1}, I_{A_2}, \ldots$ unabhängig sind.*

Die Unabhängigkeit von Ereignissen läßt sich verschieden charakterisieren.

Proposition 2.3. *Für Ereignisse $A_1, \ldots, A_n$ sind äquivalent:*

i) $A_1, \ldots, A_n$ *sind unabhängig.*

ii) Es gilt

$$\mathbf{Ws}\{A_{i_1} \cap \cdots \cap A_{i_k}\} \;=\; \mathbf{Ws}\{A_{i_1}\} \cdots \mathbf{Ws}\{A_{i_k}\}$$

für alle $1 \le i_1 < \cdots < i_k \le n$.

iii) Es gilt

$$\mathbf{Ws}\{A_1' \cap \cdots \cap A_n'\} \;=\; \mathbf{Ws}\{A_1'\} \cdots \mathbf{Ws}\{A_n'\}\ ,$$

wobei A_i' beliebig als A_i oder A_i^c gewählt werden darf.

Beweis. i) ⇒ ii): In Indikatorvariablen lassen sich die Gleichungen unter *ii)* als $\mathbf{Ws}\{I_{A_1} \in B_1, \ldots, I_{A_n} \in B_n\} = \mathbf{Ws}\{I_{A_1} \in B_1\} \cdots\ \mathbf{Ws}\{I_{A_n} \in B_n\}$ schreiben, mit $B_{i_j} = \{1\}$ für $j = 1, \ldots, k$ und $B_i = \{0, 1\}$ sonst.

ii) ⇒ iii): In den Gleichungen unter *ii)* lassen sich schrittweise die Ereignisse durch ihre Komplementärereignisse ersetzen, nach dem Schema

$$\begin{aligned}
&\mathbf{Ws}\{A_{i_1} \cap \cdots \cap A_{i_{k-1}} \cap A_{i_k}^c\} \\
&\qquad = \;\mathbf{Ws}\{A_{i_1} \cap \cdots \cap A_{i_{k-1}}\} - \mathbf{Ws}\{A_{i_1} \cap \cdots \cap A_{i_k}\} \\
&\qquad = \;\mathbf{Ws}\{A_{i_1}\} \cdots \mathbf{Ws}\{A_{i_k}\} - \mathbf{Ws}\{A_{i_1}\} \cdots \mathbf{Ws}\{A_{i_{k-1}}\} \\
&\qquad = \;\mathbf{Ws}\{A_{i_1}\} \cdots \mathbf{Ws}\{A_{i_{k-1}}\} \cdot \mathbf{Ws}\{A_{i_k}^c\}\ .
\end{aligned}$$

iii) ⇒ i): Die Gleichungen unter *iii)* lassen sich mit Indikatorvariablen als $\mathbf{Ws}\{I_{A_1} = x_1, \ldots, I_{A_n} = x_n\} = \mathbf{Ws}\{I_{A_1} = x_1\} \cdots \mathbf{Ws}\{I_{A_n} = x_n\}$ mit $x_i = 0$ oder 1 schreiben. Nach Proposition 2.2 folgt daraus die Unabhängigkeit von $I_{A_1}, \ldots, I_{A_n}$. □

Man beachte: Die Gleichung $\mathbf{Ws}\{A_1 \cap \cdots \cap A_n\} = \mathbf{Ws}\{A_1\} \cdots \mathbf{Ws}\{A_n\}$ ist im Fall $n \geq 3$ für die Unabhängigkeit der Ereignisse nicht ausreichend. Allein ist sie zur Definition von Unabhängigkeit ungeeignet, dann wäre nicht einmal garantiert, daß sich Unabhängigkeit auf Teilfolgen überträgt.

Stochastische Unabhängigkeit kann man nicht ohne weiteres mit kausaler Unverknüpftheit gleichsetzen. Daß die Verhältnisse komplizierter liegen können, zeigen die folgenden Beispiele.

Beispiele.

1. **Ziehen ohne Zurücklegen.** Aus einem (aus 32 Karten bestehenden) Skatblatt werden zwei Karten gezogen. Dann sind die Ereignisse

$$\begin{aligned} A_1 &= \{\text{die erste Karte ist ein As}\} , \\ A_2 &= \{\text{die zweite Karte ist Karo}\} \end{aligned}$$

nicht nur beim Ziehen mit Zurücklegen unabhängig, sondern auch, falls *ohne Zurücklegen* gezogen wird. Dann gilt nämlich

$$\begin{aligned} \mathbf{Ws}\{A_1\} &= \frac{1}{8} , \\ \mathbf{Ws}\{A_2\} &= \frac{24 \cdot 8 + 8 \cdot 7}{32 \cdot 31} = \frac{1}{4} , \\ \mathbf{Ws}\{A_1 \cap A_2\} &= \frac{3 \cdot 8 + 1 \cdot 7}{32 \cdot 31} = \frac{1}{32} . \end{aligned}$$

2. **Treffer beim Lotto.** Seien X, Y, U unabhängige Zufallsvariable mit Werten in der endlichen Menge S der Mächtigkeit r, seien X und Y identisch verteilt mit Verteilung $\mu = (p_x)$ und sei U uniform verteilt. Wir betrachten die Ereignisse

$$A_1 := \{X = U\} , \ A_2 := \{Y = U\} .$$

Dann gilt wegen der σ-Additivität von Wahrscheinlichkeiten und wegen Unabhängigkeit

$$\mathbf{Ws}\{A_1\} = \sum_{x \in S} \mathbf{Ws}\{U = x, X = x\} = \sum_{x \in S} \mathbf{Ws}\{U = x\}\mathbf{Ws}\{X = x\}$$

Da U uniform verteilt, hat die Summe den Wert $r^{-1} \sum_x p(x) = r^{-1}$. Dies und eine analoge Rechnung für $\mathbf{Ws}\{A_2\}$ ergibt

$$\mathbf{Ws}\{A_1\} = \mathbf{Ws}\{A_2\} = r^{-1} .$$

Entsprechend folgt

$$\begin{aligned}\mathbf{Ws}\{A_1 \cap A_2\} &= \sum_{x\in S} \mathbf{Ws}\{U = x, X = x, Y = x\} \\ &= \sum_{x\in S} \mathbf{Ws}\{U = x\}\mathbf{Ws}\{X = x\}\mathbf{Ws}\{Y = x\} \ .\end{aligned}$$

Uniformität von U ergibt

$$\mathbf{Ws}\{A_1 \cap A_2\} = r^{-1} \sum_{x\in S} p_x^2 \ .$$

Unabhängigkeit der beiden Ereignisse liegt also genau dann vor, wenn $\sum_x p_x^2 = r^{-1}$ gilt. Wegen $\sum_x (p_x - r^{-1})^2 = \sum_x p_x^2 - r^{-1}$ gilt das genau dann, wenn $p_x = r^{-1}$ für alle x gilt, wenn also auch X und Y uniform verteilt sind.

Interpretation. Seien X und Y die von zwei Lottospielern unabhängig getippten Lottozahlen und U die danach von dem Lotto–Ziehgerät ermittelten Lottozahlen. Dann sind die beiden Ereignisse, daß der eine bzw. der andere Spieler einen Hauptgewinn hat, i.a. nicht unabhängig (es sei denn, sie wählen ihre Zahlen rein zufällig). – Dieses Phänomen, daß zwei kausal unverknüpfte Größen (wie Schuhgröße und täglicher Zigarettenkonsum einer Person) dennoch über eine dritte Größe (das Alter der Person) zu stochastisch abhängigen Größen werden, gilt es etwa bei statistischen Untersuchungen zu berücksichtigen. □

Die Lemmata von Borel-Cantelli

Eine Illustration für das Rechnen mit Ereignissen und Wahrscheinlichkeiten bieten die Lemmata von Borel-Cantelli. Wir betrachten den Limes superior einer Folge $A_1, A_2, \ldots \in \mathcal{A}$, definiert als

$$\limsup_n A_n := \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n \ . \tag{2.2}$$

Anschaulich tritt dieses Ereignis ein, falls die Ereignisse $\bigcup_{n=m}^{\infty} A_n$ für alle m eintreten, und dies ist der Fall, falls es für jedes m ein $n_m \geq m$ gibt, so daß A_{n_m} eintritt. Mit anderen Worten: $\limsup_n A_n$ tritt ein, falls von den Ereignissen $A_1, A_2, \ldots$ unendlich viele eintreten. Wir schreiben daher auch

$$\limsup_n A_n = \{\infty\text{-viele } A_n \text{ treten ein}\} \ .$$

Bemerkung. Für Teilmengen $A_1, A_2, \ldots$ von Ω gilt

$$\limsup_n A_n \;=\; \{\omega \in \Omega : \omega \in A_n \text{ für unendlich viele } n\}\ .$$

□

Satz 2.4. (Erstes Borel-Cantelli-Lemma) *Aus* $\sum_{n=1}^{\infty} \mathbf{Ws}\{A_n\} < \infty$ *folgt* $\mathbf{Ws}\{\limsup_n A_n\} = 0$.

Beweis. Wegen Monotonie und σ-Subadditivität von Wahrscheinlichkeiten gilt für beliebiges $m \geq 1$

$$\mathbf{Ws}\Big\{\bigcap_{m=1}^{\infty}\bigcup_{n=m}^{\infty} A_n\Big\} \;\leq\; \mathbf{Ws}\Big\{\bigcup_{n=m}^{\infty} A_n\Big\} \;\leq\; \sum_{n=m}^{\infty} \mathbf{Ws}\{A_n\}\ .$$

Der rechte Ausdruck konvergiert nach Voraussetzung mit $m \to \infty$ gegen 0.□

Satz 2.5. (Zweites Borel-Cantelli-Lemma) *Die Ereignisse* $A_1, A_2, \ldots$ *seien stochastisch unabhängig. Aus* $\sum_{n=1}^{\infty} \mathbf{Ws}\{A_n\} = \infty$ *folgt dann* $\mathbf{Ws}\{\limsup_n A_n\} = 1$.

Beweis. Wir zeigen, daß das Komplementärereignis Wahrscheinlichkeit 0 hat. Wegen der Subadditivität von W-Maßen gilt

$$\mathbf{Ws}\{(\limsup_n A_n)^c\} \;=\; \mathbf{Ws}\Big\{\bigcup_{m=1}^{\infty}\bigcap_{n=m}^{\infty} A_n^c\Big\} \;\leq\; \sum_{m=1}^{\infty} \mathbf{Ws}\Big\{\bigcap_{n=m}^{\infty} A_n^c\Big\}\ .$$

Weiter gilt für alle $\ell \geq m$ unter Beachtung von Proposition 2.3, *iii*) und der Ungleichung $1 - x \leq e^{-x}$

$$\begin{aligned}\mathbf{Ws}\Big\{\bigcap_{n=m}^{\infty} A_n^c\Big\} \;&\leq\; \mathbf{Ws}\Big\{\bigcap_{n=m}^{\ell} A_n^c\Big\} \;=\; \prod_{n=m}^{\ell}(1 - \mathbf{Ws}\{A_n\}) \\ &\leq\; \prod_{n=m}^{\ell} \exp\big(-\mathbf{Ws}\{A_n\}\big) \;=\; \exp\Big(-\sum_{n=m}^{\ell} \mathbf{Ws}\{A_n\}\Big)\ .\end{aligned}$$

Indem wir ℓ gegen ∞ gehen lassen, folgt aus der vorausgesetzten Reihendivergenz $\mathbf{Ws}\{\bigcap_{n=m}^{\infty} A_n^c\} = 0$ und damit die Behauptung. □

Beispiel. Die Länge von Erfolgsserien. Für eine Folge $Z_1, Z_2, \ldots$ von Zufallsvariablen mit den Werten 1 oder 0 ist

$$X_n \;:=\; \min\{i \geq 0 : Z_{n+i} = 0\}$$

die Länge der Serie aus Einsen, die an der n-ten Stelle der Folge beginnt. Wie lang werden solche Erfolgsserien? Wir nehmen an, daß die Z_n unabhängig und Bernoulli-verteilt zur Erfolgswahrscheinlichkeit p sind. Dann sind Erfolgsserien höchstens von logarithmischer Länge, für $\lambda > 0$ gilt nämlich

$$\mathbf{Ws}\{X_n \geq \lambda \log n \text{ tritt } \infty\text{-oft ein}\} \;=\; \begin{cases} 1\,, & \text{falls } \lambda \log p^{-1} < 1\,, \\ 0\,, & \text{falls } \lambda \log p^{-1} > 1\,. \end{cases}$$

Dafür schreibt man auch kurz

$$\limsup_n \frac{X_n}{\log n} \;=\; \frac{1}{\log p^{-1}} \qquad \textit{fast sicher}\,.$$

Beweis. Es gilt

$$\begin{aligned} \mathbf{Ws}\{X_n \geq \lambda \log n\} \;&\leq\; p^{\lambda \log n - 1} \;=\; p^{-1}(e^{\log p})^{\lambda \log n} \\ &=\; p^{-1}(e^{\log n})^{\lambda \log p} \;=\; p^{-1} n^{-\lambda \log p^{-1}}\,. \end{aligned}$$

Im Fall $\lambda \log p^{-1} > 1$ ist also $\sum_n \mathbf{Ws}\{X_n \geq \lambda \log n\} < \infty$, und die eine Hälfte der Aussage folgt aus dem ersten Borel-Cantelli-Lemma. Für den anderen Teil der Aussage können wir das zweite Borel-Cantelli-Lemma nicht direkt anwenden, da die Ereignisse $\{X_n \geq \lambda \log n\}$ nicht unabhängig sind. Wir fixieren daher ein $\epsilon > 0$ und betrachten die Teilfolge $\{X_{k_n} \geq \lambda \log k_n\}$ mit $n^{1+\epsilon} - 1 < k_n \leq n^{1+\epsilon}$. Die zugehörigen Erfolgsserien überlappen sich nicht, falls $k_{n+1} > k_n + \lambda \log k_n$, und dies ist wegen $(n+1)^{1+\epsilon} - n^{1+\epsilon} \geq (1+\epsilon) n^{\epsilon}$ für großes n der Fall. Die Ereignisse $\{X_{k_n} \geq \lambda \log k_n\}$, $n \geq n_0$, sind deswegen für ausreichend großes n_0 unabhängig. Es gilt

$$\mathbf{Ws}\{X_{k_n} \geq \lambda \log k_n\} \;\geq\; p^{\lambda \log k_n} \;\geq\; p^{\lambda(1+\epsilon)\log n} \;=\; n^{-\lambda(1+\epsilon)\log p^{-1}}\,,$$

so daß im Fall $\lambda \log p^{-1} < 1$ die Reihe $\sum_n \mathbf{Ws}\{X_{k_n} \geq \lambda \log k_n\}$ divergiert, sofern ϵ genügend klein ist. Nach dem zweiten Borel-Cantelli-Lemma folgt $\mathbf{Ws}\{\limsup_n \{X_{k_n} \geq \lambda \log k_n\}\} = 1$, also erst recht die Behauptung. □

Dichten

Neben den Verteilungen, die durch Gewichte gegeben sind, sind Verteilungen mit Dichten von besonderer Bedeutung. Wir geben einen Überblick.

Sei S messbare Teilmenge des $\mathbb{R}^k$ (messbar heißt hier kurz gesagt, dass man S mit einem wohldefinierten Inhalt versehen kann) und sei V Zufallsvariable (Zufallsvektor) mit Werten in S. Wir betrachten nun den Fall, dass die Verteilung von V von der Gestalt

$$\mathbf{Ws}\{V \in B\} \;=\; \int\limits_B p(v)\,dv$$

für alle messbaren $B \subset S$ ist. Es handelt sich um ein k-dimensionales Integral, ausführlicher notiert als $\int \cdots \int_B p(v_1, \ldots, v_k)\,dv_1 \ldots dv_k$, mit einer Dichtefunktion $p : S \to \mathbb{R}$, also einer nicht-negativen integrierbaren Funktion mit

$$\int\limits_S p(v)\,dv \;=\; 1\ .$$

Man sagt dann, dass $p(v)\,dv$ die **Dichte von V** ist, und schreibt

$$\mathbf{Ws}\{V \in dv\} \;=\; p(v)\,dv\ .$$

Dies ist einzig als eine Kurzschreibweise aufzufassen (zumal als eine etwas ungenaue, denn dv steht hier sowohl für ein ‚infinitesimales' Raumstück an der Stelle v als auch für seinen infinitesimalen Inhalt). Ist V durch Komponenten gegeben, $V = (V_1, \ldots, V_k)$, so nennt man $p(v)\,dv$ auch die **gemeinsame Dichte** von $V_1, \ldots, V_k$ und schreibt sie als $p(v_1, \ldots, v_k)\,dv_1 \ldots dv_k$.

In diesem Abschnitt betrachten wir zwei Sachverhalte über das Rechnen mit Dichten. Erst geben wir die Resultate, dann kommen Beispiele.

Das erste Resultat betrifft die Transformation von Dichten beim Wechsel der Koordinaten. Unter einer Koordinatentransformation verstehen wir eine Bijektion $\phi : S \to S'$ zwischen messbaren Teilmengen S und S' des $\mathbb{R}^k$ mit Umkehrabbildung ψ, in beiden Richtungen stetig differenzierbar. Dann hat der Zufallsvektor $\phi(V)$ ebenfalls eine Dichte. Die genaue Form der transformierten Dichte wird durch die sog. Transformationsformel gegeben. Wir betrachten hier hauptsächlich den Fall, dass ϕ eine inhaltstreue Abbildung ist, d.h. dass B und $B' := \phi(B)$ denselben Rauminhalt haben für alle messbaren $B \subset S$. Dann nimmt die Transformationsformel eine besonders einfache Gestalt an.

Proposition 2.6. *Sei V Zufallsvariable mit Werten in $S \subset \mathbb{R}^k$ und Dichte $p(v)\,dv$, und sei $\phi : S \to S'$ inhaltstreue Koordinatentransformation. Dann hat $W := \phi(V)$ die Dichte*

$$\{W \in dw\} \;=\; q(w)\,dw \quad , \quad \textit{mit} \;\; q(w) \;:=\; p(\psi(w)) \;.$$

Den Beweis findet man in Büchern der Analysis. Das zweite Resultat ist eine zu Proposition 2.2 analoge Aussage für unabhängige Zufallsvariable mit Dichten. Statt der Gewichte werden nun die Dichten multipliziert.

Proposition 2.7. *Seien $V_1, \ldots, V_k$ reellwertige Zufallsvariable, seien $p_1, \ldots, p_k$ Dichtefunktionen auf $\mathbb{R}$ und $p(x_1, \ldots, x_k) := p_1(x_1) \cdots p_k(x_k)$. Dann sind folgende Aussagen äquivalent:*

i) $X_1, \ldots, X_k$ sind unabhängig und haben die Dichten $p_1(x)\,dx, \ldots, p_k(x)\,dx$.

ii) $X_1, \ldots, X_k$ haben die gemeinsame Dichte $p(x_1, \ldots, x_k)\,dx_1 \ldots dx_k$.

Beweis. Der Beweis benötigt Resultate der Maß- und Integrationstheorie. Nach dem Satz von Fubini gilt

$$\int\limits_{[a_1,b_1]\times\cdots\times[a_k,b_k]} \cdots \int p(x_1, \ldots, x_k)\,dx_1 \cdots dx_k \;=\; \int\limits_{a_1}^{b_1} p_1(x_1)\,dx_1 \cdots \int\limits_{a_k}^{b_k} p_k(x_k)\,dx_k \;.$$

Gilt also *i)*, so folgt *ii)* für alle Quader $B = [a_1, b_1] \times \cdots \times [a_k, b_k]$ und dann (nach dem Eindeutigkeitssatz für Maße) für alle $B \subset \mathbb{R}^k$ mit wohldefiniertem Inhalt.

Gilt umgekehrt *ii)*, so ergibt der Satz von Fubini

$$\mathbf{Ws}\{X_1 \in [a_1, b_1], \ldots, X_k \in [a_k, b_k]\} \;=\; \int\limits_{a_1}^{b_1} p_1(x_1)\,dx_1 \cdots \int\limits_{a_k}^{b_k} p_k(x_k)\,dx_k \;.$$

Wählen wir speziell $a_j = -\infty, b_j = \infty$ für alle $j \neq i$, so erkennt man, daß $p_i(x)\,dx$ die Dichte von X_i ist und daß folglich $X_1, \ldots, X_k$ unabhängig sind. □

Beispiel. Uniforme Verteilungen. Für eine Zufallsvariable $U = (U_1, \ldots, U_k)$ mit uniformer Verteilung auf der Menge $S \subset \mathbb{R}^k$ von Inhalt $|S|$ gilt für messbares $B \subset S$

$$\mathbf{Ws}\{U \in B\} \;= \frac{|B|}{|S|} \;=\; \int\limits_B \frac{1}{|S|}\,du_1 \ldots du_k \;.$$

U besitzt also auf S die Dichte $\frac{1}{|S|}\, du_1 \dots d_k$.

Ist insbesondere S ein achsenparalleles Quader, $S = [a_1, b_1] \times \cdots \times [a_k, b_k]$, so gilt

$$\mathbf{Ws}\{U \in du\} \;=\; \frac{1}{b_1 - a_1} \cdots \frac{1}{b_k - a_k}\, du_1 \dots du_k \;,$$

nach Proposition 2.7 ist also U genau dann uniform auf $[a_1, b_1] \times \cdots \times [a_k, b_k]$ verteilt, wenn $U_1, \dots, U_k$ unabhängige, uniform auf $[a_1, b_1]$, $\dots$, $[a_k, b_k]$ verteilte Zufallsvariable sind.

Seien speziell $U_1, \dots, U_k$ unabhängig und uniform auf $[0, 1]$ verteilt. Dann ist der Vektor $V = (U_{(1)}, \dots, U_{(k)})$ der Ordnungsstatistiken, der geordneten Werte von $U_1, \dots, U_k$, uniform verteilt auf dem ‚Simplex'

$$S_\Delta := \{(v_1, \dots, v_k) \in \mathbb{R}^k : 0 \le v_1 \le \cdots \le v_k \le 1\}$$

vom Inhalt $|S_\Delta| = 1/k!$. Es gilt nämlich für $B \subset S_\Delta$, indem wir über alle Permutationen π der Zahlen $1, \dots, k$ summieren,

$$\begin{aligned} \mathbf{Ws}\{(U_{(1)}, \dots, U_{(k)}) \in B\} \;&=\; \sum_\pi \mathbf{Ws}\{(U_{\pi(1)}, \dots, U_{\pi(k)}) \in B\} \\ &=\; k!\mathbf{Ws}\{(U_1, \dots, U_k) \in B\} \;=\; \frac{|B|}{|S_\Delta|} \;. \end{aligned}$$ □

Beispiel. Normalverteilte Zufallsvariable.

1. Nach der letzten Proposition sind $Z_1, \dots, Z_k$ genau dann unabhängig und standard normalverteilt, wenn $Z := (Z_1, \dots, Z_k)$ die Dichte

$$\mathbf{Ws}\{Z \in dz\} \;=\; \mathsf{n}(z)\, dz$$

hat, mit

$$\mathsf{n}(z_1, \dots, z_k) \;:=\; \prod_{i=1}^k \frac{1}{\sqrt{2\pi}} e^{-z_i^2/2} \;=\; (2\pi)^{-k/2} \exp\big(-|z|^2/2\big)$$

und $|z|^2 := z_1^2 + \cdots + z_k^2$. Diese Dichte ist wie die Euklidische Norm $|\cdot|$ invariant unter Drehungen um den Ursprung des $\mathbb{R}^k$. Nach Proposition 2.6 erkennt man daher: Sind $Z_1', \dots, Z_k'$ die Koordinaten von $Z = (Z_1, \dots, Z_k)$ in einem anderen orthonormalen Koordinatensystem des $\mathbb{R}^k$, so sind mit $Z_1, \dots, Z_k$ auch $Z_1', \dots, Z_k'$ unabhängig und standard normalverteilt.

Speziell sind mit Z_1, Z_2 auch

$$Z_1' = aZ_1 + bZ_2 \;,\; Z_2' = bZ_1 - aZ_2$$

unabhängig und standard normalverteilt, sofern $a^2 + b^2 = 1$, denn Z_1', Z_2' sind die Koordinaten von Z bezüglich der orthonormalen Basis (a, b), $(b, -a)$ des $\mathbb{R}^2$.

Diese fundamentalen Eigenschaften sind für die Normalverteilung charakteristisch. $\mathsf{n}(z)\,dz$ heißt **Dichte der multivariaten standard Normalverteilung**.

2. Wir betrachten nun Polarkoordinaten in der Ebene, genauer die Abbildungen

$$\phi(x,y) := \Big(\frac{x^2+y^2}{2}, \arctan\frac{y}{x}\Big) \quad , \quad \psi(z,u) := \big(\sqrt{2z}\cos u, \sqrt{2z}\sin u\big)$$

zwischen $\mathbb{R}^2 - \{0\}$ und $(0,\infty) \times [0, 2\pi)$. ψ überführt das Rechteck $[a,b] \times [c,d]$ in das Segment eines Kreisrings mit Innenradius $r_0 = \sqrt{2a}$ und Außenradius $r_1 = \sqrt{2b}$ zwischen den Winkeln c und d. Beide Flächen haben den Inhalt $(b-a)(d-c) = \frac{d-c}{2}(r_1^2 - r_0^2)$. Daher ist ψ und damit auch ϕ flächentreu.

Seien nun X, Y unabhängige, standard normalverteilte Zufallsvariable. Ihre gemeinsame Dichte ist dann $(2\pi)^{-1} e^{-(x^2+y^2)/2}\,dxdy$. Nach den Propositionen 2.6 und 2.7 haben daher

$$Z := \frac{X^2+Y^2}{2} \quad , \quad U := \arctan\frac{Y}{X}$$

die gemeinsame Dichte

$$q(z,u)\,dzdu = \frac{1}{2\pi} e^{-z}\,dzdu\ .$$

Nach Proposition 2.6 können wir folgern: Z und U sind stochastisch unabhängig, Z ist exponential verteilt mit Dichte $e^{-z}\,dz$ auf $\mathbb{R}_+$ und U ist uniform verteilt auf dem Intervall $[0, 2\pi)$.

Weiter ist $V := e^{-Z}$ uniform auf $[0,1]$ verteilt, denn für $0 \le a \le b \le 1$ folgt

$$\mathbf{Ws}\{a \le V \le b\} = \mathbf{Ws}\{-\ln b \le Z \le -\ln a\} = \int_{-\ln b}^{-\ln a} e^{-z}\,dz = b - a\ .$$

Man kann also unabhängige, standard normalverteilte Zufallsvariable X, Y gemäß

$$X = \sqrt{-2\ln V}\cos U \quad , \quad Y = \sqrt{-2\ln V}\sin U$$

erhalten, mit unabhängigen, uniform in $[0, 2\pi]$ bzw. $[0, 1]$ verteilten Zufallsvariablen U, V. Diesen Sachverhalt hat man auch für das Simulieren von $N(0, 1)$-verteilten Zufallsvariablen verwendet (Box-Muller Verfahren).

Bemerkung: Da sich für Z und U zwei Dichten ergeben, erkennt man, daß $\sqrt{2\pi}$ die richtige Normierungskonstante für die Normalverteilung ist. □

Beispiel. Falten von Dichten.

1. Seien X, Y unabhängige, reellwertige Zufallsvariable mit den Dichten $p_1(x)\,dx$ und $p_2(y)\,dy$. Dann hat $(Z, U) := (X + Y, Y)$ die Dichte $p_1(z - u)p_2(u)\,dzdu$, denn $\phi(x, y) := (x + y, y)$ mit Umkehrabbildung $\psi(z, u) = (z - u, u)$ ist flächentreu. Es folgt

$$\begin{aligned}\mathbf{Ws}\{a \le X + Y \le b\} &= \mathbf{Ws}\{\phi(X, Y) \in [a, b] \times \mathbb{R}\} \\ &= \iint\limits_{[a,b]\times\mathbb{R}} p_1(z - u)p_2(u)\,dzdu = \int_a^b \Big(\int_{-\infty}^{\infty} p_1(z - u)p_2(u)\,du \Big) dz\ ,\end{aligned}$$

also

$$\mathbf{Ws}\{X + Y \in dz\} = p_1 * p_2(z)\ dz$$

mit

$$p_1 * p_2(z) := \int_{-\infty}^{\infty} p_1(z - u)p_2(u)\,du\ .$$

$p_1 * p_2$ heißt **Faltung** von p_1 und p_2.

Also: Haben unabhängige, reellwertige Zufallsvariable X, Y die Dichten $p_1(x)\,dx$ und $p_2(y)\,dy$, so hat $Z := X + Y$ die Dichte $p_1 * p_2(z)\,dz$. Man vergleiche mit Formel (2.1).

2. **Gamma-Verteilungen.**

$$p_{\alpha,\lambda}(z)\,dz := c_{\alpha,\lambda} z^{\alpha-1} e^{-\lambda z}\,dz \quad , \quad z > 0$$

heißt **Dichte der $\Gamma(\alpha, \lambda)$-Verteilung** zu den Parametern $\alpha, \lambda > 0$, mit Konstanten $c_{\alpha,\lambda}$, die $p_{\alpha,\lambda}$ zur Dichtefunktion normieren. Es gilt

$$p_{\alpha,\lambda} * p_{\beta,\lambda} = p_{\alpha+\beta,\lambda}\ . \tag{2.3}$$

Denn

$$\begin{aligned}p_{\alpha,\lambda} * p_{\beta,\lambda}(z) &= c_{\alpha,\lambda} c_{\beta,\lambda} \int_0^z u^{\alpha-1} e^{-\lambda u} (z - u)^{\beta-1} e^{-\lambda(z-u)}\,du \\ &= c z^{\alpha+\beta-1} e^{-\lambda z}\end{aligned}$$

mit $c := c_{\alpha,\lambda} c_{\beta,\lambda} \int_0^1 t^{\alpha-1}(1 - t)^{\beta-1}\,dt$. Da die Faltung zweier Dichtefunktionen wieder eine Dichtefunktion ergibt, folgt $c = c_{\alpha+\beta,\lambda}$. □

Bemerkung. Ist die Koordinatentransformation $\phi : S \to S'$ nicht länger inhaltstreu, so muss man die (lokalen) Verzerrungen des Rauminhalts durch ϕ in der Transformationsformel berücksichtigen. Dann ist die Dichte in Proposition 2.6 zu ersetzen durch

$$\mathbf{Ws}\{W \in dw\} \;=\; p(\psi(w))|\psi'|(w)\, dw \;,$$

dabei bezeichnet $|\psi'|$ den Absolutbetrag der Funktionaldeterminante von ψ. Insbesondere sind ϕ und ψ genau dann inhaltstreu, wenn $|\psi'|$ identisch 1 ist, oder äquivalent, wenn $|\phi'|$ identisch 1 ist. □

2.4 Der Poisson-Prozeß*

Der **Poisson-Prozeß** gehört zu den grundlegenden Modellen der Stochastik. Es handelt sich um ein Modell für eine *zufällige diskrete Punktmenge* N (eine zufällige Menge ohne Häufungspunkte), enthalten in $\mathbb{R}$ oder einem Teilintervall von $\mathbb{R}$, das explizite Rechnungen erlaubt.

In Anwendungen repräsentieren die Punkte häufig Zeitpunkte von Geschehnissen. In der Warteschlangentheorie denkt man etwa an die Ankunftszeiten von Kunden an einem Schalter. Beim radioaktiven Zerfall bilden die Zeitpunkte, zu denen Teilchen zerfallen, einen Poisson-Prozeß.

Sei $N(J)$ die Anzahl der Punkte aus N, die im Intervall $J \subset \mathbb{R}$ enthalten sind, der ‚Zuwachs' des Prozesses in J. Die folgenden Eigenschaften sind charakteristisch für einen (stationären) Poisson-Prozeß.

P1. **Unabhängigkeit der Zuwächse.** Für disjunkte Intervalle $J_1, \ldots, J_k$ sind $N(J_1), \ldots, N(J_k)$ unabhängige Zufallsvariable.

P2. **Stationarität der Zuwächse.** $N(J)$ hat eine Verteilung, die nur von $|J|$, der Länge von J abhängig ist.

Es stellt sich heraus, daß für $N(J)$ als Verteilung nur eine Poisson-Verteilung in Frage kommt.

Proposition 2.8. *Es gibt eine reelle Zahl $\lambda \geq 0$, so daß $N(J)$ für alle Intervalle J Poisson-verteilt zum Parameter $\lambda \cdot |J|$ ist.*

λ heißt die **Rate** des Prozesses, sie ist die erwartete Anzahl von Punkten zwischen 0 und 1.

Beweis. Wir benutzen die Poisson-Approximation der Binomialverteilung und betrachten zunächst das Intervall $J = [0, 1]$, das wir in n disjunkte Intervalle $J_1, \ldots, J_n$ der Länge $1/n$ zerlegen. Sei

$$Y_n \;:=\; I_{\{N(J_1)>0\}} + \cdots + I_{\{N(J_n)>0\}}$$

die Anzahl der Intervalle $J_1, \ldots, J_n$, die mindestens einen Punkt aus N enthalten. Nach den Annahmen P1 und P2 ist dann Y_n binomialverteilt zum Parameter (n, p_n), mit

$$p_n \;:=\; \mathbf{Ws}\{N([0, n^{-1}]) > 0\} \;.$$

Insbesondere gilt $\{Y_n = 0\} = \{N([0,1]) = 0\} = \{Y_1 = 0\}$ und folglich

$$(1 - p_n)^n \;=\; 1 - p_1 \;=\; \exp(-\lambda) \qquad \text{mit} \quad \lambda := -\ln(1 - p_1) \;. \tag{2.4}$$

Die Möglichkeit $p_1 = 1$ können wir sofort ausschließen. Dann folgt nämlich $p_n = 1$, so daß N für jedes n mit Wahrscheinlichkeit 1 mindestens n Punkte in [0,1] enthält, und es ergibt sich der Widerspruch $\mathbf{Ws}\{N([0,1]) = \infty\} = 1$. Daher gilt $p_1 < 1$ und $0 \le \lambda < \infty$. Indem wir nun n gegen ∞ gehen lassen, ergibt sich aus (2.4) wegen $(1 - t/n)^n \to \exp(-t)$

$$p_n \sim \lambda/n \;.$$

Nach Satz 1.1 ist daher Y_n asymptotisch Poisson-verteilt zum Parameter λ.

Weiter gilt

$$\begin{aligned} &|\mathbf{Ws}\{N([0,1]) = x\} - \mathbf{Ws}\{Y_n = x\}| \\ &\qquad \le\; \mathbf{Ws}\{Y_n \neq N([0,1])\} \;\le\; \mathbf{Ws}\{D \le n^{-1}\} \;, \end{aligned}$$

dabei bezeichne die Zufallsvariable D die minimale Distanz zwischen den Punkten von N, die in $[0,1]$ liegen. Da $\{D \le n^{-1}\} \downarrow \{D = 0\} = \Phi$, folgt nach der σ-Stetigkeit von W-Maßen $\mathbf{Ws}\{D \le n^{-1}\} \to 0$ und folglich

$$\mathbf{Ws}\{Y_n = x\} \;\to\; \mathbf{Ws}\{N([0,1]) = x\} \;,$$

und $N([0,1])$ ist wie behauptet Poisson-verteilt. Genauso zeigt man, daß alle $N(J)$ Poisson-verteilte Zufallsvariable sind. Da sich, wie wir gesehen haben, bei der Summation von unabhängigen Poisson-verteilten Zufallsvariablen die Parameter addieren, hat $N(J)$ für $J = [0, \frac{1}{n}]$ den Parameter $\lambda\frac{1}{n}$ und für $J = [0, \frac{m}{n}]$ den Parameter $\lambda\frac{m}{n}$. Durch ein Stetigkeitsärgument folgt schließlich die Behauptung für alle Intervalle J. □

Aber existieren Poisson-Prozesse als mathematische Objekte überhaupt? Das wird aus einer 2-stufigen Konstruktion eines Poisson-Prozesses in einem endlichen Intervall klar: Erst legt man fest, wieviele Punkte insgesamt in das Intervall kommen - sie ist nach der Proposition Poisson-verteilt -, anschließend verteilt man diese Punkte uniform und unabhängig voneinander im Intervall. Wir führen dies genauer aus.

Konstruktion I: Ein Poisson-Prozeß im Intervall $[a, b]$. Wir erzeugen zuerst die gesamte Anzahl $\widetilde{N} = N([a, b])$ aller Punkte in $[a, b]$ und verteilen dann $\widetilde{N}$ Punkte unabhängig und gleichförmig auf $[a, b]$. Seien also $U_1, U_2, \ldots$ unabhängige Zufallsvariable, uniform verteilt in $[a, b]$, und sei $\widetilde{N}$ eine davon unabhängige, Poisson-verteilte Zufallsvariable zum Parameter $\lambda(b-a)$. Setze

$$N \;:=\; \{U_1, U_2, \ldots, U_{\widetilde{N}}\}\ ,$$

also für $J \subset [a, b]$

$$N(J) \;=\; \mathrm{card}\{i \le \widetilde{N} : U_i \in J\}\ .$$

Dann erfüllt N die an einen Poisson-Prozeß gestellten Bedingungen. Für disjunkte Intervalle $J_1, \ldots, J_k$ mit $J_1 \cup \ldots \cup J_k = [a, b]$, ganze Zahlen $x_1, \ldots, x_k \ge 0$ und $n = x_1 + \cdots + x_k$ gilt nämlich

$$\begin{aligned}
&\mathbf{Ws}\{N(J_1) = x_1, \ldots, N(J_k) = x_k\} \\
&\quad= \mathbf{Ws}\{\widetilde{N} = n\} \cdot \binom{n}{x_1, \ldots, x_k} \cdot \mathbf{Ws}\{U \in J_1\}^{x_1} \cdots \mathbf{Ws}\{U \in J_k\}^{x_k} \\
&\quad= e^{-\lambda(b-a)} \frac{(\lambda(b-a))^n}{n!} \cdot \binom{n}{x_1, \ldots, x_k} \cdot \Big(\frac{|J_1|}{b-a}\Big)^{x_1} \cdots \Big(\frac{|J_k|}{b-a}\Big)^{x_k} \\
&\quad= e^{-\lambda|J_1|} \frac{(\lambda|J_1|)^{x_1}}{x_1!} \cdots e^{-\lambda|J_k|} \frac{(\lambda|J_k|)^{x_k}}{x_k!}\ .
\end{aligned}$$

Nach Proposition 2.2 sind daher $N(J_1), \ldots, N(J_k)$ unabhängig und Poisson-verteilt zu den geforderten Parametern. □

Bemerkung. Diese Konstruktion funktioniert auch allgemeiner. Mit ihr kann man räumlich verteilte zufällige Punktmengen in einer Teilmenge S des $\mathbb{R}^d$ erzeugen. Dazu sei $\widetilde{N}$ gemäß $P(\lambda)$-verteilt und $U_1, U_2, \ldots$ unabhängige, identisch verteilte Zufallsvariable mit Werten in S und der Verteilung μ. Wir nehmen an, daß μ keine Atome hat, d.h. $\mathbf{Ws}\{U_1 = x\} = 0$ für alle $x \in S$ gilt. Dann fallen mit Wahrscheinlichkeit 1 keine zwei Punkte aufeinander. Die soeben durchgeführte Rechnung zeigt, daß die Anzahl der Punkte in $B \subset S$ Poisson-verteilt zum Parameter $\nu(B) := \lambda\mu(B)$ ist, und daß die Punktzahlen in disjunkten Teilmengen stochastisch unabhängig sind. Man spricht von einem **Poissonschen Punktprozeß** mit **Intensitätsmaß $\boldsymbol{\nu}$**. Poissonsche Punktprozesse hat man beispielsweise als Beschreibung dafür benutzt, wie sich einzelne Bakterienkolonien auf einem Nährboden (einer Petriplatte) verteilen. □

Es folgt eine Konstruktion für einen Poisson-Prozesses in $\mathbb{R}_+$. Sie wird sich als Variante von Konstruktion I erweisen, gleichwohl erlaubt sie tieferen Einblick in die Eigenschaften von Poisson-Prozessen.

Konstruktion II: Ein Poisson-Prozeß in $\mathbb{R}_+$. Seien $X_1, X_2, \ldots$ unabhängige, exponential verteilte Zufallsvariable zum Parameter λ. Setze

$$T_n \;:=\; X_1 + \cdots + X_n \;.$$

Dann ist

$$N \;:=\; \{T_1, T_2, \ldots\}$$

ein Poisson-Prozeß auf $\mathbb{R}_+$ mit Intensitätsrate λ. Wir zeigen dies, indem wir diese Konstruktion auf Konstruktion I zurückführen.

Dazu stellen wir fest, daß $X_1, \ldots, X_{k+1}$ nach Proposition 2.7 auf $\mathbb{R}^{k+1}$ die gemeinsame Dichte $\lambda e^{-\lambda x_1} \cdots \lambda e^{-\lambda x_{k+1}}\, dx_1 \ldots dx_{k+1}$ hat. Bei der Abbildung $\phi(x_1, \ldots, x_{k+1}) := (x_1, x_1 + x_2, \ldots, x_1 + \cdots + x_{k+1})$ von $\mathbb{R}^{k+1}$ nach $S_\le := \{(t_1, \ldots, t_{k+1}) \in \mathbb{R}^{k+1} : 0 \le t_1 \le \cdots \le t_{k+1}\}$ handelt es sich um eine flächentreue Abbildung. Nach Proposition 2.6 hat daher $T_1, \ldots, T_{k+1}$ die gemeinsame Dichte $\lambda^{k+1} e^{-\lambda t_{k+1}}\, dt_1 \ldots dt_{k+1}$ auf $S_\le$. Es folgt für messbare Teilmengen B von $S_\Delta := \{(t_1, \ldots, t_k) \in \mathbb{R}^k : 0 \le t_1 \le \cdots \le t_k \le 1\}$

$$\begin{aligned}\mathbf{Ws}\{N([0,1]) = k, (T_1, \ldots, T_k) \in B\} \;&=\; \mathbf{Ws}\{(T_1, \ldots, T_k) \in B, T_{k+1} > 1\} \\ &= \int\limits_{B \times (1,\infty)} \lambda^{k+1} e^{-\lambda t_{k+1}}\, dt_1 \ldots dt_{k+1} \;=\; \lambda^k e^{-\lambda} \int\limits_B dt_1 \ldots dt_k \;.\end{aligned}$$

Wir haben früher festgestellt, daß die Ordnungsstatistiken $U_{(1)} \le \cdots \le U_{(k)}$ von unabhängigen, uniform auf $[0,1]$ verteilten Zufallsvariablen als gemeinsame Verteilung die uniforme Verteilung auf S_Δ mit Inhalt $|S_\Delta| = 1/k!$ haben. Deswegen folgt insgesamt

$$\mathbf{Ws}\{N([0,1]) = k, (T_1, \ldots, T_k) \in B\} \;=\; e^{-\lambda} \frac{\lambda^k}{k!} \mathbf{Ws}\{(U_{(1)}, \ldots, U_{(k)}) \in B\} \;.$$

Dies bedeutet: Unsere Konstruktion leistet im Intervall $[0,1]$ dasselbe wie Konstruktion I. Die Anzahl der Punkte in $[0,1]$ ist Poisson-verteilt zum Parameter λ, und sie nehmen ihre Werte wie unabhängige, uniform verteilte Zufallsvariable, denn für die Punkte in N macht es keinen Unterschied, ob sie wie unabhängige Zufallsvariable oder wie deren Ordnungsstatistiken verteilt werden.

Diese Überlegung überträgt sich auf jedes Intervall $[0,b]$, daher leistet Konstruktion II dasselbe wie Konstruktion I. □

Bemerkung. Für die Verteilung von T_k erhalten wir

$$\mathbf{Ws}\{T_k \le t\} \;=\; \mathbf{Ws}\{N([0,t]) \ge k\} \;=\; 1 - \sum_{i=0}^{k-1} e^{\lambda t}\frac{(\lambda t)^i}{i!}\ .$$

Der Differentialquotient der rechten Seite errechnet sich als $p_{k,\lambda} = \lambda^k t^{k-1} e^{-\lambda t}/(k-1)!$, und es folgt

$$\mathbf{Ws}\{T_k \le t\} \;=\; \frac{1}{(k-1)!}\int_0^t \lambda^k s^{k-1} e^{-\lambda s}\, ds\ .$$

T_k ist also Gamma-verteilt zu den Parametern k, λ. Man erhält dies auch aus der Faltungsformel (2.3) und der Tatsache, daß T_k Summe der unabhängigen, exponential verteilten Zufallsvariablen $X_1, \ldots, X_k$ ist. □

Zum Abschluß untersuchen wir mittels Konstruktion II einen Poisson-Prozess N auf $\mathbb{R}$. Dazu wählen wir einen Bezugspunkt $a \in \mathbb{R}$ und zerlegen N in

$$N_+ := N \cap [a,\infty) \quad , \quad N_- := N \cap (-\infty, a]\ .$$

Da N mit Wahrscheinlichkeit 1 nicht a enthält (d.h. keinen Punkt im Intervall $[a,a]$ besitzt), folgt

$$N \;= N_+ \cup N_- \quad \text{mit Wahrscheinlichkeit 1}\ .$$

N_+ und analog N_- sind unabhängige Poisson-Prozesse auf reellen Halbachsen, wie sie in Konstruktion II beschrieben sind. Wir schreiben sie deswegen als

$$N_+ \;=\; \{T_1, T_2, \ldots\} \quad , \quad N_- \;=\; \{T_{-1}, T_{-2}, \ldots\}$$

mit $a \le T_1 < T_2 < \cdots$ und $a \ge T_{-1} > T_{-2} > \cdots$. Setzen wir noch $T_0 := a$, so erkennen wir aus Konstruktion II, daß die Zufallsvariablen $T_i - T_{i-1}$, $i \in \mathbb{Z}$ unabhängige, exponential verteilte Zufallsvariable zum Parameter λ sind.

Beachtenswert ist der Sachverhalt, daß der Abstand zwischen zwei Punkten in N nicht überall exponential verteilt ist: Das Intervall $[T_{-1}, T_1]$, welches a überdeckt, hat eine Länge, die wie die Summe von zwei unabhängigen, exponentialen Zufallsvariablen verteilt ist. Sie stimmt in Verteilung mit $T_2 - T_0$ überein, hat also gemäß der vorangehenden Bemerkung die Dichte

$$\mathbf{Ws}\{T_1 - T_{-1} \in dt\} \;=\; \lambda^2 t e^{-\lambda t}\, dt\ .$$

Man sagt, die Dichte ist durch **Größenverzerrung** aus der Exponentialverteilung entstanden.

Paradox erscheint, daß sich das größenverzerrte Intervall $[T_{-1}, T_1]$ in anderer Lage findet, wenn wir den Referenzpunkt a verändern. Dieses Phänomen hat aber eine gute Erklärung. Der Zufall produziert kurze und lange Intervalle, und die langen Intervalle haben eine größere Chance, a zu überdecken als die kurzen Intervalle. Man stelle sich etwa vor, daß die T_i die Zeitpunkte sind, zu denen die Glühbirne in einer Lampe ersetzt werden. Wenn man zu einem gewissen Zeitpunkt a nach der Brenndauer der gerade leuchtenden Glühbirne schaut, so ist die Chance gering, eine Birne mit besonders kurzer Brenndauer vorzufinden.

Das Phänomen ist auch unter dem Namen **Wartezeitparadox** bekannt: Fahren die Busse einer Linie im festen Takt mit gleichlangen Zeitabständen t, so muß man an einer Haltestelle bei zufälliger Ankunftszeit im Mittel nur $t/2$ Zeiteinheiten auf den nächsten Bus warten. Folgen die Ankunftszeiten dagegen einem Poisson-Prozeß, so ist die Wartezeit auf den nächsten Bus exponential verteilt, sie wird im Mittel nicht kürzer.

Kapitel 3

Erwartungswert und Varianz

Wir behandeln nun reellwertige Zufallsvariable. Erwartungswert und Varianz sind wichtige Kenngrößen ihrer Verteilungen. Beim Erwartungswert spricht man auch vom mittleren Wert der Zufallsvariablen, die Varianz ist die mittlere quadratische Abweichung der Zufallsvariablen von ihrem Erwartungswert. Die Bedeutung der beiden Größen ergibt sich aus ihren günstigen Eigenschaften als lineares bzw. quadratisches Funktional, deswegen lassen sich Erwartungswerte und Varianzen häufig explizit berechnen. Darüber hinaus spielen sie für theoretische Untersuchungen eine zentrale Rolle. Wir werden sie in diesem Abschnitt benutzen, um *Gesetze der großen Zahlen* abzuleiten.

Da wir uns auf diskrete Zufallsvariable beschränken, sind die Beweise elementar. Wie die Integrationstheorie lehrt, bleiben die Resultate über den diskreten Fall hinaus allgemein für reellwertige Zufallsvariable gültig.

3.1 Der Erwartungswert

Definition. *Sei X eine Zufallsvariable, deren Wertebereich S aus abzählbar vielen reellen Zahlen besteht. Dann ist ihr* ***Erwartungswert*** *definiert als*

$$\mathbf{E}[X] \;:=\; \sum_{x\in S} x\cdot \mathbf{Ws}\{X=x\}\ ,$$

vorausgesetzt, die Summe ist wohldefiniert in dem Sinne, daß sie von der Summationsreihenfolge unabhängig ist (∞ ist als Summationswert zugelassen). Wir schreiben für den Erwartungswert auch kurz $\mathbf{E}X$.

Diese völlig einleuchtende Voraussetzung wird später wichtig, wenn wir (stillschweigend) verschiedene Umordnungen der Summationsreihenfolge vornehmen werden.

Erläuterungen.

1. Für den Erwartungswert der Indikatorvariablen des Ereignisses A gilt

$$\mathbf{E}[I_A] \;=\; \mathbf{Ws}\{A\}\ .$$

2. Für diskrete Zufallsvariable $X_1, \ldots, X_n$ und eine Abbildung $\phi(x_1, \ldots, x_n)$ von den Wertebereichen der Zuvallsvariablen in die reellen Zahlen gilt die **Transformationsformel**

$$\mathbf{E}[\phi(X_1, \ldots, X_n)] \;=\; \sum_{x_1,\ldots,x_n} \phi(x_1, \ldots, x_n) \cdot \mathbf{Ws}\{X_1 = x_1, \ldots, X_n = x_n\},$$

vorausgesetzt, der Wert der Summe hängt nicht von der Summationsreihenfolge ab. Denn

$$\begin{aligned}
&\sum_x x \cdot \mathbf{Ws}\{\phi(X_1, \ldots, X_n) = x\} \\
&\qquad = \sum_x x \sum_{\phi(x_1,\ldots,x_n)=x} \mathbf{Ws}\{X_1 = x_1, \ldots, X_n = x_n\} \\
&\qquad = \sum_{x_1,\ldots,x_n} \phi(x_1, \ldots, x_n) \cdot \mathbf{Ws}\{X_1 = x_1, \ldots, X_n = x_n\}\ .
\end{aligned}$$

3. Liegt der Wertebereich von X in $\mathbb{R}_+$, d.h. gilt $X \geq 0$, so ist die Summationsreihenfolge ohne Auswirkung auf den Wert der Summe. Nichtnegative Zufallsvariable haben immer eine wohldefinierte Erwartung, deren Wert möglicherweise ∞ ist.

4. X hat einen endlichen Erwartungswert genau dann, wenn $\mathbf{E}|X| < \infty$ ist, und dann gilt

$$|\mathbf{E}X| \;\leq\; \mathbf{E}|X|\ .$$

Denn: $\sum x \cdot \mathbf{Ws}\{X = x\}$ ist genau dann endlich und unabhängig von der Summationsreihenfolge, wenn die Reihe absolut konvergiert, d.h. $\mathbf{E}|X| = \sum |x| \cdot \mathbf{Ws}\{X = x\}$ einen endlichen Wert hat. Die zweite Behauptung folgt aus $|\sum x \cdot \mathbf{Ws}\{X = x\}| \leq \sum |x| \cdot \mathbf{Ws}\{X = x\}$.

5. Für eine Zufallsvariable X mit Werten in $\mathbb{N}_0$ ist manchmal folgende Formel nützlich,

$$\mathbf{E}X \;=\; \sum_{t=1}^{\infty} \mathbf{Ws}\{X \geq t\}\ .$$

Sie folgt aus

$$\sum_{x=0}^{\infty} x \cdot \mathbf{Ws}\{X = x\}$$
$$= \sum_{x=1}^{\infty}\sum_{t=1}^{x} \mathbf{Ws}\{X = x\} = \sum_{t=1}^{\infty}\sum_{x=t}^{\infty} \mathbf{Ws}\{X = x\} .$$

□

Wir kommen nun zu zwei grundlegenden Eigenschaften des Erwartungswertes. Die erste, die Monotonie, erscheint offensichtlich.

Satz 3.1. *X und Y seien reellwertige Zufallsvariable mit wohldefinierten Erwartungswerten. Gilt dann $X \le Y$, so folgt*

$$\mathbf{E}X \le \mathbf{E}Y .$$

Beweis. Aufgrund der σ-Additivität gilt

$$\mathbf{Ws}\{X = x\} = \sum_{y:y\ge x} \mathbf{Ws}\{X = x, Y = y\} , \tag{3.1}$$

und eine analoge Formel für $\mathbf{Ws}\{Y = y\}$. Nach Voraussetzung ist $\{X \le Y\}$ das sichere Ereignis, so daß $\{X = x, Y = y\}$ für alle Paare $y < x$ Wahrscheinlichkeit 0 hat. Damit folgt

$$\sum_x x \cdot \mathbf{Ws}\{X = x\} = \sum_{y\le x} x \cdot \mathbf{Ws}\{X = x, Y = y\}$$
$$\le \sum_{x\le y} y \cdot \mathbf{Ws}\{X = x, Y = y\} = \sum_y y \cdot \mathbf{Ws}\{Y = y\} .$$

□

Grundlegend ist die Linearität des Erwartungswertes, wie sie im folgenden Satz erfasst ist.

Satz 3.2. *Seien X und Y reellwertige Zufallsvariable und $\lambda, \mu \in \mathbb{R}$. Falls entweder $X, Y \ge 0$ und $\lambda, \mu \ge 0$, oder falls X, Y endliche Erwartungswerte haben, dann hat auch $\lambda X + \mu Y$ einen wohldefinierten Erwartungswert und es gilt*

$$\mathbf{E}[\lambda X + \mu Y] = \lambda \mathbf{E}X + \mu \mathbf{E}Y .$$

Beweis. Aufgrund von (3.1) und der analogen Formel für $\mathbf{Ws}\{Y = y\}$ können wir folgende Rechnung durchführen:

$$\begin{aligned}
\mathbf{E}[\lambda X + \mu Y] &= \sum_{x,y} (\lambda x + \mu y) \cdot \mathbf{Ws}\{X = x, Y = y\} \\
&= \lambda \sum_x x \sum_y \mathbf{Ws}\{X = x, Y = y\} \\
&\quad + \mu \sum_y y \sum_x \mathbf{Ws}\{X = x, Y = y\} \\
&= \lambda \sum_x x \cdot \mathbf{Ws}\{X = x\} + \mu \sum_y y \cdot \mathbf{Ws}\{Y = y\} \\
&= \lambda \cdot \mathbf{E}X + \mu \cdot \mathbf{E}Y \ .
\end{aligned}$$

Es bleibt zu begründen, dass die dabei vorgenommenen Vertauschungen der Summationsreihenfolge zulässig sind. Dies ist im Fall $X, Y \geq 0$ und $\lambda, \mu \geq 0$ klar, da dann alle Summanden nichtnegativ sind. Für den zweiten Fall erhalten wir damit unter zusätzlicher Beachtung von Satz 3.1 $\mathbf{E}|\lambda X + \mu Y| \leq \mathbf{E}[|\lambda||X| + |\mu||Y|] = |\lambda|\mathbf{E}|X| + |\mu|\mathbf{E}|Y| < \infty$. Daher hat $\lambda X + \mu Y$ einen endlichen Erwartungswert und alle Summen sind absolut konvergent, so dass die Reihenumordnungen vorgenommen werden dürfen. □

Insbesondere gilt für reelles λ

$$\mathbf{E}[X + \lambda] = \mathbf{E}X + \lambda \ ,$$

man sagt, der Erwartungswert ist ein **Lageparameter**.

Die Bedeutung der Linearität des Erwartungswerts kann man nicht überschätzen. Sie erlaubt häufig, für einen Erwartungswert einen übersichtlichen Ausdruck zu finden, auch in Fällen, in denen die Verteilung nicht explizit vorliegt. Dies illustrieren die folgenden Beispiele.

Beispiele.

1. **Binomialverteilung.** Seien $Z_1, \ldots, Z_n$ unabhängige Bernoulli-verteilte Zufallsvariable zur Erfolgswahrscheinlichkeit p, dann ist

   $$X = Z_1 + \cdots + Z_n$$

   binomialverteilt zum Parameter (n, p). (Dies entspricht der Vorstellung einer binomialverteilten Zufallsvariablen als Anzahl der Erfolge bei n-facher

unabhängiger Wiederholung eines Zufallsexperiments mit Erfolgswahrscheinlichkeit p, $Z_i = 1$ bedeutet Erfolg im i-ten Experiment.) Z_i hat Erwartungswert p, daher folgt nach Satz 3.2 die uns schon bekannte Formel für den **Erwartungswert einer Binomialverteilung**

$$\mathbf{E}X \;=\; np\;.$$

2. **Runs.** Die Anzahl Y der Runs, der Maximalserien aus Nullen oder aus Einsen in einer Folge von 01-wertigen Zufallsvariablen $Z_0, Z_1, \ldots, Z_n$ haben wir bereits in einem Beispiel von Abschnitt 1.2 betrachtet. Sie läßt sich mit Indikatorvariablen darstellen als

$$Y \;=\; 1 + \sum_{i=1}^{n} I_{\{Z_i \neq Z_{i-1}\}}\;,$$

für den Erwartungswert gilt daher

$$\mathbf{E}Y \;=\; 1 + \sum_{i=1}^{n} \mathbf{Ws}\{Z_i \neq Z_{i-1}\}\;.$$

Sind die Z_i unabhängig und Bernoulli-verteilt mit Erfolgswahrscheinlichkeit p, so folgt

$$\mathbf{E}Y \;=\; 1 + 2pqn\;.$$

3. **Randomisierter Quicksort.** Verschiedene deterministische Algorithmen benötigen bei ungünstigen Eingabedaten unverhältnismäßig lange Laufzeiten. Manchmal kann man diese Schwäche kurieren, indem man in die Algorithmen ein Element des Zufalls einfügt. Die folgende randomisierte Version des Sortieralgorithmus Quicksort hat eine zufällige Laufzeit, deren Verteilung unabhängig von der Eingabe ist.

   Eingabe: Eine Menge $S = \{s_1, \ldots, s_n\}$ von n verschiedenen Zahlen.
   Ausgabe: Die Elemente $s_{(1)} < \cdots < s_{(n)}$ von S in aufsteigender Ordnung.

   1. Wähle rein zufällig ein Element Y aus S.
   2. Bestimme die Mengen $S_<$ und $S_>$, bestehend aus den Elementen kleiner bzw. größer als Y.
   3. Verfahre mit $S_<$ und $S_>$ analog etc., bis S in lauter einpunktige Mengen zerfällt, die man dann mühelos anordnen kann.

   Wie es bei Sortieralgorithmen üblich ist, nehmen wir die Gesamtanzahl der durchgeführten Vergleiche als Maß für die Laufzeit. X_n bezeichne die

Anzahl der von Quicksort durchgeführten Vergleiche. Sie liegt zwischen $n \log_2 n$ (falls stets der Median der zu sortierenden Mengen gewählt wird) und $n^2/2$ (falls der Zufall als Vergleichselement stets das kleinste oder größte Element auswählt). Mit solch ungünstigem Verhalten muß man aber nicht rechnen, die mittlere Laufzeit ist gegeben durch

$$\mathbf{E}X_n \;=\; 2(n+1)\big(1+\tfrac{1}{2}+\tfrac{1}{3}+\cdots+\tfrac{1}{n}\big) - 4n \;=\; 2n\ln n + O(n)\;.$$

Wegen $2\ln n \simeq 1,39\log_2 n$ muß man also im Mittel nur 39% mehr Vergleiche vornehmen als im günstigsten Fall.

Zum Beweis schreiben wir X_n als Summe von Indikatorvariablen,

$$X_n \;=\; \sum_{1\le i<j\le n} I_{A_{ij}}\;,$$

wobei A_{ij} das Ereignis sei, daß es zum Vergleich kommt zwischen den Elementen $s_{(i)}$ und $s_{(j)}$ von S. Es gilt

$$\mathbf{Ws}\{A_{ij}\} \;=\; \frac{2}{j-i+1}\;,\quad i<j\;,$$

denn A_{ij} ist das Ereignis, daß entweder $s_{(i)}$ oder $s_{(j)}$ als erstes Vergleichselement unter den $s_{(i)}, s_{(i+1)}, \ldots, s_{(j)}$ ausgewählt wird, und jedes dieser $j-i+1$ Elemente wird mit gleicher Wahrscheinlichkeit zuerst ausgewählt. Es folgt

$$\begin{aligned}\mathbf{E}X_n \;&=\; \sum_{i<j}\mathbf{Ws}\{A_{ij}\} \;=\; 2\sum_{j=2}^{n}\sum_{i=1}^{j-1}\frac{1}{j-i+1}\\ &=\; 2\sum_{j=2}^{n}\sum_{i=2}^{j}\frac{1}{i} \;=\; 2\sum_{i=2}^{n}\sum_{j=i}^{n}\frac{1}{i}\\ &=\; 2\sum_{i=2}^{n}\frac{n-i+1}{i} \;=\; 2(n+1)\sum_{i=2}^{n}\frac{1}{i} - 2(n-1)\;.\end{aligned}$$

Dies ergibt den ersten Teil der Behauptung, der zweite folgt aus der bekannten Asymptotik $1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{n} = \ln n + O(1)$.

4. **Austauschbare Verteilungen.** Wir zeigen nun, wie man Symmetrieeigenschaften von Verteilungen bei der Berechnung von Erwartungswerten ausnutzen kann. Man sagt, diskrete Zufallsvariable $X_1, \ldots, X_n$ mit demselben Wertebereich S haben eine **austauschbare Verteilung**, falls sich der Wert von $\mathbf{Ws}\{X_1 = x_1, \ldots, X_n = x_n\}$ beim Vertauschen der $x_1, \ldots, x_n$

untereinander nicht ändert, oder, anders ausgedrückt, falls für beliebige Permutationen π der Zahlen $1, \ldots, n$ gilt

$$\mathbf{Ws}\{X_1 = x_1, \ldots, X_n = x_n\} \;=\; \mathbf{Ws}\{X_{\pi(1)} = x_1, \ldots, X_{\pi(n)} = x_n\} \;.$$

Summiert man über alle $x_2, \ldots, x_n \in S$, so folgt

$$\mathbf{Ws}\{X_1 = x_1\} \;=\; \mathbf{Ws}\{X_{\pi(1)} = x_1\} \;.$$

Da $\pi(1)$ alle Werte $1, \ldots, n$ annehmen kann, erkennt man, daß $X_1, \ldots, X_n$ identisch verteilt sind. Insbesondere haben $X_1, \ldots, X_n$ im reellwertigen Fall gleiche Erwartungswerte. Ähnliches gilt für Paare: Summiert man über alle $x_3, \ldots, x_n \in S$, so ergibt sich, daß (X_i, X_j) für $i \neq j$ in Verteilung mit (X_1, X_2) übereinstimmt. - Zwei Beispiele: Zufallsvariable $X_1, \ldots, X_n$ haben eine austauschbare Verteilung, wenn sie unabhängig und identisch verteilt sind oder aber wenn sie untereinander gleich sind.

Eine Anwendung: Eine Urne enthält t Kugeln, davon r rote und s schwarze $(r + s = t)$. Die Kugeln werden der Reihe nach ohne Zurücklegen gezogen. Wir wollen den Erwartungswert der Anzahl X_0 der roten Kugeln, die vor der ersten schwarzen Kugel erscheinen, berechnen. Dazu betrachten wir auch die Anzahl X_i der Kugeln zwischen der i-ten und $(i{+}1)$-ten schwarzen Kugel und die Anzahl X_s von Kugeln nach der letzten schwarzen Kugel. Es gilt

$$\mathbf{Ws}\{X_0 = x_0, \ldots, X_s = x_s\} \;=\; \begin{cases} \binom{t}{s}^{-1}, & \text{falls } x_i \in \mathbb{N}_0 \text{ und } \sum_i x_i = r \\ 0 & \text{sonst.} \end{cases}$$

$X_0, \ldots, X_s$ haben also eine austauschbare Verteilung. Es folgt $\mathbf{E}X_0 = \cdots = \mathbf{E}X_s$ und mittels Linearität wegen $X_0 + \cdots + X_s = r$

$$(s+1)\mathbf{E}X_0 \;=\; \mathbf{E}X_0 + \cdots + \mathbf{E}X_s \;=\; \mathbf{E}[X_0 + \cdots + X_s] \;=\; r \;,$$

also

$$\mathbf{E}X_0 \;=\; \frac{r}{s+1} \;.$$

Zum Vergleich: Zieht man die Stichprobe mit Zurücklegen, so ist X_0 geometrisch verteilt und

$$\mathbf{E}X_0 \;=\; \frac{t}{s} - 1 \;=\; \frac{r}{s} \;.$$

5. Die **Einschluß-Ausschluß-Formel** für Ereignisse $A_1, \ldots, A_n$ besagt

$$\begin{aligned}&\mathbf{Ws}\{A_1 \cup \cdots \cup A_n\} \\ &\quad = \sum_i \mathbf{Ws}\{A_i\} - \sum_{i<j} \mathbf{Ws}\{A_i \cap A_j\} + \cdots \pm \mathbf{Ws}\{A_1 \cap \cdots \cap A_n\} \ .\end{aligned}$$

Zum Beweis gehe man in der Identität für Indikatorvariablen

$$\begin{aligned}1 - I_{A_1 \cup \cdots \cup A_n} &= (1 - I_{A_1}) \cdots (1 - I_{A_n}) \\ &= 1 - \sum_i I_{A_i} + \sum_{i<j} I_{A_i} \cdot I_{A_j} - \cdots \\ &= 1 - \sum_i I_{A_i} + \sum_{i<j} I_{A_i \cap A_j} - \cdots\end{aligned}$$

zum Erwartungswert über. □

Durch Kombination von Linearität und Monotonie erhält man wichtige Ungleichungen. Die Cauchy-Schwarz-Ungleichung läßt sich aus der Ungleichung $(u-v)^2 \geq 0$ bzw. $2uv \leq u^2+v^2$, $u, v \in \mathbb{R}$ gewinnen. Für reellwertige Zufallsvariable X und Y und reelle Zahlen α und β gilt daher $2\alpha\beta XY \leq \alpha^2 X^2 + \beta^2 Y^2$. Setzen wir $\alpha^2 = \mathbf{E}[Y^2]$, $\beta^2 = \mathbf{E}[X^2]$ und gehen mit Satz 3.1 und Satz 3.2 zum Erwartungswert über, so folgt nach Kürzen von $2\alpha\beta$ die **Cauchy-Schwarz-Ungleichung**

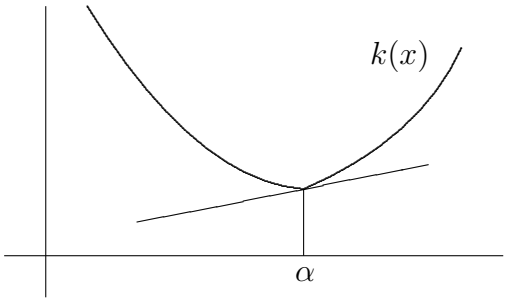
$$\mathbf{E}[XY]^2 \ \leq \ \mathbf{E}[X^2]\mathbf{E}[Y^2] \ .$$

Der Fall $\alpha = 0$ bzw. $\beta = 0$ bedarf einer gesonderten Betrachtung: Die Gleichung $\mathbf{E}X^2 = \sum x^2 \cdot \mathbf{Ws}\{X = x\} = 0$ impliziert $\mathbf{Ws}\{X = 0\} = 1$, und es folgt $\mathbf{Ws}\{XY = 0\} = 1$ und $\mathbf{E}[XY] = 0$.

Ähnlich wichtig ist die Jensensche Ungleichung für konvexe Funktionen. Konvexe Funktionen $k : \mathbb{R} \to \mathbb{R}$ sind dadurch charakterisiert, daß sie in jedem Punkt von unten durch eine Gerade gestützt werden können, d.h. daß es für alle $\alpha \in \mathbb{R}$ ein $\beta \in \mathbb{R}$ gibt, so daß für alle x

$$k(\alpha) + \beta(x - \alpha) \ \leq \ k(x) \ .$$

Die folgende Graphik zeigt, daß die stützende Gerade im allgemeinen keine Tangente von $k(x)$ zu sein braucht.



Für eine reellwertige Zufallsvariable X gilt folglich $k(\alpha)+\beta(X-\alpha) \le k(X)$. Nach Satz 3.1 folgt $k(\alpha)+\beta(\mathbf{E}X-\alpha) \le \mathbf{E}k(X)$. Wählen wir speziell $\alpha = \mathbf{E}X$ so erhalten wir die **Jensen-Ungleichung**

$$k(\mathbf{E}X) \;\le\; \mathbf{E}k(X)\ .$$

Wichtige Spezialfälle sind $|\mathbf{E}X| \le \mathbf{E}|X|$, $\mathbf{E}[X]^2 \le \mathbf{E}[X^2]$ und allgemeiner

$$\mathbf{E}[|X|^p]^{1/p} \;\le\; \mathbf{E}[|X|^q]^{1/q}$$

für $0 < p \le q$ (wähle $k(x) = |x|^{q/p}$ und ersetze X durch $|X|^p$).

Wir definieren nun noch den Erwartungswert für reellwertige Zufallsvariable mit einer Dichte.

Definition. *Für eine Zufallsvariable X mit Werten in $\mathbb{R}$ und der Dichte $\mathbf{Ws}\{X \in dx\} = f(x)\,dx$ definiert man den Erwartungswert als*

$$\mathbf{E}X \;:=\; \int_{-\infty}^{\infty} x \cdot f(x)\ dx,$$

vorausgesetzt, das Integral hat einen wohldefinierten Wert (d.h. $\int_0^\infty xf(x)\,dx$ und $\int_{-\infty}^0 xf(x)\,dx$ nehmen nicht gleichzeitig den Wert ∞ bzw. $-\infty$ an).

Beispiele.

1. Für die $N(\mu,\sigma)$-Verteilung ergibt sich als Erwartungswert μ,

$$\int_{-\infty}^{\infty} x \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big)\ dx \;=\; \mu.$$

   Dies folgt daraus, daß die Dichte symmetrisch um μ ist.

2. Die Exponential-Verteilung hat den Erwartungswert

$$\int_0^{\infty} x \cdot \lambda e^{-\lambda x}\ dx \;=\; \lambda^{-1}.$$

□

3.2 Die Varianz

Definition. *Sei X reellwertige, diskrete Zufallsvariable mit endlichem Erwartungswert. Dann ist die* ***Varianz*** *von X definiert als*

$$\mathbf{Var}[X] \;:=\; \mathbf{E}[(X-\mathbf{E}X)^2] \;=\; \sum_x (x-\mathbf{E}X)^2 \cdot \mathbf{Ws}\{X=x\}\ .$$

Wir schreiben auch kürzer $\mathbf{Var}X$*.*

Die Quadratwurzel $s(X) := (\mathbf{Var}X)^{1/2}$, die **Streuung** oder **Standardabweichung** von X, wird als Maßzahl für die mittlere Abweichung der Werte von X vom Erwartungswert benutzt. $s(X)$ ist ein **Skalenparameter**, d.h. für $\mu, \sigma \in \mathbb{R}$ gilt

$$s(\sigma X + \mu) \;=\; |\sigma| s(X)\ .$$

Die Varianz ist ein quadratisches Funktional. Das bedeutet, daß für das Rechnen mit Varianzen ein bilineares Funktional zur Verfügung steht. Dieses Funktional ist die Covarianz.

Definition. *Seien X und Y reellwertige Zufallsvariable mit endlichen Varianzen. Dann ist die* ***Covarianz*** *von X und Y gegeben durch*

$$\begin{aligned}\mathbf{Cov}[X,Y] \;&:=\; \mathbf{E}[(X-\mathbf{E}X)(Y-\mathbf{E}Y)] \\ &=\; \sum_{x,y}(x-\mathbf{E}X)(y-\mathbf{E}Y)\cdot \mathbf{Ws}\{X=x, Y=y\}\ .\end{aligned}$$

Die Covarianz ist wohldefiniert und endlich, wie sich aus der Cauchy-Schwarz-Ungleichung ergibt. Die Regeln für das Rechnen mit Varianzen und Covarianzen leiten sich aus der Linearität des Erwartungswertes ab, sie sind in der folgenden Proposition zusammengestellt.

Proposition 3.3. *Es gilt*

i) $\mathbf{Var}X \;=\; \mathbf{E}[X^2] - \mathbf{E}[X]^2\ , \quad \mathbf{Cov}[X,Y] \;=\; \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y]\ .$

ii) $\mathbf{Var}[\sigma X + \mu] \;=\; \sigma^2 \mathbf{Var}X \qquad$ *für alle* $\mu\ , \sigma \in \mathbb{R}$ *.*

iii) Die Covarianz ist symmetrisch, bilinear und nichtnegativ definit, d.h. es gilt

$$\begin{aligned}\mathbf{Cov}[X,Y] \;&=\; \mathbf{Cov}[Y,X]\ , \\ \mathbf{Cov}[\sigma X + \tau Y, Z] \;&=\; \sigma \mathbf{Cov}[X,Z] + \tau \mathbf{Cov}[Y,Z]\ , \\ \mathbf{Cov}[X,X] \;&=\; \mathbf{Var}X \;\geq 0\ .\end{aligned}$$

iv) $\mathbf{Cov}[X,Y]^2 \;\le\; \mathbf{Var}X\mathbf{Var}Y$.

v) $\mathbf{Var}[X+Y] \;=\; \mathbf{Var}X + \mathbf{Var}Y + 2\mathbf{Cov}[X,Y]$.

Beweis. i) - *iii)* sind Konsequenz von Satz 3.2, *iv)* ist eine Variante der Cauchy-Schwarz-Ungleichung. *v)* ergibt sich aus *iii)* unter Berücksichtigung von

$$\mathbf{Var}[X+Y] \;=\; \mathbf{Cov}[X,X] + \mathbf{Cov}[X,Y] + \mathbf{Cov}[Y,X] + \mathbf{Cov}[Y,Y] \ .$$

□

Die Formeln unter *i)* für Varianzen und Covarianzen erlauben eine zweite Lesart,

$$\mathbf{E}[X^2] \;=\; \mathbf{Var}X + \mathbf{E}[X]^2 \ , \quad \mathbf{E}[XY] \;=\; \mathbf{Cov}[X,Y] + \mathbf{E}[X]\mathbf{E}[Y] \ . \tag{3.2}$$

Sie sind hilfreich, wenn man $\mathbf{E}[X^2]$, das **2. Moment** von X, bzw. $\mathbf{E}[XY]$, das **gemischte Moment** von X und Y bestimmen möchte.

Der Fall, daß die Covarianz zweier Zufallsvariablen verschwindet, verdient besondere Beachtung.

Definition. *Zwei reellwertige Zufallsvariable X und Y von endlicher Varianz heißen* ***unkorreliert****, falls* $\mathbf{Cov}[X,Y] = 0$.

Für unkorrelierte Zufallsvariable wird das Rechnen mit Varianzen besonders übersichtlich. Sind $X_1, \ldots, X_n$ paarweise unkorreliert, so gilt

$$\mathbf{Var}\Big[\sum_{i=1}^n X_i\Big] \;=\; \sum_{i=1}^n \mathbf{Var}X_i \ . \tag{3.3}$$

Diese Gleichung ist insbesondere für unabhängige Zufallsvariable gültig.

Satz 3.4. *Sind X und Y stochastisch unabhängige, reellwertige Zufallsvariable mit endlichen Erwartungswerten, so gilt*

$$\mathbf{E}[XY] \;=\; \mathbf{E}[X]\mathbf{E}[Y] \ .$$

Im Fall endlicher Varianzen sind X und Y also unkorreliert.

Beweis.

$$\begin{aligned}\sum_{x,y} xy \cdot \mathbf{Ws}\{X=x, Y=y\} \;&=\; \sum_{x,y} x \cdot \mathbf{Ws}\{X=x\} \cdot y \cdot \mathbf{Ws}\{Y=y\} \\ &=\; \sum_x x \cdot \mathbf{Ws}\{X=x\} \cdot \sum_y y \cdot \mathbf{Ws}\{Y=y\} \ .\end{aligned}$$

□

Beispiele.

1. **Binomialverteilung.** Zur Berechnung der Varianz einer $\mathrm{B}(n,p)$-verteilten Zufallsvariablen X gehen wir wie in Abschnitt 3.1, Beispiel 1 von der Darstellung
$$X \;=\; Z_1 + \cdots + Z_n$$
aus, mit unabhängigen, $\mathrm{B}(p)$-verteilten Zufallsvariablen $Z_1, \ldots, Z_n$. Es gilt $\mathbf{Var}Z_i = \mathbf{E}Z_i^2 - (\mathbf{E}Z_i)^2 = p - p^2 = pq$, daher erhalten wir nach (3.3) für die **Varianz der Binomialverteilung** die schon in Abschnitt 1.4 angegebene Formel
$$\mathbf{Var}X \;=\; npq\ .$$

2. **Hypergeometrische Verteilung.** Nun sei
$$X \;=\; Z_1 + \cdots + Z_n$$
die Anzahl der roten Kugeln in einer Stichprobe vom Umfang n, die einer Urne mit r roten und s schwarzen Kugeln (insgesamt $t = r + s$ Kugeln) entnommen ist. Dabei bezeichne Z_i die Indikatorvariable des Ereignisses, daß die i-te gezogene Kugel rot ist. Wird die Stichprobe ohne Zurücklegen gezogen, so ist X hypergeometrisch verteilt zum Parameter (n,r,t). Dann sind die Z_i nicht länger unabhängige Zufallsvariable, zur Berechnung der Varianz kann man dennoch ähnlich wie bei der Binomialverteilung vorgehen. Unser Ausgangspunkt ist die Beobachtung, daß die Z_i immer noch eine austauschbare Verteilung besitzen (vgl. Beispiel 4 in Abschnitt 3.1), wie man der Formel
$$\begin{aligned}&\mathbf{Ws}\{Z_1 = z_1, \ldots, Z_n = z_n\}\\ &\quad= \frac{r(r-1)\cdots(r-x+1)\cdot s(s-1)\cdots(s-y+1)}{t(t-1)\cdots(t-n+1)}\ ,\end{aligned}$$
entnimmt (mit $x = z_1 + \cdots + z_n$ und $y = n - x$). Insgesamt gibt es nämlich $t(t-1)\cdots(t-n+1)$ verschiedene Stichproben ohne Zurücklegen (in vorgegebener Reihenfolge), und an die x Positionen in der Stichprobe, gegeben durch $z_i = 1$, können wir auf $r(r-1)\cdots(r-x+1)$ Weisen verschiedene rote Kugeln unterbringen (mit den Positionen für schwarze Kugeln verhält es sich ähnlich). Wie wir in Beispiel 4, Abschnitt 3.1 festgestellt haben, folgt $\mathbf{E}Z_1 = \cdots = \mathbf{E}Z_n$. Wegen $\mathbf{E}Z_1 = r/t$ erhalten wir für den **Erwartungswert der hypergeometrischen Verteilung** die uns schon bekannte Formel
$$\mathbf{E}X \;=\; np\ , \qquad \text{mit } p \;:= r/t\ .$$

Genauso gilt für $i = 1, \ldots, n$ und $j \neq i$

$$\begin{aligned} \mathbf{Var}Z_i &= \mathbf{Var}Z_1 = pq \\ \mathbf{Cov}[Z_i, Z_j] &= \mathbf{Cov}[Z_1, Z_2] = \mathbf{E}[Z_1 Z_2] - \mathbf{E}[Z_1]\mathbf{E}[Z_2] \\ &= \frac{r(r-1)}{t(t-1)} - \frac{r^2}{t^2} = -\frac{pq}{t-1} \end{aligned}$$

mit $q = 1 - p = s/t$. Nach Proposition 3.3 v) folgt für die **Varianz der hypergeometrischen Verteilung** die Gleichung

$$\mathbf{Var}X = npq - n(n-1)\frac{pq}{t-1} = npq \cdot \frac{t-n}{t-1} ,$$

sie ist um den Faktor $(t-n)/(t-1)$ kleiner als die Varianz der Binomialverteilung. Für $n = t$ ist die Varianz 0. Dies ist offensichtlich, denn dann hat X den festen Wert r. (Aus $0 = \mathbf{Var}X = t\mathbf{Var}Z_1 + t(t-1)\mathbf{Cov}[Z_1, Z_2]$ hätten wir $\mathbf{Cov}[Z_1, Z_2]$ ebenfalls bestimmen können.)

3. **Schätzen mit Stichproben.** Die Ergebnisse über die Binomial- und die hypergeometrische Verteilung werden noch plastischer, wenn man sie auf ein Schätzproblem anwendet: Um die Zusammensetzung einer Population S aus t Individuen zweier Typen kennenzulernen, entnimmt man ihr eine zufällige Stichprobe der Länge n. Enthält die Stichprobe X Individuen vom Typ 1, so ist $\widehat{p} := X/n$ ein plausibler Schätzer für die relative Häufigkeit $p = r/t$ der Individuen vom Typ 1 in der Population. Wie wir gesehen haben, gilt $\mathbf{E}[\widehat{p}] = p$, gleichgültig, ob die Stichprobe mit oder ohne Wiederholungen (‚Zurücklegen') gezogen wird. Im Mittel liegt man in beiden Fällen richtig, jedoch rechnet man beim Ziehen ohne Zurücklegen mit einer genaueren Schätzung. Dies spiegelt sich in der um den Faktor $(t-n)/(t-1)$ verkleinerten Varianz wieder.

   Dieser Sachverhalt läßt sich verallgemeinern. Wir betrachten ein quantitatives Merkmal, das für das Individuum $u \in S$ den Wert $\phi(u) \in \mathbb{R}$ annehme. Um seinen mittleren Wert

$$\mu := \frac{1}{t}\sum_{u \in S} \phi(u)$$

   zu schätzen, wird der Population rein zufällig eine Stichprobe $U_1, \ldots, U_n$ entnommen. Dann ist

$$\widehat{\mu} := \frac{1}{n}\sum_{i=1}^{n} Z_i , \qquad \text{mit } Z_i := \phi(U_i)$$

ein plausibler Schätzer für μ. Aufgrund der Linearität des Erwartungswertes und der Austauschbarkeit von $Z_1, \ldots, Z_n$ gilt

$$\mathbf{E}[\widehat{\mu}] \;=\; \mu \;,$$

gleichgültig ob die Stichprobe mit oder ohne Zurücklegen gezogen ist. Man sagt, $\widehat{\mu}$ ist ein *erwartungstreuer Schätzer*. Für die Varianz ergibt sich

$$\mathbf{Var}[\widehat{\mu}] \;=\; \frac{\sigma^2}{n} \quad \text{bzw.} \quad \frac{\sigma^2}{n} \cdot \frac{t-n}{t-1} \;,$$

je nachdem ob die Stichprobe mit oder ohne Zurücklegen gezogen ist, mit

$$\sigma^2 \;:=\; \frac{1}{t} \sum_{u \in S} (\phi(u) - \mu)^2 \;.$$

Der Beweis wird wie zuvor geführt.

4. **Runs.** Die Anzahl der Runs in $Z_0, Z_1, \ldots, Z_n$ ist, wie in Abschnitt 3.1, Beispiel 2 festgestellt, $Y = 1 + \sum_{i=1}^{n} I_{\{Z_i \neq Z_{i-1}\}}$. Unter der Annahme, daß die Z_i unabhängige, Bernoulli-verteilte Zufallsvariable zum Parameter p sind, folgt

$$\begin{aligned} \mathbf{Var} I_{\{Z_i \neq Z_{i-1}\}} &= 2pq(1-2pq) \\ \mathbf{Cov}[I_{\{Z_{i+1} \neq Z_i\}},\; I_{\{Z_i \neq Z_{i-1}\}}] &= p^2 q + q^2 p - (2pq)^2 \\ &= pq(1-4pq) \end{aligned}$$

Da alle anderen Covarianzen verschwinden, erhalten wir

$$\mathbf{Var} Y \;=\; 2npq(1-2pq) + 2(n-1)pq(1-4pq) \;.$$

□

Bemerkung. Für eine reellwertige Zufallsvariable X mit Dichte $f(x)\,dx$ ist die Varianz gegeben durch

$$\mathbf{Var} X \;:=\; \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)\; dx \;,$$

wobei μ ihr Erwartungswert sei. Die Varianz der $N(\mu, \sigma^2)$-Verteilung ist σ^2,

$$\int_{-\infty}^{\infty} (x-\mu)^2 \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2} \Big)\; dx \;=\; \sigma^2 \;.$$

Der Korrelationskoeffizient

Der Wert der Covarianz zweier Zufallsvariablen läßt sich nicht anschaulich deuten. Die Situation ist wie beim Skalarprodukt $\langle x, y\rangle$ zweier Vektoren, das bekanntlich erst nach Normierung zu einer geometrisch interpretierbaren Größe wird: $\langle x, y\rangle \,/\, \|x\| \cdot \|y\|$ ist der Cosinus des Winkels zwischen x und y. Formal unterscheidet sich die Covarianz nicht vom Skalarprodukt, beide sind symmetrische, nicht-negativ definite Bilinearformen. An die Stelle des Normquadrats $\|x\|^2 = \langle x, x\rangle$ für Vektoren tritt die Varianz $\mathbf{Var}X = \mathbf{Cov}[X, X]$.

Definition. *Der **Korrelationskoeffizient** zweier Zufallsvariablen X und Y mit positiven, endlichen Varianzen ist definiert als*

$$\kappa \;=\; \kappa[X,Y] \;:=\; \frac{\mathbf{Cov}[X,Y]}{\sqrt{\mathbf{Var}[X]\mathbf{Var}[Y]}}\;.$$

Nach Proposition 3.3 gilt $-1 \le \kappa \le 1$.

Der Korrelationskoeffizient läßt sich anschaulich interpretieren: κ ist ein Maß dafür, um wieviel besser Y durch eine Zufallsvariable der Gestalt $aX+b$ angenähert werden kann als durch einen festen Wert b. Es gilt nämlich

$$\min_{a,b}\; \mathbf{E}[(Y-aX-b)^2] \;=\; (1-\kappa^2)\cdot \min_{b}\; \mathbf{E}[(Y-b)^2]\;.$$

Beweis. Wir benutzen (3.2). $\mathbf{E}[(Y-b)^2] = \mathbf{Var}Y + (\mathbf{E}Y - b)^2$ hat $\mathbf{Var}Y$ als minimalen Wert, während

$$\begin{aligned}\mathbf{E}[(Y-aX-b)^2] \;&=\; \mathbf{Var}[Y-aX-b] + (\mathbf{E}Y - a\mathbf{E}X - b)^2\\ &=\; \mathbf{Var}Y - 2a\mathbf{Cov}[X,Y] + a^2\mathbf{Var}X + (\mathbf{E}Y - a\mathbf{E}X - b)^2\end{aligned}$$

als Minimum den Wert $(1-\kappa^2)\mathbf{Var}Y$ hat (setze $b = \mathbf{E}Y - a\mathbf{E}X$ und minimiere dann in a). □

Insbesondere folgt:

- Es gilt $|\kappa| = 1$ genau dann, wenn man reelle Zahlen $a,\ b$ wählen kann, so daß $\mathbf{E}[(Y-aX-b)^2] = 0$ und damit $Y = aX + b$ f.s. gilt.

- Im Fall $\kappa = 0$ ist kein affiner Zusammenhang zwischen X und Y festzustellen. Ein nicht-linearer Zusammenhang kann dagegen sehr wohl bestehen. Sei etwa X eine reellwertige Zufallsvariable mit $\mathbf{E}[X] = \mathbf{E}[X^3] = 0$ (etwa eine symmetrisch um 0 verteilte Zufallsvariable), dann gilt
$$\mathbf{Cov}[X, X^2] \;=\; \mathbf{E}[X^3] - \mathbf{E}[X]\mathbf{E}[X^2] \;=\; 0\;.$$
Korrelation 0 impliziert also im allgemeinen nicht stochastische Unabhängigkeit.

3.3 Erzeugende Funktionen

Manchmal ist die Verteilung einer Zufallsvariablen mit Werten in $\mathbb{N}_0$ besonders gut über ihre erzeugende Funktion zugänglich.

Definition. *Sei X eine Zufallsvariable mit Werten in $\mathbb{N}_0$. Dann ist ihre erzeugende Funktion ϕ definiert als*

$$\phi(t) \;:=\; \sum_{x=0}^{\infty} t^x \cdot \mathbf{Ws}\{X = x\} \;=\; \mathbf{E}[t^X]\;, \quad |t| \;\le\; 1\;.$$

Wegen $\phi(1) = 1$ ist der Konvergenzradius der Potenzreihe mindestens 1 und ϕ auf dem Intervall $[-1, 1]$ eine wohldefinierte Funktion. Die Verteilung von X ist eindeutig durch die erzeugende Funktion gegeben, wenn auch in recht indirekter Weise. Es kann ein kompliziertes Geschäft sein, Kenntnisse über $\phi(t)$ in explizite Aussagen über die Verteilung zu übertragen. Erwartungswert und Varianz erhält man durch Differentiation.

Proposition 3.5. *Für die k-ten Ableitungen der erzeugenden Funktion $\phi(t)$ von X an der Stelle $t = 1$ gilt*

$$\phi^{(k)}(1) \;=\; \mathbf{E}[X(X-1)\cdots(X-k+1)]$$

und folglich

$$\mathbf{E}X \;=\; \phi'(1)\;, \quad \mathbf{Var}X \;=\; \phi''(1) + \phi'(1) - \phi'(1)^2\;.$$

Beweis. Für die k-te Ableitung von $\phi(t)$ gilt

$$\phi^{(k)}(t) \;=\; \sum_{x=0}^{\infty} x(x-1)\cdots(x-k+1)t^{x-k} \cdot \mathbf{Ws}\{X = x\}$$

für alle t innerhalb des Konvergenzbereichs von ϕ. Ist also der Konvergenzradius von ϕ größer als 1, so folgt die Behauptung, indem wir $t = 1$ setzen. Im Fall, daß der Konvergenzradius gleich 1 ist, muß man etwas sorgfältiger argumentieren. Dann setzt man die Ableitung $\phi^{(k)}(1)$ als den linksseitigen Grenzwert

$$\phi^{(k)}(1) \;:=\; \lim_{t\uparrow 1} \phi^{(k)}(t)\;.$$

Er existiert, da $\phi^{(k)}(t)$ auf $[0, 1)$ monoton wächst, möglicherweise ist sein Wert ∞. Die Behauptung folgt nun aus der Abschätzung

$$t^{n-k} \sum_{x=0}^{n} x(x-1)\cdots(x-k+1) \cdot \mathbf{Ws}\{X = x\}$$
$$\le\; \phi^{(k)}(t) \;\le\; \sum_{x=0}^{\infty} x(x-1)\cdots(x-k+1) \cdot \mathbf{Ws}\{X = x\}$$

für $0 \le t < 1$ und $n \in \mathbb{N}$ durch die Grenzübergänge $t \to 1$ und $n \to \infty$. □

Beispiele.

1. Eine Poisson-verteilte Zufallsvariable X hat die erzeugende Funktion

$$\phi(t) \;=\; \sum_{x=0}^{\infty} t^x e^{-\lambda}\,\frac{\lambda^x}{x!} \;=\; e^{\lambda(t-1)}\;.$$

   Durch Differenzieren folgt $\mathbf{E}X = \mathbf{Var}X = \lambda$.

2. Eine geometrisch verteilte Zufallsvariable X mit den Gewichten $\mathbf{Ws}\{X = x\} = pq^{x-1}$, $x = 1, 2, \ldots$ hat die erzeugende Funktion

$$\phi(t) \;=\; \sum_{x=1}^{\infty} t^x p q^{x-1} \;=\; \frac{pt}{1-qt}\;.$$

   Damit erhält man $\mathbf{E}X = 1/p$, $\mathbf{Var}X = q/p^2$. □

Die Bedeutung von erzeugenden Funktionen erklärt sich zu einem Großteil daraus, daß man die erzeugenden Funktionen von Summen unabhängiger Zufallsvariablen leicht berechnen kann.

Proposition 3.6. *Sind X_1, X_2 unabhängige, $\mathbb{N}_0$-wertige Zufallsvariable mit erzeugenden Funktionen $\phi_1(t)$ und $\phi_2(t)$, so hat $X_1 + X_2$ das Produkt $\phi_1(t) \cdot \phi_2(t)$ als erzeugende Funktion.*

Der Beweis ergibt sich aus Satz 3.4:

$$\mathbf{E}[t^{X_1+X_2}] \;=\; \mathbf{E}[t^{X_1}] \cdot \mathbf{E}[t^{X_2}]\;.$$

Beispiel. Eine Bernoulli-verteilte Zufallsvariable Z mit Erfolgswahrscheinlichkeit p hat die erzeugende Funktion $pt + q$. Daher ist für eine $\mathrm{B}(n, p)$-verteilte Zufallsvariable die erzeugende Funktion

$$(pt + q)^n\;.$$

□

Der Kupon-Sammler

Wir zeigen nun anhand eines Beispiels, wie man sich erzeugende Funktionen zunutze macht.

Das Problem: Eine Person sammelt Bildchen, wie man sie in verpackter Schokolade findet. Wir nehmen an, daß es r verschiedene Bildchen gibt. Was läßt sich über die Anzahl X von Schokoladentafeln sagen, die der Sammler kaufen muß, damit er jedes der r Bildchen mindestens einmal hat? Bezeichnen wir mit X_i die Anzahl von Käufen nach dem $(i-1)$ten neuen Bildchen, bis er das nächste neue Bildchen bekommt, so gilt

$$X \;=\; X_1 + X_2 + \cdots + X_r \;.$$

Da er beim ersten Kauf sofort erfolgreich ist, gilt $X_1 = 1$. Um die Verteilung der anderen Summanden festzulegen, machen wir die Annahme, daß bei jedem Kauf einer Tafel Schokolade das vorgefundene Bild rein zufällig unter den r Bildchen verteilt ist und daß die Bildchen in unterschiedlichen Tafeln voneinander unabhängig sind. Dann ist die Erfolgswahrscheinlichkeit für ein neues Bildchen, wenn man schon $i-1$ verschiedene Bildchen hat, gleich $p_i = 1-(i-1)/r$, und die X_i sind unabhängige, geometrisch verteilte Zufallsvariable zu den Erfolgswahrscheinlichkeiten p_i, Gegenwahrscheinlichkeiten q_i und Erwartungswerten $1/p_i$, $i = 1, \ldots, r$. Damit ist die Verteilung von X festgelegt. Für den Erwartungwert folgt

$$\begin{aligned} \mathbf{E}X \;&=\; p_1^{-1} + \cdots + p_r^{-1} \\ &=\; r\big(1 + \tfrac{1}{2} + \tfrac{1}{3} + \cdots + \tfrac{1}{r}\big) \;\sim\; r \log r \;. \end{aligned}$$

Eine ähnliche Formel gilt für die Varianz. Die Verteilung von X läßt sich mit der Einschluß-Ausschluß-Formel berechnen. Dazu betrachten wir das Ereignis A_{lx}, daß das Bildchen Nr. l nicht in den ersten x gekauften Tafeln vorgefunden wurde. Es folgt

$$\begin{aligned} \mathbf{Ws}\{X > x\} \;&=\; \mathbf{Ws}\{A_{1x} \cup \cdots \cup A_{rx}\} \\ &=\; \sum_l \mathbf{Ws}\{A_{lx}\} - \sum_{l<m} \mathbf{Ws}\{A_{lx} \cap A_{mx}\} \pm \cdots \end{aligned}$$

Weiter gilt $\mathbf{Ws}\{A_{lx}\} = (1-1/r)^x$, $\mathbf{Ws}\{A_{lx} \cap A_{mx}\} = (1-2/r)^x$, …, deswegen erhält man

$$\mathbf{Ws}\{X > x\} \;=\; \sum_{i=1}^{r-1} (-1)^{i-1} \binom{r}{i} \Big(1 - \frac{i}{r}\Big)^x$$

und wegen $\mathbf{Ws}\{X = x\} = \mathbf{Ws}\{X > x-1\} - \mathbf{Ws}\{X > x\}$ nach einer Umformung

$$\mathbf{Ws}\{X = x\} \;=\; \sum_{i=1}^{r-1}(-1)^{i-1}\binom{r-1}{i-1}\Big(1-\frac{i}{r}\Big)^{x-1} \;.$$

Wir wollen dieses Resultat nun mit erzeugenden Funktionen ableiten. Der Weg ist etwas aufwendiger, wie erhalten aber auch ein allgemeineres Resultat, das für Summen von unabhängigen, geometrisch verteilten Zufallsvariablen mit beliebigen Erfolgswahrscheinlichkeiten $1 \geq p_1 > p_2 > \cdots > p_r$ gilt, und das nicht mehr mit der Einschluß-Ausschluß-Formel erhalten werden kann. X_i hat als geometrisch verteilte Zufallsvariable die erzeugende Funktion $\phi_i(t) = p_i t/(1-q_i t)$ (dies schließt den Grenzfall $p_1 = 1$ ein, für den $X_1 = 1$ gilt). Wegen der Unabhängigkeit der X_i gilt daher für die erzeugende Funktion $\phi(t)$ von X die Formel

$$\phi(t) \;=\; \phi_1(t)\cdots\phi_r(t) \;=\; \frac{p_1\cdots p_r t^r}{\prod_{i=1}^r(1-q_i t)} \;.$$

Um daraus die Gewichte der Verteilung von X zu erhalten, formen wir diese Funktion in eine Potenzreihe in t um. Dazu machen wir uns zunutze, daß $\phi(t)$ eine gebrochen rationale Funktion ist, die folglich eine *Partialbruchzerlegung* erlaubt. Da die Nullstellen des Nenners von $\phi(t)$ einfach sind, hat sie, wie die Algebra lehrt, die Gestalt

$$\phi(t) \;=\; t^r \sum_{j=1}^{r} \frac{b_j}{1-q_j t} \;.$$

Die Koeffizienten b_j errechnen sich als

$$b_j \;=\; \lim_{t\to q_j^{-1}} (1-q_j t)\phi(t)/t^r \;=\; \frac{p_1\cdots p_r q_j^{r-1}}{\prod_{k\neq j}(q_j - q_k)} \;.$$

Indem wir noch $1/(1-q_j t)$ als geometrische Reihe $\sum_{y=0}^{\infty}(q_j t)^y$ schreiben, erhalten wir insgesamt die Formel

$$\phi(t) \;=\; \sum_{y=0}^{\infty}\sum_{j=1}^{r} b_j q_j^y t^{y+r} \;.$$

Ein Koeffizientenvergleich mit $\phi(t) = \sum_x \mathbf{Ws}\{X = x\}\cdot t^x$ ergibt schließlich umgeformt die (den obigen Spezialfall umfassende) Formel

$$\mathbf{Ws}\{X = x\} \;=\; \sum_{j=1}^{r} \frac{p_1\cdots p_r}{\prod_{k\neq j}(p_k - p_j)}\; q_j^{x-1} \;, \quad x \geq r \;.$$

Bemerkung. Erzeugende Funktionen sind ein bewährtes Hilfsmittel zur Untersuchung von Verteilungen, die es erlauben, Methoden der Analysis heranzuziehen. Ähnlich betrachtet man für eine Zufallsvariable X mit Werten in den nicht-negativen reellen Zahlen die **Laplace-Transformierte**

$$\psi(\lambda) \;:=\; \mathbf{E}[\exp(-\lambda X)]\;, \quad \lambda \;\geq\; 0\;,$$

die man als Verallgemeinerung einer erzeugenden Funktion ansehen kann, und für eine beliebige reellwertige Zufallsvariable X die **charakteristische Funktion** (die **Fourier-Transformierte**)

$$\eta(\lambda) \;:=\; \mathbf{E}[\exp(\mathrm{i}\lambda X)] \;=\; \mathbf{E}[\cos(\lambda X)] \;+\mathrm{i}\;\mathbf{E}[\sin(\lambda X)]\;, \quad \lambda \in \mathbb{R}\;.$$

Ihre Eigenschaften sind den von erzeugenden Funktionen analog. Tieferliegend ist die Tatsache, daß die Laplace-Transformierte bzw. die charakteristische Funktion die Verteilung von X eindeutig bestimmt. Damit eröffnet sich die Möglichkeit, Verteilungen durch das Studium ihrer Transformierten in den Griff zu bekommen und dabei Methoden der Analysis ins Spiel zu bringen. Dies ist ein klassisches, ausführlich behandeltes Kapitel der Stochastik. □

3.4 Gesetze der großen Zahlen und die Tschebyschev-Ungleichung

Wiederholt man ein Zufallsexperiment mit Erfolgswahrscheinlichkeit p, so stabilisiert sich die relative Häufigkeit $H = X/n$ der Erfolge mit wachsender Versuchszahl n bei p. Allgemeiner gilt, daß das arithmetische Mittel von n identisch verteilten, unabhängigen Zufallsvariablen mit wachsendem n gegen den Erwartungswert strebt. Ein erstes Resultat diesen Typs stammt von Jacob Bernoulli (1645 -1705), Poisson (1781-1840) sprach dann von einem „Gesetz der großen Zahlen“. Später war es der Ausgangspunkt für den Versuch, Wahrscheinlichkeiten als Grenzwert von relativen Häufigkeiten zu definieren. Nach anfänglichen Fehlschlägen ist dieser „frequentistische Ansatz“ durch Ideen von Kolmogorov legitimiert worden. In der Wahrscheinlichkeitstheorie haben solche Ansätze kaum Spuren hinterlassen, dort betont man vielmehr den begrifflichen Unterschied, der zwischen der Erfolgswahrscheinlichkeit p als einer Zahl und der relativen Anzahl H von Erfolgen als einer Zufallsvariablen besteht - und die Gesetze der großen Zahlen sind *Sätze* über die Wahrscheinlichkeit bestimmter Ereignisse.

Eine einfache Version eines Gesetzes der großen Zahlen ist das folgende Resultat.

Satz 3.7. *Die Zufallsvariablen $X_1, X_2, \ldots$ seien reellwertig, unkorreliert und identisch verteilt mit endlichem Erwartungswert μ und endlicher Varianz. Dann gilt für alle $\epsilon > 0$*

$$\lim_{n\to\infty} \mathbf{Ws}\Big\{\Big|\frac{X_1 + \cdots + X_n}{n} - \mu\Big| \geq \epsilon\Big\} \;=\; 0 \;.$$

Der Beweis läßt sich leicht mit der Tschebyschevschen Ungleichung führen. Wir benutzen nun für reellwertige Zufallsvariable X und Ereignisse A die Schreibweise

$$\mathbf{E}[X; A] \;:=\; \mathbf{E}[X \cdot I_A] \;.$$

Proposition 3.8.

i) Für jede Zufallsvariable $X \geq 0$ und jedes $\epsilon > 0$ gilt die ***Markov-Ungleichung***

$$\mathbf{Ws}\{X \geq \epsilon\} \;\leq\; \epsilon^{-1}\mathbf{E}[X; X \geq \epsilon] \;\leq\; \epsilon^{-1}\mathbf{E}X \;.$$

ii) Für eine reellwertige Zufallsvariable X mit endlichem Erwartungswert gilt für beliebiges $\epsilon > 0$ die ***Tschebyschev-Ungleichung***

$$\mathbf{Ws}\{|X - \mathbf{E}X| \geq \epsilon\} \;\leq\; \epsilon^{-2} \cdot \mathbf{Var}X$$

Beweis. Für Zufallsvariable $X \geq 0$ gilt $\epsilon I_{\{X\geq\epsilon\}} \leq X I_{\{X\geq\epsilon\}} \leq X$. Die erste Behauptung folgt, indem man nach Satz 3.1 zum Erwartungswert übergeht. *ii*) ergibt sich, indem man *i*) auf $Y = (X - \mathbf{E}X)^2$ anwendet,

$$\mathbf{Ws}\{|X - \mathbf{E}X| \geq \epsilon\} \;=\; \mathbf{Ws}\{Y \geq \epsilon^2\} \;\leq\; \epsilon^{-2}\mathbf{E}Y \;=\; \epsilon^{-2}\mathbf{Var}X \;.$$

□

Vielleicht noch übersichtlicher sind folgende Versionen der beiden Ungleichungen für standardisierte Zufallsvariable,

$$\mathbf{Ws}\{X/\mathbf{E}X \geq \epsilon\} \;\leq\; \epsilon^{-1} \quad \text{bzw.} \quad \mathbf{Ws}\{|X - \mathbf{E}X|/\sqrt{\mathbf{Var}X} \geq \epsilon\} \;\leq\; \epsilon^{-2} \;.$$

Beweis von Satz 3.7. Nach der Tschebyschev-Ungleichung und Proposition 3.3 gilt

$$\mathbf{Ws}\Big\{\Big|\frac{X_1 + \cdots + X_n}{n} - \mu\Big| \geq \epsilon\Big\} \;\leq\; \epsilon^{-2}\mathbf{Var}\Big[\frac{X_1 + \cdots + X_n}{n}\Big] \;=\; \frac{\mathbf{Var}X_1}{n\epsilon^2} \;,$$

woraus die Behauptung folgt. □

Gesetze der großen Zahlen gelten auch unter anderen Bedingungen als denen in Satz 3.7 genannten. Dazu ein Beispiel:

Beispiel. Runs. Sei Y_n die Anzahl der Runs in einer Serie $Z_0, Z_1, \ldots, Z_n$ von 01-wertigen Zufallsvariablen, von denen wir wieder annehmen, daß sie Bernoulli-verteilt mit Erfolgswahrscheinlichkeit p sind. Wie in Abschnitt 3.1 (Beispiel 2) und Abschnitt 3.2 (Beispiel 4) gezeigt, gilt $\mathbf{E}Y_n = 1 + 2npq$ und $\mathbf{Var}Y_n = O(n)$. Nach der Markov-Ungleichung, angewandt auf $(Y_n/n - 2pq)^2$ und (3.2) folgt

$$\begin{aligned}\mathbf{Ws}\Big\{\Big|\frac{Y_n}{n} - 2pq\Big| \geq \epsilon\Big\} \;&\leq\; \epsilon^{-2}\mathbf{E}\Big[\Big(\frac{Y_n}{n} - 2pq\Big)^2\Big] \\ &=\; \epsilon^{-2}\Big(\mathbf{Var}\Big[\frac{Y_n}{n}\Big] + \Big(\mathbf{E}\Big[\frac{Y_n}{n}\Big] - 2pq\Big)^2\Big) \;=\; O(n^{-1})\;.\end{aligned}$$

Als Resultat erhalten wir ein **Gesetz der großen Zahlen für Runs**,

$$\lim_{n\to\infty} \mathbf{Ws}\Big\{\Big|\frac{Y_n}{n} - 2pq\Big| \geq \epsilon\Big\} \;=\; 0 \quad \text{für alle } \epsilon > 0\;.$$

□

Wahrscheinlichkeiten wie in der Markov- oder Tschebyschev-Ungleichung mit Hilfe von Erwartungswerten abzuschätzen, ist eine wirkungsvolle Methode. Wir wollen dieses Thema am Beispiel der Binomialverteilung weiterverfolgen. Für eine $\mathrm{B}(n,p)$-verteilte Zufallsvariable X lautet die Tschebyschev-Ungleichung

$$\mathbf{Ws}\Big\{\Big|\frac{X}{n} - p\Big| \geq \epsilon\Big\} \;\leq\; \epsilon^{-2}\mathbf{Var}\Big[\frac{X}{n}\Big] \;=\; \frac{pq}{n\epsilon^2}\;. \tag{3.4}$$

In den folgenden Beispielen behandeln wir Konsequenzen bzw. Verschärfungen dieser Abschätzung.

Beispiele. Binomialverteilung.

1. **Der Weierstraßsche Approximationssatz.** Nach Weierstrass läßt sich jede stetige Funktion $f : [0,1] \to \mathbb{R}$ gleichmäßig durch Polynome approximieren. Bernstein hat bemerkt, daß dafür die Polynome

$$f_n(t) \;:=\; \sum_{x=0}^{n} f\left(\frac{x}{n}\right)\binom{n}{x} t^x(1-t)^{n-x}$$

geeignet sind. Die Formel

$$f_n(t) \;=\; \mathbf{E}[f\big(X/n\big)]$$

mit einer $\mathrm{B}(n,t)$-verteilten Zufallsvariablen X macht dies plausibel: Nach dem Gesetz der großen Zahlen nimmt X/n Werte nahe bei t an, daher wird auch $f_n(t)$ nahe bei $f(t)$ liegen. Genauer gilt

$$\sup_{0\le t\le 1} |f_n(t) - f(t)| \;\to\; 0 \qquad \text{für } \; n\to\infty \; .$$

Beweis. Sei $m := \max_t |f(t)|$. Wähle $\epsilon > 0$. In der Analysis wird gezeigt, daß f als stetige Funktion auf einem kompakten Intervall gleichmäßig stetig ist, d.h. es gibt zu jedem $\epsilon > 0$ ein $\delta > 0$, so daß $|f(s) - f(t)| \le \epsilon$, falls $|s - t| \le \delta$. Daher gilt

$$\big|f\big(\tfrac{X}{n}\big) - f(t)\big| \;\le\; \epsilon + 2m \cdot I_{\{|\frac{X}{n}-t|\ge\delta\}} \; ,$$

und es folgt

$$\big|\mathbf{E}[f\big(\tfrac{X}{n}\big)] - f(t)\big| \;\le\; \mathbf{E}\big|f\big(\tfrac{X}{n}\big) - f(t)\big| \;\le\; \epsilon + 2m\cdot \mathbf{Ws}\{\big|\tfrac{X}{n} - t\big| \ge \delta\} \; .$$

Mit (3.4) folgt für alle $t \in [0,1]$

$$|f_n(t) - f(t)| \;\le\; \epsilon + \frac{2mt(1-t)}{n\delta^2} \;\le\; 2\epsilon \; ,$$

falls $n \ge 2m/\delta^2\epsilon$ (denn $t(1-t) \le 1$). Dies war zu zeigen.

2. **Die Chernoff-Schranke.** Nach (3.4) gilt $\mathbf{Ws}\{|n^{-1}X - p| \ge \epsilon\} = O(n^{-1})$ für $\mathrm{B}(n,p)$-verteilte Zufallsvariable X. Wir wollen zeigen, daß diese Wahrscheinlichkeit in n sogar exponentiell schnell fällt. Dazu wenden wir die Markov-Ungleichung auf die Zufallsvariable $\exp(\lambda X)$ mit $\lambda \in \mathbb{R}$ an,

$$\mathbf{Ws}\{X \ge n(p+\epsilon)\} \;=\; \mathbf{Ws}\{e^{\lambda X} \ge e^{\lambda n(p+\epsilon)}\} \;\le\; e^{-\lambda n(p+\epsilon)}\mathbf{E}[e^{\lambda X}] \; .$$

Um den Erwartungswert zu berechnen, benutzen wir die Darstellung $X = Z_1 + \cdots + Z_n$ mit unabhängigen, Bernoulli-verteilten Zufallsvariablen $Z_1, \ldots, Z_n$,

$$\mathbf{E}[e^{\lambda X}] \;=\; \mathbf{E}[e^{\lambda Z_1}] \cdots \mathbf{E}[e^{\lambda Z_n}] \;=\; (pe^{\lambda} + q)^n \; ,$$

und folglich

$$\mathbf{Ws}\{X \ge n(p+\epsilon)\} \;\le\; e^{-\lambda n(p+\epsilon)}(pe^{\lambda} + q)^n \;=\; \big(pe^{\lambda(q-\epsilon)} + qe^{-\lambda(p+\epsilon)}\big)^n \; .$$

Die rechte Seite ist minimal für dasjenige λ, das $e^{\lambda} = q(p+\epsilon)/p(q-\epsilon)$ erfüllt. Es folgt für $0 < \epsilon < q$

$$\begin{aligned} \mathbf{Ws}\{X \ge n(p+\epsilon)\} \;&\le\; \Big[p\Big(\frac{q(p+\epsilon)}{p(q-\epsilon)}\Big)^{q-\epsilon} + q\Big(\frac{q(p+\epsilon)}{p(q-\epsilon)}\Big)^{-(p+\epsilon)}\Big]^n \\ &=\; \Big(\frac{p}{p+\epsilon}\Big)^{n(p+\epsilon)}\Big(\frac{q}{q-\epsilon}\Big)^{n(q-\epsilon)} \end{aligned}$$

oder mit $h(t) = t \ln \frac{t}{p} + (1-t) \ln \frac{1-t}{q}$

$$\mathbf{Ws}\Big\{ \frac{X}{n} \geq p + \epsilon \Big\} \;\leq\; \exp(-n \cdot h(p+\epsilon)) \;. \tag{3.5}$$

$n \cdot h(t)$ ist die Entropiefunktion der Binomialverteilung, die uns bereits in (1.7) begegnet ist. Es gilt $h(t) > 0$ für $t \neq p$, und dies ergibt die behauptete exponentielle Konvergenzgeschwindigkeit. Analog beweist man für $0 < \epsilon < p$

$$\mathbf{Ws}\Big\{ \frac{X}{n} \leq p - \epsilon \Big\} \;\leq\; \exp(-n \cdot h(p-\epsilon)) \;. \tag{3.6}$$

3. **Ein unfaires Spiel.** Mit diesem Beispiel wollen wir deutlich machen, daß der Erwartungswert eine völlig falsche Vorstellung über typische Werte einer Zufallsvariablen bzw. die Lage einer Verteilung vermitteln kann.

   Bei einem Glücksspiel mit vorteilhafter Gewinnwahrscheinlichkeit $p \geq 1/2$ sei die Bedingung, daß ein Mitspieler immer einen festen Anteil $\delta \cdot K$ seines aktuellen Spielkapitals K setzen muß, mit fest vorgegebenem $\delta \in (0,1)$. Nach n Runden hat der Spieler bei einem Startkapital der Größe $K_0 = 1$ also das Kapital

   $$K_n \;=\; Z_1 \cdots Z_n \;,$$

   wobei $Z_1, Z_2, \ldots$ unabhängige Zufallsvariable mit $\mathbf{Ws}\{Z_n = 1+\delta\} = p$ und $\mathbf{Ws}\{Z_n = 1-\delta\} = q$ bezeichnen. Der Erwartungswert ist nach Satz 3.4

   $$\mathbf{E}K_n \;=\; \mathbf{E}Z_1 \cdots \mathbf{E}Z_n \;=\; (1 + (p-q)\delta)^n \;.$$

   Für $p > q$ verspricht dies exponentielles Wachstum von K_n, aber das täuscht. Um die Formeln übersichtlich zu gestalten, beschränken wir uns nun auf den Fall $p = 1/2$. Dann gilt

   $$\mathbf{E}K_n \;=\; 1 \;.$$

   so daß ein faires Spiel vorzuliegen scheint. Der Erwartungswert führt hier jedoch in die Irre, denn nach dem Gesetz der großen Zahlen ist damit zu rechnen, daß von den Z_k ungefähr $\frac{n}{2}$ den Wert $1+\delta$, die anderen $\frac{n}{2}$ den Wert $(1-\delta)$ annehmen. Daher hat K_n die Größenordnung

   $$(1+\delta)^{n/2}(1-\delta)^{n/2} \;=\; \Big(\sqrt{1-\delta^2}\Big)^n \;,$$

   ein Ausdruck, der exponentiell schnell fällt! Genauer gilt mit $X_k = \ln Z_k$, also $\mathbf{E}X_k = \ln\sqrt{1-\delta^2}$,

   $$\mathbf{Ws}\{\ln K_n > n \ln\sqrt{1-\delta^2} + \epsilon n\} \;=\; \mathbf{Ws}\Big\{ \frac{X_1 + \cdots + X_n}{n} > \mathbf{E}X_1 + \epsilon \Big\}$$

und nach Satz 3.7 folgt

$$\mathbf{Ws}\Big\{K_n > \Big(\sqrt{1-\delta^2}\Big)^n e^{\epsilon n}\Big\} \;\to\; 0 \tag{3.7}$$

für alle $\epsilon > 0$, was ein exponentiell fallendes Guthaben bestätigt. (Für $p > 1/2$ gibt es einen analogen Effekt, dann ist K_n von der Größenordnung $((1+\delta)^p(1-\delta)^q)^n$.) Es gibt also

$$\mathbf{E}[\ln K_n] \;=\; n\;\mathbf{E}[\ln X_1] \;=\; \ln\Big(\sqrt{1-\delta^2}\Big)^n$$

die richtige Vorstellung von der Größe von $\ln K_n$, jedoch $\mathbf{E}K_n$ eine völlig falsche von K_n. Der Sachverhalt erklärt sich aus Jensens Ungleichung: Der Logarithmus ist eine strikt konkave Funktion, deswegen gilt

$$\mathbf{E}[\ln K_n] \;<\; \ln \mathbf{E}[K_n]\;.$$

Nach (3.7) konzentriert sich Verteilung von K_n mit wachsendem n bei 0. Es müssen sich daher Werte von K_n im Erwartungswert $\mathbf{E}K_n$ durchsetzen, die eine untypische Größe besitzen. Welche Werte sind dies? Wie im vorigen Beispiel hilft ein Maßwechsel. Wir betrachten die Anzahl Y_n der Erfolge in den ersten n Spielen. Es gilt $K_n = f(Y_n)$ mit $f(y) = (1+\delta)^y(1-\delta)^{n-y}$. Y_n ist binomialverteilt zum Parameter $\big(n, \frac{1}{2}\big)$, also folgt

$$\begin{aligned}
\mathbf{E}\Big[K_n\;;\;\Big|\frac{Y_n}{n} - \frac{1+\delta}{2}\Big| \ge \epsilon\Big] \;&=\; \sum_{\substack{y\\ \left|\frac{y}{n}-\frac{1+\delta}{2}\right|\ge\epsilon}} f(y)\binom{n}{y}2^{-n} \\
&=\; \sum_{\substack{y\\ \left|\frac{y}{n}-\frac{1+\delta}{2}\right|\ge\epsilon}} \binom{n}{y}\Big(\frac{1+\delta}{2}\Big)^y\Big(\frac{1-\delta}{2}\Big)^{n-y} \\
&=\; \mathbf{Ws}\Big\{\Big|\frac{Y'_n}{n} - \frac{1+\delta}{2}\Big| \ge \epsilon\Big\}\;,
\end{aligned}$$

wobei Y'_n eine binomialverteilte Zufallsvariable zum Parameter $\big(n, \frac{1+\delta}{2}\big)$ bezeichne. Nach (3.4) folgt für alle $\epsilon > 0$

$$\mathbf{E}\Big[K_n\;;\;\Big|\frac{Y_n}{n} - \frac{1+\delta}{2}\Big| \ge \epsilon\Big] \;\to\; 0\;,$$

bzw. wegen $\mathbf{E}K_n = 1$

$$\mathbf{E}\Big[K_n\;;\;\Big|\frac{Y_n}{n} - \frac{1+\delta}{2}\Big| < \epsilon\Big] \;\to\; 1\;.$$

Zu $\mathbf{E}K_n$ tragen daher wesentlich nur solche Spielverläufe bei, bei denen Y_n/n einen untypischen Wert nahe bei $(1+\delta)/2$ annimmt. Wir wissen aus dem vorangegangenen Beispiel, daß diese Ereignisse von exponentiell kleiner Wahrscheinlichkeit sind. □

Nach Satz 3.7 hat $\left\{\left|\frac{X_1+\cdots+X_n}{n}-\mu\right|\geq\epsilon\right\}$ asymptotisch die Wahrscheinlichkeit 0. Offen bleibt, wie dieses Ereignis mit n variiert. Zunächst einmal ist nicht ausgeschlossen, daß mit positiver Wahrscheinlichkeit ∞-viele dieser Ereignisse eintreten. Wie wir sehen werden, gilt aber auch die stärkere Aussage, daß die Wahrscheinlichkeit des Ereignisses

$$\begin{aligned}&\left\{\left|\frac{X_1+\cdots+X_m}{m}-\mu\right|\geq\epsilon \text{ für ein } m\geq n\right\}\\ &\qquad:=\bigcup_{m=n}^{\infty}\left\{\left|\frac{X_1+\cdots+X_m}{m}-\mu\right|\geq\epsilon\right\},\end{aligned}$$

für $n\to\infty$ verschwindet. Dabei langt es dann allerdings nicht mehr, die X_i lediglich als unkorreliert anzunehmen.

Definition. *Seien X und $X_1, X_2, \ldots$ reellwertige Zufallsvariable. Man sagt:*

*i) X_n konvergiert **stochastisch** (oder **in Wahrscheinlichkeit**) gegen X, falls für alle $\epsilon>0$*

$$\lim_{n\to\infty}\mathbf{Ws}\{|X_n-X|\geq\epsilon\} \;=\; 0\ .$$

*ii) X_n konvergiert **fast sicher** gegen X, falls für alle $\epsilon>0$*

$$\lim_{m\to\infty}\mathbf{Ws}\Big\{\bigcup_{n=m}^{\infty}\{|X_n-X|\geq\epsilon\}\Big\} \;=\; 0\ .$$

Die fast sichere Konvergenz ist offenbar der stärkere Konvergenzbegriff. Da für $m\to\infty$

$$\begin{aligned}&\bigcup_{n=m}^{\infty}\{|X_n-X|\geq\epsilon\}\\ &\quad\downarrow\ \bigcap_{m=1}^{\infty}\bigcup_{n=m}^{\infty}\{|X_n-X|\geq\epsilon\} \;=:\; \{|X_n-X|\geq\epsilon \text{ für } \infty\text{-viele } n\}\end{aligned}$$

(der Limes superior der Ereignisse $\{|X_n-X|\geq\epsilon\}$, vgl. (2.2)), läßt sich wegen der σ-Stetigkeit von Wahrscheinlichkeiten fast sichere Konvergenz auch durch die Forderung

$$\mathbf{Ws}\{|X_n-X|\geq\epsilon \text{ für } \infty\text{-viele } n\} \;=\; 0 \qquad \text{für alle } \epsilon>0$$

charakterisieren. Fast sichere Konvergenz heißt deswegen auch **Konvergenz mit Wahrscheinlichkeit 1**. Die Grenzvariable X ist für stochastische wie für fast sichere Konvergenz fast sicher eindeutig bestimmt.

Gesetze der großen Zahlen werden danach unterschieden, was für Konvergenzaussagen sie treffen. In Satz 3.7 geht es um stochastische Konvergenz, man spricht dann von einem **schwachen Gesetz der großen Zahlen**. Bei dem folgenden Satz, der fast sichere Konvergenz beinhaltet, handelt es sich um ein **starkes Gesetz der großen Zahlen**.

Satz 3.9. *Seien $X_1, X_2, \ldots$ unabhängige, identisch verteilte Zufallsvariable mit Werten in den reellen Zahlen und einem endlichen Erwartungswert μ. Dann konvergiert $n^{-1}(X_1 + \cdots + X_n)$ für $n \to \infty$ fast sicher gegen μ.*

Beweis. Wir führen den Beweis unter der zusätzlichen Voraussetzung, daß die X_i ein endliches viertes Moment besitzen, d.h. $\mathbf{E}[X_i^4] < \infty$ gilt. Ohne Einschränkung sei $\mu = 0$ (sonst ersetze man die X_i durch $X_i - \mu$). Nach Satz 3.4 gilt dann $\mathbf{E}[X_i X_j X_k X_l] = 0$, es sei denn, die i, j, k, l sind paarweise gleich. Man kann solche Paare auf 3 Weisen bilden, daher folgt unter Berücksichtigung der Cauchy-Schwarz Ungleichung

$$\begin{aligned}\mathbf{E}[(X_1+\cdots+X_n)^4] \;&=\; \sum_i \sum_j \sum_k \sum_l \mathbf{E}[X_i X_j X_k X_l] \\ &\le\; 3\sum_{i,j} \mathbf{E}[X_i^2 X_j^2] \;\le\; 3\sum_{i,j} \mathbf{E}[X_i^4]^{1/2}\mathbf{E}[X_j^4]^{1/2} \;=\; 3n^2\,\mathbf{E}[X_1^4]\;.\end{aligned}$$

Nach der Markov-Ungleichung folgt für $\epsilon > 0$

$$\begin{aligned}\mathbf{Ws}\Big\{\Big|\frac{X_1+\cdots+X_n}{n}\Big| \ge \epsilon\Big\} \;&=\; \mathbf{Ws}\Big\{\Big(\frac{X_1+\cdots+X_n}{n}\Big)^4 \ge \epsilon^4\Big\} \\ &\le\; (\epsilon n)^{-4}\,\mathbf{E}[(X_1+\cdots+X_n)^4] \;=\; O(n^{-2})\;,\end{aligned}$$

so daß die Summe dieser Wahrscheinlichkeiten konvergiert. Nach dem ersten Borel-Cantelli Lemma (Satz 2.4) folgt daher

$$\mathbf{Ws}\Big\{\Big|\frac{X_1+\cdots+X_n}{n}\Big| \ge \epsilon \text{ für } \infty\text{-viele } n\Big\} \;=\; 0$$

und damit die Behauptung. □

Ohne endliches viertes Moment ist der Sachverhalt komplizierter. Ein vergleichsweise kurzer Beweis stammt von N. Etemadi, er findet sich in dem Lehrbuch *Probability: Theory and Examples* von R. Durrett.

3.5 Der Satz von der monotonen Konvergenz

Vorbereitend auf das Kapitel über Markov-Ketten betrachten wir nun noch diskrete Zufallsvariable X, die neben Werten in den nicht-negativen Zahlen auch den Wert ∞ annehmen dürfen. Wir schreiben dann $X \geq 0$ und setzen

$$\mathbf{E}X \;:=\; \sum_x x \cdot \mathbf{Ws}\{X = x\} + \infty \cdot \mathbf{Ws}\{X = \infty\}$$

mit $\infty \cdot 0 := 0$ und $\infty \cdot w := \infty$ für $w > 0$ (dabei erstreckt sich die Summe über alle reellen Zahlen $x \geq 0$, die X als Wert annehmen kann).

Die Rechenregeln des Erwartungswertes bleiben weitgehend erhalten, z.B. gilt für Zufallsvariablen $X, Y \geq 0$

$$\mathbf{E}[X + Y] \;=\; \mathbf{E}X + \mathbf{E}Y \;.$$

Gilt nämlich $\mathbf{Ws}\{X+Y = \infty\} = 0$ bzw. $\mathbf{Ws}\{X = \infty\} = \mathbf{Ws}\{Y = \infty\} = 0$, so sind wir in dem uns bereits bekannten Fall reellwertiger Zufallsvariabler. Andernfalls steht ∞ auf beiden der Gleichung. Ähnlich überzeugt man sich von

$$\mathbf{E}X \;\leq\; \mathbf{E}Y$$

für Zufallsvariablen $0 \leq X \leq Y$.

Das Hauptresultat über die Erwartungswerte nicht-negativer Zufallsvariabler ist der **Satz von der monotonen Konvergenz** (Satz von Beppo Levi).

Satz 3.10. *Sei $0 \leq X_1 \leq X_2 \leq \cdots$ und sei $X = \lim_n X_n$ fast sicher. Dann gilt*

$$\mathbf{E}X \;=\; \lim_n \mathbf{E}X_n \;.$$

Beweis. Einerseits gilt $X_n \leq X$ und damit $0 \leq \mathbf{E}X_1 \leq \mathbf{E}X_2 \leq \cdots \leq \mathbf{E}X$ sowie $\lim_n \mathbf{E}X_n \leq \mathbf{E}X$.

Seien andererseits $x_1, x_2, \ldots$ die reellen Werte von X, in irgendeiner Reihenfolge aufgezählt. Dann gilt für $\epsilon > 0$ und natürliche Zahlen k, n

$$\sum_{i=1}^{k} x_i \mathbf{Ws}\{X = x_i, X < X_n + \epsilon\} \;=\; \mathbf{E}\Big[\sum_{i=1}^{k} x_i I_{\{X=x_i, X<X_n+\epsilon\}}\Big] \;\leq\; \mathbf{E}[X_n + \epsilon]\,.$$

Da nach Annahme X_n fast sicher monoton gegen X konvergiert, folgt $\mathbf{Ws}\{X = x_i, X < X_n + \epsilon\} \to \mathbf{Ws}\{X = x_i\}$ für $n \to \infty$ und damit

$$\sum_{i=1}^{k} x_i \mathbf{Ws}\{X = x_i\} \;\leq\; \lim_n \mathbf{E}X_n + \epsilon \;.$$

Nehmen wir auch noch $\mathbf{Ws}\{X = \infty\} = 0$ an, so folgt mit $k \to \infty$ und $\epsilon \to 0$ nun $\mathbf{E}X \le \lim_n \mathbf{E}X_n$, also die Behauptung.

Es bleibt der Fall $\mathbf{Ws}\{X = \infty\} > 0$. Dann gilt nach der Markov-Ungleichung für $c > 0$

$$c \cdot \mathbf{Ws}\{X_n > c\} \;\le\; \mathbf{E}X_n \;.$$

Nach Annahme gilt $\mathbf{Ws}\{X_n > c\} \to \mathbf{Ws}\{X > c\}$ und folglich

$$c \cdot \mathbf{Ws}\{X = \infty\} \;\le\; c \cdot \mathbf{Ws}\{X > c\} \;\le\; \lim_n \mathbf{E}X_n \;.$$

Mit $c \to \infty$ erhalten wir $\lim_n \mathbf{E}X_n = \infty = \mathbf{E}X$ und damit die Behauptung. □

Später werden wir diesen Satz in der folgenden Version benutzen.

Satz 3.11. *Gilt für Zufallsvariablen $X, X_1, X_2, \ldots \ge 0$ fast sicher die Beziehung $X = \sum_{n=1}^{\infty} X_n$, so folgt*

$$\mathbf{E}X \;=\; \sum_{n=1}^{\infty} \mathbf{E}X_n \;.$$

Beweis. Für die Zufallsvariablen $Y_n = X_1 + \cdots + X_n$ gilt $0 \le Y_1 \le Y_2 \le \cdots$ und $\lim_n Y_n = X$ fast sicher, daher folgt aus dem Satz von der monotonen Konvergenz

$$\mathbf{E}X \;=\; \lim_n \mathbf{E}Y_n \;=\; \lim_n (\mathbf{E}X_1 + \cdots + \mathbf{E}X_n) \;=\; \sum_{n=1}^{\infty} \mathbf{E}X_n \;.$$

□

Kapitel 4

Folgen von Zufallsentscheidungen und bedingte Wahrscheinlichkeiten

Mehrstufige Zufallsexperimente sind Experimente, die aus einer Folge von Zufallsentscheidungen bestehen. Auf diese Weise lassen sich komplexe stochastische Modelle aus einfachen Bestandteilen aufbauen, die Abfolge der Einzelexperimente kann man anhand von Graphen veranschaulichen. Umgekehrt lassen sich Zufallsexperimente in vielfältiger Weise in Stufen zerlegen. Solche Konstruktionen benutzen bedingte Wahrscheinlichkeiten und den Satz von der totalen Wahrscheinlichkeit.

4.1 Ein Beispiel: Suchen in Listen

Zugriffszeiten auf Informationen kann man verkürzen, indem man mit mehreren Listen arbeitet. Diese Idee wird etwa für die Verwaltung von Daten in Computern verwandt (Informatiker sprechen vom *Hashing*). Wir stellen uns eine Situation vor, bei der n Namen nach einem bestimmten Gesichtspunkt (etwa alphabetisch) auf k Listen verteilt werden, es ist also bei jedem Namen klar, in welche Liste er einsortiert wird. Die Frage ist, wie lange es dauert, bis man einen Namen in den Listen findet bzw. feststellt, daß er nicht in den Listen steht.

Zur Beantwortung dieser Frage gehen wir davon aus, daß es sich bei den Namen, die in den Listen aufgeführt sind, um *zufällige* Namen handelt. Wir nehmen an, daß sie unabhängig voneinander jeweils mit Wahrscheinlichkeit p_i in die i-te Liste gelangen ($p_1 + \cdots + p_k = 1$, vgl. die Maxwell-Boltzmann Statistik aus Abschnitt 1.3). Mit Y_i bezeichnen wir die Anzahl

der Namen in der i-ten Liste. Dann ist Y_i binomialverteilt zum Parameter (n, p_i) und $Y = (Y_1, \ldots, Y_k)$ ein multinomialverteilter Zufallsvektor zum Parameter $(n, p_1, \ldots, p_k)$.

Genauso nehmen wir an, daß der Name, den wir in den Listen suchen, zufällig ist, also in den Listen an einem zufälligen Platz steht, wenn er sich überhaupt in den Listen findet. Wir haben es also mit einer Situation zu tun, in dem der Zufall in zwei Stufen ins Spiel kommt,

erstens werden zufällige Namen in die Listen eingeordnet,

zweitens wird unter diesen Namen ein weiterer zufälliger Name gesucht.

Wir wollen die erwartete Suchzeit berechnen. Dazu bietet es sich an, in zwei Schritten vorzugehen. Zunächst betrachte man nicht-zufällige Listen der Längen $y_1, \ldots, y_k$ und berechne den Erwartungswert der Suchdauer X für einen zufälligen Namen in Abhängigkeit von $y = (y_1, \ldots, y_k)$. Man spricht von einem *bedingten Erwartungswert* und schreibt ihn als $\mathbf{E}[X \,\big|\, Y = y]$. Im zweiten Schritt kann man dann zu zufälligen Listen übergehen. Die bedingte Erwartung wird zu einer zufälligen reellwertigen Größe, für die man die Bezeichnung $\mathbf{E}[X \,\big|\, Y]$ benutzt. Die gesuchte erwartete Suchdauer $\mathbf{E}X$ erhält man durch Bilden eines weiteren Erwartungswertes,

$$\mathbf{E}X \;=\; \mathbf{E}[\mathbf{E}[X \,\big|\, Y]] \;.$$

In Abschnitt 4.5 werden wir diese naheliegende Vorgehensweise genauer begründen (vgl. Formel (4.7)).

Für unser Beispiel müssen wir noch präzisieren, wie der Zufall auf der zweiten Stufe wirkt. Dazu unterscheiden wir, ob der gesuchte Name in der Liste auftaucht oder nicht.

Fall 1. Der Name steht nicht in der Liste. Dies ist der einfachere Fall. Wir nehmen an, daß der neue Name ebenfalls mit Wahrscheinlichkeit p_i in die i-te Liste gehört. Man muß dann alle Einträge dieser Liste durchgehen, um festzustellen, daß der Name nicht eingetragen wurde. Enthält die i-te Liste die feste Anzahl von y_i Einträgen, so muß man im Mittel

$$\mathbf{E}[X \,\big|\, Y = y] \;=\; \sum_{i=1}^{k} y_i p_i$$

Eintragungen überprüfen, bis man feststellt, daß der Name noch nicht in die Listen aufgenommen wurde.

Im zweiten Schritt ersetzen wir nun die y_i durch die Zufallsvariablen Y_i und erhalten

$$\mathbf{E}[X \,\big|\, Y] \;=\; \sum_{i=1}^{k} Y_i p_i \;.$$

Der gesuchte Erwartungswert ist

$$\mathbf{E}X \;=\; \mathbf{E}\Big[\sum_{i=1}^{k} p_i Y_i\Big] \;.$$

Da Y_i den Erwartungswert np_i hat, folgt

$$\mathbf{E}X \;=\; n(p_1^2 + \cdots + p_k^2) \;.$$

Im uniformen Fall ergibt sich die plausible Formel

$$\mathbf{E}X \;=\; \frac{n}{k} \;.$$

Fall 2. Der Name steht in der Liste. Wir nehmen nun an, daß der gesuchte Name sich an einer rein zufälligen Stelle unter allen n Eintragungen in den Listen befindet. Damit ist die Wahrscheinlichkeit, daß er in einer bestimmten Liste steht, davon abhängig, wie lang die Liste ist: Falls die i-te Liste y_i Einträge enthält, ist der Name mit Wahrscheinlichkeit y_i/n in ihr enthalten. Steht er dort an der j-ten Stelle, so sind j Vergleiche erforderlich, die mittlere Anzahl der Vergleiche ist also

$$\mathbf{E}[X \,\big|\, Y = y] \;=\; \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{y_i} j \;=\; \frac{1}{n}\sum_{i=1}^{k}\frac{(y_i+1)y_i}{2} \;.$$

Anders ausgedrückt: Die i-te Liste wird mit Wahrscheinlichkeit y_i/n durchsucht und dafür werden im Mittel $(y_i+1)/2$ Vergleiche benötigt.

Gehen wir nun zu zufälligen Listen über, so erhalten wir

$$\mathbf{E}[X \,\big|\, Y] \;=\; \frac{1}{n}\sum_{i=1}^{k}\frac{(Y_i+1)Y_i}{2}$$

und den gesuchten Erwartungswert als

$$\mathbf{E}X \;=\; \mathbf{E}\Big[\frac{1}{n}\sum_{i=1}^{k}\frac{(Y_i+1)Y_i}{2}\Big] \;.$$

Unter Beachtung von (3.2) erhalten wir

$$\mathbf{E}[(Y_i+1)Y_i] \;=\; \mathbf{Var}[Y_i] + \mathbf{E}[Y_i]^2 + \mathbf{E}[Y_i] \;=\; np_i(1-p_i) + n^2p_i^2 + np_i \;,$$

und es folgt

$$\mathbf{E}X \;=\; \frac{n-1}{2}(p_1^2 + \cdots + p_k^2) + 1 \;.$$

Im uniformen Fall gilt

$$\mathbf{E}X \;=\; \frac{n-1}{2k} + 1 \;.$$

4.2 Bedingte Wahrscheinlichkeiten

In diesem Abschnitt behandeln wir 2-stufige Experimente von einem formalen Standpunkt. Es stellt sich heraus, daß man jedem Zufallsexperiment in recht beliebiger Weise eine 2-stufige Gestalt geben kann. Wir benötigen dazu den Begriff der bedingten Wahrscheinlichkeit.

Definition. *Für Ereignisse A, A' ist die* ***bedingte Wahrscheinlichkeit von A bzgl. A'*** *gegeben durch*

$$\mathbf{Ws}\{A \,\big|\, A'\} \;:=\; \frac{\mathbf{Ws}\{A \cap A'\}}{\mathbf{Ws}\{A'\}} \;.$$

Dabei setzt man üblicherweise $0/0 := 0$ (dies ist der Fall $\mathbf{Ws}\{A'\} = 0$).

Erläuterungen.

1. Sind A_1 und A_2 unabhängige Ereignisse, so folgt

   $$\mathbf{Ws}\{A_1 \,\big|\, A_2\} \;=\; \mathbf{Ws}\{A_1\} \;.$$

   Allgemeiner gilt

   $$\mathbf{Ws}\{A_1 \,\big|\, A_2 \cap A_3\} \;=\; \mathbf{Ws}\{A_1 \,\big|\, A_2\} \;,$$

   falls A_3 von den Ereignissen $A_1 \cap A_2$ und A_2 unabhängig ist. Dies ergibt sich direkt aus der Definition.

2. Die Zuordnung $A \mapsto \mathbf{Ws}\{A \,\big|\, A'\}$ ist bei festem A' ein neues W-Maß auf den Ereignissen,

   i) $0 \;\le\; \mathbf{Ws}\{A \,\big|\, A'\} \;\le\; 1 \;,\quad \mathbf{Ws}\{\Phi \,\big|\, A'\} \;=\; 0 \;,\quad \mathbf{Ws}\{\Omega \,\big|\, A'\} \;=\; 1 \;,$

   ii) $\mathbf{Ws}\{\bigcup_n A_n \,\big|\, A'\} \;=\; \sum_n \mathbf{Ws}\{A_n \,\big|\, A'\}$ für paarweise disjunkte Ereignisse $A_1, A_2, \ldots$

   Der Beweis ergibt sich unmittelbar aus der Definition.

3. Sind X und Y diskrete Zufallsvariable mit Wertebereichen S und S', so nennt man die Familie $k_y(\cdot)$, $y \in S'$ von W-Verteilungen auf S, gegeben durch

   $$k_y(B) \;:=\; \mathbf{Ws}\{X \in B \,\big|\, Y = y\} \;,\quad B \subset S \;,$$

   die **bedingte Verteilung von *X*, gegeben *Y***. Sie und die Verteilung von Y legen die gemeinsame Verteilung von X und Y fest, gemäß der Formel

   $$\mathbf{Ws}\{X = x, Y = y\} \;=\; \mathbf{Ws}\{X = x \,\big|\, Y = y\}\, \mathbf{Ws}\{Y = y\} \;.$$

□

Grundlegend für das Rechnen mit bedingten Wahrscheinlichkeiten ist der **Satz von der totalen Wahrscheinlichkeit**. Im diskreten Fall lautet er folgendermaßen.

Satz 4.1. *Sei X eine diskrete Zufallsvariable mit Werten in S. Dann gilt für jedes Ereignis A*

$$\mathbf{Ws}\{A\} \;=\; \sum_{x\in S} \mathbf{Ws}\{A \,\big|\, X=x\}\cdot \mathbf{Ws}\{X=x\}\;.$$

Der Beweis folgt aus

$$\mathbf{Ws}\{A\} \;=\; \mathbf{Ws}\{A\cap\{X\in S\}\} \;=\; \sum_{x} \mathbf{Ws}\{A\cap\{X=x\}\}$$

und der Definition von bedingten Wahrscheinlichkeiten. Anders ausgedrückt lautet der Satz von der totalen Wahrscheinlichkeit

$$\mathbf{Ws}\{A\} \;=\; \sum_{x\in S} \mathbf{Ws}\{A \,\big|\, A_x\}\cdot \mathbf{Ws}\{A_x\}$$

für jede Partition A_x, $x\in S$, des sicheren Ereignisses (vgl. Proposition **??**).

Wir betrachten nun irgendein Zufallsexperiment, formal gegeben durch einen W-Raum. Aus den zu dem Experiment gehörigen diskreten Zufallsvariablen wählen wir eine Zufallsvariable X aus. Damit können wir das folgende **2-stufige Ersatzexperiment** bilden:

1. Beobachte den Wert von X.
2. Nimmt X den Wert x an, so führe ein Experiment durch, bei dem das Ereignis A mit Wahrscheinlichkeit $P_x\big(A\big) := \mathbf{Ws}\{A \,\big|\, X=x\}$ eintritt.

Von einem formalen Standpunkt aus handelt es sich um ein wohldefiniertes Experiment, denn P_x ist, wie wir oben festgestellt haben, ein W-Maß. Insgesamt erhält A in dem neuen Experiment die Wahrscheinlichkeit

$$\widetilde{\mathbf{Ws}}\{A\} \;=\; \sum_{x} \mathbf{Ws}\{X=x\}P_x\big(A\big)\;,$$

die sich nach dem Satz von der totalen Wahrscheinlichkeit als

$$\widetilde{\mathbf{Ws}}\{A\} \;=\; \sum_{x} \mathbf{Ws}\{X=x\}\mathbf{Ws}\{A \,\big|\, X=x\} \;=\; \mathbf{Ws}\{A\}$$

erweist. Damit wird deutlich, was das Ersatzexperiment leistet: Es ist zu dem Ausgangsexperiment äquivalent in dem Sinne, daß in beiden Experimenten

alle Ereignisse die gleiche Wahrscheinlichkeit haben. Außerdem wird der Wert von X in der zweiten Stufe mit Wahrscheinlichkeit 1 nicht mehr revidiert, denn es gilt

$$P_x\big(X = y\big) \;=\; \frac{\mathbf{Ws}\{X = y, X = x\}}{\mathbf{Ws}\{X = x\}} \;=\; \begin{cases} 1\,, & y = x \\ 0\,, & y \neq x\,. \end{cases}$$

Das Ausgangsexperiment wird in dem 2-stufigen Experiment also nicht nur dupliziert, es erhält eine zusätzliche Struktur: Erst wird festgestellt, welchen Wert X annimmt, danach entscheidet sich, welche der anderen Ereignisse eintreten. Das folgende Beispiel gibt eine typische Anwendung.

Beispiel. Sind $X_1, \ldots, X_k$ unabhängige Poisson-verteilte Zufallsvariable mit Erwartungswerten $\lambda_1, \ldots, \lambda_k$, so ist ihre Summe $X = X_1 + \cdots + X_k$ Poisson-verteilt mit Erwartungswert $\lambda = \lambda_1 + \cdots + \lambda_k$. Eine kurze Rechnung ergibt

$$\begin{aligned} \mathbf{Ws}\{X_1 = x_1, \ldots, X_k = x_k \,\big|\, X = x\} \;&=\; \frac{\mathbf{Ws}\{X_1 = x_1\} \cdots \mathbf{Ws}\{X_k = x_k\}}{\mathbf{Ws}\{X = x\}} \\ &=\; \binom{x}{x_1, \ldots, x_k} p_1^{x_1} \cdots p_k^{x_k} \end{aligned}$$

mit $x = x_1 + \cdots + x_k$ und $p_i := \lambda_i/\lambda$. Unabhängige, Poisson-verteilte Zufallsvariable $X_1, \ldots, X_k$ lassen sich daher folgendermaßen generieren (man vergleiche dazu auch Konstruktion I aus Abschnitt 2.4.):

1. Wähle eine zufällige Zahl X aus $\mathbb{N}_0$ gemäß einer Poisson-Verteilung mit Erwartungswert λ.

2. Nimmt X den Wert x an, so verteile x Kugeln nach Art der Maxwell-Boltzmann Statistik auf k Schachteln, so daß jede Kugel mit Wahrscheinlichkeit p_i in die i-te Schachtel gelangt. $X_1, \ldots, X_k$ seien die Besetzungszahlen der k Schachteln. □

Die 2-stufige Konstruktion führt ganz zwanglos zu der üblichen *Interpretation von bedingten Wahrscheinlichkeiten*: $\mathbf{Ws}\{A \,|\, X = x\}$ ist die Wahrscheinlichkeit von A, wenn schon bekannt ist, daß X den Wert x angenommen hat. Wählt man speziell $X = I_{A'}$, so ergibt sich folgende Deutung:

> $\mathbf{Ws}\{A \,|\, A'\}$ *ist die Wahrscheinlichkeit von* A*, wenn man bereits weiß, daß* A' *eingetreten ist.*

Die Interpretation macht Sinn, solange keine anderen Informationen vorliegen, die für das Eintreten von A belangvoll wären. Stimmig ist, daß $\mathbf{Ws}\{A \,|\, A'\} = \mathbf{Ws}\{A\}$ genau dann gilt, wenn A und A' stochastisch unabhängige Ereignisse sind.

Beispiele.

1. **Die Formel von Bayes.** Ein medizinischer Test zeige bei einer kranken Person in 95% aller Fälle, bei einer gesunden in 2% aller Fälle eine positive Reaktion. In einer Population sei der Anteil der erkrankten Individuen 1%. Wie groß ist die Wahrscheinlichkeit, daß ein zufällig ausgewähltes Individuum mit positivem Testresultat krank ist? Sei A das Ereignis, daß diese Person krank ist, und A' das Ereignis, daß sie positiv getestet wird. Gegeben sind die Wahrscheinlichkeiten $\mathbf{Ws}\{A\} = 0,01$, $\mathbf{Ws}\{A' \,|\, A\} = 0,95$, $\mathbf{Ws}\{A' \,|\, A^c\} = 0,02$, gefragt ist nach $\mathbf{Ws}\{A \,|\, A'\}$. Die Umrechnung leistet die für ein Ereignis A' und eine diskrete Zufallsvariable X gültige Gleichung

$$\mathbf{Ws}\{X = x_0 \,|\, A'\} \;=\; \frac{\mathbf{Ws}\{A' \,|\, X = x_0\}\mathbf{Ws}\{X = x_0\}}{\sum_{x\in S} \mathbf{Ws}\{A' \,|\, X = x\}\mathbf{Ws}\{X = x\}} \;,$$

eine direkte Konsequenz des Satzes von der totalen Wahrscheinlichkeit. Sie heißt **Formel von Bayes** und hat in den Anfängen der Wahrscheinlichkeitsrechnung besondere Aufmerksamkeit auf sich gezogen. Wählen wir speziell $X = I_A$ und $x_0 = 1$, so erhalten wir

$$\begin{aligned}\mathbf{Ws}\{A \,|\, A'\} \;&=\; \frac{\mathbf{Ws}\{A' \,|\, A\}\mathbf{Ws}\{A\}}{\mathbf{Ws}\{A' \,|\, A\}\mathbf{Ws}\{A\} + \mathbf{Ws}\{A' \,|\, A^c\}\mathbf{Ws}\{A^c\}} \\ &=\; \frac{0,95 \cdot 0,01}{0,95 \cdot 0,01 + 0,02 \cdot 0,99} \;=\; 0,32 \;.\end{aligned}$$

In einer zufälligen Stichprobe aus der Population (man denke etwa an eine Reihenuntersuchung) ist also nur jedes dritte positiv getestete Individuum wirklich krank.

2. **Runs.** Seien $Z_1, Z_2, \ldots$ unabhängige, Bernoulli-verteilte Zufallsvariable mit Erfolgswahrscheinlichkeit p, und seien k, l natürliche Zahlen. Wir wollen die Wahrscheinlichkeit von

$$\begin{aligned}A \;:=\; \{&\text{der erste Run aus } k \text{ Einsen} \\ &\text{tritt vor dem ersten Run aus } l \text{ Nullen auf}\}\end{aligned}$$

bestimmen. Nach dem Satz von der totalen Wahrscheinlichkeit gilt

$$\mathbf{Ws}\{A\} \;=\; p\,\mathbf{Ws}\{A \,|\, Z_1 = 1\} + q\,\mathbf{Ws}\{A \,|\, Z_1 = 0\} \;, \tag{4.1}$$

mit $q = 1 - p$, und ebenfalls

$$\begin{aligned}\mathbf{Ws}\{A\} \;=\; &\sum_{m=1}^{k} \mathbf{Ws}\{A \,|\, M = m\} \cdot \mathbf{Ws}\{M = m\} \\ &+ \mathbf{Ws}\{A \,|\, M > k\} \cdot \mathbf{Ws}\{M > k\} \;,\end{aligned}$$

wobei wir $M = \min\{m : Z_m = 0\}$ wählen, den Zeitpunkt der ersten Null. Um diese Gleichung weiter umzuformen, betrachten wir auch die Ereignisse A_m, daß in $Z_m, Z_{m+1}, \ldots$ der erste Run aus k Einsen vor dem ersten Run aus l Nullen eintritt. Aufgrund der Unabhängigkeit gilt für $m \le k$

$$\begin{aligned}\mathbf{Ws}\{A \,\big|\, M = m\} &= \mathbf{Ws}\{A_m \,\big|\, Z_1 = \cdots = Z_{m-1} = 1, Z_m = 0\} \\ &= \mathbf{Ws}\{A_m \,\big|\, Z_m = 0\} \;=\; \mathbf{Ws}\{A \,\big|\, Z_1 = 0\}\ .\end{aligned}$$

Außerdem gilt $\mathbf{Ws}\{A \,\big|\, M > k\} = 1$ und $\mathbf{Ws}\{M > k\} = p^k$, und es folgt

$$\mathbf{Ws}\{A\} \;=\; (1 - p^k)\,\mathbf{Ws}\{A \,\big|\, Z_1 = 0\} + p^k\ . \tag{4.2}$$

Umgekehrt gilt auch $\mathbf{Ws}\{A^c\} = (1 - q^l)\,\mathbf{Ws}\{A^c \,\big|\, Z_1 = 1\} + q^l$, bzw.

$$\mathbf{Ws}\{A\} \;=\; (1 - q^l)\,\mathbf{Ws}\{A \,\big|\, Z_1 = 1\}\ , \tag{4.3}$$

denn Runs aus Einsen oder Nullen treten mit Wahrscheinlichkeit 1 auf, und A^c ist daher bis auf ein Nullereignis gleich dem Ereignis, daß der erste Run aus l Nullen vor dem ersten Run aus k Einsen eintritt.

Mit (4.2) und (4.3) lassen sich aus (4.1) die bedingten Wahrscheinlichkeiten eliminieren, und wir erhalten

$$\mathbf{Ws}\{A\} \;=\; \frac{qp^k/(1-p^k)}{qp^k/(1-p^k) + pq^l/(1-q^l)}\ .$$

Man bemerke: Für großes k wird ein Run aus k Einsen normalerweise durch eine 0 angeführt (denn dann wird nur ausnahmsweise der Run gleich mit den ersten k Werten realisiert), und die entsprechende Wahrscheinlichkeit ist qp^k.

Bedingte Wahrscheinlichkeiten werden dazu benutzt, um Zufallsexperimente stufenweise aufzubauen oder aber in neu gewählte Stufen zu zerlegen. Beide Aspekte kommen im nächsten Abschnitt zum Tragen.

4.3 Das Urnenmodell von Pólya

Wir betrachten nun ein Zufallsexperiment mit mehr als 2 Stufen. Einer Urne, die rote und schwarze Kugeln enthält, werden sukzessive Kugeln entnommen, und zwar nach der folgenden, von PÓLYA vorgeschlagenen Regel:

Jede gezogene Kugel wird vor dem nächsten Ziehen in die Urne zurückgelegt, zusammen mit einer weiteren Kugel derselben Farbe.

Der Inhalt der Urne wächst also fortwährend. Pólya dachte an ein Modell für die Ausbreitung einer Infektion in einer Population, die verschiedenen Kugeln stehen dabei für infizierte bzw. immunisierte Individuen. Heute findet die Pólya-Urne und verfeinerte Urnenmodelle Beachtung in der Populationsgenetik.

Die Zusammensetzung der Urne ändert sich zufällig. Was läßt sich darüber aussagen? Wir betrachten die 01-wertigen Zufallsvariablen

$$Z_i := \begin{cases} 1, & \text{falls die } i\text{-te gezogene Kugel rot ist,} \\ 0, & \text{falls die } i\text{-te gezogene Kugel schwarz ist.} \end{cases}$$

Dann ist Pólyas Mechanismus formal in den Forderungen

$$\mathbf{Ws}\{Z_1 = 1\} = \frac{r}{t}\,, \quad \mathbf{Ws}\{Z_1 = 0\} = \frac{s}{t}$$

sowie

$$\begin{aligned} \mathbf{Ws}\{Z_{n+1} = 1 \,\big|\, Z_1 = z_1, \ldots, Z_n = z_n\} &= \frac{r+x}{t+n}\,, \\ \mathbf{Ws}\{Z_{n+1} = 0 \,\big|\, Z_1 = z_1, \ldots, Z_n = z_n\} &= \frac{s+y}{t+n} \end{aligned}$$

erfaßt, mit

$$\begin{array}{lcll} t & = & r+s & = \text{anfängliche Kugelzahl, mit} \\ & & & r \text{ roten und } s \text{ schwarzen Kugeln,} \\ x & = & z_1 + \cdots + z_n & = \text{Zahl der nach } n\text{-maligem Ziehen} \\ & & & \text{hinzugelegten roten Kugeln,} \\ y & = & n - x & = \text{Zahl der nach } n\text{-maligem Ziehen} \\ & & & \text{hinzugelegten schwarzen Kugeln.} \end{array}$$

Durch diese Annahmen ist die gemeinsame Verteilung von $Z_1, \ldots, Z_n$ bereits festgelegt, denn es folgt

$$\begin{aligned} &\mathbf{Ws}\{Z_1 = z_1, \ldots, Z_n = z_n\} \\ &\quad = \frac{r(r+1)\cdots(r+x-1)\cdot s(s+1)\cdots(s+y-1)}{t(t+1)\cdots(t+n-1)}\,. \end{aligned} \tag{4.4}$$

Der Beweis läßt sich leicht durch Induktion nach n führen, mit Hilfe der Gleichung

$$\begin{aligned} &\mathbf{Ws}\{Z_1 = z_1, \ldots, Z_{n+1} = z_{n+1}\} = \\ &\qquad \mathbf{Ws}\{Z_{n+1} = z_{n+1} \,\big|\, Z_1 = z_1, \ldots, Z_n = z_n\} \cdot \mathbf{Ws}\{Z_1 = z_1, \ldots, Z_n = z_n\}\,. \end{aligned}$$

Für

$$X \;:=\; Z_1 + \cdots + Z_n \;,$$

der Zahl der nach n-maliger Wiederholung gezogenen roten Kugeln, ergibt sich aus (4.4) die Formel

$$\begin{aligned}\mathbf{Ws}\{X = x\} \;&=\; \sum_{z_1+\cdots+z_n=x} \mathbf{Ws}\{Z_1 = z_1, \ldots, Z_n = z_n\} \\ &=\; \binom{n}{x}\, \mathbf{Ws}\{Z_1 = z_1, \ldots, Z_n = z_n\} \\ &=\; \binom{r+x-1}{x}\binom{s+y-1}{y}\Big/\binom{t+n-1}{n} \;.\end{aligned}$$

Im Fall $r = s = 1$ erhält man recht übersichtlich

$$\mathbf{Ws}\{Z_1 = z_1, \ldots, Z_n = z_n\} \;=\; \frac{x!y!}{(n+1)!} \tag{4.5}$$

und die bemerkenswerte Formel

$$\mathbf{Ws}\{X = x\} \;=\; \frac{1}{n+1} \;, \quad x = 0, \ldots, n \;. \tag{4.6}$$

X ist dann uniform verteilt, es gibt keine bevorzugte Belegung der Urne.

Eine zweistufige Zerlegung austauschbarer Zufallsvariabler

In die zufällige Entwicklung des Inhalts der Pólya-Urne bieten diese Formeln nur einen vorläufigen Einblick. Entscheidend für ein vertieftes Verständnis ist ein struktureller Gesichtspunkt: Nach (4.4) erzeugt Pólyas Mechanismus eine Folge $Z_1, \ldots, Z_n$ mit einer *austauschbaren Verteilung*.

Wir betrachten zunächst für ein fest vorgegebenes n beliebige 01-wertige Zufallsvariable $Z_1, \ldots, Z_n$ mit einer austauschbaren Verteilung. Dies bedeutet (vgl. Abschnitt 3.1, Beispiel 4), daß $\mathbf{Ws}\{Z_1 = z_1, \ldots, Z_n = z_n\}$ unverändert bleibt, wenn man die z_i untereinander vertauscht. Für

$$X \;:=\; Z_1 + \cdots + Z_n$$

und $x = 0, 1, \ldots, n$ gilt

$$\mathbf{Ws}\{X = x\} \;=\; \sum_{z_1+\cdots+z_n=x} \mathbf{Ws}\{Z_1 = z_1, \ldots, Z_n = z_n\} \;.$$

Die Summe enthält $\binom{n}{x}$ Summanden, die nach Voraussetzung alle gleich sind, daher folgt

$$\mathbf{Ws}\{Z_1 = z_1, \ldots, Z_n = z_n\} \;=\; \frac{\mathbf{Ws}\{X = x\}}{\binom{n}{x}} \;,$$

bzw.

$$\mathbf{Ws}\{Z_1 = z_1, \ldots, Z_n = z_n \,\big|\, X = x\} \;=\; \frac{1}{\binom{n}{x}} \;, \quad \text{mit } x = z_1 + \cdots + z_n \;.$$

Diese Formeln zeigen: Die gemeinsame Verteilung von $Z_1, \ldots, Z_n$ ist durch die Verteilung von X, d.h. durch die Gewichte

$$p_x \;:=\; \mathbf{Ws}\{X = x\}$$

vollständig festgelegt. An die Verteilung von X bestehen keine Einschränkungen. Wir erhalten so einen Überblick über die austauschbaren Verteilungen. Gleichzeitig eröffnet sich eine Möglichkeit, wie man 01-wertige, austauschbare Zufallsvariable $Z_1, \ldots, Z_n$ in einem 2-stufigen Experiment erzeugen kann:

1. Wähle ein zufälliges Element X aus $\{0, 1, \ldots, n\}$, und zwar x mit Wahrscheinlichkeit p_x.

2. Ist das Ereignis $\{X = x\}$ eingetreten, so ziehe aus einer Urne mit x roten und $y = n - x$ schwarzen Kugeln alle Kugeln der Reihe nach ohne Zurücklegen heraus. Setze $Z_i = 1$ oder 0, je nachdem ob die i-te gezogene Kugel rot oder schwarz ist.

Für dieses Experiment gilt

$$\mathbf{Ws}\{Z_1 = z_1, \ldots, Z_n = z_n\} \;=\; \frac{p_x}{\binom{n}{x}} \;=\; \frac{\mathbf{Ws}\{X = x\}}{\binom{n}{x}} \;, \quad x = z_1 + \cdots + z_n \;,$$

$Z_1, \ldots, Z_n$ hat also die gewünschte Verteilung.

Diese 2-stufige Zerlegung wenden wir nun auf das n-malige Ziehen von Kugeln nach Pólyas Schema an. Der Einfachheit halber sei $r = s = 1$. Nach (4.6) erhalten wir folgendes **Ersatzexperiment 1 für n-maliges Ziehen aus einer Pólya-Urne**:

A. Wähle ein Element X aus $\{0, 1, \ldots, n\}$, mit uniformer Verteilung.

B. Nimmt X den Wert x an, so ziehe nacheinander alle Kugeln aus einer Urne mit x roten und $n - x$ schwarzen Kugeln. Setze $Z_i = 1$ oder 0, je nachdem ob die i-te Kugel rot oder schwarz ist.

In seiner Durchführung unterscheidet es sich deutlich vom ursprünglichen Urnenexperiment: Zunächst wird die Zusammensetzung der Urne nach n Schritten bestimmt, erst danach wird entschieden, in welcher Reihenfolge die Kugeln gezogen werden. In den stochastischen Eigenschaften bestehen aber keine Unterschiede, beide Zufallsexperimente führen zu derselben gemeinsamen Verteilung von $Z_1, \ldots, Z_n$.

Ein Nachteil des Ersatzexperiments ist, daß man (anders als bei der Pólya-Urne) ganz von vorn anfangen muß, wenn man etwa $n+1$ statt n Züge aus der Pólya-Urne simulieren will. Um diesen Nachteil zu beseitigen, betrachten wir für jedes $l \geq n$ das **Ersatzexperiment 2 für n-maliges Ziehen aus einer Pólya-Urne**:

A′. Ziehe rein zufällig ein Element U_l aus $\{0, \frac{1}{l}, \frac{2}{l}, \ldots, 1\}$.

B′. Hat $X_l := l \cdot U_l$ den Wert x angenommen, so ziehe ohne Zurücklegen n Kugeln aus einer Urne mit x roten und $l-x$ schwarzen Kugeln. Setze $Z_i = 1$ oder 0, falls die i-te Kugel rot oder schwarz ist, $i = 1, \ldots, n$.

Es ist leicht zu sehen daß dieses Experiment das gewünschte leistet. Der erste Schritt bereitet analog zum Ersatzexperiment 1 daß l-malige Ziehen aus einer Pólya-Urne vor. Im zweiten Schritt wird dementsprechend eine Urne mit insgesamt l Kugeln betrachtet, die relative Häufigkeit der roten Kugel ist U_l. Es werden dann aber statt l nur n Züge realisiert. - Wir können nun l sehr viel größer als n wählen, dann spielt es kaum noch eine Rolle, ob man im zweiten Schritt die n Kugeln mit oder ohne Zurücklegen zieht. Im Limes $l \to \infty$ nimmt das Ersatzexperiment eine besonders attraktive Gestalt an. Offenbar ist dann U_l asymptotisch uniform in $[0,1]$ verteilt, und wir erhalten das **Ersatzexperiment 3 für die Pólya-Urne**:

A″. Wähle eine zufällige reelle Zahl U, mit uniformer Verteilung in $[0,1]$.

B″. Nimmt U den Wert p an, so erzeuge unabhängige, Bernoulli-verteilte Zufallsvariable $Z_1, Z_2, \ldots$ mit Erfolgswahrscheinlichkeit p.

Die erste Stufe ist nun nicht mehr von n abhängig, dieses Experiment ist geeignet, eine beliebige Anzahl von Zügen aus einer Pólya-Urne simulieren. Es ermöglicht bemerkenswerte Einsichten in die asymptotische Zusammensetzung der Urne, die dem ursprünglichen Urnenexperiment nicht so leicht anzusehen sind: Nimmt U den Wert p an, so konvergiert $(Z_1 + \cdots + Z_n)/n$ nach dem Gesetz der großen Zahlen für $n \to \infty$ fast sicher gegen p, d.h. es gilt

$$\frac{Z_1 + \cdots + Z_n}{n} \;\to\; U \text{ f.s..}$$

Eine eingehendere Untersuchung der fast sicheren Konvergenz von Folgen reellwertiger Zufallsvariablen (die in der Höheren Stochastik nachgeholt wird) zeigt, daß sich dieses Resultat von dem Ersatzexperiment auf das ursprüngliche Urnenexperiment überträgt. Für die Pólya-Urne stabilisiert sich der relative Anteil der roten Kugeln mit wachsendem n fast sicher bei einem Wert U. Dieser Wert ist zufällig und uniform in $[0,1]$ verteilt. Man kann das Ersatzexperiment 3 also so verstehen, daß zunächst die asymptotische relative Häufigkeit U der roten Kugeln in der Urne festgelegt wird. Bedingt auf diesen Wert erweisen sich die Zufallsvariablen $Z_1, Z_2, \ldots$ als unabhängig.

Unser letztes 2-stufiges Experiment unterscheidet sich von den vorherigen dadurch, daß die in der ersten Stufe betrachtete Zufallsvariable nicht mehr diskret ist. Dementsprechend nimmt der **Satz von der totalen Wahrscheinlichkeit** eine andere Gestalt an. Er lautet nun (mit $x = z_1 + \cdots + z_n$)

$$\begin{aligned}&\mathbf{Ws}\{Z_1 = z_1, \ldots, Z_n = z_n\}\\ &\quad = \int_0^1 \mathbf{Ws}\{Z_1 = z_1, \ldots, Z_n = z_n \,\big|\, U = p\}\, dp \;=\; \int_0^1 p^x (1-p)^{n-x}\, dp\;.\end{aligned}$$

Die allgemeine Begründung solcher Formeln wird in der Maßtheorie geleistet. In unserem Fall können wir sie direkt bestätigen: Der Wert des Integrals berechnet sich mittels partieller Integration als $x!(n-x)!/(n+1)!$, in Übereinstimmung mit Formel (4.5).

Analoge Resultate sind gültig, wenn die Urne anfänglich mehr als 2 Kugeln enthält, r rote und s schwarze. Mit etwas Mehraufwand läßt sich zeigen, daß das äquivalente Ersatzexperiment nun die folgende Gestalt besitzt:

A″. Ziehe U zufällig aus $[0,1]$, so daß $\mathbf{Ws}\{x \le U \le x + dx\} = f(x)\, dx$ mit der Dichte $f(x) := cx^{r-1}(1-x)^{s-1}$ gilt.

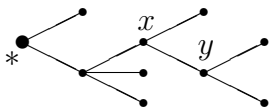
B″. Nimmt U den Wert p an, so erzeuge unabhängige, Bernoulli-verteilte Zufallsvariable Z_1, $Z_2, \ldots$ mit Erfolgswahrscheinlichkeit p.

Die Konstante c ist so zu wählen, daß $f(x)$ sich auf $[0,1]$ zu 1 aufintegriert. $f(x)$ ist die Dichte der **Beta-Verteilung**, die wir schon im Zusammenhang mit Ordnungsstatistiken kennengelernt haben (vgl. Abschnitt 2.3).

Sätze, die Folgen von austauschbaren Zufallsvariablen wie in Experiment 3 auf unabhängige Zufallsvariablen mit einer zufälligen Verteilung zurückführen, heißen nach ihrem Entdecker **de Finetti-Theoreme**.

4.4 Mehrstufige Experimente

Manche Experimente setzen sich aus einer zufälligen Anzahl von Stufen zusammen. Man kann sich die Abfolge der Teilexperimente durch baumartige Graphen (Graphen ohne geschlossene Wege) veranschaulichen.



Eine Realisation des Gesamtexperimentes entspricht einem zufälligen Weg von der Wurzel $*$ in eines der Blätter des Baumes (in der Zeichnung von links nach rechts). Jeder interne Knoten steht für ein Teilexperiment, in dem entschieden wird, über welche Kante man den Knoten verläßt. Die Anzahl der Teilexperimente ist gleich der Länge des Weges, also der Anzahl der Kanten zwischen Wurzel und dem erreichten Blatt (man spricht von der *Tiefe* des Blattes).

Sei

$$A_x \;=\; \{\text{der Weg geht durch } x\}$$

das Ereignis, daß auf dem zufälligen Weg durch den Graphen der Knoten x passiert wird. Wir leiten nun eine allgemeine Formel für seine Wahrscheinlichkeit ab. Dazu beachten wir, daß jeder Knoten $y \neq *$ einen eindeutig bestimmten *Vorgänger* x hat, der Knoten, der auf dem Weg von $*$ nach y am nächsten bei y liegt. Umgekehrt heißt y *direkter Nachfolger* von x. Ist x der Vorgänger von y, so definieren wir die **Übergangswahrscheinlichkeit** von x nach y als

$$P_{xy} \;:=\; \mathbf{Ws}\{A_y \,\big|\, A_x\}\ .$$

Offenbar gilt dann $A_y \subset A_x$ und damit

$$P_{xy} \;=\; \frac{\mathbf{Ws}\{A_y\}}{\mathbf{Ws}\{A_x\}}\ .$$

Aus $\bigcup_{y\in N(x)} A_y = A_x$ folgt

$$\sum_{y\in N(x)} P_{xy} \;=\; 1\ ,$$

dabei sei $N(x)$ die Menge aller direkten Nachfolger von x. Es sind daher $(P_{xy})_y$ die Wahrscheinlichkeiten für das Teilexperiment, das im Knoten x

ausgeführt wird. Sind $x_1, \ldots, x_n$ die Knoten auf dem Weg von $*$ nach x, so folgt unter Beachtung von $\mathbf{Ws}\{A_*\} = 1$

$$\mathbf{Ws}\{A_x\} \;=\; P_{*x_1} \cdot P_{x_1x_2} \cdots P_{x_nx} \;.$$

Die Übergangswahrscheinlichkeiten legen also alle anderen Wahrscheinlichkeiten fest. - Allgemeiner besteht für beliebige Ereignisse $A_1, \ldots, A_n$ die leicht zu verifizierende **Multiplikationsformel**

$$\begin{aligned}&\mathbf{Ws}\{A_1 \cap \cdots \cap A_n\}\\ &\quad = \;\mathbf{Ws}\{A_1\} \cdot \mathbf{Ws}\{A_2 \,\big|\, A_1\} \cdots \mathbf{Ws}\{A_n \,\big|\, A_1 \cap \cdots \cap A_{n-1}\} \;.\end{aligned}$$

Beispiele.

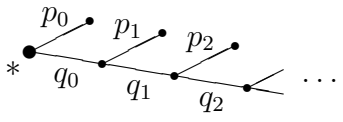
1. **Gedächtnislosigkeit der geometrischen Verteilung.** Sei T eine Zufallsvariable mit Werten in $\{0, 1, 2, \ldots\}$ und einer beliebigen Verteilung. Dann gilt

   $$\mathbf{Ws}\{T = n\} \;=\; q_0 \cdot q_1 \cdots q_{n-1} \cdot p_n \;,$$

   mit

   $$p_n \;:=\; \mathbf{Ws}\{T = n \,\big|\, T \geq n\} \;, \quad q_n \;:=\; \mathbf{Ws}\{T \geq n+1 \,\big|\, T \geq n\} \;.$$

   Wir können also T auffassen als die Anzahl der Mißerfolge vor dem ersten Erfolg in einer Reihe von Experimenten mit variablen Erfolgswahrscheinlichkeiten p_n.

   Stellt man sich vor, daß T die Lebensdauer (in Tagen) einer Komponente in einem technischen System ist, so drückt die Produktdarstellung von $\mathbf{Ws}\{T = n\}$ aus, wie ihre Funktionstüchtigkeit von der Zeit abhängt. Wenn p_n sich mit wachsendem n verändert, haben wir es mit einem Alterungsprozeß der Komponente zu tun. p_n ist genau dann von n unabhängig, falls T geometrisch verteilt ist,

   $$\mathbf{Ws}\{T = n\} \;=\; q^n p \;,$$

   man spricht daher von der **Gedächtnislosigkeit der geometrischen Verteilung**. Diese Verteilungsannahme bedeutet also, daß der Ausfall

der Komponente nicht auf Abnutzung oder Alterung zurückzuführen ist, sondern eher auf äußere Einflüsse.

Diese Eigenschaft der Gedächtnislosigkeit drückt sich in vielfältiger Weise aus. Z.B. gilt $\mathbf{Ws}\{T \geq n\} = q^n$ für eine geometrisch verteilte Zufallsvariable ($\{T \geq n\}$ ist das Ereignis, daß anfangs n Misserfolge eintreten) und folglich

$$\mathbf{Ws}\{T = n+m \,\big|\, T \geq m\} \;=\; \frac{\mathbf{Ws}\{T = n+m\}}{\mathbf{Ws}\{T \geq m\}} \;=\; q^n p \;= \mathbf{Ws}\{T = n\}\;.$$

In Worten ausgedrückt: Die Kenntnis, daß eine Komponente schon n Tage in Betrieb ist, hat keinen Einfluss auf ihre verbleibende Funktionsdauer.

Ähnliches trifft für eine $\mathbb{R}^+$-wertige Zufallsvariable X mit einer Exponentialverteilung zu,

$$\mathbf{Ws}\{x \leq X \leq x + dx\} = \lambda e^{-\lambda x}\, dx\;, \quad x \geq 0\;.$$

Dann gilt $\mathbf{Ws}\{X \geq t\} = \int_t^\infty \lambda e^{-\lambda x}\, dx = e^{-\lambda t}$ und

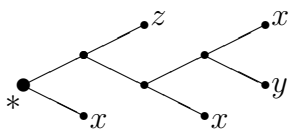
$$\mathbf{Ws}\{X \geq t+s \,\big|\, X \geq t\} \;=\; \frac{\mathbf{Ws}\{X \geq t+s\}}{\mathbf{Ws}\{X \geq t\}} \;=\; e^{-\lambda s}\;,$$

und es folgt

$$\mathbf{Ws}\{X \geq t+s \,\big|\, X \geq t\} \;=\; \mathbf{Ws}\{X \geq s\}$$

für alle $s, t \geq 0$. Man spricht von der **Gedächtnislosigkeit der Exponentialverteilung**, sie ist für die Exponentialverteilung charakteristisch.

2. **Binäre Simulation von Verteilungen.** Aus $S = \{x, y, \ldots\}$ soll zufällig ein Element ausgewählt werden, und zwar x mit Wahrscheinlichkeit p_x ($\sum_x p_x = 1$). Man kann dies durch eine Folge von Münzwürfen realisieren, unter Benutzung geeigneter binärer Bäume mit beschrifteten Blättern. Der Baum



gehört zu der Verteilung $p_x = \frac{1}{2} + \frac{1}{8} + \frac{1}{16} = \frac{11}{16}$, $p_y = \frac{1}{16}$, $p_z = \frac{1}{4}$. Ausgehend von der Wurzel $*$ wählt man per fairem Münzwurf einen zufälligen Weg nach rechts durch den Baum, bis ein Blatt erreicht ist. Seine Beschriftung ist dann das aus S ausgewählte Element.

In Abschnitt 7.3 zeigen wir, daß sich jede Verteilung $\mu = (p_x)$ mit Hilfe eines geeigneten binären Baums simulieren läßt. □

4.5 Bedingte Erwartungen

Die Vorgehensweise, Wahrscheinlichkeiten nach dem Satz von der totalen Wahrscheinlichkeit in bedingte Wahrscheinlichkeiten zu zerlegen, erweist sich ganz ähnlich auch für Erwartungswerte als fruchtbar.

Definition. *Sei X eine diskrete, reellwertige Zufallsvariable mit endlichem Erwartungswert. Dann ist die* ***bedingte Erwartung von X bzgl. A*** *definiert als*

$$\mathbf{E}[X \,|\, A] \;:=\; \sum_x x \cdot \mathbf{Ws}\{X = x \,|\, A\} \;.$$

Wegen $\mathbf{Ws}\{X = x \,|\, A\} \le \mathbf{Ws}\{X = x\}/\mathbf{Ws}\{A\}$ ist mit dem Erwartungswert $\mathbf{E}X$ auch $\mathbf{E}[X \,|\, A]$ wohldefiniert und endlich. Es ist klar, daß sich die Eigenschaften des gewöhnlichen Erwartungswertes auf bedingte Erwartungen übertragen. Sind X und A unabhängig, so gilt $\mathbf{E}[X \,|\, A] = \mathbf{E}X$. Der Satz von der totalen Wahrscheinlichkeit nimmt nun die folgende Gestalt an.

Proposition 4.2. *Sei X eine reellwertige Zufallsvariable mit endlicher Erwartung und sei Y eine diskrete Zufallsvariable mit Werten in S. Dann gilt*

$$\mathbf{E}X \;=\; \sum_{y \in S} \mathbf{E}[X \,|\, Y = y] \cdot \mathbf{Ws}\{Y = y\} \;.$$

Beweis. Nach Satz 4.1 gilt

$$\begin{aligned} \sum_x x \cdot \mathbf{Ws}\{X = x\} \;&=\; \sum_x \sum_y x \cdot \mathbf{Ws}\{X = x \,|\, Y = y\} \cdot \mathbf{Ws}\{Y = y\} \\ &=\; \sum_y \mathbf{E}[X \,|\, Y = y] \cdot \mathbf{Ws}\{Y = y\} \;. \end{aligned}$$

□

Beispiel. Einen Algorithmus mit deterministischem Endresultat, aber zufälliger Laufzeit T nennt man einen **Las Vegas Algorithmus**. Wenn ihn ein ungeduldiger Benutzer immer nach spätestens t Zeiteinheiten unterbricht und neu startet, wird er erst nach N Fehlversuchen und einer Gesamtlaufzeit X ein Endresultat liefern. Wir berechnen den Erwartungswert von X, und zwar unter der Annahme, daß die Laufzeiten bei verschiedenen Anwendungen des Algorithmus unabhängige Zufallsvariable sind. Es gilt

$$\mathbf{E}[X \,|\, N = n] \;=\; nt + \mathbf{E}[T \,|\, T \le t]$$

und daher nach Proposition 4.2

$$\mathbf{E}X \;=\; t\,\mathbf{E}N + \mathbf{E}[T \,|\, T \le t]$$

Nach der Unabhängigkeitsannahme ist N geometrisch verteilt mit Erwartungswert $\mathbf{Ws}\{T > t\}/\mathbf{Ws}\{T \le t\}$, daher folgt

$$\mathbf{E}X \;=\; \big(t\,\mathbf{Ws}\{T > t) + \mathbf{E}[T \,\big|\, T \le t]\mathbf{Ws}\{T \le t\}\big)/\mathbf{Ws}\{T \le t\}\;.$$

Nach Proposition 4.2 gilt weiter

$$\mathbf{E}[\min(T,t)] \;=\; t\,\mathbf{Ws}\{T > t) + \mathbf{E}[T \,\big|\, T \le t]\mathbf{Ws}\{T \le t\}\;,$$

daher erhalten wir insgesamt

$$\mathbf{E}X \;=\; \frac{\mathbf{E}[\min(T,t)]}{\mathbf{Ws}\{T \le t\}}\;.$$

Dieser Wert kann kleiner, aber auch größer als $\mathbf{E}T$ sein. Man bemerke: Bei einem Las Vegas Algorithmus ist es unproblematisch, den Algorithmus zu unterbrechen und neu zu starten. Dies gilt nicht für **Monte Carlo Algorithmen**, also Algorithmen, deren Endresultat R zufällig ist. Unterbricht man einen solchen Algorithmus immer nach t Zeiteinheiten, so wird das Endresultat R' im allgemeinen nicht wie R verteilt sein, sondern die bedingte Verteilung $\mathbf{Ws}\{R \in \cdot \,\big|\, T \le t\}$ als Verteilung besitzen. Nur wenn R und T unabhängige Zufallsvariable sind, kann man sicher sein, daß das Endresultat durch die Unterbrechungen nicht verfälscht wird. □

Proposition 4.2 läßt sich eine kompakte Gestalt geben, indem wir in die bedingten Erwartungen wieder den Zufall einführen (wir haben uns schon in Abschnitt 4.1 von diesem Gedanken leiten lassen.) Dazu definieren wir mit Hilfe von

$$e(y) \;:=\; \mathbf{E}[X \,\big|\, Y = y]$$

die **bedingte Erwartung von X bzgl. Y** als

$$\mathbf{E}[X \,\big|\, Y] \;:=\; e(Y)\;.$$

Man beachte, daß es sich bei dieser bedingten Erwartung um eine *Zufallsvariable* handelt. Sie hat den Erwartungswert

$$\mathbf{E}[\mathbf{E}[X \,\big|\, Y]] \;=\; \sum_y e(y)\cdot\mathbf{Ws}\{Y = y\} \;=\; \sum_y \mathbf{E}[X \,\big|\, Y = y]\;\mathbf{Ws}\{Y = y\}\;,$$

Proposition 4.2 können wir daher in der Gleichung

$$\mathbf{E}X \;=\; \mathbf{E}[\mathbf{E}[X \,\big|\, Y]] \tag{4.7}$$

zusammenfassen. Dies ist die Formel, die wir in Abschnitt 4.1 zum Berechnen der Erwartungswerte benutzt haben. Sind X und Y unabhängig, so gilt $\mathbf{E}[X \,|\, Y] = \mathbf{E}X$.

Wir wenden nun Proposition 4.2 auf die Berechnung von Varianzen an. Die **bedingte Varianz von X, gegeben das Ereignis A**, definieren wir als

$$\mathbf{Var}[X \,|\, A] \;:=\; \mathbf{E}[\big(X - \mathbf{E}[X \,|\, A]\big)^2 \,|\, A] \;.$$

Proposition 4.3. *Sei X reellwertig mit endlichem zweiten Moment $\mathbf{E}X^2$, und sei Y eine diskrete Zufallsvariable. Dann gilt*

$$\begin{aligned}\mathbf{E}[X^2] \;&=\; \sum_y \mathbf{Var}[X \,|\, Y = y] \cdot \mathbf{Ws}\{Y = y\} \\ &\qquad +\; \sum_y \mathbf{E}[X \,|\, Y = y]^2 \cdot \mathbf{Ws}\{Y = y\} \;.\end{aligned}$$

Beweis. Nach Proposition 4.2 gilt

$$\mathbf{E}[X^2] \;=\; \sum_y \mathbf{E}[X^2 \,|\, Y = y] \cdot \mathbf{Ws}\{Y = y\} \;,$$

und nach (3.2)

$$\mathbf{E}[X^2 \,|\, Y = y] \;=\; \mathbf{Var}[X \,|\, Y = y] + \mathbf{E}[X \,|\, Y = y]^2 \;.$$

□

Indem wir X durch $X - \mathbf{E}X$ ersetzen, erhalten wir folgende Version von Proposition 4.3,

$$\begin{aligned}\mathbf{Var}X \;&=\; \sum_y \mathbf{Var}[X \,|\, Y = y] \cdot \mathbf{Ws}\{Y = y\} \\ &\qquad +\; \sum_y \big(\mathbf{E}[X \,|\, Y = y] - \mathbf{E}X\big)^2 \cdot \mathbf{Ws}\{Y = y\} \;. \qquad (4.8)\end{aligned}$$

Also: Zwar können die bedingten Varianzen $\mathbf{Var}[X \,|\, Y = y]$ im allgemeinen kleiner oder auch größer als die unbedingte Varianz $\mathbf{Var}X$ sein. Im Mittel verkleinert sich jedoch die Varianz von X bei Kenntnis von Y. Der Differenzbetrag ist nach (4.8) die mittlere quadratische Abweichung zwischen $\mathbf{E}[X \,|\, Y = y]$ und $\mathbf{E}X$.

Definieren wir die **bedingte Varianz von X bzgl. Y** als die Zufallsvariable

$$\mathbf{Var}[X \,|\, Y] \;:=\; v(Y)$$

mit

$$v(y) \;:=\; \mathbf{Var}[X \,\big|\, Y = y] \;,$$

so lassen sich Proposition 4.3 und Gleichung (4.8) zu den Formeln

$$\mathbf{E}[X^2] \;=\; \mathbf{E}[\mathbf{Var}[X \,\big|\, Y]] + \mathbf{E}[\mathbf{E}[X \,\big|\, Y]^2] \tag{4.9}$$

bzw.

$$\mathbf{Var}X \;=\; \mathbf{E}[\mathbf{Var}[X \,\big|\, Y]] + \mathbf{Var}[\mathbf{E}[X \,\big|\, Y]] \tag{4.10}$$

zusammenfassen.

Beispiel. Summen von zufälliger Länge. Seien $Z_1, Z_2, \ldots$ unabhängige Kopien einer reellwertigen Zufallsvariablen Z mit endlicher Erwartung, und sei Y eine davon unabhängige Zufallsvariable mit Werten in $\mathbb{N}$ und endlicher Erwartung. Dann gilt für den Erwartungswert von

$$X \;:=\; Z_1 + \cdots + Z_Y$$

die Gleichung

$$\mathbf{E}X \;=\; \mathbf{E}Y \cdot \mathbf{E}Z \;.$$

Denn wegen der Unabhängigkeit von Y und $X_1, X_2, \ldots$ gilt

$$\begin{aligned} \mathbf{Ws}\{X = x \,\big|\, Y = y\} \;&=\; \mathbf{Ws}\{Z_1 + \cdots + Z_y = x \,\big|\, Y = y\} \\ &=\; \mathbf{Ws}\{Z_1 + \cdots + Z_y = x\} \;, \end{aligned}$$

also

$$\mathbf{E}[X \,\big|\, Y = y] \;=\; \mathbf{E}[Z_1 + \cdots + Z_y] \;=\; y \cdot \mathbf{E}Z \;.$$

Es folgt $\mathbf{E}[X \,\big|\, Y] = Y \cdot \mathbf{E}Z$ und damit nach (4.7) die Behauptung.

Die Varianz von X setzt sich aus zwei Anteilen zusammen. Zum einen geht darin die Varianz der Summanden ein, zum anderen die Variabilität in der Länge Y der Summe. Genauer gilt aufgrund von Unabhängigkeit

$$\mathbf{Var}[X \,\big|\, Y = y] \;=\; \mathbf{Var}[Z_1 + \cdots + Z_y] \;=\; y \cdot \mathbf{Var}Z$$

und damit $\mathbf{Var}[X \,\big|\, Y] = Y \cdot \mathbf{Var}Z$. Nach (4.10) folgt

$$\mathbf{Var}X \;= \mathbf{E}Y \cdot \mathbf{Var}Z + \mathbf{E}[Z]^2 \cdot \mathbf{Var}Y \;.$$

Sofern es sich bei Z um eine Zufallsvariable mit Werten in $\mathbb{N}_0$ handelt, kann man diese Formeln auch mittels erzeugender Funktionen ableiten. Seien

$\phi(t) = \mathbf{E}[t^Z]$ und $\psi(s) = \mathbf{E}[s^Y]$ die erzeugenden Funktionen von Z und Y. Aufgrund der Unabhängigkeitsannahmen gilt

$$\mathbf{E}[t^X \,\big|\, Y = y] \;=\; \mathbf{E}[t^{Z_1+\cdots+Z_y} \,\big|\, Y = y] \;=\; \mathbf{E}[t^{Z_1+\cdots+Z_y}] \;=\; \phi(t)^y \;.$$

Die erzeugende Funktion von X ergibt sich also nach (4.7) als

$$\mathbf{E}[t^X] \;=\; \mathbf{E}[\phi(t)^Y] \;=\; \psi(\phi(t)) \;.$$

Durch zweimaliges Differenzieren kann man nun erneut Erwartung und Varianz berechnen. □

Bemerkung. Bedingte Erwartungen als Prädiktoren. Unter der Prädiktion von Zufallsvariablen versteht man die Aufgabe, den Wert einer Zufallsvariablen X aufgrund der Beobachtung des Wertes einer anderen Zufallsvariablen Y möglichst gut vorherzusagen. Diese Aufgabe stellt sich zum Beispiel bei der Steuerung von zufällig gestörten Systemen. Wir setzen voraus, daß X reellwertige Zufallsvariable ist, und fragen wie man die reellwertige Funktion ϕ wählen sollte, damit $\phi(Y)$ ein geeigneter Prädiktor für X wird. Als Kriterium für seine Güte benutzen wir ihre mittlere quadratische Differenz

$$\mathbf{E}[(X - \phi(Y))^2] \;,$$

deren Wert wir minimieren wollen. Es gilt

$$\begin{aligned}
\mathbf{E}[X - \phi(Y) \,\big|\, Y = y] \;&=\; \mathbf{E}[X - \phi(y) \,\big|\, Y = y] \;=\; \mathbf{E}[X \,\big|\, Y = y] - \phi(y) \;, \\
\mathbf{Var}[X - \phi(Y) \,\big|\, Y = y] \;&=\; \mathbf{Var}[X - \phi(y) \,\big|\, Y = y] \;=\; \mathbf{Var}[X \,\big|\, Y = y] \;,
\end{aligned}$$

daher folgt nach(4.9)

$$\mathbf{E}[(X - \phi(Y))^2] \;=\; \mathbf{E}\,\mathbf{Var}[X \,\big|\, Y] + \mathbf{E}[\big(\mathbf{E}[X \,\big|\, Y] - \phi(Y)\big)^2] \;.$$

Dieser Ausdruck wird minimal, wenn der zweite Summand verschwindet, was bei der Wahl $\phi(y) = \mathbf{E}[X \,\big|\, Y = y]$ bzw. $\phi(Y) = \mathbf{E}[X \,\big|\, Y]$ der Fall ist. Im Sinne einer minimalen mittleren quadratischen Abweichung ist also die bedingte Erwartung $\mathbf{E}[X \,\big|\, Y]$ der beste Prädiktor von X. □

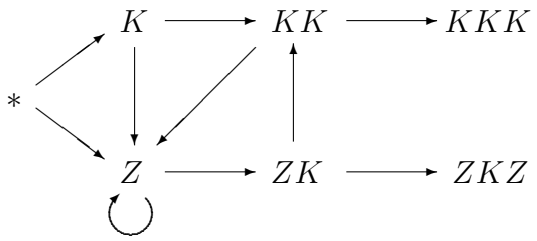
Kapitel 5

Markov-Ketten

In diesem Abschnitt betrachten wir Folgen von Teilexperimenten, deren Ablauf sich als Veränderungen in einem Raum von Zuständen beschreiben läßt. Von einer Markov-Kette spricht man, falls die Zustandswechsel in gedächtnisloser Weise stattfinden, d.h. der Fortgang des Geschehens vom vergangenen Verlauf unbeeinflußt bleibt. In Verallgemeinerung der in Abschnitt 4.4 beschriebenen Situation lassen sich Markov-Ketten als zufällige Wanderungen durch einen Graphen veranschaulichen. Die Theorie der Markov-Ketten ist umfangreich, wir behandeln sie in ein paar Ansätzen und illustrieren sie an einer Anzahl von Beispielen.

5.1 Grundlegende Eigenschaften

Beispiel. Nach A. Engel betrachten wir folgendes Spiel zwischen A und B: Eine Münze wird so lange geworfen, bis entweder 3 Köpfe oder Zahl, Kopf, Zahl in Serie gefallen sind. A gewinnt im ersten, B im zweiten Fall. Die Spieler brauchen sich nicht den gesamten vergangenen Spielablauf zu merken, sondern nur die Resultate der letzten ein oder zwei Würfe. So kann man sieben verschiedene Zustände unterscheiden, die während des Spiels angenommen werden können. Die möglichen Wechsel zwischen Zuständen sind im folgenden Diagramm veranschaulicht.



$*$ ist die Startposition, KKK und ZKZ die Gewinnpositionen für A bzw. B. Die Pfeile stellen die möglichen Übergänge dar, die Übergangswahrscheinlichkeiten entlang eines Pfeils sind jeweils $\frac{1}{2}$.

Wie groß sind die Chancen von A, wie groß ist also die Wahrscheinlichkeit, von $*$ aus schließlich KKK zu erreichen (und nicht ZKZ). Wir betrachten dazu für alle Zustände x die Wahrscheinlichkeit $w(x)$, ausgehend von x in den Zustand KKK zu gelangen. Durch Zerlegung der Wahrscheinlichkeiten nach dem ersten Schritt aus x heraus können wir das folgende Gleichungssystem aufstellen,

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

Durch Auflösen folgt $w(K) = \frac{1}{2}$, $w(KK) = \frac{2}{3}$, $w(Z) = w(ZK) = \frac{1}{3}$. Die Gewinnwahrscheinlichkeit von A ist also

$$w(*) = \frac{1}{2}\, w(K) + \frac{1}{2}\, w(Z) = \frac{5}{12}\ .$$

Die Gewinnwahrscheinlichkeit von B ist $\frac{7}{12}$. □

Dieses Gleichungsystem ist naheliegend, gleichwohl stellt sich die Frage, unter welchen Bedingungen solche Gleichungen gelten. Dies führt uns zum Begriff der Markov-Kette. Dabei handelt es sich um einen einfachen Ansatz zur Beschreibung der zeitlichen Entwicklung eines zufälligen Systems, das unterschiedliche Zustände in einer abzählbaren Menge S annehmen kann. Die Zustandsänderungen erfolgen schrittweise, so daß die angenommenen Zustände im zeitlichen Ablauf eine Folge $X_0, X_1, \ldots$ von S-wertigen Zufallsvariablen bilden. Im Beispiel sind die Zustände mit Bedacht so gewählt, daß die Grundannahme an eine Markov-Kette erfüllt ist. Sie besteht darin, daß sich das System gedächtnislos entwickelt: Die Wahrscheinlichkeit, nach n Schritten von x nach y überzuwechseln, bleibt von der Vorgeschichte unbeeinflußt. Es sind diese Übergangswahrscheinlichkeiten P_{xy}, die das System in seinen wesentlichen Eigenschaften bestimmen. Sie erfüllen

$$P_{xy} \geq 0 \ \text{ für alle } \ x, y \in S\ , \qquad \sum_y P_{xy} = 1 \ \text{ für alle } \ x \in S\ .$$

Eine reelle Matrix $P = (P_{xy})_{x,y\in S} = (P_{xy})$, die diesen Forderungen genügt, heißt **stochastisch**.

In der folgenden Definition erhält die Vorstellung der Gedächtnislosigkeit einen mathematisch präzisen Sinn.

Definition. *Sei S abzählbar und $P = (P_{xy})$ eine stochastische Matrix. Eine Folge von S-wertigen Zufallsvariablen $X_0, X_1, \ldots$ heißt* ***Markov-Kette*** *mit* ***Zustandsraum*** *S und* ***Übergangsmatrix*** *P, falls*

$$\mathbf{Ws}\{X_{n+1} = y \,\big|\, X_n = x, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0\} \;=\; P_{xy}$$

für $n \geq 0$, $x_0, \ldots, x_{n-1}, x, y \in S$ gilt, sofern das bedingende Ereignis strikt positive Wahrscheinlichkeit hat.

n wird häufig als diskreter Zeitparameter aufgefaßt.

Bemerkung. Da die Übergangswahrscheinlichkeiten nicht von n abhängen, handelt es sich genauer um eine **zeitlich homogene** Markov-Kette. Den allgemeineren Fall, daß die Übergangswahrscheinlichkeiten (wie bei der Pólya–Urne) auch von n abhängig sind, lassen wir beiseite und bemerken nur, daß die Forderung aus der Definition dann durch die Bedingung

$$\mathbf{Ws}\{X_{n+1} = y \,\big|\, X_n = x, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0\} \\ = \mathbf{Ws}\{X_{n+1} = y \,\big|\, X_n = x\}$$

ersetzt werden kann. □

Die Verteilungen von $X_0, X_1, \ldots$ sind durch die Übergangsmatrix noch nicht eindeutig bestimmt. Für Markov-Ketten gilt definitionsgemäß (und auch in dem Fall, dass die betrachteten Ereignisse Wahrscheinlichkeit 0 haben)

$$\mathbf{Ws}\{X_0 = x_0, \ldots, X_n = x_n\} \;=\; \mathbf{Ws}\{X_0 = x_0, \ldots, X_{n-1} = x_{n-1}\}\, P_{x_{n-1}x_n} \;,$$

und durch Iteration

$$\mathbf{Ws}\{X_0 = x_0, \ldots, X_n = x_n\} \;=\; \mathbf{Ws}\{X_0 = x_0\} \cdot P_{x_0x_1} \cdots P_{x_{n-1}x_n} \quad . \qquad (5.1)$$

Umgekehrt folgt aus diesen Gleichungen unmittelbar die in der Definition geforderte Eigenschaft. Wir halten diese Charakterisierung von Markov-Ketten fest.

Proposition 5.1. *Eine Folge $X_0, X_1, \ldots$ von Zufallsvariablen mit Werten in S ist genau dann eine Markov-Kette mit Übergangsmatrix (P_{xy}), wenn (5.1) gilt.*

Um die Verteilung einer Markov-Kette festzulegen, muß man daher neben den Übergangswahrscheinlichkeiten die Verteilung μ von X_0 angeben, die **Start-** oder **Anfangsverteilung** der Markov-Kette. Häufig erweist es sich als sinnvoll, μ nicht von vornherein zu fixieren. Möchte man μ hervorheben, so führt man sie als Index an Wahrscheinlichkeiten und Erwartungswerten mit und schreibt $\mathbf{Ws}_\mu\{\cdot\}$ bzw. $\mathbf{E}_\mu[\cdot]$. Startet man mit Wahrscheinlichkeit 1 im Zustand $x \in S$, so schreibt man $\mathbf{Ws}_x\{\cdot\}$ und $\mathbf{E}_x[\cdot]$.

Die Gedächtnislosigkeit einer Markov-Kette ergibt sich daraus, daß die in der Definition angegebenen bedingten Wahrscheinlichkeiten von $x_0, \ldots, x_{n-1}$ unabhängig sind. Diese **Markov-Eigenschaft** läßt sich verallgemeinern. Anschaulich gesprochen bleibt, gegeben die Gegenwart $\{X_n = x\}$, das Eintreten eines **Ereignisses der Zukunft** $\{X_n, \ldots, X_{n+m}) \in B'\}$ von einem **Ereignis der Vergangenheit** $\{(X_0, \ldots, X_n) \in B\}$ unbeeinflußt.

Proposition 5.2. *Für eine (zeitlich homogene) Markov-Kette $X_0, X_1, \ldots$ gilt*

$$\mathbf{Ws}\{(X_n, \ldots, X_{n+m}) \in B' \,\big|\, X_n = x,\, (X_0, \ldots, X_n) \in B\} \\ = \mathbf{Ws}_x\{(X_0, \ldots, X_m) \in B'\}$$

für alle $m, n \geq 0$, $x \in S$, $B \subset S^{n+1}$, $B' \subset S^{m+1}$ (sofern die bedingte Wahrscheinlichkeit wohldefiniert ist).

Beweis. Aus (5.1) folgt für $x_0, \ldots, x_n, x, y_0, \ldots, y_m \in S$

$$\mathbf{Ws}\{X_0 = x_0, \ldots, X_n = x_n, X_n = x, X_n = y_0, \ldots, X_{n+m} = y_m\} \\ = \mathbf{Ws}\{X_0 = x_0, \ldots, X_n = x_n, X_n = x\} \cdot \mathbf{Ws}_x\{X_0 = y_0, \ldots, X_m = y_m\}$$

(für $x_n \neq x$ oder $y_0 \neq x$ sind beide Seiten 0). Durch Summation über alle $(x_0, \ldots, x_n) \in B$ und $(y_0, \ldots, y_m) \in B'$ folgt

$$\mathbf{Ws}\{(X_0, \ldots, X_n) \in B, X_n = x, (X_n, \ldots, X_{n+m}) \in B'\} \\ = \mathbf{Ws}\{(X_0, \ldots, X_n) \in B, X_n = x\} \cdot \mathbf{Ws}_x\{(X_0, \ldots, X_m) \in B'\}$$

und damit die Behauptung. □

Aus der Markov-Eigenschaft leiten wir nun eine Formel für die ***n*-Schritt-Übergangswahrscheinlichkeiten**

$$P_{xy}^n \;:=\; \mathbf{Ws}_x\{X_n = y\}$$

$n \geq 0$, ab. Nach dem Satz von der totalen Wahrscheinlichkeit gilt

$$\mathbf{Ws}_x\{X_{n+m} = z\} \;=\; \sum_y \mathbf{Ws}_x\{X_n = y\} \cdot \mathbf{Ws}_x\{X_{n+m} = z \,\big|\, X_n = y\}\;.$$

Unter Beachtung von Proposition 5.2 erhalten wir die folgende unter der Bezeichnung **Chapman-Kolmogorov-Gleichung** bekannte Formel

$$\mathbf{Ws}_x\{X_{n+m} = z\} \;=\; \sum_y \mathbf{Ws}_x\{X_n = y\} \cdot \mathbf{Ws}_y\{X_m = z\} \;, \quad m, n \geq 0 \;,$$

bzw. $P_{xz}^{n+m} = \sum_y P_{xy}^n P_{yz}^m$. Für die ***n*-Schritt-Übergangsmatrizen** $P^n := (P_{xy}^n)$ bedeutet dies, dass sie durch Matrixmultiplikation auseinander hervorgehen,

$$P^{n+m} \;=\; P^n \cdot P^m \;.$$

P^0 ist die Einheitsmatrix. Da außerdem P^1 die vorgegebene Übergangsmatrix P ist, ergibt sich P^n als das n-maliges Matrixprodukt von P mit sich selbst,

$$P^n \;=\; P \cdots P \quad n\text{-mal} \;.$$

Beispiel. Explizite Formeln für die n-Schritt-Übergangswahrscheinlichkeiten stehen nur ausnahmsweise zur Verfügung. Ein solcher Ausnahmefall ist der **Riffle-Shuffle**, das in Abschnitt 1.5 beschriebene Modell für das Kartenmischen. Wir können ihn als Markov-Kette $X_0, X_1, \ldots$ auf dem Raum der Permutationen $S = \{\pi : K \to K \;:\; \pi \text{ ist Bijektion}\}$ auffassen. Wie wir sahen (vgl. (1.14)), sind die Übergangswahrscheinlichkeiten bei k Spielkarten durch

$$P_{xy}^n \;=\; \binom{k + 2^n - s}{k}\, 2^{-kn} \;, \quad x, y \in S \;,$$

gegeben, dabei ist s die Anzahl der wachsenden Sequenzen der Permutation π, die x in y überführt ($\pi := y \circ x^{-1}$). □

Treffwahrscheinlichkeiten und erwartete Eintrittszeiten

Wir leiten nun im Rahmen der Markov-Ketten das oben im Beispiel aufgestellte Gleichungssystem für Treffwahrscheinlichkeiten ab. Sei B eine nichtleere Teilmenge des Zustandsraumes S und $X_0, X_1, \ldots$ Markov-Kette mit Übergangsmatrix (P_{xy}). Wir betrachten den **Zeitpunkt des ersten Eintritts** in B,

$$M \;:=\; \min\{m \geq 0 \;:\; X_m \in B\} \;,$$

und das Ereignis $\{M < \infty, X_M = z\} = \bigcup_m \{M = m, X_m = z\}$, daß die Markov-Kette dann ein vorgegebenes $z \in B$ trifft. Die Wahrscheinlichkeiten

$$w(x) \;:=\; \mathbf{Ws}_x\{M < \infty, X_M = z\} \;,$$

von dem Startpunkt x aus bei Eintritt in B den Zustand z zu erreichen, heißen **Treffwahrscheinlichkeiten** oder **Absorbtionswahrscheinlichkeiten**. Sie erfüllen die Gleichungen

$$w(x) \;=\; \sum_y P_{xy} w(y) \;, \quad \text{falls } x \notin B \;, \tag{5.2}$$

wie wir sie oben im Beispiel aufgestellt haben (mit $B = \{KKK, ZKZ\}$ und $z = KKK$). Für $x \in B$ gilt $M = 0$ und folglich

$$w(x) \;=\; \begin{cases} 1 \;, & \text{falls } x = z \;, \\ 0 \;, & \text{falls } x \in B,\, x \neq z \;. \end{cases} \tag{5.3}$$

(5.2) ist eine Konsequenz des Satzes von der totalen Wahrscheinlichkeit, der **Zerlegung nach dem ersten Schritt**

$$\begin{aligned} &\mathbf{Ws}_x\{M < \infty, X_M = z\} \\ &\qquad = \sum_y \mathbf{Ws}_x\{X_1 = y\} \cdot \mathbf{Ws}_x\{M < \infty, X_M = z \,\big|\, X_1 = y\} \;, \end{aligned}$$

sowie der Markov-Eigenschaft (Proposition 5.2),

$$\begin{aligned} &\mathbf{Ws}_x\{M < \infty,\, X_M = z \,\big|\, X_1 = y\} \\ &\qquad = \sum_{l=0}^{\infty} \mathbf{Ws}_x\{X_{l+1} = z, X_l, \ldots, X_1 \notin B \,\big|\, X_1 = y\} \\ &\qquad = \sum_{l=0}^{\infty} \mathbf{Ws}_y\{X_l = z, X_{l-1}, \ldots, X_0 \notin B\} \;=\; w(y) \end{aligned}$$

(für $x \notin B$ gilt mit Wahrscheinlichkeit 1 $M \geq 1$).

In den folgenden Beispielen spezifizieren wir für $x \in B$ nicht immer die Übergangswahrscheinlichkeiten P_{xy}, denn in (5.2) werden sie nicht benötigt. Wenn man will, kann man die Zustände $x \in B$ als **absorbierend** annehmen, d.h. $P_{xx} = 1$ setzen.

Beispiele.

1. **Die symmetrische Irrfahrt auf $\mathbb{Z}$.** Ein Irrfahrer taumelt durch $\mathbb{Z}$. Er macht jeweils mit Wahrscheinlichkeit $\frac{1}{2}$ unabhängige Schritte nach rechts oder links. Mit welcher Wahrscheinlichkeit erreicht er die 0? Die Übergangswahrscheinlichkeiten sind

$$P_{xy} \;:=\; \begin{cases} \frac{1}{2} \;, & \text{falls } \; y = x \pm 1 \;\; , \\ 0 & \text{sonst} \;\; . \end{cases}$$

Die Wahrscheinlichkeiten $w(x)$, ausgehend von x irgendwann 0 zu treffen, erfüllen nach (5.2) die Gleichungen

$$\begin{aligned} w(x) &= \frac{1}{2}\, w(x+1) + \frac{1}{2}\, w(x-1)\ , \quad x \neq 0 \\ w(0) &= 1\ . \end{aligned}$$

Die $w(x)$ liegen folglich für $x \geq 0$ auf einer Geraden durch $w(0) = 1$. Wegen $0 \leq w(x) \leq 1$ kommt nur eine einzige Gerade in Frage:

$$w(x) = 1 \quad \text{für alle } x\ .$$

Der Irrfahrer erreicht also 0 mit Wahrscheinlichkeit 1, gleichgültig, von wo er startet. Man sagt, die Irrfahrt ist rekurrent. (Taumelt der Irrfahrer dagegen rein zufällig durch den $\mathbb{Z}^m$, so ist das Resultat ab $m = 3$ ein anderes, wie wir im nächsten Abschnitt sehen werden.)

2. **Des Spielers Ruin.** Jemand beteiligt sich solange an einem Glücksspiel, bis er über ein Kapital von a Euro verfügt oder aber sein gesamtes Spielkapital verloren hat. Wenn er vorsichtig ist und pro Spiel nur einen Euro setzt, mit welcher Wahrscheinlichkeit verspielt er sein gesamtes Kapital?

   Wir nehmen a als ganzzahlig an und modellieren den Spielverlauf als Markov-Kette mit Zustandsraum $S = \{0, 1, \ldots, a\}$ und den Übergangswahrscheinlichkeiten

$$P_{xy} := \begin{cases} p & \text{für } y = x+1 \\ q & \text{für } y = x-1 \end{cases}\ , \qquad x = 1, \ldots, a-1\ .$$

   p ist die Gewinnwahrscheinlichkeit für ein Einzelspiel und $q = 1 - p$ die Verlustwahrscheinlichkeit. Gefragt ist nach der Wahrscheinlichkeit $v(x)$, ausgehend von x den Zustand 0 vor dem Zustand a zu erreichen. Nach (5.2) gilt

$$v(x) = p \cdot v(x+1) + q \cdot v(x-1)\ , \quad x \neq 0, a\ .$$

   Dieses Gleichungssystem läßt sich ohne weiteres auflösen. Gegeben $v(0)$ und $v(1)$ lassen sich der Reihe nach $v(2), v(3), \ldots$ bestimmen, der Lösungsraum ist also 2-dimensional. Man rechnet leicht nach, daß die allgemeine Lösung im Fall $p \neq q$ durch $c + d(q/p)^x$ und im Fall $p = q = 1/2$ durch $c + dx$ gegeben ist, mit reellen Zahlen c, d. Unter Beachtung der Randbedingungen $v(0) = 1$ und $v(a) = 0$ erhalten wir insgesamt

$$v(x) = \frac{\left(\frac{q}{p}\right)^a - \left(\frac{q}{p}\right)^x}{\left(\frac{q}{p}\right)^a - 1} \quad \text{für } p \neq 1/2 \tag{5.4}$$

bzw.

$$v(x) \;=\; \frac{a-x}{a} \qquad \text{für } p = 1/2\ . \qquad \square$$

Das Gleichungssystem (5.2) ist nicht immer eindeutig lösbar. Die Treffwahrscheinlichkeiten sind dadurch charakterisiert, daß sie *minimal* sind unter allen nicht-negativen Lösungen von (5.2), die die Nebenbedingung (5.3) erfüllt. Allgemeiner gilt die folgende Aussage.

Proposition 5.3. *Sei $f : S \to \mathbb{R}_+$ eine nicht-negative Funktion, die (5.3) sowie*

$$\sum_y P_{xy} f(y) \;\le\; f(x)$$

für alle $x \notin B$ erfüllt. Dann gilt $w(x) \le f(x)$ für alle $x \in S$.

Beweis. Wir zeigen $\mathbf{Ws}_x\{X_M = z, M \le l\} \le f(x)$. Für $x \in B$ ist die Aussage offenbar, für $x \notin B$ führen wir eine Induktion nach l durch. Der Induktionsanfang ergibt sich aus $\mathbf{Ws}_x\{X_M = z, M \le 0\} = 0$ für $x \notin B$, und der Induktionsschritt von l nach $l+1$ mit einer Zerlegung nach dem ersten Schritt,

$$\begin{aligned} \mathbf{Ws}_x\{X_M = z, M \le l+1\} \;&=\; \sum_y P_{xy}\, \mathbf{Ws}_y\{X_M = z, M \le l\} \\ &\le\; \sum_y P_{xy} f(y) \;\le\; f(x)\ . \end{aligned}$$

Wegen $\mathbf{Ws}_x\{X_M = z, M \le l\} \to \mathbf{Ws}_x\{X_M = z, M < \infty\} = w(x)$ für $l \to \infty$ folgt die Behauptung. $\square$

Beispiele.

1. **Das Rot-Schwarz-Spiel.** Die Gewinnwahrscheinlichkeit p ist bei Glücksspielen normalerweise kleiner als $1/2$ - etwa, wenn man beim Roulette auf Rot oder Schwarz setzt. Fährt ein Spieler besser, wenn er nicht so vorsichtig wie im vorangehenden Beispiel vorgeht und pro Spiel mehr als einen Euro setzt? Wenn er über das Startkapital $a/2$ verfügt (a geradzahlig), läßt sich die Frage leicht beantworten. Setzt er sofort sein gesamtes Kapital, so benötigt er ein einziges Spiel, um a Euro zu erlangen oder aber pleite zu gehen, und seine Ruin-Wahrscheinlichkeit ist dann q. Für $a > 2$ und $p < q$ gilt (vgl. (5.4))

$$v(a/2) \;=\; \frac{\left(\frac{q}{p}\right)^a - \left(\frac{q}{p}\right)^{a/2}}{\left(\frac{q}{p}\right)^a - 1} \;=\; \frac{\left(\frac{q}{p}\right)^{a/2}}{1 + \left(\frac{q}{p}\right)^{a/2}} \;>\; \frac{\frac{q}{p}}{1+\frac{q}{p}} \;=\; q\ ,$$

(denn $x/(1+x)$ ist monoton wachsend), d.h. es lohnt sich für den Spieler, sofort sein gesamtes Kapital zu setzen. Dies ist plausibel: Das Spiel ist dann schnell beendet, und die unfaire Spielbedingung $p < q$ hat wenig Gelegenheit, sich negativ auszuwirken.

Wir wollen nun für ein beliebiges Startkapital die *vorsichtige Strategie* des vorigen Beispiels vergleichen mit der Strategie, möglichst viel pro Spiel zu setzen, aber nur soviel, daß der anvisierte Betrag von a Euro nicht übertroffen wird. Der Einsatz bei dieser *kühnen Strategie* beträgt also $\min(x, a-x)$, falls man über x Euro verfügt. Der Spielverlauf wird beschrieben durch eine Markov-Kette auf $\{0, 1, \ldots, a\}$ mit den Übergangswahrscheinlichkeiten

$$Q_{xy} \;:=\; \begin{cases} p & \text{für } \; y = 2x \\ q & \text{für } \; y = 0 \end{cases} , \quad \text{falls } \; 0 < x \le \frac{a}{2} ,$$

und

$$Q_{xy} \;:=\; \begin{cases} p & \text{für } \; y = a \\ q & \text{für } \; y = 2x - a \end{cases} , \quad \text{falls } \; \frac{a}{2} \le x < a .$$

Sei $k(x)$ die Ruinwahrscheinlichkeit für die kühne Strategie bei einem Startkapital x. Wir wollen zeigen, daß unter der Annahme $p < q$ die kühne Strategie vorteilhaft ist, d. h. für alle $x = 0, 1, \ldots, a$ die Ungleichung

$$k(x) \;\le\; v(x)$$

gilt. Auf direktem Wege ist der Beweis nicht mehr möglich, denn für das nach (5.2) gültige Gleichungssystem

$$k(x) \;=\; \sum_y Q_{xy} k(y) , \quad x \ne 0, a ,$$

mit $k(0) = 1$ und $k(a) = 0$ hat man keine allgemeine explizite Lösung. Ein Beweis ergibt sich stattdessen aus Proposition 5.3, denn es gilt die Ungleichung

$$\sum_y Q_{xy} v(y) \;\le\; v(x) , \quad x \ne 0, a .$$

Für $x \le a/2$ folgt sie aus der für $x \ge 1$ und $p < q$ gültigen Abschätzung

$$\begin{aligned} p\big(\big(\tfrac{q}{p}\big)^a - \big(\tfrac{q}{p}\big)^{2x}\big) + q\big(\big(\tfrac{q}{p}\big)^a - 1\big) \;&=\; \big(\tfrac{q}{p}\big)^a - 1 - p\big(\big(\tfrac{q}{p}\big)^x + 1\big)\big(\big(\tfrac{q}{p}\big)^x - 1\big) \\ &\le\; \big(\tfrac{q}{p}\big)^a - 1 - p\big(\tfrac{q}{p} + 1\big)\big(\big(\tfrac{q}{p}\big)^x - 1\big) \;=\; \big(\tfrac{q}{p}\big)^a - \big(\tfrac{q}{p}\big)^x , \end{aligned}$$

und für $x \ge a/2$ analog aus der für $x \le a - 1$ gültigen Ungleichung

$$\begin{aligned} q\big(\big(\tfrac{q}{p}\big)^a - \big(\tfrac{q}{p}\big)^{2x-a}\big) \;&=\; q\big(\tfrac{p}{q}\big)^a\big(\big(\tfrac{q}{p}\big)^a + \big(\tfrac{q}{p}\big)^x\big)\big(\big(\tfrac{q}{p}\big)^a - \big(\tfrac{q}{p}\big)^x\big) \\ &\le\; q\big(\tfrac{p}{q}\big)^a\big(\big(\tfrac{q}{p}\big)^a + \big(\tfrac{q}{p}\big)^{a-1}\big)\big(\big(\tfrac{q}{p}\big)^a - \big(\tfrac{q}{p}\big)^x\big) \;=\; \big(\tfrac{q}{p}\big)^a - \big(\tfrac{q}{p}\big)^x . \end{aligned}$$

2. **Ein Warteschlangenmodell.** An einem Skilift, der pro Zeiteinheit eine Person befördert, steht eine Warteschlange von zufällig wechselnder Länge. Unter welchen Bedingungen wird sie sich mit Wahrscheinlichkeit 1 auflösen?

   Sei X_n die Länge der Warteschlange nach n Zeiteinheiten. Wir modellieren $X_0, X_1, \ldots$ als Markov-Kette mit Werten in $\mathbb{N}_0$. Dazu nehmen wir an, daß sich mit Wahrscheinlichkeit p_x pro Zeiteinheit x neue Skiläufer an die Warteschlange anstellen. Die Übergangswahrscheinlichkeiten sind dann

$$P_{xy} \;=\; \begin{cases} p_{y-x+1}\;, & \text{falls } x \geq 1\;, \\ p_y\;, & \text{falls } x = 0\;. \end{cases}$$

   In diesem Modell betrachten wir nun die Wahrscheinlichkeit

$$w(x) \;:=\; \mathbf{Ws}_x\{X_n = 0 \text{ für ein } n \geq 0\}\;,$$

   daß sich eine Warteschlange der Anfangslänge x schließlich auflöst. Aufgrund der Gedächtnislosigkeit der Markov-Kette gilt

$$w(x) \;=\; w^x$$

   mit $w := w(1)$ (denn zum Auflösen der Schlange ist es nötig, daß sie sich x-mal um 1 Person verringert, und dies geschieht jeweils mit Wahrscheinlichkeit w). Damit wird (5.2) zu

$$w^x \;=\; \sum_y P_{xy} w(y) \;=\; \sum_{y=x-1}^{\infty} p_{y-x+1} w^y$$

   bzw.

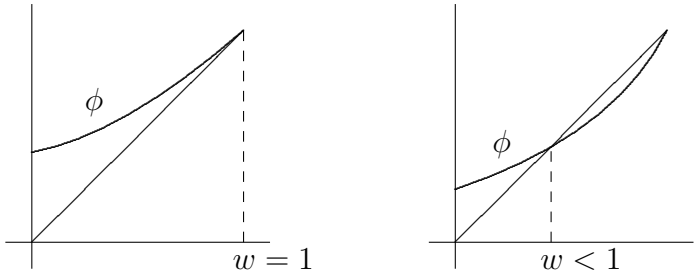
$$w \;=\; \phi(w)$$

   mit der erzeugenden Funktion

$$\phi(t) \;:=\; \sum_{x=0}^{\infty} p_x t^x\;.$$

   Nach Proposition 5.3 ist also w die *minimale* positive Lösung der Gleichung $t = \phi(t)$.

   Eine Lösungen der Gleichung läßt sich sofort angeben: Es gilt $1 = \phi(1)$. Damit stellt sich die Frage, ob es unterhalb 1 weitere positive Lösungen gibt. Wie aus den folgenden Abbildungen ersichtlich sind 2 Fälle zu unterscheiden: Aufgrund der Konvexität von ϕ gibt es im Intervall $[0, 1]$ entweder eine oder zwei Lösungen. Der erste Fall tritt für $\phi'(1) \leq 1$ ein, der zweite für $\phi'(1) > 1$.



Es gilt $\phi'(1) = \mu$ mit

$$\mu \;:=\; \sum_{x=1}^{\infty} x p_x \;,$$

der mittleren Anzahl von Skiläufern, die pro Zeiteinheit am Lift ankommen. Insgesamt erhalten wir folgendes einleuchtende Ergebnis: Stellt sich pro Zeiteinheit im Mittel höchstens 1 Person am Lift an, so löst sich die Warteschlange mit Wahrscheinlichkeit 1 schließlich auf, kommt dagegen im Mittel mehr als 1 Person, so wird sich die Schlange mit positiver Wahrscheinlichkeit nie auflösen. □

Nach demselben Schema läßt sich auch die **erwartete Eintrittszeit** in die Menge B berechnen, der Erwartungswert von

$$M \;:=\; \min\{m \geq 0 \;:\; X_m \in B\} \;.$$

M kann den Wert ∞ annehmen, also gilt

$$e(x) \;:=\; \mathbf{E}_x M \;=\; \sum_{m=0}^{\infty} m \cdot \mathbf{Ws}_x\{M = m\} + \infty \cdot \mathbf{Ws}_x\{M = \infty\}$$

(mit $\infty \cdot 0 = 0$ und $\infty \cdot w = \infty$ für $w > 0$). In Analogie zu (5.2) gilt dann

$$e(x) \;=\; 1 + \sum_{y} P_{xy} e(y) \;, \quad \text{falls } x \notin B \;, \tag{5.5}$$

sowie

$$e(x) \;=\; 0 \quad \text{für alle } x \in B \;.$$

Zum Beweis von (5.5) benutzen wir die Markov-Eigenschaft: Für $x \notin B$ und $l \geq 0$ gilt

$$\begin{aligned} &\mathbf{Ws}_x\{M = l+1 \,\big|\, X_1 = y\} \\ &\quad= \mathbf{Ws}_x\{X_{l+1} \in B, X_l, \ldots, X_1 \notin B \,\big|\, X_1 = y\} \\ &\quad= \mathbf{Ws}_y\{X_l \in B, X_{l-1}, \ldots, X_0 \notin B\} \;=\; \mathbf{Ws}_y\{M = l\} \;, \end{aligned}$$

und ähnlich $\mathbf{Ws}_x\{M = \infty \,\big|\, X_1 = y\} = \mathbf{Ws}_y\{M = \infty\}$. Es folgt

$$\begin{aligned}
&\mathbf{E}_x[M \,\big|\, X_1 = y] \\
&= \sum_{l=0}^{\infty}(l+1)\,\mathbf{Ws}_x\{M = l+1 \,\big|\, X_1 = y\} \\
&\qquad + \infty \cdot \mathbf{Ws}_x\{M = \infty \,\big|\, X_1 = y\} \\
&= \sum_{l=0}^{\infty}(l+1)\,\mathbf{Ws}_y\{M = l\} + \infty \cdot \mathbf{Ws}_y\{M = \infty\} \;=\; 1 + e(y)\ .
\end{aligned}$$

Nach Proposition 4.2 folgt mit einer Zerlegung nach dem ersten Schritt wie behauptet

$$e(x) \;=\; \sum_y \mathbf{Ws}_x\{X_1 = y\}\,\mathbf{E}_x[M \,\big|\, X_1 = y] \;=\; 1 + \sum_y P_{xy}e(y)\ .$$

Bemerkung. Das Gleichungssystem (5.5) für Eintrittszeiten ist im Allgemeinen nicht eindeutig lösbar, insbesondere muß man beachten, daß $e(x)$ auch ∞ sein kann. Hier gilt wie bei den Treffwahrscheinlichkeiten, daß die erwarteten Eintrittszeiten eine *minimale* nicht-negative Lösung ergeben. Allgemeiner gilt eine zu Proposition 5.3 analoge Aussage (Übung). □

Beispiele.

1. **Wartezeiten für Runs.** Wie lange braucht man im Mittel, bis bei unabhängiger Wiederholung eines Zufallsexperiments mit Erfolgswahrscheinlichkeit p eine Serie von r aufeinanderfolgenden Einsen (Erfolgen) gelingt? Wir betrachten dazu Zustände in $S = \{0, 1, \ldots, r\}$ und vereinbaren, daß wir uns momentan im Zustand x befinden, falls das zuletzt durchgeführte Experiment das Ende einer Erfolgsserie der Länge x bildet (so werden in der Serie 110111 nacheinander die Zustände 1,2,0,1,2,3 eingenommen). Bei Wiederholung des Zufallsexperimentes ändert sich der Zustand nach Art einer Markov-Kette mit den Übergangswahrscheinlichkeiten

$$P_{xy} \;=\; \begin{cases} p\ , & \text{falls } y = x+1\ , \\ q\ , & \text{falls } y = 0\ , \\ 0 & \text{sonst} \end{cases}$$

für $x < r$ (mit $q = 1 - p$). Der Startpunkt ist 0, wir fragen also nach der erwarteten Zeit, von 0 nach r zu gelangen. Dazu betrachten wir auch die

Erwartungswerte $e(x)$ der Anzahl der Schritte, um vom Zustand x aus r zu erreichen. Nach (5.5) gilt

$$e(x) \;=\; 1 + pe(x+1) + qe(0)\;, \quad x = 0, 1, \ldots, r-1\;,$$

sowie $e(r) = 0$. Wenn wir diese Gleichungen mit p^x multiplizieren und anschließend aufsummieren, heben sich die Terme $e(x)p^x$ für $x = 1, \ldots, r-1$ weg, und es ergibt sich

$$e(0) \;=\; \sum_{x=0}^{r-1} p^x + p^r e(r) + qe(0) \sum_{x=0}^{r-1} p^x\;.$$

Unter Beachtung von $e(r) = 0$ und $\sum_{x=0}^{r-1} p^x = (1-p^r)/q$ erhalten wir für den gesuchten Erwartungswert die Formel

$$e(0) \;=\; \frac{1}{p^r q} - \frac{1}{q}\;.$$

Diese Formel ist plausibel: Sieht man einmal von dem Fall ab, daß gleich am Anfang r Erfolge gelingen, so wird die erste Serie von r Einsen von einer Null angeführt, und eine solche verlängerte Serie hat die Eintrittswahrscheinlichkeit $p^r q$. (Übung: Die mittlere Wartezeit für einen Run aus einer Null und r anschließenden Einsen ist $1/(p^r q)$.)

Geht man davon aus, daß pro Sekunde ein Experiment durchgeführt wird, so ergeben sich für $e(0)$ folgende Werte (nach Feller, *An Introduction to Probability Theory and its Applications*, Band 1):

r	5	10	15	20
p=1/2	1 Minute	34 Minuten	18 Stunden	34 Tage
p=1/6	2,6 Stunden	28 Monate	$1,8 \cdot 10^4$ Jahre	$1,4 \cdot 10^8$ Jahre

2. **Ein Algorithmus zur Bestimmung von Maxima.** $a(1), a(2), \ldots, a(r)$ seien paarweise verschiedene Zahlen. Um ihr Maximum zu bestimmen, kann man so vorgehen: Erst vergleicht man $a(1)$ mit den Zahlen $a(2), a(3), \ldots$, bis man einen Index $y_2 > y_1 := 1$ gefunden hat, so daß $a(y_2)$ größer als $a(1)$ ist. Dann ersetzt man $a(1)$ durch $a(y_2)$ und vergleicht diese Zahl mit $a(y_2+1), a(y_2+2), \ldots$ So geht man $a(1), \ldots, a(r)$ der Reihe nach durch und erhält eine wachsende Folge $a(y_1), \ldots, a(y_j)$ von Vergleichszahlen, deren letzte das gesuchte Maximum ist. Möchte man das Verfahren im Computer implementieren, so wird man die Vergleichszahlen an einem speziellen Platz speichern. Zu den Zeitpunkten $y_1 < y_2 < \cdots < y_j$

muß man den Inhalt dieses Speichers austauschen. Mit wievielen solcher Speicherwechsel muß man rechnen?

Wir wollen eine *Average-Analyse* durchführen und nehmen dazu an, daß $a(1), \ldots, a(r)$ eine rein zufällige Permutation von Zahlen $b(1) < \cdots < b(r)$ ist. Die Zeitpunkte der Speicherwechsel bilden dann einen Zufallsvektor $(Y_1, \ldots, Y_J)$ von zufälliger Länge J. Es ist praktisch $Y_n = Y_J$ für $n > J$ zu setzen. Wir können dann die Zufallsvariablen $0 := X_0 \leq X_1 \leq \cdots$ betrachten, gegeben durch die Gleichung $b(X_i) = a(Y_i)$ für $i \geq 1$. Die Zahlen, die nacheinander als Vergleichsgrößen abgespeichert werden, sind also gerade $b(X_1) < b(X_2) < \cdots < b(X_J) = b(r)$.

Bei $X_0, X_1, \ldots$ handelt es sich um eine Markov-Kette: Ist das Ereignis $\{X_0 = x_0, \ldots, X_n = x_n\}$ eingetreten, so hat der Vergleich mit $b(x_n)$ bereits stattgefunden, die Vergeiche mit $b(x_n + 1), \ldots, b(r)$ stehen dagegen noch aus (wie möglicherweise auch noch andere Vergleiche, die aber zu keinem Austausch führen). Diese Zahlen besitzen nach Annahme eine rein zufällige Reihenfolge, sie haben daher alle dieselbe Chance, zuerst mit $b(x_n)$ verglichen zu werden. Daher nimmt X_{n+1} mit gleicher Wahrscheinlichkeit einen der Werte $x_n + 1, \ldots, r$ an, völlig unabhängig von dem bisherigen Geschehen, wir haben es also mit einer Markov-Kette mit dem Zustandsraum $S = \{0, 1, \ldots, r\}$ und den Übergangswahrscheinlichkeiten

$$P_{xy} \;=\; \begin{cases} \frac{1}{r-x}\,, & \text{falls } y > x\,, \\ 0 & \text{sonst} \end{cases}$$

zu tun.

Wir können nun unsere Fragestellung umformulieren und die mittlere Anzahl $e(0)$ von Schritten untersuchen, ausgehend vom Zustand 0 in den Zustand r zu gelangen. Dazu betrachten wir auch die mittlere Anzahl $e(x)$ von Schritten, um von x nach r zu kommen. Nach (5.5) gelten die Gleichungen

$$e(x) \;=\; 1 + \tfrac{1}{r-x}(e(x+1) + \cdots + e(r))\,.$$

Indem wir die Gleichung mit $r-x$ multiplizieren und dann diese Ausdrücke für x und $x+1$ voneinander abziehen, folgt

$$(r-x)e(x) - (r-x-1)e(x+1) \;=\; 1 + e(x+1)\,.$$

Durch Auflösen nach $e(x)$ und Iterieren der Gleichung ergibt sich

$$e(x) \;=\; \tfrac{1}{r-x} + \tfrac{1}{r-x-1} + \cdots + 1 + e(r)\,.$$

Unter Beachtung von $e(r) = 0$ folgt schließlich, daß die mittlere Anzahl von Wechseln des Speicherinhalts von logarithmischer Größe ist,

$$e(0) \;=\; 1 + \frac{1}{2} + \cdots + \frac{1}{r} \;\sim\; \ln r \;.$$

Eine eingehendere Analyse des Algorithmus findet sich in D. KNUTH, *The Art of Computer Programming*, Band 1, Abschnitt I.2.10. □

5.2 Rekurrenz und Transienz

Die Zustände einer Markov-Kette unterscheidet man danach, ob sie im Verlauf der Zeit immer wieder besucht werden, oder ob es schließlich einen letzten Zeitpunkt der Rückkehr gibt. Sei, bei Start in x,

$$T_x \;:=\; \min\{n \geq 1 \;:\; X_n = x\}$$

der **Zeitpunkt der ersten Rückkehr** nach x (er hat möglicherweise den Wert ∞).

Definition. *Der Zustand x heißt **rekurrent**, falls $\mathbf{Ws}_x\{T_x < \infty\} = 1$, und **transient**, falls $\mathbf{Ws}_x\{T_x < \infty\} < 1$. Eine Markov-Kette heißt rekurrent (transient), falls alle ihre Zustände rekurrent (transient) sind.*

Beispiel. Warteschlangen. In dem schon behandelten Warteschlangenmodell untersuchen wir nun, wann der Zustand 0 rekurrent ist. Bei dieser Markov-Kette macht es keinen Unterschied, ob man im Zustand 0 oder 1 startet (ob in der Warteschlange anfangs keine Person steht, oder eine, die dann sofort abtransportiert wird). Deswegen ist die Wahrscheinlichkeit $\mathbf{Ws}_0\{T_0 < \infty\}$ einer Rückkehr nach 0 gleich der Wahrscheinlichkeit w, daß sich eine Warteschlange der Länge 1 auflöst. In diesem Lichte können wir unsere früheren Resultate so ausdrücken: Gilt $\mu \leq 1$ für mittlere Anzahl μ der ankommenden Personen, so ist der Zustand 0 rekurrent, ist dagegen $\mu > 1$, so ist 0 transient. □

Die folgende Aussage macht die Terminologie verständlich.

Proposition 5.4. *Für einen transienten Zustand x gilt bei beliebiger Startverteilung $\mathbf{Ws}\{X_n = x \text{ für } \infty\text{-viele } n\} = 0$. Für einen rekurrenten Zustand x gilt $\mathbf{Ws}_x\{X_n = x \text{ für } \infty\text{-viele } n\} = 1$.*

Beweis. Sei $C_x = \text{card}\{n \geq 1 : X_n = x\}$ die Anzahl der Besuche in x. Nach der Markov-Eigenschaft gilt für $m \geq 1$

$$\begin{aligned}
&\mathbf{Ws}\{C_x \geq m\} \\
&\quad = \sum_{l=1}^{\infty} \mathbf{Ws}\{X_1, \ldots, X_{l-1} \neq x, X_l = x, \\
&\qquad\qquad X_n = x \text{ noch mindestens } m-1 \text{ mal}\} \\
&\quad = \sum_{l=1}^{\infty} \mathbf{Ws}\{X_1, \ldots, X_{l-1} \neq x, X_l = x\} \\
&\qquad\qquad \cdot \mathbf{Ws}_x\{X_n = x \text{ mindestens } m-1 \text{ mal}\} \\
&\quad = \mathbf{Ws}\{C_x \geq 1\} \cdot \mathbf{Ws}_x\{C_x \geq m-1\} \ .
\end{aligned}$$

Iteration führt zu der Gleichung

$$\begin{aligned}
\mathbf{Ws}\{C_x \geq m\} &= \mathbf{Ws}\{C_x \geq 1\} \cdot \mathbf{Ws}_x\{C_x \geq 1\}^{m-1} \\
&= \mathbf{Ws}\{T_x < \infty\} \cdot \mathbf{Ws}_x\{T_x < \infty\}^{m-1} \ .
\end{aligned}$$

Mit $m \to \infty$ folgt im transienten Fall $\mathbf{Ws}\{C_x = \infty\} = 0$ und damit die erste Behauptung. Im rekurrenten Fall erhalten wir $\mathbf{Ws}_x\{C_x \geq m\} = \mathbf{Ws}_x\{T_x < \infty\}^m = 1$ und wegen $\{C_x \geq m\} \downarrow \{C_x = \infty\}$ die zweite Behauptung: $\mathbf{Ws}_x\{X_n = x \ \infty\text{-oft}\} = \mathbf{Ws}_x\{C_x = \infty\} = 1$. □

Oft ist das folgende Kriterium hilfreich.

Satz 5.5. *$x \in S$ ist transient genau dann, wenn*

$$\sum_{n=1}^{\infty} \mathbf{Ws}_x\{X_n = x\} \ < \ \infty \ .$$

Beweis. Falls die Reihe konvergiert, hat das Ereignis $\{X_n = x \ \infty\text{-oft}\}$ nach dem ersten Borel-Cantelli Lemma Wahrscheinlichkeit 0, so daß nach Proposition 5.4 Transienz vorliegt. Sei umgekehrt x transient. Das zweite Borel-Cantelli Lemma können wir nicht benutzen, denn die Ereignisse $\{X_n = x\}$ sind nicht unabhängig. Stattdessen machen wir von der Tatsache Gebrauch, daß die Anzahl C_x der Rückkünfte nach x geometrisch verteilt ist:

$$\begin{aligned}
\mathbf{Ws}_x\{C_x = m\} &= \mathbf{Ws}_x\{C_x \geq m\} - \mathbf{Ws}_x\{C_x \geq m+1\} \\
&= q^m - q^{m+1} \ = \ q^m p
\end{aligned}$$

mit $q = \mathbf{Ws}_x\{T_x < \infty\} < 1$ (vgl. den Beweis von Prop 5.4). Wegen

$$\sum_{n=1}^{\infty} I_{\{X_n = x\}} \ = \ C_x$$

folgt nun die Behauptung nach Satz 3.11,

$$\sum_{n=1}^{\infty} \mathbf{Ws}_x\{X_n = x\} \;=\; \mathbf{E}_x C_x \;=\; \frac{1}{p} - 1 \;< \infty. \qquad \square$$

Beispiel. Symmetrische Irrfahrten. Wir betrachten nun Markov-Ketten $X_0, X_1, \ldots$ auf dem d-dimensionalen Gitter $\mathbb{Z}^d$ mit den Übergangswahrscheinlichkeiten

$$P_{xy} \;=\; \begin{cases} \frac{1}{2d}\,, & \text{falls } |x-y| = 1\,, \\ 0 & \text{sonst}\,. \end{cases}$$

Ein Wechsel zu einem neuen Gitterpunkt ergibt sich also, indem man zu einem der $2d$ benachbarten Gitterpunkte übergeht, den man rein zufällig auswählt. Diese Markov-Ketten heißen **einfache, *d*-dimensionale, symmetrische Irrfahrten**.

Behauptung. *Diese Irrfahrten sind für $d = 1, 2$ rekurrent und für $d \geq 3$ transient.*

Zum Beweis stellen wir eine Formel für $\mathbf{Ws}_x\{X_n = x\}$ auf. Offenbar kann man nur in einer geraden Anzahl von Schritten mit positiver Wahrscheinlichkeit nach x zurückkehren. Geht man dabei n_i Schritte in die positive Richtung des i-ten Einheitsvektors, so muß man auch n_i Schritte in die entgegengesetzte Richtung gehen, deshalb gilt

$$\mathbf{Ws}_x\{X_{2n} = x\} \;=\; \sum_{n_1+\cdots+n_d=n} \binom{2n}{n_1, n_1, \ldots, n_d, n_d} (2d)^{-2n}$$

oder nach einer einfachen Umformung

$$\mathbf{Ws}_x\{X_{2n} = x\} \;=\; \binom{2n}{n} \sum_{n_1+\cdots+n_d=n} \binom{n}{n_1, \ldots, n_d}^2 (2d)^{-2n}\,.$$

Für $d = 1$ erhalten wir

$$\mathbf{Ws}_x\{X_{2n} = x\} \;=\; \binom{2n}{n} 2^{-2n} \;\sim\; \frac{1}{\sqrt{\pi n}}\,,$$

daher ist die 1-dimensionale Irrfahrt (in Übereinstimmung mit Ergebnissen des vorigen Abschnitts) nach Satz 5.5 rekurrent. Für $d = 2$ folgt (vgl. (1.4))

$$\mathbf{Ws}_x\{X_{2n} = x\} \;=\; \binom{2n}{n} \sum_{n_1=0}^{n} \binom{n}{n_1}\binom{n}{n-n_1} 4^{-2n} \;=\; \binom{2n}{n}^2 4^{-2n} \;\sim\; \frac{1}{\pi n}\,,$$

die 2-dimensinale Irrfahrt ist also nach Satz 5.5 ebenfalls rekurrent. Für $d \geq 3$ schätzen wir zunächst Multinomialkoeffizienten ab. Gilt $n_i \leq n_j - 2$, so folgt

$$\binom{n}{n_1, \ldots, n_i, \ldots, n_j, \ldots, n_d} \;\leq\; \binom{n}{n_1, \ldots, n_i + 1, \ldots, n_j - 1, \ldots, n_d} ,$$

für den maximalen Multinomialkoeffizient gilt daher $n_i = m-1$ oder $n_i = m$, wobei m die kleinste ganze Zahl $\geq n/d$ bezeichne. Indem wir gegebenenfalls n_i noch von $m-1$ auf m vergrößern, und dann schrittweise auch n, erhalten wir insgesamt die Abschätzung

$$\binom{n}{n_1, \ldots, n_d} \;\leq\; \binom{dm}{m, \ldots, m} .$$

Es folgt

$$\mathbf{Ws}_x\{X_{2n} = x\} \;\leq\; \binom{2n}{n} 2^{-2n} \binom{dm}{m, \ldots, m} d^{-n} \sum_{n_1 + \cdots + n_d = n} \binom{n}{n_1, \ldots, n_d} d^{-n} .$$

Die Multinomialgewichte summieren sich zu 1 auf, deswegen vereinfacht sich die Abschätzung zu

$$\mathbf{Ws}_x\{X_{2n} = x\} \;\leq\; \binom{2n}{n} 2^{-2n} \binom{dm}{m, \ldots, m} d^{-n} .$$

Der rechte Ausdruck ist nach der Stirling Formel asymptotisch gleich $(\pi n)^{-1/2} (2\pi dm)^{1/2} (2\pi m)^{-d/2} d^{dm-n}$. Wegen $dm \leq n + d$ erhalten wir insgesamt für $n \to \infty$

$$\mathbf{Ws}_x\{X_{2n} = x\} \;=\; O(n^{-d/2}) .$$

Folglich ist nach Satz 5.5 die Irrfahrt für $d \geq 3$ transient. □

5.3 Gleichgewichtsverteilungen

Wir betrachten nun die Situation, daß eine Markov-Kette (X_n) sich im ‚Zustand des Gleichgewichts' befindet. Gemeint ist, daß sich die Verteilung von X_n nicht mit n ändert.

Definition. *Eine Wahrscheinlichkeitsverteilung $\pi = (\pi_x)$ auf S heißt* ***Gleichgewichtsverteilung*** *(oder* ***stationäre Verteilung****) für eine Markov-Kette mit Übergangswahrscheinlichkeiten P_{xy}, falls gilt*

$$\pi_x \;=\; \sum_{y \in S} \pi_y P_{yx} \quad \textit{für alle } x \in S .$$

Wählt man für eine Markov-Kette (X_n) die Anfangsverteilung als Gleichgewichtsverteilung π, so folgt nach dem Satz von der totalen Wahrscheinlichkeit

$$\begin{aligned}\mathbf{Ws}_\pi\{X_1 = x\} &= \sum_y \mathbf{Ws}_\pi\{X_1 = x \,\big|\, X_0 = y\}\mathbf{Ws}_\pi\{X_0 = y\} \\ &= \sum_y P_{yx}\pi_y \;=\; \pi_x \;=\; \mathbf{Ws}_\pi\{X_0 = x\}\ .\end{aligned}$$

Durch Iteration ergibt sich für alle $x \in S$, $n \in \mathbb{N}$

$$\mathbf{Ws}_\pi\{X_n = x\} \;=\; \mathbf{Ws}_\pi\{X_0 = x\}$$

bzw.

$$\sum_y \pi_y P^n_{yx} \;=\; \pi_x\ ,$$

die Verteilung von X_n ist also in der Tat von n unabhängig.

Gleichgewichtsverteilungen brauchen nicht immer zu existieren. Der folgende Satz gibt ein Kriterium mittels

$$T_x \;:=\; \min\{n \geq 1 \;:\; X_n = x\}\ ,$$

dem Zeitpunkt des ersten Besuchs von x bzw. - falls man in x startet - der ersten Rückkehr nach x. T_x kann mit positiver Wahrscheinlichkeit den Wert ∞ annehmen, daher gilt

$$\mathbf{E}T_x \;=\; \sum_{t=1}^{\infty} t \cdot \mathbf{Ws}\{T_x = t\} + \infty \cdot \mathbf{Ws}\{T_x = \infty\}$$

(mit $\infty \cdot 0 = 0$ und $\infty \cdot w = \infty$ für $w > 0$).

Satz 5.6. *Sei $x \in S$. Dann sind äquivalent:*

i) $\mathbf{E}_x T_x < \infty$ *(und damit* $\mathbf{Ws}_x\{T_x < \infty\} = 1$*),*

ii) es gibt eine stationäre W-Verteilung π mit $\pi_x > 0$.

Beweis. i) $\Rightarrow$ ii): Wir betrachten die Erwartungswerte

$$\rho_y \;:=\; \mathbf{E}_x\Big[\sum_{n=0}^{T_x-1} I_{\{X_n=y\}}\Big]\ .$$

Nach Voraussetzung ist $\rho_y < \infty$ für alle y, denn es gilt $\sum_{n=0}^{T_x-1} I_{\{X_n=y\}} \leq T_x$. Wir zeigen

$$\rho_y \;=\; \sum_z \rho_z P_{zy}\ .$$

Dies ist plausibel: Links steht die mittlere Anzahl der Besuche in y zwischen den Zeitpunkten 0 und $T_x - 1$, rechts die mittlere Anzahl von Besuchen in y zwischen den Zeitpunkten 1 und T_x (z ist der vor y besuchte Zustand), außerdem gilt mit Wahrscheinlichkeit 1 $X_0 = X_{T_x} = x$. Eine formale Rechnung bestätigt dies. Nach der Markov Eigenschaft gilt

$$\begin{aligned}\mathbf{Ws}_x\{T_x > n, X_n = y\}P_{yz} &= \mathbf{Ws}_x\{X_1, \ldots, X_n \neq x, X_n = y\}P_{yz} \\ &= \mathbf{Ws}_x\{X_1, \ldots, X_n \neq x, X_n = y, X_{n+1} = z\} \\ &= \mathbf{Ws}_x\{T_x > n, X_n = y, X_{n+1} = z\}\end{aligned}$$

und folglich nach Satz 3.11

$$\begin{aligned}\sum_y \rho_y P_{yz} &= \sum_z \mathbf{E}_x\Big[\sum_{n=0}^{\infty} I_{\{T_x > n, X_n = y\}}\Big] P_{yz} \\ &= \sum_z \sum_{n=0}^{\infty} \mathbf{Ws}_x\{T_x > n, X_n = y\}P_{yz} \\ &= \sum_y \sum_{n=0}^{\infty} \mathbf{Ws}_x\{T_x > n, X_n = y, X_{n+1} = z\} \\ &= \sum_{n=0}^{\infty} \mathbf{Ws}_x\{T_x > n, X_{n+1} = z\} \\ &= \mathbf{E}_x\Big[\sum_{n=0}^{\infty} I_{\{T_x > n, X_{n+1} = z\}}\Big] = \mathbf{E}_x\Big[\sum_{m=1}^{T_x} I_{\{X_m = z\}}\Big] = \rho_z\ .\end{aligned}$$

Ähnlich gilt nach Satz 3.11

$$\sum_y \rho_y = \sum_y \mathbf{E}_x\Big[\sum_{n=0}^{T_x - 1} I_{\{X_n = y\}}\Big] = \mathbf{E}_x\Big[\sum_y \sum_{n=0}^{T_x - 1} I_{\{X_n = y\}}\Big] = \mathbf{E}_x T_x\ .$$

Daher ist

$$\pi_y = \rho_y \,/\, \mathbf{E}_x T_x\ , \quad y \in S$$

stationäre W-Verteilung, zudem gilt $\pi_x > 0$ wegen $\rho_x = 1$.

ii) $\Rightarrow$ i): Um endliche Erwartungswerte zu garantieren, betrachten wir zunächst $\mathbf{E}_y[\min(T_x, l)]$ mit $l \in \mathbb{N} \cup \{\infty\}$. Ähnlich wie (5.5) zeigt man mittels einer Zerlegung nach dem ersten Schritt die Gleichung

$$\mathbf{E}_y[\min(T_x, l+1)] = 1 + \sum_{z \neq x} P_{yz}\, \mathbf{E}_z[\min(T_x, l)]\ . \tag{5.6}$$

Insbesondere ergibt sich für $l = \infty$

$$\mathbf{E}_y T_x \;=\; 1 + \sum_{z \neq x} P_{yz}\, \mathbf{E}_z T_x \;. \tag{5.7}$$

Ist nun π eine stationäre W-Verteilung, so folgt unter Beachtung von $\mathbf{E}_y[\min(T_x, l)] \leq \mathbf{E}_y[\min(T_x, l+1)]$

$$\sum_y \pi_y\, \mathbf{E}_y[\min(T_x, l)] \;\leq\; 1 + \sum_{z \neq x} \pi_z\, \mathbf{E}_z[\min(T_x, l)] \;.$$

Wegen $\mathbf{E}_y[\min(T_x, l)] \leq l$ sind diese Reihen für $l < \infty$ konvergent, daher folgt

$$\pi_x\, \mathbf{E}_x[\min(T_x, l)] \;\leq\; 1$$

oder, da nach Annahme $\pi_x > 0$ gilt,

$$\sum_{t=1}^{l-1} t \cdot \mathbf{Ws}_x\{T_x = t\} + l \cdot \mathbf{Ws}_x\{T_x \geq l\} \;\leq\; 1 \,\big/\, \pi_x \;.$$

Der Grenzübergang $l \to \infty$ gibt $\mathbf{Ws}_x\{T_x = \infty\} = 0$ und

$$\mathbf{E}_x T_x \;=\; \sum_{t=1}^{\infty} t \cdot \mathbf{Ws}_x\{T_x = t\} \;\leq\; 1 \,\big/\, \pi_x \;<\; \infty \;,$$

also die Behauptung. □

Korollar 5.7. *Besitzt eine Markov-Kette eine eindeutig bestimmte Gleichgewichtsverteilung π, so gilt*

$$\pi_x \;=\; 1 \,\big/\, \mathbf{E}_x T_x$$

für alle $x \in S$ ($1/\infty$ ist 0 zu setzen).

Beweis. Ist $\mathbf{E}_x T_x < \infty$, so gilt wegen der Eindeutigkeit von π

$$\pi_y \;=\; \rho_y \,\big/\, \mathbf{E}_x T_x \;,$$

mit dem im letzten Beweis konstruierten ρ_y. Die Behauptung folgt dann wegen $\rho_x = 1$. Ist dagegen $\mathbf{E}_x T_x = \infty$, so gilt $\pi_x = 0$ nach Satz 5.6. □

Bemerkungen.

1. Der reziproke Erwartungswert im Korollar hat eine natürliche Bedeutung. Seien $Y_{1x}, Y_{2x}, \ldots$ die Wartezeiten zwischen den Momenten, in denen sich die Markov-Kette in x aufhält. Sie sind, wie man sich leicht überzeugt, unabhängige Kopien von T_x, deswegen strebt $k^{-1}(Y_{1x} + \cdots + Y_{kx})$ nach dem Gesetz der großen Zahlen fast sicher gegen $\mathbf{E}_x T_x$. Anders ausgedrückt bedeutet dies, daß die relative Häufigkeit H_{nx} der Besuche in x bis zum Zeitpunkt n fast sicher gegen $1/\mathbf{E}_x T_x$ konvergiert,

$$H_{nx} \;:=\; \frac{1}{n+1}\mathrm{card}\{m \le n \;:\; X_m = x\} \;\to\; \frac{1}{\mathbf{E}_x T_x}\;.$$

   Das Korollar stellt den Zusammenhang zur Gleichgewichtsverteilung her,

$$H_{nx} \;\to\; \pi_x\;.$$

   Für eine Funktion $f : S \to \mathbb{R}_+$ folgt (zumindest für endliches S)

$$\frac{1}{n+1}\sum_{m=0}^{n} f(X_m) \;=\; \sum_x f(x) H_{nx} \;\to\; \sum_x f(x)\pi_x \;=\; \mathbf{E}_\pi f(X_0)\;.$$

   Diese Aussage faßt man kurz in der Formel **Zeitliches Mittel = räumliches Mittel** zusammen.

2. **Stationäre Verteilungen als Eigenvektoren.** In der Sprache der Matrizentheorie ist eine Gleichgewichtsverteilung ein linker Eigenvektor der Übergangsmatrix P zum Eigenwert 1. Wir zeigen, wie man für einen endlichen Zustandsraum S die Existenz einer Gleichgewichtsverteilung mit den Hilfsmitteln der Linearen Algebra beweisen kann. Ausgangspunkt ist die Beobachtung, daß eine stochastische Matrix P immer den Eigenwert 1 hat, zum rechten Eigenvektor $r = (r_x)_{x\in S}$ mit $r_x = 1$ für alle $x \in S$. Also besitzt P, wie die Lineare Algebra lehrt, auch einen linken nichtverschwindenden Eigenvektor $l = (l_x)_{x\in S}$ zum Eigenwert 1. Die Komponenten von l können unterschiedliches Vorzeichen haben, deswegen betrachten wir $\pi_x := |l_x|$. Es folgt

$$\pi_x \;=\; |l_x| \;=\; \Big|\sum_y l_y P_{yx}\Big| \;\le\; \sum_y \pi_y P_{yx}\;.$$

   Die Annahme $\pi_x < \sum_y \pi_y P_{yx}$ für ein $x \in S$ führt zum Widerspruch, dann würde nämlich

$$\sum_x \pi_x \;<\; \sum_x\sum_y \pi_y P_{yx} \;=\; \sum_y \pi_y$$

folgen. Deswegen gilt

$$\pi_x \;=\; \sum_y \pi_y P_{yx}$$

für alle $x \in S$, so daß auch π linker Eigenvektor ist. Da S endlich ist, läßt sich π zur stationären W-Verteilung normieren. □

Notwendig für die Existenz einer Gleichgewichtsverteilung mit $\pi_x > 0$ ist nach Satz 5.6, daß $\mathbf{Ws}_x(T_x < \infty\} = 1$ gilt, d.h. daß x ein rekurrenter Zustand ist. Gilt zusätzlich $\mathbf{E}_x T_x < \infty$, so heißt x **positiv rekurrent**. Gilt dagegen $\mathbf{E}_x T_x = \infty$ für einen rekurrenten Zustand x, so heißt x **null rekurrent**.

Beispiel. Warteschlangen. Wir kehren zu unserem Warteschlangenmodell zurück, bei dem pro Zeiteinheit 1 Person abgefertigt wird und sich im Mittel

$$\mu \;:=\; \sum_{x=1}^{\infty} x p_x$$

neue Personen an die Schlange anstellen. Wir stellen eine Gleichung auf für die mittlere Rückkehrzeit der Markov-Kette in den Zustand 0 (die mittlere Dauer, in der die Warteschlange sich wieder auflöst). Es gilt: $\mathbf{E}_x T_{x-1} = \mathbf{E}_1 T_0$ (dies ist die mittlere Zeit, in der sich die Warteschlange um eine Person vermindert) und $\mathbf{E}_x T_0 = \mathbf{E}_x T_{x-1} + \mathbf{E}_{x-1} T_{x-2} + \cdots + \mathbf{E}_1 T_0 = x\,\mathbf{E}_1 T_0$ (soll sich eine Warteschlange der Länge x auflösen, so müssen der Reihe nach die Zustände $x-1, x-2, \ldots$, durchlaufen werden), daher folgt nach (5.7)

$$\mathbf{E}_0 T_0 \;=\; 1 + \sum_{x \neq 0} P_{0x} \cdot \mathbf{E}_x T_0 \;=\; 1 + \sum_x p_x \cdot x\,\mathbf{E}_1 T_0 \;=\; 1 + \mu\,\mathbf{E}_1 T_0\ .$$

Außerdem gilt $\mathbf{E}_1 T_0 = \mathbf{E}_0 T_0$ (denn es macht keinen Unterschied, ob anfangs eine oder keine Person in der Schlange steht), daher erhalten wir

$$\mathbf{E}_0 T_0 \;=\; 1 + \mu\,\mathbf{E}_0 T_0\ .$$

Für $\mu \geq 1$ kommt als Lösung nur $\mathbf{E}_0 T_0 = \infty$ in Frage, denn negative Werte sind für $\mathbf{E}_0 T_0$ ausgeschlossen. Ist dagegen $\mu < 1$, so hat das Gleichungssystem zwei positive Lösungen, $\mathbf{E}_0 T_0 = \infty$ und $\mathbf{E}_0 T_0 = (1-\mu)^{-1}$, und es bedarf einer Zusatzüberlegung, welches der richtige Wert ist. Zunächst zeigen wir induktiv $\mathbf{E}_x[\min(T_0, l)] \leq x/(1-\mu)$ für $\mu < 1$ und $x \neq 0$. Für $l = 0$ ist die Behauptung

evident, der Induktionsschritt ergibt sich mit Hilfe von (5.6),

$$\begin{aligned}\mathbf{E}_x[\min(T_0, l+1)] &= 1 + \sum_{y\neq 0} P_{xy}\, \mathbf{E}_y \min[(T_0, l)] \\ &\leq 1 + \sum_y p_{y-x+1} \frac{y}{1-\mu} = 1 + \frac{\mu + x - 1}{1-\mu} = \frac{x}{1-\mu}\,.\end{aligned}$$

Der Grenzübergang $l \to \infty$ ergibt für $x \neq 0$ (wie im Beweis von Satz 5.6) $\mathbf{E}_x T_0 \leq x/(1-\mu)$, und es folgt

$$\mathbf{E}_0 T_0 = 1 + \sum_{x\neq 0} P_{0x} \mathbf{E}_x T_0 \leq 1 + \sum_x p_x \frac{x}{1-\mu} = 1 + \frac{\mu}{1-\mu} < \infty\,.$$

Daher ist im Fall $\mu < 1$ die Möglichkeit $\mathbf{E}_0 T_0 = \infty$ ausgeschlossen.

Zusammenfassend stellen wir fest: Im Fall $\mu < 1$ gilt $\mathbf{E}_0 T_0 = (1-\mu)^{-1}$, mit anderen Worten, 0 ist ein positiv rekurrenter Zustand. Die Markov-Kette besitzt dann nach Satz 5.6 eine Gleichgewichtsverteilung π mit

$$\pi_0 = 1 - \mu\,. \tag{5.8}$$

Für $\mu = 1$ ist nach früheren Resultaten 0 ebenfalls rekurrent, nun aber mit einer unendlichen erwarteten Rückkehrzeit. In diesem Fall ist 0 ein null rekurrenter Zustand. □

Wir kommen nun zu der Frage der Eindeutigkeit von Gleichgewichtsverteilungen. Für eine wichtige Klasse von Markov-Ketten ist sie gewährleistet.

Definition. *Eine Markov-Kette mit Übergangsmatrix (P_{xy}) heißt **irreduzibel**, falls für beliebige Zustände x, y ein $n \in \mathbb{N}$ existiert, so daß $P^n_{xy} > 0$.*

Mit anderen Worten: Bei einer irreduziblen Markov-Kette ist jeder Zustand y von jedem Zustand x mit positiver Wahrscheinlichkeit erreichbar.

Proposition 5.8. *Im irreduziblen Fall gibt es höchstens eine stationäre Wahrscheinlichkeitsverteilung π. Sie hat dann die Eigenschaft $\pi_x > 0$ für alle x.*

Beweis. Seien π und ν stationäre W-Verteilungen. Dann gilt $\pi_x - \nu_x = \sum_y (\pi_y - \nu_y) P^n_{yx}$ für alle $x \in S$, $n \in \mathbb{N}$, also

$$|\pi_x - \nu_x| \leq \sum_y |\pi_y - \nu_y| P^n_{yx}\,.$$

Ist nun $\pi \neq \nu$, so gibt es $u, v \in S$, so daß $\pi_u < \nu_u$ und $\pi_v > \nu_v$. Wähle n so, daß $P^n_{uv} > 0$. Es folgt

$$|\pi_v - \nu_v| \;=\; \pi_v - \nu_v \;=\; \sum_y (\pi_y - \nu_y) P^n_{yv} \;<\; \sum_y |\pi_y - \nu_y| P^n_{yv} \;.$$

Summation dieser Ungleichungen führt zum Widerspruch:

$$\sum_x |\pi_x - \nu_x| \;<\; \sum_x \sum_y |\pi_y - \nu_y| P^n_{yx} \;=\; \sum_y |\pi_y - \nu_y| \;.$$

Sei nun $\pi_x = 0$ für ein $x \in S$. Für $y \in S$ folgt $\pi_y P^n_{yx} \leq \sum_z \pi_z P^n_{zx} = \pi_x = 0$. Bei passender Wahl von n ist $P^n_{yx} > 0$, es müßte also $\pi_y = 0$ für alle y gelten. Das ist für eine W-Verteilung ausgeschlossen, so daß die Behauptung folgt. □

Beispiele.

1. **Warteschlangen.** Unser früheres Warteschlangenmodell ist irreduzibel, falls $p_0 > 0$ und $p_0 + p_1 < 1$ (falls sich also die Warteschlange mit positiver Wahrscheinlichkeit sowohl verkürzt als auch verlängert). Unter dieser Annahme können wir die im vorigen Beispiel vorgenommene Analyse folgendermaßen abrunden: Im Fall $\mu < 1$ besitzt die Markov-Kette eine eindeutige Gleichgewichtsverteilung, deren Gewichte alle strikt positiv sind. Im Fall $\mu \geq 1$ gibt es dagegen keine Gleichgewichtsverteilung (denn andernfalls hätte sie nach Proposition 5.8 strikt positive Gewichte, was im Widerspruch zu $\mathbf{E}_0 T_0 = \infty$ stünde).

2. **Endlicher Zustandsraum.** Die Existenz einer Gleichgewichtsverteilung haben wir für Markov-Ketten mit endlichem Zustandsraum S bereits gesichert. Im irreduziblen Fall ist sie eindeutig mit strikt positiven Gewichten. Nach Korollar 5.7 folgt $\mathbf{E}_x T_x < \infty$ für alle $x \in S$, ein Resultat, was sich auch direkt beweisen läßt. □

Stationäre Verteilungen π lassen sich wegen $\sum_y P_{xy} = 1$ auch durch die Forderung

$$\sum_y \pi_x P_{xy} \;=\; \sum_y \pi_y P_{yx} \quad \text{für alle } x \in S$$

charakterisieren. Dies ist eine globale Gleichgewichtsforderung: Die Wahrscheinlichkeit, von x in irgendeinen Zustand überzuwechseln, ist gleich der Wahrscheinlichkeit, von irgendwoher nach x zu gelangen. Stärker ist die Bedingung aus der folgenden Definition, daß für alle Paare x, y Übergänge von

x nach y und von y nach x mit gleicher Wahrscheinlichkeit stattfinden, das System sich also auch lokal im Gleichgewicht befindet.

Definition. *Eine Markov-Kette mit Übergangswahrscheinlichkeiten P_{xy} heißt **reversibel** bzgl. der W-Verteilung π, falls für alle Zustände x, y gilt*

$$\pi_x P_{xy} \;=\; \pi_y P_{yx} \;.$$

π ist dann Gleichgewichtsverteilung der Markov-Kette. Man nennt solche Markov-Ketten reversibel, da ihre Eigenschaften unter Zeitumkehr erhalten bleiben. Gemeint ist, daß bei Startverteilung π die Zufallsvektoren $(X_0, X_1, \ldots, X_n)$ und $(X_n, X_{n-1}, \ldots, X_0)$ identisch verteilt sind:

$$\begin{aligned}
\mathbf{Ws}_\pi\{X_n = x_0, \ldots, X_0 = x_n\} \;&=\; \pi_{x_n} P_{x_n x_{n-1}} \cdots P_{x_1 x_0} \\
&=\; \frac{\pi_{x_n} P_{x_n x_{n-1}}}{\pi_{x_{n-1}}} \cdot \frac{\pi_{x_{n-1}} P_{x_{n-1} x_{n-2}}}{\pi_{x_{n-2}}} \cdots \frac{\pi_{x_1} P_{x_1 x_0}}{\pi_{x_0}} \cdot \pi_{x_0} \\
&=\; \pi_{x_0} P_{x_0 x_1} \cdots P_{x_{n-1} x_n} \\
&=\; \mathbf{Ws}_\pi\{X_0 = x_0, \ldots, X_n = x_n\} \;.
\end{aligned}$$

Beispiele.

1. Sei S die Menge aller Knoten eines endlichen (ungerichteten) Graphen. Man spricht von einer **Irrfahrt** auf S, falls man von jedem Knoten zu einem rein zufällig ausgewählten Nachbarknoten überwechselt (zwei Knoten heißen benachbart, falls sie durch eine Kante verbunden sind). Bezeichnet also $n(x)$ die Anzahl der Nachbarknoten von x, so sind die Übergangswahrscheinlichkeiten durch

$$P_{xy} \;=\; \begin{cases} \frac{1}{n(x)} \;, & \text{falls } x \text{ und } y \text{ benachbart sind,} \\ 0 & \text{sonst} \end{cases}$$

   gegeben. Man überzeugt sich unmittelbar, daß die Irrfahrt zusammen mit der W-Verteilung

$$\pi_x \;=\; \frac{n(x)}{c}$$

   reversibel ist. Die Normierungskonstante ist $c = \sum_x n(x) = 2k$, wobei k die Anzahl der Kanten im Graphen sei.

2. **Das Modell von P. und T. Ehrenfest.** Dies ist ein Modell für die Fluktuationen eines Gases. Das Gas sei in einem Behälter eingeschlossen, den wir in zwei gleichgroße Teilbereiche A und B zerlegt denken. Zwischen diesen Bereichen wechseln die Gasteilchen in zufälliger Weise hin und her.

Wir stellen uns vereinfachend vor, daß pro Zeiteinheit ein rein zufälliges Teilchen von seinem Teilbereich in den anderen Bereich gelangt, das unabhängig von den vorangegangenen Fluktuationen ausgewählt ist. Nach n Zeitschritten befindet sich dann eine zufällige Anzahl X_n der Teilchen im Bereich A. Unser Ansatz besagt, daß $X_0, X_1, \ldots$ eine Markov-Kette bildeen. Der Zustandsraum ist $S = \{0, 1, \ldots, r\}$, wobei r die Gesamtzahl aller Teilchen bezeichne, und die Übergangswahrscheinlichkeiten sind

$$P_{xy} \;=\; \begin{cases} \frac{(r-x)}{r}\,, & \text{falls } y = x+1\,, \\ \frac{x}{r}\,, & \text{falls } y = x-1\,. \end{cases}$$

Das Ehrenfest-Modell ist zusammen mit der Binomial-Verteilung

$$\pi_x \;=\; \binom{r}{x} 2^{-r}$$

reversibel, denn es gilt $\pi_x P_{x,x+1} = \pi_{x+1} P_{x+1,x}$. Insbesondere ist π Gleichgewichtsverteilung. Dies ist nicht schwer zu verstehen: Man wird erwarten, daß sich im stationären Zustand die Teilchen unabhängig voneinander auf die beiden Teilbereiche verteilen und sich mit Wahrscheinlichkeit $1/2$ jeweils in A oder in B befinden. Da das Ehrenfest-Modell irreversibel ist, ist die Gleichgewichtsverteilung eindeutig.

Das Ehrenfestsche Modell hat dazu gedient, den in der statistischen Physik diskutierten *Wiederkehr-Einwand* zu entkräften. Es ist nicht schwer zu zeigen, daß man in dem Modell von jedem Zustand aus jeden anderen Zustand mit Wahrscheinlichkeit 1 erreicht. Insbesondere sagt das Modell voraus, daß das Gas mit Wahrscheinlichkeit 1 immer wieder in den Zustand 0 zurückkehrt, in den Zustand, daß alle Teilchen sich in B befindet. Dieser Befund, den das Ehrenfestsche Modell mit verwandten Modellen der statistischen Physik teilt, widerspricht jeglicher Erfahrung. Sind diese Modelle deswegen unbrauchbar? Das Gegenargument ist, daß diese Rückkehrzeiten so gewaltig groß sind, daß das Phänomen keine praktische Bedeutung besitzt.

Wir wollen diese Behauptung am Ehrenfestschen Modell bestätigen und betrachten dazu die Rückkehrzeit

$$T_x \;:=\; \min\{n \geq 1 \,:\, X_n = x\}$$

in den Zustand x. Nach Korollar 5.7 gilt

$$\mathbf{E}_x T_x \;=\; 2^r \Big/ \binom{r}{x}\,.$$

Wählen wir $r = 10^{23}$, physikalisch gesehen eine realistische Größe, so hat die erwartete Rückkehrzeit in den Zustand 0 (alle Teilchen sind wieder in B) den jenseits jeglicher Vorstellung liegenden Wert

$$2^{10^{23}} .$$

Zum Vergleich: Nach der Stirling-Approximation gilt für geradzahliges r

$$\mathbf{E}_{r/2} T_{r/2} \;\approx\; \sqrt{\frac{\pi r}{2}} .$$

In den Zustand, daß sich die Teilchen gleichmäßig auf A und B verteilen, kehrt das Gas also vergleichsweise ganz schnell zurück. □

5.4 Konvergenz ins Gleichgewicht

Wir haben stationäre W-Verteilungen als Gleichgewichtszustände von Markov-Ketten beschrieben. Diese Sprechweise gewinnt Berechtigung durch den Nachweis, daß eine Markov-Kette auch bei einer nicht-stationären Startverteilung in den Gleichgewichtszustand strebt. Wir konzentrieren uns auf den einfachsten Fall. Situationen wie beim Ehrenfestschen Urnenmodell, wo man von geradzahligen Zuständen nur in ungerade Zustände, und von ungeraden nur in gerade Zustände wechseln kann, bedürfen zusätzlicher Überlegungen. Wir wollen hier solche Periodizitäten ausschließen.

Definition. *Eine Markov-Kette mit Übergangswahrscheinlichkeiten P_{xy} heißt **aperiodisch irreduzibel**, falls für alle $x, y, y' \in S$ ein $m \in \mathbb{N}$ existiert mit $P^m_{xy} > 0$ und $P^m_{xy'} > 0$.*

Jede aperiodisch irreduzible Markov-Kette ist offenbar irreduzibel.

Satz 5.9. *Sei $X_0, X_1, \ldots$ eine aperiodisch irreduzible Markov-Kette, die eine stationäre W-Verteilung π besitzt. Dann gilt bei beliebiger Startverteilung für $n \to \infty$*

$$\mathbf{Ws}\{X_n \in B\} \;\to\; \pi(B)$$

für alle $B \subset S$.

Der Satz heißt der **Ergodensatz für Markov-Ketten**, und Markov-Ketten, die in dem Satz formulierte Konvergenzeigenschaft erfüllen, nennt man **ergodisch**. Der Beweis ergibt, daß diese Konvergenz gleichmäßig in B stattfindet.

Beweis. Wir führen den Beweis durch ein *Kopplungsargument*. Dazu betrachten wir noch eine weitere Markov-Kette $X'_0, X'_1, \ldots$ mit derselben Übergangsmatrix P und Startverteilung π, die unabhängig von der gegebenen Kette $X_0, X_1, \ldots$ sei. Setze

$$T \;:=\; \min\{n \geq 0 \;:\; X_n = X'_n\}$$

und durch Kopplung der beiden Ketten

$$Y_n \;:=\; \begin{cases} X'_n\,, & \text{falls } n \leq T\,, \\ X_n\,, & \text{falls } n > T\,. \end{cases}$$

Die naheliegende Vermutung, daß dann auch $Y_0, Y_1, \ldots$ eine Markov-Kette mit Übergangsmatrix P ist, ist leicht bestätigt: Für $y_0, \ldots, y_n \in S$ gilt aufgrund von Unabhängigkeit

$$\begin{aligned}
&\mathbf{Ws}\{Y_0 = y_0, \ldots, Y_n = y_n\} \\
&\quad= \sum_{m=0}^{n} \mathbf{Ws}\{Y_0 = y_0, \ldots, Y_n = y_n, T = m\} \\
&\qquad\qquad + \mathbf{Ws}\{Y_0 = y_0, \cdots, Y_n = y_n, T > n\} \\
&\quad= \sum_{m=0}^{n} \mathbf{Ws}\{X'_0 = y_0, \ldots, X'_m = y_m\} \\
&\qquad\qquad \cdot \mathbf{Ws}\{X_0 \neq y_0, \ldots, X_{m-1} \neq y_{m-1}, X_m = y_m, \ldots, X_n = y_n\} \\
&\qquad\qquad + \mathbf{Ws}\{X'_0 = y_0, \ldots, X'_n = y_n\}\mathbf{Ws}\{X_0 \neq y_0, \ldots, X_n \neq y_n\}
\end{aligned}$$

und unter Beachtung von (5.1) und der Markov-Eigenschaft

$$\begin{aligned}
&\mathbf{Ws}\{Y_0 = y_0, \ldots, Y_n = y_n\} \\
&\quad= \sum_{m=0}^{n} \pi_{y_0} P_{y_0 y_1} \cdots P_{y_{m-1} y_m} \\
&\qquad \cdot \mathbf{Ws}\{X_0 \neq y_0, \ldots, X_{m-1} \neq y_{m-1}, X_m = y_m\} P_{y_m y_{m+1}} \cdots P_{y_{n-1} y_n} \\
&\qquad\quad + \pi_{y_0} P_{y_0 y_1} \cdots P_{y_{n-1} y_n} \mathbf{Ws}\{X_0 \neq y_0, \ldots, X_n \neq y_n\} \\
&\quad= \pi_{y_0} P_{y_0 y_1} \cdots P_{y_{n-1} y_n}\,.
\end{aligned}$$

Nach Proposition 5.1 ist also $Y_0, Y_1, \ldots$ tatsächlich eine Markov-Kette mit stationärer Anfangsverteilung. Es folgt

$$\begin{aligned}
\big|\mathbf{Ws}\{X_n \in B\} - \pi(B)\big| \;&=\; \big|\mathbf{Ws}\{X_n \in B\} - \mathbf{Ws}\{Y_n \in B\}\big| \\
&\leq\; \mathbf{Ws}\{X_n \neq Y_n\} \;=\; \mathbf{Ws}\{T > n\}\,,
\end{aligned}$$

und es bleibt zu zeigen, daß $\lim_n \mathbf{Ws}\{T > n\} = \mathbf{Ws}\{T = \infty\}$ gleich 0 ist.

Dazu bemerken wir, daß auch $Z_n := (X_n, X'_n)$, $n = 0, 1, \ldots$ eine Markov-Kette im Zustandsraum $S \times S$ ist, denn für $z_0 = (x_0, x'_0), \ldots, z_n = (x_n, x'_n)$ gilt aufgrund von Unabhängigkeit und (5.1)

$$\begin{aligned}
&\mathbf{Ws}\{Z_0 = z_0, \ldots, Z_n = z_n\} \\
&\quad = \mathbf{Ws}\{X_0 = x_0, \ldots, X_n = x_n\}\mathbf{Ws}\{X'_0 = x'_0, \ldots, X'_n = x'_n\} \\
&\quad = \mathbf{Ws}\{X_0 = x_0\}P_{x_0x_1} \cdots P_{x_{n-1}x_n} \cdot \mathbf{Ws}\{X'_0 = x'_0\}P_{x'_0x'_1} \cdots P_{x'_{n-1}x'_n} \\
&\quad = \mathbf{Ws}\{Z_0 = z_0\}Q_{z_0z_1} \cdots Q_{z_{n-1}z_n}
\end{aligned}$$

mit $Q_{zu} := P_{xy}P_{x'y'}$, $z = (x, x'), u = (y, y')$. Da für beliebiges $x \in S$

$$\begin{aligned}
\mathbf{Ws}\{T = \infty\} &\le \mathbf{Ws}\{Z_n \neq (x, x) \text{ für alle } n \ge 0\} \\
&= \sum_{y,y'} \mathbf{Ws}\{Z_0 = (y, y')\}\mathbf{Ws}_{(y,y')}\{Z_n \neq (x, x) \text{ für alle } n \ge 0\} \ ,
\end{aligned}$$

bleibt zu zeigen, daß $\mathbf{Ws}_{(y,y')}\{Z_n \neq (x, x) \text{ für alle } n \ge 0\} = 0$ gilt für alle x, y, y'. Den trivialen Fall $x = y = y'$ lassen wir beiseite.

Zum Beweis bemerken wir, daß die Übergangsmatrix $Q = (Q_{zu})$ die stationäre Verteilung ρ mit den Gewichten $\rho_{(x,x')} := \pi_x\pi_{x'}$ besitzt. Nach Proposition 5.8 sind die Gewichte von π und damit von ρ alle strikt positiv, und nach Satz 5.6 sind folglich alle Zustände $z \in S \times S$ bzgl. Q (positiv) rekurrent. Nun bringen wir die Annahme aperiodischer Irreduzibilität ins Spiel. Danach ist $\mathbf{Ws}_{(x,x)}\{Z_m = (y, y')\} = Q^m_{(x,x)(y,y')} = P^m_{xy}P^m_{xy'}$ zu vorgegebenen x, y, y' für geeignetes $m \in \mathbb{N}$ strikt positiv. Es folgt, daß umgekehrt wie behauptet $\mathbf{Ws}_{(y,y')}\{Z_n \neq (x, x) \text{ für alle } n \ge 1\}$ für alle x, y, y' gleich 0 ist – denn andernfalls würde (Z_n) mit positiver Wahrscheinlichkeit nicht nach (x, x) zurückkehren, und (x, x) wäre kein rekurrenter Zustand. □

Der Konvergenzsatz hat vielfältige Anwendungen.

Beispiele.

1. **Kartenmischen.** Kann man einen Mangel an Übung beim Mischen von Spielkarten dadurch kompensieren, daß man ausreichend lange mischt? Um den Vorgang mathematisch zu beschreiben, identifizieren wir das Blatt mit der Menge $B := \{1, 2, \ldots, b\}$; die 1 steht für die Karte oben auf dem Stapel und b für die Karte ganz unten. Einmaliges Mischen entspricht dann (wie wir schon in Abschnitt 1.5 beschrieben haben) einer zufälligen Permutation Π von B, einem zufälligen Element der Menge

$$S := \{\pi : B \to B \ : \ \pi \text{ ist eine Bijektion}\} \ ,$$

und mehrfaches Mischen einer Hintereinanderausführung

$$X_n \;:=\; \Pi_n \circ \Pi_{n-1} \circ \cdots \circ \Pi_1$$

von mehreren zufälligen Permutationen Π_1, $\Pi_2, \ldots$ Wir nehmen an, daß es sich um unabhängige Kopien von Π handelt, dann ist $id = X_0, X_1, \ldots$ eine Markov-Kette mit den Übergangswahrscheinlichkeiten

$$P_{xy} \;=\; \mathbf{Ws}\{\Pi \circ x = y\}\;, \quad x, y \in S\;.$$

Diese Übergangsmatrix die besondere Eigenschaft, daß neben $\sum_y P_{xy} = 1$ auch

$$\sum_{x\in S} P_{xy} \;=\; \sum_{x\in S} \mathbf{Ws}\{\Pi = y \circ x^{-1}\} \;=\; \sum_{z\in S} \mathbf{Ws}\{\Pi = z\} \;=\; 1$$

gilt, man sagt, die Matrix ist **doppelt-stochastisch**. Damit sind $\pi_x = \frac{1}{r!}$ die Gewichte einer Gleichgewichtsverteilung auf S. Im aperiodisch irreduziblen Fall strebt daher $\mathbf{Ws}\{X_n = x\}$ gegen $\frac{1}{r!}$, d.h. X_n ist asymptotisch uniform verteilt auf der Menge aller möglichen Anordnungen des Kartenspiels. Der gewünschte Mischeffekt stellt sich also wirklich ein, vorausgesetzt, man mischt das Blatt ausreichend lange.

2. **Der Metropolis-Algorithmus.** Hier handelt es sich um eine besonders wichtige Anwendung des Konvergenzsates. Der Algorithmus wurde von Physikern erfunden zum Zwecke der Simulation von W-Verteilungen π mit Gewichten von der Gestalt

$$\pi_x \;=\; c\rho_x\;, \quad x \in S\;.$$

Dabei ist insbesondere an die Situation gedacht, daß nur die Zahlen ρ_x bekannt sind. Natürlich ist dann auch die Normierungskonstante $c = \big(\sum_x \rho_x\big)^{-1}$ festgelegt, in vielen wichtigen Fällen läßt sie sich aber nicht einmal näherungsweise berechnen (ein typisches Beispiel sind die Gibbs-Verteilungen der Statistischen Physik).

Für den Algorithmus benötigt man eine (gut auf dem Computer zu simulierende) Markov-Kette mit Zustandsraum S (etwa einer Irrfahrt, wenn S die Struktur eines Graphen besitzt). Aus ihrer Übergangsmatrix $Q = (Q_{xy})$ bildet man eine neue Übergangsmatrix P nach der Vorschrift

$$P_{xy} \;:=\; \pi_x^{-1} \min(\pi_x Q_{xy}, \pi_y Q_{yx}) \;=\; \min\left(Q_{xy}, \frac{\rho_y}{\rho_x}\, Q_{yx}\right)\;, \quad \text{falls} \;\; x \neq y\;,$$

$$P_{xx} \;:=\; 1 - \sum_{y\neq x} P_{xy}\;.$$

P ist ebenfalls stochastische Matrix, denn wegen $P_{xy} \le Q_{xy}$ für $x \neq y$ gilt $P_{xx} \ge Q_{xx} \ge 0$. Man bemerke, daß man zur Berechnung von Q_{xy} nur die ρ_x, nicht aber die Normierungskonstante zu kennen braucht. Aus

$$\pi_x P_{xy} \;=\; \pi_y P_{yx} = \min(\pi_x Q_{xy}, \pi_y Q_{yx}) \;, \quad x \neq y$$

folgt, daß die Markov-Kette reversibel und π Gleichgewichtsverteilung bzgl. P ist. Für eine Markov-Kette $X_0, X_1, \ldots$ mit Übergangsmatrix P wird nach Satz 5.9 X_n daher (unter der Annahme von Ergodizität) approximativ die Verteilung π besitzen, wenn n nur ausreichend groß ist. Die Idee des Metropolis-Algorithmus ist daher, eine solche Markov-Kette auf einem Rechner zu simulieren. Dabei kann man so vorgehen:

1. Befindet man sich im Zustand x, so wähle zufällig einen neuen Zustand, und zwar y mit Wahrscheinlichkeit Q_{xy}.
2. Davon unabhängig wähle man rein zufällig eine Zahl U aus dem Intervall $[0, 1]$.
3. Ist $y \neq x$ und $U \le \frac{\rho_y Q_{yx}}{\rho_x Q_{xy}}$, so vollziehe man den Übergang nach y, anderfalls verharre man in x.

Ein Wechsel von x nach y findet nach diesem Rezept wie gewünscht mit der Wahrscheinlichkeit

$$Q_{xy} \cdot \min\left(1, \frac{\rho_y Q_{yx}}{\rho_x Q_{xy}}\right) \;=\; Q_{xy} \cdot \frac{P_{xy}}{Q_{xy}} \;=\; P_{xy}$$

statt.

Der Metropolis-Algorithmus wird gern zum Simulieren komplexer W-Verteilungen π benutzt. Das Hauptproblem bei seiner Anwendung besteht darin zu entscheiden, wie lange die Markov-Kette laufen muß, um eine ausreichende Genauigkeit zu erreichen.

3. **Erneuerungstheorie.** Ein (diskreter) **Erneuerungsprozeß** ist eine Folge $Y_1, Y_2, \ldots$ von unabhängigen Kopien einer Zufallsvariablen Y mit Werten in der Menge $\mathbb{N}$ der natürlichen Zahlen und Gewichten

$$p_y \;=\; \mathbf{Ws}\{Y = y\} \;, \quad y = 1, 2, \ldots$$

Man stellt sich vor, daß $Y_1 + \cdots + Y_k$, $k \in \mathbb{N}$, die Zeitpunkte sind, zu denen ein bestimmter Baustein (eine ‚Glühbirne') in einer technischen Anlage erneuert werden muß, daß also Y_1 die Funktionsdauer der ursprünglichen Komponente und Y_{k+1} die Funktionsdauer der k-ten Ersatzkomponente angibt.

Eine Möglichkeit zur Analyse des Erneuerungsprozesses besteht darin, ihn in eine Markov-Kette $X_0, X_1, \ldots$ einzubetten. Als Zustandsraum wählen wir $S = \{0, 1, \ldots, r-1\}$, wobei r die kleinste Zahl mit $p_1 + \cdots + p_r = 1$ bezeichne, bzw. $S = \mathbb{N}_0$ im Fall $r = \infty$, und als Übergangswahrscheinlichkeiten

$$P_{0x} \;=\; p_{x+1} \quad \text{für } x \geq 0\ , \quad P_{x,x-1} \;=\; 1 \quad \text{für } x \geq 1\ , \quad P_{xy} \;=\; 0 \quad \text{sonst}\ .$$

Sei Y_k die Anzahl der Schritte, die die Markov-Kette nach dem k-ten Aufenthalt im Zustand 0 benötigt, um die 0 erneut zu erreichen. Bei Start im Zustand 0 gilt

$$\begin{aligned}&\{X_0 = 0, Y_1 = y_1, \ldots, Y_k = y_k\}\\ &\quad = \; \{X_0 = 0\} \cap \bigcap_{i=1}^{k} \{X_{y_0+\cdots+y_{i-1}+j} = y_i - j \text{ für } j = 1, \ldots, y_i\}\end{aligned}$$

(mit $y_0 := 0$) und folglich

$$\mathbf{Ws}_0\{Y_1 = y_1, \ldots, Y_k = y_k\} \;=\; \prod_{i=1}^{k} P_{0,y_i-1} P_{y_i-1,y_i-2} \cdots P_{10} \;=\; p_{y_1} \cdots p_{y_k}\ ,$$

die Rückkehrdauern $Y_1, Y_2, \ldots$ sind also unabhängige Kopien von Y und stellen damit den ursprünglichen Erneuerungsprozess dar. X_n gibt, vom Zeitpunkt n aus gesehen, die Dauer bis zur nächsten Erneuerung an. Erneuerungen finden zu den Zeitpunkten n mit $X_n = 0$ statt.

Wir untersuchen die Markov-Kette auf ihre Gleichgewichtsverteilungen π. Stationarität bedeutet im vorliegenden Fall

$$\pi_x \;=\; \pi_0 p_{x+1} + \pi_{x+1}$$

für alle $x \in S$ (mit $\pi_r = 0$). Summiert man die Gleichungen von 0 bis $x-1$, so folgt

$$\pi_0 \;=\; \pi_0(p_1 + \ldots + p_x) + \pi_x$$

oder

$$\pi_x \;=\; \pi_0(1 - p_1 - \ldots - p_x) \;=\; \pi_0\, \mathbf{Ws}\{Y \geq x+1\}\ .$$

Wegen $\sum_{y=1}^{r} \mathbf{Ws}\{Y \geq y\} = \mathbf{E}Y$ kann man diese Gewichte nur dann zu einer Wahrscheinlichkeitsverteilung normieren, wenn Y endlichen Erwartungswert hat. Wir setzen dies nun voraus und erhalten, daß mit dieser Annahme die Existenz einer eindeutigen Gleichgewichtsverteilung gesichert ist, deren Gewichte sich nach der Formel

$$\pi_x \;=\; \frac{\mathbf{Ws}\{Y \geq x+1\}}{\mathbf{E}Y}$$

bestimmen. Um den Konvergenzsatz anwenden zu können, nehmen wir weiter an, daß die Markov-Kette aperiodisch irreduzibel ist. (Es ist eine Aufgabe der elementaren Zahlentheorie zu zeigen, daß dies genau dann der Fall ist, wenn der größte gemeinsame Teiler der natürlichen Zahlen y mit $p_y > 0$ gleich 1 ist.)

Eine erste Folgerung betrifft die Wahrscheinlichkeit

$$u_n \;:=\; \mathbf{Ws}\{X_n = 0\} \;=\; \mathbf{Ws}\{Y_k = n \text{ für ein } k \geq 1\} \;,$$

daß genau zum Zeitpunkt n eine Erneuerung stattfindet. Nach dem Konvergenzsatz hat u_n den Grenzwert π_0, für $n \to \infty$ gilt also

$$u_n \;\to\; \frac{1}{\mathbf{E}Y} \;.$$

Diese Aussage nennt man den **Erneuerungssatz**.

Im Folgenden betrachten wir die Zufallsvariablen

$$\begin{aligned} L_n \;&:=\; \min\{k \geq 0 \,:\, X_{n-k} = 0\} \;, \\ R_n \;&:=\; \min\{l \geq 1 \,:\, X_{n+l} = 0\} \;, \\ G_n \;&:=\; L_n + R_n \;, \end{aligned}$$

das **aktuelle Lebensalter**, die **Restlebenszeit** und die **Gesamtlebensdauer** der zum Zeitpunkt n arbeitenden (bzw. der im Fall einer Erneuerung soeben neu eingesetzten) Komponente.

Für die Gesamtlebensdauer gilt

$$\mathbf{Ws}\{G_n = m\} \;=\; \sum_{k=0}^{m-1} \mathbf{Ws}\{X_{n-k} = 0, X_{n-k+1} = m-1\} \;=\; \sum_{k=0}^{m-1} u_{n-k} p_m \;,$$

daher folgt nach dem Erneuerungssatz

$$\mathbf{Ws}\{G_n = m\} \;\to\; \frac{m p_m}{\mathbf{E}Y} \;, \quad m = 1, 2, \ldots$$

Man sagt, die W-Verteilung mit den Gewichten

$$\widehat{p}_m \;:=\; \frac{m p_m}{\mathbf{E}Y}$$

ergibt sich durch **Größenverzerrung** aus der Verteilung von Y. Das Resultat ist gut zu verstehen: Die Chance vergrößert sich, zu einem fernen Zeitpunkt n ein Bauteil mit langer Gesamtlebenszeit vorzufinden, und

zwar proportional zur seiner Lebensdauer. Dies schlägt sich auch im Erwartungswert der Grenzverteilung nieder, wegen $\mathbf{E}Y^2 \geq (\mathbf{E}Y)^2$ gilt

$$\sum_m m\hat{p}_m \;=\; \frac{\mathbf{E}Y^2}{\mathbf{E}Y} \;\geq\; \mathbf{E}Y \;.$$

Als nächstes betrachten wir, wie sich das aktuelle Lebensalter zur Gesamtlebensdauer verhält. Es gilt

$$\mathbf{Ws}\{L_n = k, G_n = m\} \;=\; \mathbf{Ws}\{X_{n-k} = 0, X_{n-k+1} = m-1\} \;=\; u_{n-k}p_m \;,$$

daher folgt mit einer weiteren Anwendung des Erneuerungssatzes

$$\mathbf{Ws}\{L_n = k \,\big|\, G_n = m\} \;\to\; \frac{1}{m} \;, \quad 1 \leq k \leq m \;.$$

Asymptotisch ist L_n also uniform verteilt auf $\{1, \ldots, G_n\}$.

Damit wird klar, wie das Lebensalter der zum Zeitpunkt n funktionierenden Komponente asymptotisch verteilt ist:

$$\mathbf{Ws}\{L_n = k\} \;=\; \sum_{m=k+1}^{\infty} \mathbf{Ws}\{L_n = k \,\big|\, G_n = m\}\, \mathbf{Ws}\{G_n = m\}$$

hat den Grenzwert

$$\sum_{m=k+1}^{\infty} \frac{1}{m}\, \widehat{p}_m \;=\; \frac{\mathbf{Ws}\{Y \geq k+1\}}{\mathbf{E}Y} \;.$$

Dieses Resultat (das im Fall einer unendlichen Summe genaugenommen einer zusätzlichen Begründung für Vertauschung von Grenzwertbildung und Summation bedarf) läßt sich auch direkt ableiten: Es gilt

$$\mathbf{Ws}\{L_n = k\} = \mathbf{Ws}\{X_{n-k} = 0, X_{n-k+1} \geq k\} = u_{n-k}\mathbf{Ws}\{Y \geq k+1\} \;,$$

daher folgt aus dem Erneuerungssatz

$$\mathbf{Ws}\{L_n = k\} \;\to\; \frac{\mathbf{Ws}\{Y \geq k+1\}}{\mathbf{E}Y} \;, \quad k = 0, 1, \ldots$$

Für die Restlebenszeit gilt ähnlich die Asymptotik

$$\mathbf{Ws}\{R_n = l\} \;\to\; \frac{\mathbf{Ws}\{Y \geq l\}}{\mathbf{E}Y} \;, \quad l = 1, 2, \ldots \;,$$

die aus $\mathbf{Ws}\{R_n = l\} = \mathbf{Ws}\{X_{n+1} = l-1\} \to \pi_{l-1}$ folgt. □

Kapitel 6

Die Normalverteilung

Die Normalverteilung nimmt in der Stochastik einen prominenten Platz ein. Das liegt hauptsächlich an ihren günstigen strukturellen Eigenschaften, die erst in einer mehrdimensionalen Betrachtungsweise zutage treten: Ein standard normalverteilter Zufallsvektor mit Werten in einem endlich dimensionalen Euklidischen Vektorraum hat unabhängige Komponenten, außerdem ist seine Verteilung unter Drehungen invariant. Dieser Sachverhalt hat wichtige Anwendungen, wir behandeln die Varianzanalyse, ein zentrales Kapitel der Statistik. - Außerdem beweisen wir den zentralen Grenzwertsatz, der eine weitere Begründung für den Stellenwert der Normalverteilung bietet. Grob gesprochen besagt er folgendes: Eine reellwertige Zufallsvariable, die sich aus vielen kleinen, unabhängigen Summanden zusammensetzt, ist approximativ normalverteilt.

Technisch gesehen hat man es bei der Normalverteilung mit mehrdimensionalen Integralen zu tun. Da wir keine speziellen Kenntnisse der Integrationstheorie voraussetzen wollen (wie den Satz von Fubini und die Transformationsformel für mehrdimensionale Integrale) gestatten wir uns, an ein paar Stellen mit infinitesimalen Größen zu rechnen.

6.1 Standard normalverteilte Zufallsvektoren

In diesem Kapitel ist es zweckmäßig, Zufallsvektoren als Spaltenvektoren aufzufassen.

Definition. *Ein Zufallsvektor $Z = (Z_1, \dots, Z_n)^t$ mit Werten im $\mathbb{R}^n$ heißt* ***standard normalverteilt****, falls seine reellwertigen Komponenten $Z_1, \dots, Z_n$ unabhängige, standard normalverteilte Zufallsvariable sind.*

Die Unabhängigkeit der Komponenten ist das eine wichtige Merkmal von standard normalverteilten Zufallsvektoren, das andere wird aus ihrer Dichte ersichtlich. In Verallgemeinerung der Dichte der 1-dimensionalen standard Normalverteilung (der Gaußschen Glockenkurve) definieren wir die **Dichte der multivariaten standard Normalverteilung** als

$$\mathsf{n}(z) \;:=\; (2\pi)^{-n/2} \exp\big(-\tfrac{1}{2}|z|^2\big)\;, \qquad z = (z_1, \dots, z_n)^t\;,$$

mit der quadrierten Euklidischen Norm $|z|^2 = z_1^2 + \cdots + z_n^2$ von z.

Proposition 6.1. *Z mit Werten in $\mathbb{R}^n$ ist genau dann standard normalverteilt, wenn für alle $B \subset \mathbb{R}^n$ mit wohldefiniertem Volumen*

$$\mathbf{Ws}\{Z \in B\} \;=\; \int_B \mathsf{n}(z)\, dz$$

gilt.

Wie in (**??**) schreiben wir

$$\mathbf{Ws}\{Z \in dv\} \;=\; \mathsf{n}(z)\, dz$$

und lassen uns dabei von der Vorstellung leiten, daß dv ein ‚Volumenelement' an der Stelle $z \in \mathbb{R}^n$ vom infinitesimalen Volumen dz ist, also ein infinitesimaler Quader oder allgemeiner ein Parallelepiped, daß von n linear unabhängigen infinitesimalen Vektoren aufgespannt ist.

Beweis. Die Behauptung folgt aus Proposition 2.7, denn $\mathsf{n}(z)$ läßt sich als Produkt der Dichten 1-dimensionaler Normalverteilungen darstellen,

$$\mathsf{n}(z) \;=\; \prod_{i=1}^{n} (2\pi)^{-1/2} \exp\big(-z_i^2/2\big)\;.$$

□

Die Dichte der standard Normalverteilung hat die grundlegende Eigenschaft, daß sie unter Drehungen des Koordinatensystems invariant ist, denn $\mathsf{n}(z)$ hängt nur von der Euklidischen Norm $|z|$ ab, die bekanntlich bei Drehungen um 0 unverändert bleibt. In Zufallsvariablen ausgedrückt bedeutet dies das Folgende.

Proposition 6.2. *Ist $Z = (Z_1, \dots, Z_n)^t$ standard normalverteilt, und ist $O = (o_{ij})$ eine orthogonale $n \times n$-Matrix, dann ist auch $\widetilde{Z} := O \cdot Z$ standard normalverteilt. Mit anderen Worten: Die Komponenten $\widetilde{Z}_i := \sum_j o_{ij} Z_j$ sind unabhängige, standard normalverteilte Zufallsvariable.*

Beispiel. Seien $X_1, \ldots, X_n$ unabhängige, normalverteilte Zufallsvariable mit Erwartungswerten $\mu_1, \ldots, \mu_n$ und Varianzen $\sigma_1^2, \ldots, \sigma_n^2$. Wir wollen zeigen, daß

$$X \;:=\; X_1 + \cdots + X_n$$

ebenfalls normalverteilt ist, mit Erwartung $\mu := \mu_1 + \cdots + \mu_n$ und Varianz $\sigma^2 := \sigma_1^2 + \cdots + \sigma_n^2$. Wir gehen von der Darstellung $X_i = \mu_i + \sigma Z_i$ aus, mit unabhängigen, standardnormalverteilten Zufallsvariablen $Z_1, \ldots, Z_n$. Dann gilt

$$X \;=\; \mu + \sigma\big(\frac{\sigma_1}{\sigma} Z_1 + \cdots + \frac{\sigma_n}{\sigma} Z_n\big) \;.$$

Zu zeigen bleibt, daß der Klammerausdruck eine standard normalverteilte Zufallsvariable ist. Zum Beweis erinnern wir daran, daß eine Matrix genau dann orthogonal ist, wenn ihre Zeilen, aufgefaßt als Vektoren, eine orthonormale Basis des $\mathbb{R}^n$ bilden. Wie man in der Linearen Algebra lernt, kann man jeden Vektor der Länge 1 zu einer orthonormalen Basis ergänzen, also gibt es eine orthogonale Matrix O, deren erste Zeile aus den Zahlen $\sigma_1/\sigma, \ldots, \sigma_n/\sigma$ besteht. Die Behauptung folgt nun aus Proposition 6.2, denn $(\sigma_1/\sigma)Z_1 + \cdots + (\sigma_n/\sigma)Z_n$ ist dann die erste Komponente von $\widetilde{Z}$. - Man kann diese Behauptung auch mit der Faltungsformel (**??**) nachrechnen. □

Damit sind die beiden fundamentalen Eigenschaften von standard normalverteilten Zufallsvektoren genannt: Sie haben unabhängige Komponenten, und sie haben eine unter Drehungen invariante Verteilung. Man kann beweisen, daß diese Eigenschaften für die standard Normalverteilung charakteristisch sind. Sie ermöglichen eine Reihe konkreter Rechnungen, die die Normalverteilung gerade auch für Anwendungszwecke interessant macht. Wir zeigen, wie man sie zur Konstruktion von Konfidenzintervallen nutzt.

Ein exaktes Konfidenzintervall

Ein gebräuchliches statistisches Modell für wiederholte Messungen $X_1, \ldots, X_n$ einer reellen Größe μ besteht in der Annahme, daß die X_i unabhängige $\mathrm{N}(\mu, \sigma^2)$-verteilte Zufallsvariablen sind, daß also

$$X_i \;=\; \mu + \sigma Z_i \;, \quad i = 1, \ldots, n \;, \tag{6.1}$$

gilt mit unabhängigen, standard normalverteilten Zufallsvariablen $Z_1, \ldots, Z_n$. Der Skalenparameter $\sigma > 0$ bestimmt die Meßgenauigkeit, wir nehmen ihn als unbekannt an.

Die Konstruktion von exakten Konfidenzintervallen für μ gehört zu den klassischen Aufgaben der Statistik. Es liegt nahe, Intervalle der Gestalt $\widehat{\mu} \pm \gamma$

zu betrachten, mit

$$\widehat{\mu} \;:=\; n^{-1}\sum_{i=1}^{n} X_i \;,$$

denn $\widehat{\mu}$ ist ein natürlicher Schätzer von μ. Es gilt $\mathbf{E}[(\widehat{\mu}-\mu)^2] = \mathbf{Var}[\widehat{\mu}] = \sigma^2/n$, zwischen $\widehat{\mu}$ und μ ist also mit Abweichungen von der Größenordnung $\sigma/\sqrt{n}$ zu rechnen. Diese Größe läßt sich durch $\widehat{\sigma}/\sqrt{n}$ schätzen, mit

$$\widehat{\sigma}^2 \;:=\; (n-1)^{-1}\sum_{i=1}^{n}(X_i-\widehat{\mu})^2$$

(die Normierung mit $(n-1)^{-1}$ anstelle von n^{-1} findet später ihre Erklärung), für das Konfidenzintervall ist daher der Ansatz

$$\widehat{\mu} \pm c\widehat{\sigma}/\sqrt{n}$$

plausibel. Die Aufgabe bei der Konstruktion von Konfidenzintervallen besteht nun darin (vgl. Abschnitt 1.4), zu vorgegebenem Signifikanzniveau $\alpha \in (0,1)$ eine positive Zahl c_α zu bestimmen, so daß für alle μ, σ

$$\mathbf{Ws}\{\widehat{\mu} - c_\alpha\widehat{\sigma}n^{-1/2} \le \mu \le \widehat{\mu} + c_\alpha\widehat{\sigma}n^{-1/2}\} \;=\; 1-\alpha$$

bzw.

$$\mathbf{Ws}\{-c_\alpha \le T_n \le c_\alpha\} \;=\; 1-\alpha \tag{6.2}$$

gilt, mit

$$T_n \;:=\; \frac{\sqrt{n}(\widehat{\mu}-\mu)}{\widehat{\sigma}} \;.$$

Eine kurze Rechnung ergibt

$$T_n \;=\; \frac{\sqrt{n}\,\overline{Z}}{\sqrt{Y/(n-1)}} \;,$$

dabei setzen wir

$$\overline{Z} \;:=\; n^{-1}\sum_{i=1}^{n} Z_i \;, \quad Y \;:=\; \sum_{i=1}^{n}(Z_i-\overline{Z})^2 \;.$$

Insbesondere gehen μ und σ^2 nicht in T_n ein, so daß der Konstruktion eines Konfidenzintervalls grundsätzlich nichts im Wege steht. Unsere Verteilungsannahmen an $X_1, \ldots, X_n$ werden wichtig, wenn man c_α explizit bestimmen möchte.

Es stellt sich heraus, daß für unabhängige, standard normalverteilte $Z_1, \ldots, Z_n$ Zähler und Nenner von T_n unabhängige Zufallsvariable sind, deren

Dichten man explizit angeben kann. Zum Beweis wählen wir eine orthogonale Matrix O, deren erste Zeile der Vektor $(n^{-1/2}, \ldots, n^{-1/2})$ der Länge 1 ist. Indem wir wieder $\widetilde{Z} := O \cdot Z$ setzen, folgt

$$\overline{Z} \;=\; n^{-1/2}\widetilde{Z}_1$$

und unter Beachtung von $|\widetilde{Z}|^2 = |Z|^2$

$$Y \;=\; \sum_i Z_i^2 - n\overline{Z}^2 \;=\; \sum_i \widetilde{Z}_i^2 - \widetilde{Z}_1^2 \;=\; \widetilde{Z}_2^2 + \cdots + \widetilde{Z}_n^2 \,.$$

Nach Proposition 6.2 sind $\widetilde{Z}_1, \ldots, \widetilde{Z}_n$ unabhängig und standard normalverteilt, so daß auch Y und $\overline{Z}$ unabhängig sind. Der Erwartungswert von Y erweist sich als $(n-1)\mathbf{E}[\widetilde{Z}_1^2] = n-1$ (dies ist der Grund, wieso man Y bzw. $\widehat{\sigma}^2$ mit dem Faktor $(n-1)^{-1}$ normiert). Diese Resultate sind für uns Anlaß, folgende Verteilungen der Statistik einzuführen.

Definition.

i) Die reellwertige Zufallsvariable

$$Y \;:=\; Z_1^2 + \cdots + Z_n^2$$

heißt ***χ^2-verteilt mit n Freiheitsgraden****, falls $Z_1, \ldots, Z_n$ unabhängige, standard-normalverteile Zufallsvariable sind.*

ii) Die reellwertige Zufallsvariable

$$T \;:=\; \frac{W}{\sqrt{Y/n}}$$

heißt ***t-verteilt mit n Freiheitsgraden****, falls W standard normalverteilt und Y χ^2-verteilt mit n Freiheitsgraden ist und falls W und Y unabhängige Zufallsvariable sind.*

Zusammenfassend können wir feststellen, daß T_n eine t-verteilte Zufallsvariable mit $n-1$ Freiheitsgraden ist. Damit kann man c_α aus (6.2) erhalten (man vergleiche die Tabellen der t-Verteilung in Lehrbüchern der Statistik). Wir bestimmen nun noch die Dichten der χ^2- und t-Verteilungen (womit sich c_α durch numerische Integration berechnen läßt).

Ergänzungen.

1. **Die Dichte der χ^2-Verteilung.** Bei der χ^2-Verteilung handelt es sich um die Verteilung des quadrierten Abstands $|Z|^2 = Z_1^2 + \cdots + Z_n^2$ eines standard normalverteilten Zufallsvektors Z. Sei $K(r, dr)$ die infinitesimale Kugelschale um 0 mit innerem und äußerem Radius r und $r + dr$. Ihr Volumen ist proportional zu $r^{n-1}\,dr$. Da $\mathsf{n}(z)$ auf $K(r, dr)$ einen festen Wert proportional zu $\exp(-r^2/2)$ annimmt, gilt

$$\mathbf{Ws}\{|Z| \in (r, r + dr)\} \;=\; \mathbf{Ws}\{Z \in K(r, dr)\} \;=\; c\,\exp(-r^2/2) r^{n-1}\,dr$$

mit einer Normierungkonstante $c > 0$. Für $Y := |Z|^2$ folgt unter Beachtung von $\sqrt{y + dy} = \sqrt{y} + dy/2\sqrt{y}$ (vgl. (**??**))

$$\mathbf{Ws}\{Y \in (y, y + dy)\} \;=\; \mathbf{Ws}\{|Z| \in (\sqrt{y}, \sqrt{y} + dy/2\sqrt{y})\} \;=\; \mathsf{g}_n(y)\;dy$$

mit

$$\mathsf{g}_n(y) \;:=\; c_n\, y^{n/2-1} \exp\Big(-\frac{y}{2}\Big)\;.$$

Dies ist die **Dichte der χ^2-Verteilung mit n Freiheitsgraden**. Offenbar handelt es sich um eine spezielle Γ-Verteilung. Für die Normierungskonstante erhalten wir

$$c_n^{-1} \;=\; \int_0^\infty y^{\frac{n}{2}-1} e^{-\frac{y}{2}}\,dy \;=\; 2^{n/2}\Gamma(n/2)\;.$$

Die Dichte einer χ^2-Verteilung läßt sich auch mit der Faltungsformel (**??**) berechnen.

2. **Die Dichte der t-Verteilung.** Wir benutzen den Satz von der totalen Wahrscheinlichkeit in folgender infinitesimaler Version,

$$\begin{aligned}&\mathbf{Ws}\{T \in (x, x + dx)\}\\ &\qquad= \int_0^\infty \mathbf{Ws}\Big\{\frac{W}{\sqrt{Y/n}} \in (x, x + dx) \;\Big|\; Y = y\Big\}\mathsf{g}_n(y)\,dy\;,\end{aligned}$$

dabei sei $\mathsf{g}_n(y)$ wieder die Dichte der χ^2-Verteilung mit n Freiheitsgraden. Da W und Y unabhängig sind und W standard normalverteilt ist, gilt

$$\begin{aligned}&\mathbf{Ws}\Big\{\frac{W}{\sqrt{Y/n}} \in (x, x + dx) \;\Big|\; Y = y\Big\}\\ &\qquad= \mathbf{Ws}\{x\sqrt{y/n} < W < x\sqrt{y/n} + \sqrt{y/n}\,dx\}\\ &\qquad= (2\pi)^{-1/2} \exp\big(-x^2 y/2n\big)\sqrt{y/n}\,dx\;.\end{aligned}$$

Unter Beachtung der Formel für $\mathsf{g}_n(y)$ läßt sich nun das Integral ohne weiteres auswerten, indem man es durch die Substitution $z = y(1 + x^2/n)/2$ auf die Γ-Funktion zurückführt. Das Endresultat lautet

$$\mathbf{Ws}\{T \in (x, x + dx)\} \;=\; \mathsf{t}_n(x)\, dx$$

mit

$$\mathsf{t}_n(x) \;:=\; c'_n \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2} ,$$

der **Dichte der t-Verteilung mit n Freiheitsgraden**. Die Normierungskonstante ist

$$c'_n \;=\; \frac{\Gamma((n+1)/2)}{\sqrt{\pi n}\,\Gamma(n/2)} \,.$$

Mit wachsender Zahl der Freiheitsgrade konvergiert die t-Verteilung gegen die standard Normalverteilung. Dies kann man aus der Dichte erkennen, die im Limes proportional zu $\exp(-x^2/2)$ ist, oder direkt aus der Definition, denn nach dem Gesetz der großen Zahlen konvergiert

$$Y/n \;=\; n^{-1} \sum_{i=1}^{n} Z_i^2$$

für $n \to \infty$ f.s. gegen $\mathbf{E}[Z_1^2] = 1$. □

6.2 Die Varianzanalyse

Ein typischer Anwendungsfall einer Varianzanalyse sieht so aus.

Beispiel. Man möchte die Gerinnungszeiten der Blutproben von 24 Tieren vergleichen, die jeweils eine von 4 Diäten erhalten haben. Als Meßwerte liegen vor:

Diät	Gerinnungszeiten	Gruppenmittel
1	62, 60, 63, 59	61
2	63, 67, 71, 69, 65, 66	66
3	68, 66, 71, 67, 68, 68	68
4	56, 62, 60, 61, 63, 64, 63, 59	61

Die Gruppenmittel sind deutlich verschieden, ist dies aber bereits ein ausreichender Hinweis, daß die Gerinnungszeit signifikant von der Diät abhängt? □

Es geht also um eine statistische Fragestellung. Den Datensatz bezeichnen wir mit x, er ist Element einer Menge S, dem Beobachtungsraum. Im Beispiel ist x ein Tupel aus 24 reellen Zahlen und $S = \mathbb{R}^{24}$.

Eine statistische Analyse bedeutet ganz allgemein gesprochen, daß man die Daten x einem Gedankenexperiment unterwirft. Der Ausgangspunkt ist die Vorstellung, daß x Realisation einer S-wertigen Zufallsvariablen X ist (eine Annahme, die manchmal mehr, manchmal auch weniger gerechtfertigt erscheinen mag). In der Varianzanalyse nimmt man an, daß X ein normalverteilter Zufallsvektor ist. Wir betrachten den Fall einer **einfachen Varianzanalyse**, man stellt sich dann vor, daß der Datensatz in d Gruppen zerfällt und die Beobachtungswerte aus der i-ten Gruppe sich aus einem systematischen Anteil μ_i und einem zufälligen Anteil von unbekannter Varianz σ^2 (deren Größe nicht von Gruppe zu Gruppe variiere) zusammensetzen. X setzt sich also aus Komponenten X_{ij} zusammen, für die man den Ansatz

$$\begin{aligned} X_{1j} &= \mu_1 + \sigma Z_{1j} \qquad j = 1, \ldots, n_1 \\ X_{2j} &= \mu_2 + \sigma Z_{2j} \qquad j = 1, \ldots, n_2 \\ &\vdots \\ X_{dj} &= \mu_d + \sigma Z_{dj} \qquad j = 1, \ldots, n_d \end{aligned}$$

macht, mit reellen Zahlen μ_i, einem Skalenparameter σ und unabhängigen, standard normalverteilten Zufallsvariablen Z_{ij}. Vektoriell geschrieben bedeutet dies

$$X \;=\; \mu + \sigma Z \tag{6.3}$$

mit den Spaltenvektoren

$$\begin{aligned} X &:= (X_{11}, \ldots, X_{1n_1}, \ldots, X_{d1}, \ldots, X_{dn_d})^t \;, \\ \mu &:= (\; \mu_1 \;, \ldots, \;\; \mu_1 \;, \ldots, \; \mu_d \;, \ldots, \; \mu_d \;)^t \;, \\ Z &:= (\; Z_{11}, \ldots, \; Z_{1n_1}, \ldots, \; Z_{d1}, \ldots, \; Z_{dn_d})^t \end{aligned}$$

und dem Wertebereich $S = \mathbb{R}^n$ mit $n := n_1 + \ldots + n_d$. μ ist Element des d-dimensionalen linearen Teilraums

$$L \;:=\; \{(x_{ij}) \in S \,:\, x_{ij} = x_{ik} \text{ für alle } i, j, k\} \;,$$

bestehend aus den Datensätzen, deren Komponenten x_{ij} gruppenweise gleich sind.

Wir betrachtet nun die **Hypothese** H_0, daß es keine systematischen Unterschiede zwischen den verschiedenen Gruppen gibt,

$$H_0 : \; \mu_1 = \mu_2 = \cdots = \mu_d \;.$$

Anders ausgedrückt heißt dies

$$H_0 : \ \mu \ \in \ L^0 \tag{6.4}$$

mit dem 1-dimensionalen Teilraum

$$L^0 \ := \ \{(x_{ij}) \in S \ : \ x_{ij} = x_{kl} \text{ für alle } i, j, k, l\}$$

von L, bestehend aus denjenigen Datenvektoren x, deren Komponenten x_{ij} alle untereinander gleich sind. In dem Beispiel würde man diese Hypothese gern wiederlegen. Wie kann man vorgehen?

Bemerkung. Der Aufgabenstellung nach ist eine Varianzanalyse ein **statistischer Test**. Diesen Grundtypus eines statistischen Entscheidungsproblems wollen wir kurz umreißen. Zugrunde liegt ein statistisches Modell, das die Verteilung von X auf eine mehr oder weniger weite, für eine statistische Analyse geeignete Klasse von Verteilungen einschränkt. Anders ausgedrückt geht man davon aus, daß die Wahrscheinlichkeiten

$$\mathbf{Ws}_\theta\{X \in B\} \ ,$$

von einem Parameter $\theta \in \Theta$ abhängen. Θ heißt der Parameterraum. In unserem Beispiel gilt $\theta = (\mu_1, \ldots, \mu_d, \sigma^2)$ und $\Theta = \mathbb{R}^d \times \mathbb{R}_+$.

Bei einem Testproblem ist der Parameterraum in zwei disjunkte Teilmengen Θ_0 und Θ_1 zerlegt. Aus den Beobachtungswerten soll Aufschluß darüber gewonnen werden, ob die **Hypothese** $H_0 : \theta \in \Theta_0$ oder aber die **Alternative** $H_1 : \theta \in \Theta_1$ zutrifft, man spricht auch von der **Nullhypothese** und der **Gegenhypothese**. Günstig wäre es, wenn man den Beobachtungsraum S so in einen **Annahmebereich** S_0 und einen **Ablehnbereich** S_1 zerlegen könnte, daß die Zahlen

$$\begin{aligned} \beta_0 \ &:= \ \sup\{\mathbf{Ws}_\theta\{X \in S_1\} : \theta \in \Theta_0\} \\ \beta_1 \ &:= \ \sup\{\mathbf{Ws}_\theta\{X \in S_0\} : \theta \in \Theta_1\} \end{aligned}$$

beide klein sind. Dann verfügte man nämlich über ein statistisches Entscheidungsverfahren mit kleiner Irrtumswahrscheinlichkeit, gleichgültig, welchen Wert man für θ annimmt: *Entscheide bei Eintreten von $X \in S_0$ für die Nullhypothese und bei Eintreten von $X \in S_1$ für die Gegenhypothese* (die Irrtumswahrscheinlichkeit ist höchstens $\max(\beta_0, \beta_1)$).

Normalerweise gibt es jedoch keine Zerlegung $S = S_0 \cup S_1$ mit solch vorteilhaften Eigenschaften, es gelingt nicht, alle Irrtumswahrscheinlichkeiten gleichzeitig klein zu halten. Die übliche Strategie zur Konstruktion statistischer Testverfahren ist daher eine andere: Man gibt sich eine maximale

Irrtumswahrscheinlichkeit $\alpha > 0$ für einen **Fehler erster Art** vor, für eine Fehlentscheidung für die Gegenhypothese H_1, obwohl die Nullhypothese H_0 zutrifft. Es soll also $\beta_0 \leq \alpha$ gelten, man spricht vom **Signifikanzniveau** α des Tests (gängige Werte sind $\alpha = 0,05$ oder 0,01). Unter dieser Maßgabe kann man dann versuchen, den Annahmebereich S_0 und den Ablehnbereich S_1 so zu wählen, daß im Rahmen des Möglichen auch die Wahrscheinlichkeiten von **Fehlern zweiter Art** klein werden, von Entscheidungen für H_0, falls $\theta \in \Theta_1$ gilt. Offenbar entsteht eine Asymmetrie, die man bei der Testentscheidung berücksichtigen muß. Das Verfahren lautet nun so: *Bei Eintreten von $X \in S_1$ entscheide gegen die Nullhypothese H_0, im Fall $X \in S_0$ sehe aber von einer Entscheidung zwischen H_0 und H_1 ab.* Also: Nur wenn das Ereignis $\{X \in S_1\}$ eintritt, ist im Allgemeinen eine Entscheidung mit einer ausreichend kleinen Fehlerwahrscheinlichkeit möglich. Im Jargon der Statistik sagt man dann, daß *die Nullhypothese auf dem Signifikanzniveau α abgelehnt wird.* Dies führt dazu, daß bei statistischen Tests die Nullhypothese häufig die Rolle eines ‚Strohmanns' spielt: Man würde sie gern widerlegen. □

In der Varianzanalyse testet man die Nullhypothese $H_0 : \mu_1 = \cdots = \mu_d$ mit dem sogenannten F-Test. Vorbereitend wollen wir den Datenvektor in verschiedene Anteile zerlegen, die es erlauben, den systematischen und den zufälligen Anteil in den Daten voneinander zu trennen. Dazu betrachten wir für jedes $x = (x_{ij}) \in S$ das Gesamtmittel und die Gruppenmittel,

$$\overline{x}_{..} \;:=\; n^{-1} \sum_i \sum_j x_{ij} \qquad \text{und} \qquad \overline{x}_{i\cdot} \;:=\; n_i^{-1} \sum_j x_{ij} \;,$$

sowie die Teilräume von S

$$\begin{aligned} L^1 \;&:=\; \{(x_{ij}) \in S \,:\, x_{ij} = x_{ik} \text{ für alle } i,j,k \text{ und } \overline{x}_{..} = 0\} \;, \\ L^{res} \;&:=\; \{(x_{ij}) \in S \,:\, \overline{x}_{i\cdot} = 0 \text{ für alle } i\} \;. \end{aligned}$$

L^1 hat die Dimension $d-1$, er ist das orthogonale Komplement von L^0 in L. Für das Skalarprodukt von $(x_{ij}) \in L^0$ und $(y_{ij}) \in L^1$ gilt nämlich

$$\sum_{i,j} x_{ij} y_{ij} \;=\; \overline{x}_{..} \sum_{i,j} y_{ij} \;=\; \overline{x}_{..}\, n \overline{y}_{..} \;=\; 0 \;.$$

L^{res} hat die Dimension $n - d$, er ist das orthogonale Komplement von L in S, denn für das Skalarprodukt von $(x_{ij}) \in L$ und $(y_{ij}) \in L^{res}$ erhält man

$$\sum_{i,j} x_{ij} y_{ij} \;=\; \sum_i \overline{x}_{i\cdot} \sum_j y_{ij} \;=\; \sum_i \overline{x}_{i\cdot}\, n_i \overline{y}_{i\cdot} \;=\; 0 \;.$$

Insgesamt erhalten wir die Zerlegung

$$S \;=\; L^0 \oplus L^1 \oplus L^{res}$$

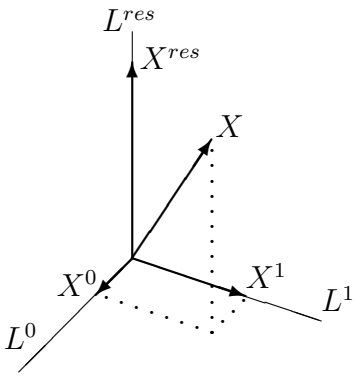
von S in orthogonale Teilräume. Dementsprechend läßt sich jedes $x \in S$ in die Anteile

$$x \;=\; x^0 + x^1 + x^{res}\;, \qquad x^0 \;\in\; L^0\;,\; x^1 \;\in\; L^1\;,\; x^{res} \;\in\; L^{res}$$

aufspalten. Die Komponenten berechnen sich als

$$x^0_{ij} \;:=\; \overline{x}_{\cdot\cdot}\;, \quad x^1_{ij} \;:=\; \overline{x}_{i\cdot} - \overline{x}_{\cdot\cdot}\;, \quad x^{res}_{ij} \;:=\; x_{ij} - \overline{x}_{i\cdot}\;,$$

man rechnet leicht nach, daß die durch diese Formeln gegebenen Vektoren in den entsprechenden Teilräumen liegen.



Auf unseren Modellansatz (6.3) angewendet erhalten wir die Zerlegung

$$X \;=\; X^0 + X^1 + X^{res}\;.$$

Der systematische Anteil μ, der sich entsprechend zerlegen läßt, schlägt sich in den drei Komponenten ganz verschieden nieder.

- Nach Annahme gilt $\mu \in L = L^0 \oplus L^1$ bzw. $\mu^{res} = 0$, deswegen geht μ in X^{res} überhaupt nicht ein. X^{res} ist also nur von den Zufallsstörungen Z abhängig und kann dafür benutzt werden, die Größe der Zufallsschwankungen abzuschätzen. X^{res} heißt der Vektor der **Residuen**, er ist ein Schätzer für σZ, denn $X_{ij} - \overline{X}_{i\cdot}$ läßt sich als Schätzer für $X_{ij} - \mu_i$ auffassen.

- Ist die Nullhypothese (6.4) gültig, so geht μ auch nicht in X^1 ein, unter der Gegenhypothese setzt sich dagegen X^1 aus einem systematischen und einem zufälligen Anteil zusammen. Dies bedeutet, daß unter der Gegenhypothese X^1 tendenziell größere Werte annimmt als unter der Nullhypothese.

- X^0 besteht in jedem Fall aus einem systematischen und einem zufälligen Anteil, die sich nicht voneinander trennen lassen, er ist für unsere Problemstellung deswegen nicht aussagekräftig.

Es liegt also nahe, den Zufallsvektor X^1 zur Testentscheidung zu benutzen und sich für die Gegenhypothese zu entscheiden, sofern seine Länge einen gewissen Wert überschreitet. Als Vergleichsgröße bietet sich die Länge von X^{res} an, denn dann hat die Größe von σ keinen Einfluß mehr auf die Testentscheidung. Dies motiviert, daß man bei einer einfachen Varianzanalyse den Test auf der Testgröße

$$F(X) \;:=\; \frac{(d-1)^{-1}|X^1|^2}{(n-d)^{-1}|X^{res}|^2} \;=\; \frac{(n-d)\sum_i n_i\big(\overline{X}_{i\cdot}-\overline{X}_{\cdot\cdot}\big)^2}{(d-1)\sum_{i,j}\big(X_{ij}-\overline{X}_{i\cdot}\big)^2}$$

aufbaut, der sogenannten ***F*-Statistik**. Um ihre Verteilung unter der Nullhypothese zu berechnen, wird nun die Annahme entscheidend, daß die Z_{ij} unabhängige normalverteilte Zufallsvariable sind.

Definition. *Seien Y_1 und Y_2 unabhängige, χ^2-verteilte Zufallsvariable mit den Freiheitsgraden m bzw. n. Man sagt dann, daß die Zufallsvariable*

$$F \;:=\; \frac{Y_1/m}{Y_2/n}$$

***F*-verteilt mit den Freiheitsgraden m, n** *(kurz $F(m,n)$-verteilt) ist.*

Proposition 6.3. *Unter der Nullhypothese (6.4) ist $F(X)$ eine F-verteilte Zufallsvariable mit den Freiheitsgraden $d-1, n-d$.*

Beweis. Wegen (6.4) gilt $\mu^1 = \mu^{res} = 0$ und $X^1 = \sigma Z^1$, $X^{res} = \sigma Z^{res}$. Sei $e_1, \ldots, e_n$ eine orthogonormale Basis von S mit $e_1 \in L^0$, $e_2, \ldots, e_d \in L^1$ und $e_{d+1}, \ldots, e_n \in L^{res}$. Bezeichnen $\widetilde{Z}_1, \ldots, \widetilde{Z}_n$ die Koordinaten von Z in dieser Basis, so folgt

$$\begin{aligned}\sigma^{-2}|X^1|^2 \;&=\; |Z^1|^2 \;=\; \widetilde{Z}_2^2 + \cdots + \widetilde{Z}_d^2\;,\\ \sigma^{-2}|X^{res}|^2 \;&=\; |Z^{res}|^2 \;=\; \widetilde{Z}_{d+1}^2 + \cdots + \widetilde{Z}_n^2\;.\end{aligned}$$

Nach Proposition 6.2 sind mit Z_{ij} auch $\widetilde{Z}_1, \ldots, \widetilde{Z}_n$ unabhängige, standard normalverteilte Zufallsvariable. Daher sind $\sigma^{-2}|X^1|^2$ und $\sigma^{-2}|X^{res}|^2$ unabhängig und χ^2-verteilt mit den Freiheitsgraden $d-1$ bzw. $n-d$. Dies ergibt die Behauptung. □

Die Zufallsvariable F ist so normiert, daß der Zähler Y_1/m und der Nenner Y_2/n Erwartungswert 1 haben. Unter der Nullhypothese ist daher damit zu rechnen, daß die F-Statistik einen Wert um 1 herum annimmt, sehr viel größere Werte sind ein Hinweis darauf, daß man die Nullhypothese verwerfen sollte. Der ***F*-Test** präzisiert dieses Vorgehen: Man bestimme zu vorgegebenem Signifikanzniveau $\alpha > 0$ eine positive Zahl c so, daß $\mathbf{Ws}\{F \le c\} = 1-\alpha$ gilt, mit $F(d-1, n-d)$-verteiltem F (c heißt das $(1-\alpha)$-Quantil der Verteilung, wichtige Werte sind in Statistikbüchern tabelliert). Wenn das Ereignis $F(X) > c$ eintritt, so verwirft man die Nullhypothese. Der Ablehnbereich hat hier also die Gestalt $S_1 = \{x \in S : F(x) > c\}$.

Diese Methode heißt Varianzanalyse, weil die F-Statistik durch Zerlegen (‚Analyse') der Varianz innerhalb der Beobachtungen entsteht: Wegen der Orthogonalität von X^1 und X^{res} gilt

$$|X^1 + X^{res}|^2 \;=\; |X^1|^2 + |X^{res}|^2$$

bzw.

$$\sum_{i,j}(X_{ij} - \overline{X}_{..})^2 \;=\; \sum_i n_i(\overline{X}_{i.} - \overline{X}_{..})^2 + \sum_{i,j}(X_{ij} - \overline{X}_{i.})^2 \;.$$

In Worten kann man das ausdrücken als

$$\begin{aligned}\text{empirische Varianz} \;=\;& \text{Varianz zwischen den Gruppen} \\ &+ \text{Varianz innerhalb der Gruppen} \;.\end{aligned}$$

Es ist üblich, das Resultat einer Varianzanalyse in einer ANOVA-Tabelle (‚Analysis of Variance') niederzulegen. Sie hat folgende Gestalt.

Quelle der Variabilität	Quadrat-summe	Freiheits-grad	mittlere Quadratsumme	F-Wert
zwischen den Gruppen	$\|X^1\|^2$	$d-1$	$\|X^1\|^2/(d-1)$	
innerhalb der Gruppen	$\|X^{res}\|^2$	$n-d$	$\|X^{res}\|^2/(n-d)$	$F(X)$

Beispiel. Für unser Ausgangsbeispiel sieht die ANOVA-Tabelle so aus.

Variabilität	QS	FG	mittlere QS	F-Wert
zwischen	228	3	76	
innerhalb	112	20	5,6	13,6

Das 99% Quantil der $F(3,20)$-Verteilung ist 4,94 und wird durch den F-Wert deutlich übertroffen. Man kann damit die Nullhypothese, daß die verschiedenen Diäten keinen Einfluß auf die Blutgerinnungszeiten haben, auf dem Signifikanzniveau von 0,01 verwerfen. □

Ergänzungen.

1. Man kann nach demselben Schema auch kompliziertere Modelle behandeln. In der **zweifachen Varianzanalyse** macht man für die Beobachtungsdaten den Ansatz

$$X_{ijk} \;=\; \mu_i + \nu_j + \sigma Z_{ijk}$$

mit reellen Zahlen μ_i, ν_j und unabhängigen standard normalverteilten Zufallsvariablen Z_{ijk}. Dahinter steckt die Vorstellung, daß zwei systematische Effekte sowie Meßfehler in die Beobachtungswerte eingehen. Man kann dann die Nullhypothese testen, daß alle μ_i (oder alle ν_j) gleich sind. Die Vorgehensweise ist analog zur einfachen Varianzanalyse.

2. Die Dichte einer $F(m,n)$-Verteilung ist gegeben durch

$$\mathbf{Ws}\{F \in (x, x+dx)\} \;=\; cx^{(m-2)/2}\Big(1+\frac{mx}{n}\Big)^{-(m+n)/2}\,dx$$

mit der Normierungskonstante

$$c \;:=\; \Big(\frac{m}{n}\Big)^{m/2}\frac{\Gamma((m+n)/2)}{\Gamma(m/2)\Gamma(n/2)}\;.$$

Sie läßt sich auf dieselbe Weise erhalten, wie wir im letzten Abschnitt die Dichte der t-Verteilung abgeleitet haben.

3. Den F-Test kann man auch anwenden, wenn die Annahme von normalverteilten Meßfehlern nicht mehr gewährleistet ist. Wie sich zeigt, hält der Test sein Signifikanzniveau auch unter anderen Verteilungsannahmen ein, vorausgesetzt, die Unabhängigkeit der Z_{ij} steht außer Zweifel. Allerdings kann dann das Verfahren deutlich an Trennschärfe verlieren. Man hat deswegen alternative Testverfahren für das Testen der Nullhypothese $\mu_1 = \cdots = \mu_d$ entworfen, die die Normalitätsannahme nicht mehr benötigen. Die modernen ‚Bootstrap-Methoden' machen dabei entscheidenden Gebrauch vom Computer.

6.3 Der zentrale Grenzwertsatz

Wir zeigen nun, daß Summen von unabhängigen reellwertigen Zufallsvariablen unter recht allgemeinen Bedingungen approximativ normalverteilt sind. Spezielle Annahmen über die Verteilung der einzelnen Summanden sind nicht nötig, sie brauchen nicht identisch verteilt zu sein und können sich mit wachsender Zahl der Summanden verändern. Um diesen Sachverhalt klar herauszuarbeiten, geht man von einem zeilenweise unabhängigen Dreiecksschema X_{ni}, $1 \le i \le n$, von reellwertigen Zufallsvariablen aus, also von Zufallsvariablen

$$\begin{array}{l} X_{11} \\ X_{21}, X_{22} \\ X_{31}, X_{32}, X_{33} \\ \quad \vdots \qquad\qquad \ddots \end{array} \tag{6.5}$$

von dem wir annehmen, daß für jedes n jeweils $X_{n1}, \ldots, X_{nn}$ stochastisch unabhängig sind. Wir wollen die asymptotische Verteilung von

$$S_n \;:=\; X_{n1} + \cdots + X_{nn}$$

betrachten. Um sie normieren zu können, betrachten wir auch die ‚gestutzte' Summe

$$\widetilde{S}_n \;:=\; X_{n1} I_{\{|X_{n1}| \le 1\}} + \cdots + X_{nn} I_{\{|X_{nn}| \le 1\}} \;.$$

$\widetilde{S}_n$ hat als beschränkte Zufallsvariable endliche Erwartung und Varianz, von S_n brauchen wir diesbezüglich nichts vorauszusetzen.

Es gilt dann folgender **allgemeiner zentraler Grenzwertsatz**.

Satz 6.4. *Sei (X_{ni}) ein zeilenweise unabhängiges Dreiecksschema von reellwertigen Zufallsvariablen, so daß*

i) $\mathbf{E}\widetilde{S}_n \;\to\; 0$, $\mathbf{Var}\widetilde{S}_n \;\to\; 1$

für $n \to \infty$. Gilt dann für alle $\epsilon > 0$ die Bedingung

ii) $\mathbf{Ws}\{\max_{i=1,\ldots,n} |X_{ni}| > \epsilon\} \;\to\; 0$

für $n \to \infty$, so ist S_n asymptotisch standard normalverteilt, d.h. für alle $-\infty \le a \le b \le \infty$ gilt

$$\mathbf{Ws}\{a \le S_n \le b\} \;\to\; \mathbf{Ws}\{a \le Z \le b\}$$

für $n \to \infty$. Dabei bezeichne Z eine reellwertige standard normalverteilte Zufallsvariable.

Beispiel. Binomialverteilung.

1. Seien $Y_1, Y_2, \ldots$ unabhängige, Bernoulli-verteilte Zufallsvariable mit Erfolgswahrscheinlichkeit p und $X_n := Y_1 + \cdots + Y_n$. Wir betrachten das Dreiecksschema $X_{ni} := (Y_i - p)/\sqrt{npq}$, $1 \le i \le n$. Offenbar gilt $|X_{ni}| \le 1$ für ausreichend großes n, und damit $\widetilde{S}_n = S_n$. Wegen $\mathbf{E}S_n = 0$, $\mathbf{Var}S_n = 1$ ist Bedingung *i)* erfüllt, und wegen $\max_i |X_{ni}| \le 1/\sqrt{npq}$ gilt *ii)*. Nach dem zentralen Grenzwertsatz ist also $S_n = (X_n - np)/\sqrt{npq}$ asymptotisch normalverteilt. Dies ist der Satz von de Moivre-Laplace für binomialverteilte Zufallsvariable X_n. Der allgemeinere Satz 1.3 folgt analog.

2. Sind dagegen $Y_1, \ldots, Y_n$ unabhängig und Bernoulli-verteilt mit Erfolgswahrscheinlichkeit p_n, so daß $np_n \to 1$ für $n \to \infty$, so ist X_n nach Satz 1.1 asymptotisch Poisson-verteilt. Der zentrale Grenzwertsatz kann nicht angewandt werden. Für $S_n := X_n - np_n$ gilt zwar $\mathbf{E}S_n = 0$ und $\mathbf{Var}S_n = np_nq_n \to 1$, jedoch ist für $X_{ni} := Y_i - p_n$ Bedingung *ii)* nicht erfüllt: Für $0 < \epsilon < 1$ und ausreichend großes n gilt $\mathbf{Ws}\{\max_{i=1,\ldots,n} |X_{ni}| > \epsilon\} = \mathbf{Ws}\{\max_{i=1,\ldots,n} Y_i \ge 1\} = \mathbf{Ws}\{X_n \ge 1\} = 1 - \mathbf{Ws}\{X_n = 0\}$, und dieser Ausdruck konvergiert nach der Poisson-Approximation gegen $1 - e^{-1}$. □

Den Beweis des zentralen Grenzwertsatzes führen wir am Ende des Abschnitts. Bedingung *i)* besagt, daß die Summen S_n geeignet normiert sind. Die entscheidende Annahme des Satzes ist also in Bedingung *ii)* enthalten, die man auch so ausdrücken kann, daß $\max_{i=1,\ldots,n} |X_{ni}|$ stochastisch gegen 0 konvergiert. Sie bedeutet, daß asymptotisch alle Summanden von S_n verschwinden und keiner auf S_n im Grenzwert einen bestimmenden Einfluß hat (man kann zeigen, daß sie sich substantiell nicht weiter abschwächen läßt). Kurz zusammengefaßt besagt der zentrale Grenzwertsatz also: *Setzt sich eine reellwertige Zufallsvariable aus vielen kleinen unabhängigen Summanden zusammen, so ist sie annähernd normal verteilt.*

Für ein Dreiecksschema mit unabhängigen Zeilen ist *ii)* äquivalent zu folgender Bedingung, die leichter nachprüfbar ist,

$$\textit{ii')} \quad \sum_{i=1}^{n} \mathbf{Ws}\{|X_{ni}| > \epsilon\} \;\to\; 0 \quad \textit{für alle } \epsilon > 0 \;.$$

Die Implikation *ii')* $\Rightarrow$ *ii)* ergibt sich aus der Ungleichung

$$\mathbf{Ws}\{\max_{i=1,\ldots,n} |X_{ni}| > \epsilon\} \;\le\; \sum_{i=1}^{n} \mathbf{Ws}\{|X_{ni}| > \epsilon\} \;.$$

Umgekehrt gilt wegen $1 - x \le \exp(-x)$ die Ungleichung

$$\begin{aligned}\mathbf{Ws}\{\max_{i=1,\dots,n} |X_{ni}| \le \epsilon\} &= \prod_{i=1}^{n} \mathbf{Ws}\{|X_{ni}| \le \epsilon\} \\ &\le \exp\Big(-\sum_{i=1}^{n} \mathbf{Ws}\{|X_{ni}| > \epsilon\}\Big) .\end{aligned}$$

Gilt *ii*), so konvergiert die linke Seite gegen 1, und damit auch die rechte Seite, was *ii'*) nach sich zieht. Damit gilt auch die Implikation *ii*) $\Rightarrow$ *ii'*).

Um festzustellen, ob S_n geeignet normiert ist, wird in Bedingung *i*) die gestutzte Summe $\widetilde{S}_n$ betrachtet. Dies hat nicht nur den Grund, daß damit für das Dreiecksschema (X_{ni}) keine Annahmen über Erwartungswerte und Varianzen erforderlich sind. Auch im Fall, daß alle X_{ni} endliche Erwartungswerte und Varianzen besitzen, sind $\mathbf{E}S_n$ und $\mathbf{Var}S_n$ nicht immer zur Normierung von S_n brauchbar. Dazu muß man zusätzlich Bedingung *ii*) durch eine stärkere Forderung ersetzen, die sich nicht mehr so einleuchtend interpretieren läßt. Dies ist der **zentrale Grenzwertsatz mit Lindeberg-Bedingung**.

Korollar 6.5. *Sei (X_{ni}) ein zeilenweise unabhängiges Dreiecksschema reellwertiger Zufallsvariabler mit endlichen Erwartungswerten und Varianzen, so daß*

iii) $\mathbf{E}S_n \to 0$, $\mathbf{Var}S_n \to 1$

für $n \to \infty$. Gilt dann für alle $\epsilon > 0$

iv) $\sum_{i=1}^{n} \mathbf{E}[X_{ni}^2\,;\, |X_{ni}| > \epsilon] \to 0$

für $n \to \infty$, so ist S_n asymptotisch standard normalverteilt.

Beweis. Wir zeigen, daß die Bedingungen von Satz 6.4 erfüllt sind. Aus *iv*) folgt

$$\begin{aligned}|\mathbf{E}\widetilde{S}_n - \mathbf{E}S_n| &= \Big|\sum_i \mathbf{E}[X_{ni} I_{\{|X_{ni}|>1\}}]\Big| \\ &\le \sum_i \mathbf{E}[X_{ni}^2\,;\, |X_{ni}| > 1] \to 0 , \\ \mathbf{E}[(\widetilde{S}_n - S_n)^2] &= \mathbf{Var}\Big[\sum_i X_{ni} I_{\{|X_{ni}|>1\}}\Big] + (\mathbf{E}\widetilde{S}_n - \mathbf{E}S_n)^2 \\ &\le \sum_i \mathbf{E}[X_{ni}^2\,;\, |X_{ni}| > 1] + (\mathbf{E}\widetilde{S}_n - \mathbf{E}S_n)^2 \to 0\end{aligned}$$

sowie wegen $\widetilde{S}_n^2 - S_n^2 = (\widetilde{S}_n - S_n)^2 + 2S_n(\widetilde{S}_n - S_n)$ unter Beachtung der Cauchy-Schwarz Ungleichung und *iii*)

$$\begin{aligned} |\mathbf{E}[\widetilde{S}_n^2] - \mathbf{E}[S_n^2]| &\le \mathbf{E}[(\widetilde{S}_n - S_n)^2] + 2\mathbf{E}[|S_n||\widetilde{S}_n - S_n|] \\ &\le \mathbf{E}[(\widetilde{S}_n - S_n)^2] + 2\big(\mathbf{E}[S_n^2] \cdot \mathbf{E}[(\widetilde{S}_n - S_n)^2]\big)^{1/2} \;\to\; 0\,. \end{aligned}$$

Wegen *iii*) ist deshalb *i*) erfüllt. *ii*) (bzw. *ii'*)) folgt aus *iv*) und der Markov-Ungleichung

$$\mathbf{Ws}\{|X_{ni}| > \epsilon\} \;\le\; \epsilon^{-2}\,\mathbf{E}[X_{ni}^2\,;\,|X_{ni}| > \epsilon]\,.$$

□

Beispiel: Identisch verteilte Summanden. Seien $X_1, X_2, \ldots$ unabhängige Kopien einer reellwertigen Zufallsvariablen X mit Erwartungswert 0 und endlicher Varianz σ^2, dann ist

$$S_n \;:=\; \frac{1}{\sqrt{\sigma^2 n}} \sum_{i=1}^{n} X_i$$

asymptotisch standard normalverteilt. Dies ist der klassische **zentrale Grenzwertsatz für identisch verteilte Zufallsvariable**. Er folgt aus dem Korollar, indem wir $X_{ni} = X_i/\sqrt{\sigma^2 n}$ setzen. S_n hat Erwartungswert 0 und Varianz 1, außerdem gilt

$$\sum_{i=1}^{n} \mathbf{E}[X_{ni}^2\,;\,|X_{ni}| > \epsilon] \;= \sigma^{-2}\mathbf{E}[X^2\,;\,|X| > \epsilon\sqrt{\sigma^2 n}]\,,$$

und dieser Ausdruck konvergiert gegen 0, da X endliche Varianz besitzt. □

Es folgt der Beweis des zentralen Grenzwertsatzes nach der Methode von Lindeberg-Breiman. Sie ist der elementar in dem Sinne, daß sie nur die einfachsten Eigenschaften von Erwartungswerten benutzt. Der Beweis hebt sich von anderen dadurch ab, daß er mit Zufallsvariablen arbeitet, und weniger mit Verteilungen. Die Grundidee ist einfach: Sind die X_{ni} normalverteilte Zufallsvariable, so ist auch S_n nach dem Beispiel aus Abschnitt 6.1 normalverteilt. Wir führen den allgemeinen Fall auf diesen Spezialfall zurück, indem wir zeigen, daß asymptotisch die Verteilung unverändert bleibt, sofern wir schrittweise die X_{ni} durch normalverteilte Zufallsvariable der gleichen Erwartung und Varianz ersetzen.

Beweis von Satz 6.4. Der Hauptschritt besteht im Nachweis der Aussage

$$\mathbf{E}[\phi(S_n)] \;\to\; \mathbf{E}[\phi(Z)]\;, \tag{6.6}$$

wobei $\phi : \mathbb{R} \to \mathbb{R}$ eine 3-mal stetig differenzierbare Funktion bezeichne, die samt ihren Ableitungen durch eine Konstante $\alpha > 0$ beschränkt sei. Zunächst gilt nach *ii*)

$$\begin{aligned} |\,\mathbf{E}[\phi(S_n)] - \mathbf{E}[\phi(\widetilde{S}_n)]\,| \;&=\; |\,\mathbf{E}[(\phi(S_n) - \phi(\widetilde{S}_n))I_{\{S_n \neq \widetilde{S}_n\}}]\,| \\ &\leq\; 2\alpha\;\mathbf{Ws}\{\max_i |X_{ni}| > 1\} \;\to\; 0\;, \end{aligned}$$

wir können also $\widetilde{S}_n$ anstelle von S_n betrachten bzw. ohne Einschränkung der Allgemeinheit annehmen, daß $|X_{ni}| \leq 1$ gilt.

Wir erweitern nun das Setting und ergänzen $X_{n1}, \ldots, X_{nn}$ durch unabhängige, standard normalverteilte Zufallsvariable $Z_{n1}, \ldots, Z_{nn}$ (auf formaler Ebene klärt die Maßtheorie, daß dies immer möglich ist). Sei

$$Y_{ni} \;:=\; \mu_{ni} + \sigma_{ni} Z_{ni}\;, \qquad \mu_{ni} \;:=\; \mathbf{E}X_{ni}\;,\; \sigma_{ni}^2 \;:=\; \mathbf{Var}X_{ni}\;,$$

und

$$S_{ni} \;:=\; Y_{n1} + \cdots + Y_{ni} + X_{n,i+1} + \cdots + X_{nn}\;.$$

Dann ist einerseits $S_{nn} = Y_{n1} + \cdots + Y_{nn}$ nach den Resultaten aus Abschnitt 6.1 eine normalverteilte Zufallsvariable mit Erwartungswert und Varianz

$$\mu_n \;:=\; \mu_{n1} + \cdots + \mu_{nn}\;, \quad \sigma_n^2 \;:=\; \sigma_{n1}^2 + \cdots + \sigma_{nn}^2\;,$$

es gilt also

$$\mathbf{E}[\phi(S_{nn})] \;=\; \mathbf{E}[\phi(\mu_n + \sigma_n Z)]\;.$$

Nach *i*) gilt $\mu_n \to 0$ und $\sigma_n^2 \to 1$, aus der Stetigkeit von $\phi(x)$ folgt daher (nach bekannten Sätzen der Integrationstheorie)

$$\mathbf{E}[\phi(S_{nn})] \;=\; \int_{-\infty}^{\infty} \phi(\mu_n + \sigma_n z)(2\pi)^{-1/2} \exp(-z^2/2)\, dz \;\to\; \mathbf{E}[\phi(Z)]\;.$$

Andererseits gilt $S_{n0} = S_n$, daher bleibt

$$\mathbf{E}[\phi(S_{nn})] - \mathbf{E}[\phi(S_{n0})] \;\to\; 0$$

zu beweisen. Dazu benutzen wir die Abschätzung

$$\begin{aligned} |\mathbf{E}[\phi(S_{nn})] - \mathbf{E}[\phi(S_{n0})]| \;&=\; |\sum_{i=1}^{n} \mathbf{E}[\phi(S_{ni})] - \mathbf{E}[\phi(S_{n,i-1})]| \\ &\leq\; \sum_{i=1}^{n} |\mathbf{E}[\phi(S_{ni})] - \mathbf{E}[\phi(S_{n,i-1})]|\;. \end{aligned}$$

Um die einzelnen Summanden abzuschätzen, entwickeln wir $\phi(S_{n,i-1})$ und $\phi(S_{ni})$ um

$$T_{ni} \;:=\; Y_{n1} + \cdots + Y_{n,i-1} + \mu_{ni} + X_{n,i+1} + \cdots + X_{nn}$$

herum gemäß der Taylor-Formel,

$$\begin{aligned}
\phi(S_{n,i-1}) \;&=\; \phi(T_{ni}) + \phi'(T_{ni})(X_{ni}-\mu_{ni}) + \frac{1}{2}\phi''(T_{ni})(X_{ni}-\mu_{ni})^2 \\
&\quad + \frac{1}{6}\phi'''(U_{ni})(X_{ni}-\mu_{ni})^3 \\
\phi(S_{ni}) \;&=\; \phi(T_{ni}) + \phi'(T_{ni})(Y_{ni}-\mu_{ni}) + \frac{1}{2}\phi''(T_{ni})(Y_{ni}-\mu_{ni})^2 \\
&\quad + \frac{1}{6}\phi'''(V_{ni})(Y_{ni}-\mu_{ni})^3
\end{aligned}$$

mit geeigneten Zufallsvariablen U_{ni}, V_{ni}. Nach Konstruktion sind X_{ni}, Y_{ni} und T_{ni} unabhängig, daher gilt nach Satz 3.4

$$\begin{aligned}
\mathbf{E}[\phi'(T_{ni})(X_{ni}-\mu_{ni})] \;&=\; \mathbf{E}[\phi'(T_{ni})(Y_{ni}-\mu_{ni})] \;=\; 0\;, \\
\mathbf{E}[\phi''(T_{ni})(X_{ni}-\mu_{ni})^2] \;&=\; \mathbf{E}[\phi''(T_{ni})(Y_{ni}-\mu_{ni})^2] \;=\; \sigma_{ni}^2\mathbf{E}[\phi''(T_{ni})]\;,
\end{aligned}$$

außerdem gilt

$$\begin{aligned}
|\mathbf{E}[\phi'''(U_{ni})(X_{ni}-\mu_{ni})^3]| \;&\le\; \alpha\mathbf{E}[|X_{ni}-\mu_{ni}|^3]\;, \\
|\mathbf{E}[\phi'''(V_{ni})(Y_{ni}-\mu_{ni})^3]| \;&\le\; \alpha\mathbf{E}[|Y_{ni}-\mu_{ni}|^3]\;.
\end{aligned}$$

Indem wir in den Taylor-Entwicklungen zum Erwartungswert übergehen, folgt

$$|\mathbf{E}[\phi(S_{ni})] - \mathbf{E}[\phi(S_{n,i-1})]| \;\le\; \alpha(\mathbf{E}[|X_{ni}-\mu_{ni}|^3] + \mathbf{E}[|Y_{ni}-\mu_{ni}|^3)]\;.$$

Mit $|X_{ni}| \le 1$ gilt auch

$$|X_{ni}-\mu_{ni}|^3 \le (\epsilon + |\mu_{ni}|)(X_{ni}-\mu_{ni})^2 + (1+|\mu_{ni}|)^3 I_{\{|X_{ni}|>\epsilon\}}\;.$$

Nach Satz 3.1 folgt

$$\mathbf{E}[|X_{ni}-\mu_{ni}|^3] \;\le\; (\epsilon + |\mu_{ni}|)\sigma_{ni}^2 + (1+|\mu_{ni}|)^3\mathbf{Ws}\{|X_{ni}| > \epsilon\}\;,$$

außerdem gilt

$$\mathbf{E}[|Y_{ni}-\mu_{ni}|^3] \;=\; \sigma_{ni}^3\mathbf{E}[|Z|^3]\;.$$

Insgesamt können wir unsere Aussagen in der Abschätzung

$$|\mathbf{E}[\phi(S_{nn})] - \mathbf{E}[\phi(S_{n0})]| \le \alpha\Big((\epsilon + \max_i |\mu_{ni}|)\sum_{i=1}^n \sigma_{ni}^2 + (1+\max_i|\mu_{ni}|)^3 \sum_{i=1}^n \mathbf{Ws}\{|X_{ni}|>\epsilon\} + \mathbf{E}[|Z|^3]\max_i \sigma_{ni}\sum_{i=1}^n \sigma_{ni}^2\Big)$$

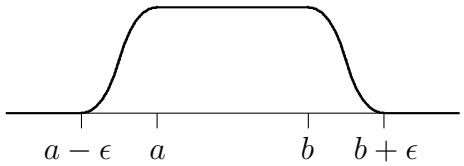
zusammensetzen. Diese Abschätzung gilt für alle $\epsilon > 0$, außerdem gilt $\sum_i \sigma_{ni}^2 = \mathbf{Var}S_n$. Um also $\mathbf{E}[\phi(S_{nn})] - \mathbf{E}[\phi(S_{n0})] \to 0$ zu erhalten, langt es in Anbetracht von *i*) und *ii'*) zu zeigen, daß $\max_i |\mu_{ni}|$ und $\max_i \sigma_{ni}$ gegen 0 konvergieren. Dies folgt aus *ii'*): Es gilt $|X_{ni}| \le \epsilon + I_{\{|X_{ni}|>\epsilon\}}$, deswegen folgt nach Satz 3.1 die für alle $\epsilon > 0$ gültigen, in i gleichmäßigen Abschätzungen

$$|\mu_{ni}| \;\le\; \epsilon + \mathbf{Ws}\{|X_{ni}|>\epsilon\} \;\le\; \epsilon + \sum_{j=1}^n \mathbf{Ws}\{|X_{nj}|>\epsilon\}\ ,$$

und ähnlich folgt aus $(X_{ni}-\mu_{ni})^2 \le (\epsilon+|\mu_{ni}|)^2 + (1+|\mu_{ni}|)^2 I_{\{|X_{ni}|>\epsilon\}}$ die Abschätzung

$$\begin{aligned}\sigma_{ni}^2 &\le (\epsilon+|\mu_{ni}|)^2 + (1+|\mu_{ni}|)^2 \mathbf{Ws}\{|X_{ni}|>\epsilon\} \\ &\le (\epsilon + \max_j|\mu_{nj}|)^2 + (1+\max_j|\mu_{nj}|)^2 \sum_{j=1}^n \mathbf{Ws}\{|X_{nj}|>\epsilon\}\ .\end{aligned}$$

(6.6) ist damit bewiesen. Die Aussage des Satzes folgt nun, indem wir Wahrscheinlichkeiten durch Erwartungswerte approximieren. Sei $a \le b$. Zu vorgegebenem $\epsilon > 0$ wähle man eine 3-mal stetig differenzierbare Funktion $0 \le \phi(x) \le 1$ so daß $\phi(x)$ auf $[a,b]$ den Wert 1 und außerhalb von $(a-\epsilon, b+\epsilon)$ den Wert 0 annimmt.



Wegen der Monotonie von Erwartungswerten gilt $\mathbf{Ws}\{a \le S_n \le b\} \le \mathbf{E}[\phi(S_n)]$ und $\mathbf{E}[\phi(Z)] \le \mathbf{Ws}\{a-\epsilon \le Z \le b+\epsilon)$, und es folgt

$$\begin{aligned}\limsup_{n\to\infty}\ \mathbf{Ws}\{a \le S_n \le b\} &\le \lim_{n\to\infty} \mathbf{E}[\phi(S_n)] \\ &= \mathbf{E}[\phi(Z)] \;\le\; \mathbf{Ws}\{a-\epsilon \le Z \le b+\epsilon\}\ .\end{aligned}$$

Da Z eine Dichte hat, ist der rechte Ausdruck stetig in ϵ. Mit $\epsilon \to 0$ folgt daher

$$\limsup_{n\to\infty} \mathbf{Ws}\{a \le S_n \le b\} \;\le\; \mathbf{Ws}\{a \le Z \le b\} \;.$$

Ähnlich zeigt man

$$\liminf_{n\to\infty} \mathbf{Ws}\{a \le S_n \le b\} \;\ge\; \mathbf{Ws}\{a \le Z \le b\} \;,$$

mit Hilfe von Funktionen $\phi(x)$, die auf $[a+\epsilon, b-\epsilon]$ den Wert 1 und außerhalb von (a,b) den Wert 0 annehmen. Dies ergibt die Behauptung des Satzes. Die Fälle $a = -\infty$ und $b = \infty$ werden analog behandelt. □

6.4 Gauß-Vektoren

Wir untersuchen nun die Verteilung des Bildes $X = \phi(Z)$ eines standard normalverteilten Zufallsvektors Z unter einer affinen Abbildung $\phi : \mathbb{R}^n \to \mathbb{R}^m$. Dazu benötigen wir den Begriff der Covarianzmatrix.

Definition. *Sei $X = (X_1, \ldots, X_m)^t$ ein Zufallsvektor, dessen reellwertige Komponenten endliche Varianz haben. Dann heißt die $m \times m$ - Matrix*

$$\mathbf{Cov}[X] \;:=\; \big(\mathbf{Cov}[X_i, X_j]\big)_{i,j}$$

*die **Covarianzmatrix** von X.*

Covarianzmatrizen sind symmetrisch, außerdem gilt für beliebige Vektoren $\lambda = (\lambda_1, \ldots, \lambda_m)$

$$\lambda \cdot \mathbf{Cov}[X] \cdot \lambda^t \;=\; \sum_i \sum_j \lambda_i \,\mathbf{Cov}[X_i, X_j]\, \lambda_j \;=\; \mathbf{Var}\big[\sum_i \lambda_i X_i\big] \;\ge\; 0 \;.$$

Symmetrische Matrizen mit dieser Eigenschaft nennt man bekanntlich nicht-negativ definit. Es ist nicht schwer zu zeigen, daß jede nicht-negativ definite Matrix Covarianzmatrix eines geeigneten Zufallsvektors ist.

Sei nun Z ein Zufallsvektor mit Werten in $\mathbb{R}^n$ und

$$X \;:=\; \phi(Z) \;=\; A \cdot Z + \mu$$

mit einer beliebigen reellen $m \times n$ - Matrix $A = (a_{ij})$ und einem Vektor $\mu = (\mu_1, \ldots, \mu_m)^t$ reeller Zahlen, also

$$X_i \;:=\; \sum_{k=1}^{n} a_{ik} Z_k + \mu_i \;, \quad i = 1, \ldots, m \;.$$

Dann gilt

$$\mathbf{Cov}[X_i, X_j] \;=\; \sum_k \sum_l a_{ik} a_{jl} \mathbf{Cov}[Z_k, Z_l] \;=\; \sum_k \sum_l a_{ik} \mathbf{Cov}[Z_k, Z_l] a^t_{lj}$$

bzw. in der Notation der Matrizenrechnung

$$\mathbf{Cov}[X] \;=\; A \cdot \mathbf{Cov}[Z] \cdot A^t \;,$$

wobei A^t die transponierte Matrix von A mit den Einträgen $a^t_{ji} := a_{ij}$ bezeichne.

Handelt es sich bei den $Z_1, \ldots, Z_n$ speziell um unkorrelierte Zufallsvariablen der Varianz 1, so ist $\mathbf{Cov}[Z]$ die Einheitsmatrix, und es folgt

$$\mathbf{Cov}[X] \;=\; A \cdot A^t \;. \tag{6.7}$$

Dies gilt insbesondere, wenn $Z = (Z_1, \ldots, Z_n)^t$ standard normalverteilt ist. Unter dieser Voraussetzung bestimmen wir nun die Verteilung von $X = \phi(Z)$, und zwar zunächst im Fall, daß A eine invertierbare $m \times m$ - Matrix ist. Die Umkehrabbildung von ϕ ist dann $\psi(x) := A^{-1} \cdot (x - \mu)$. Wegen $Z = \psi(X)$ besteht zwischen den Dichten von X und Z die Beziehung

$$\mathbf{Ws}\{X \in dv\} \;=\; \mathbf{Ws}\{Z \in \psi(dv)\}$$

mit einem infinitesimalen Volumenelement dv. Unter der affinen Abbildung ψ werden Volumina um den Faktor $\det(A^{-1}) = (\det A)^{-1}$ gestreckt. Ist also dx das infinitesimale Volumen von dv, so hat $\psi(dv)$ das infinitesimale Volumen $dx/\det A$, und wir gelangen zu der Gleichung

$$\mathbf{Ws}\{X \in dv\} \;=\; (\det A)^{-1} \mathsf{n}(\psi(x))\, dx \;.$$

Wir zeigen, daß dieser Ausdruck von A nur über die Kovarianzmatrix

$$\Sigma \;:=\; \mathbf{Cov}[X] \;=\; A \cdot A^t$$

abhängt. $\psi(x)$ ist Spaltenvektor, daher gilt

$$\begin{aligned} |\psi(x)|^2 \;&=\; \psi(x)^t \cdot \psi(x) \\ &=\; (x-\mu)^t \cdot (A^{-1})^t \cdot A^{-1} \cdot (x-\mu) \;=\; (x-\mu)^t \cdot \Sigma^{-1} \cdot (x-\mu) \;, \end{aligned}$$

mit der Inversen Σ^{-1} der Kovarianzmatrix, außerdem folgt

$$\det(\Sigma) \;=\; \det(A) \cdot \det(A^t) \;=\; \det(A)^2 \;.$$

Insgesamt erhalten wir

$$\mathbf{Ws}\{X \in dv\} \;=\; \mathsf{n}_{\mu,\Sigma}(x)\,dx$$

mit

$$\mathsf{n}_{\mu,\Sigma}(x) \;:=\; (2\pi)^{-m/2}\det(\Sigma)^{-1/2}\,\exp\big(-\tfrac{1}{2}(x-\mu)^t\cdot\Sigma^{-1}\cdot(x-\mu)\big)\;.$$

Dies ist die **Dichte der multivariaten Normalverteilung mit Erwartung $\boldsymbol{\mu=(\mu_1,\ldots,\mu_m)^t}$ und Kovarianzmatrix $\boldsymbol{\Sigma}$**.

Die Annahme, daß ϕ bijektiv ist, läßt sich abschwächen. Wir zeigen nun, daß die Formel für die Dichte von X unter der schwächeren Bedingung bestehen bleibt, daß $\phi : \mathbb{R}^n \to \mathbb{R}^m$ surjektiv ist. Dann hat der Kern von A die Dimension $n-m$, und es gibt eine orthogonale Matrix O, die die Vektoren $(z_1,\ldots,z_n)^t$ aus dem Kern genau in die Vektoren der Gestalt $(0,\ldots,0,\widetilde{z}_{m+1},\ldots,\widetilde{z}_n)^t$ transformiert. Mit Z ist nach Proposition 6.2 auch $\widetilde{Z} := O\cdot Z$ standard normalverteilt. Nach Konstruktion ist $O^{-1}\cdot(0,\ldots,0,\widetilde{Z}_{m+1},\ldots,\widetilde{Z}_n)^t$ ein Zufallsvektor mit Werten im Kern von A, deshalb folgt

$$X \;=\; A\cdot Z+\mu \;=\; A\cdot O^{-1}\cdot\widetilde{Z}+\mu \;=\; A'\cdot(\widetilde{Z}_1,\ldots,\widetilde{Z}_m)^t+\mu$$

mit der $m\times m$ - Matrix A', die aus $A\cdot O^{-1}$ durch Streichen der letzten $n-m$ Spalten entsteht. Ihr Kern ist der Nullraum, daher ist es uns gelungen, den Fall einer surjektiven Abbildung ϕ zurückzuführen auf den Fall, daß ϕ bijektiv ist.

Für nicht-surjektives ϕ liegt der Fall etwas anders. Dann hat mit A auch $\Sigma = A^t\cdot A$ einen Rang kleiner als m, so daß die Inverse von Σ nicht mehr existiert. In diesem Fall hat die Verteilung von X keine Dichte auf dem $\mathbb{R}^m$. Indem wir jedoch $\mathbb{R}^m$ auf den Bildraum von ϕ einschränken, können wir unsere bisherigen Resultate ohne weiteres auf beliebige affine Abbildungen $\phi : \mathbb{R}^n \to \mathbb{R}^m$ übertragen. Insbesondere bleibt Sachverhalt richtig, daß die Verteilung nur von μ und Σ abhängt.

Wir fassen unsere Diskussion in der folgenden Definition zusammen.

Definition. *Ein $\mathbb{R}^m$-wertiger Zufallsvektor $X=(X_1,\ldots,X_m)^t$ mit endlicher Erwartung $\mu=(\mathbf{E}X_1,\ldots,\mathbf{E}X_m)^t$ und Kovarianzmatrix Σ heißt* ***multivariat normalverteilt****, kurz* ***$\boldsymbol{N(\mu,\Sigma)}$-verteilt****, falls es eine affine Abbildung $\phi : \mathbb{R}^n \to \mathbb{R}^m$ und einen standard normalverteilten Zufallsvektor Z mit Werten im $\mathbb{R}^n$ gibt, so daß $X=\phi(Z)$ gilt.*

Zufallsvektoren mit einer multivariaten Normalverteilung nennt man auch **Gaußsche Zufallsvektoren**. Da die Verteilung durch die Erwartungswerte,

Varianzen und Covarianzen ihrer Komponenten bestimmt ist, haben sie übersichtliche Eigenschaften. Besonders wichtig ist, daß man bei Gauß-Vektoren nicht länger zwischen Unkorreliertheit und Unabhängigkeit der Komponenten zu unterscheiden braucht (ein Sachverhalt, der sonst für Zufallsvektoren nicht stimmt).

Proposition 6.6. *Sei $(X_1, \ldots, X_m)^t$ ein Gaußscher Zufallsvektor mit unkorrelierten Komponenten. Dann sind $X_1, \ldots, X_m$ unabhängige Zufallsvariable.*

Beweis. Seien $\sigma_1^2, \ldots, \sigma_m^2$ die Varianzen von $X_1, \ldots, X_m$. Im unkorrelierten Fall gilt $\det(\Sigma) = \sigma_1^2 \cdots \sigma_m^2$, und Σ^{-1} ist eine Diagonalmatrix mit den Diagonalelementen $\sigma_1^{-2}, \ldots, \sigma_m^{-2}$. Es folgt

$$\mathsf{n}_{\mu,\Sigma}(x_1, \ldots, x_m) \;=\; \prod_{i=1}^{m} \big(2\pi\sigma_i^2\big)^{-1/2} \exp\Big(-\frac{1}{2\sigma_i^2}(x_i - \mu_i)^2\Big)\ .$$

Nach Proposition 2.7 sind daher $X_1, \ldots, X_m$ unabhängige, normalverteilte Zufallsvariable. □

Bemerkung. Zwei unkorrelierte normalverteilte Zufallsvariable X und Y brauchen nicht unabhängig zu sein. Ein Beispiel: Sei X standard normalverteilt und $Y := V \cdot X$, wobei V von X unabhängig sei und die Werte ± 1 jeweils mit Wahrscheinlichkeit $1/2$ annehme. Dann ist offenbar auch Y standard normalverteilt, und es gilt $\mathbf{E}[XY] = \mathbf{E}[X^2]\mathbf{E}[V] = 0$ und folglich $\mathbf{Cov}[X, Y] = 0$. Andererseits sind X und Y nicht unabhängig, denn $\mathbf{Ws}\{X > x, Y > x\} = \frac{1}{2}\mathbf{Ws}\{X > x\} > \mathbf{Ws}\{X > x\}\mathbf{Ws}\{Y > x\}$ für $x > 0$. Zur Proposition besteht kein Wiederspruch, denn die gemeinsame Verteilung von X, Y ist nicht multivariat normalverteilt. (X, Y) nimmt nur Werte auf den Diagonalen $x = \pm y$ in $\mathbb{R}^2$ an und hat infolgedessen keine Dichte. □

Kapitel 7

Entropie und Information

Die Entropie wird als Maßzahl für den **Grad der Ungewißheit** über den Ausgang eines Zufallsexperimentes benutzt, positiv ausgedrückt für seinen **Informationsgehalt**. Information ist hier nicht in einem inhaltlichen, sondern einem statistischen Sinn gemeint: Führt man ein Zufallsexperiment mit Erfolgswahrscheinlichkeit p durch, so erfährt man wenig, wenn p nahe bei 0 oder 1 liegt, denn dann ist man sich über den Versuchsausgang schon von vornherein relativ sicher. So gesehen ist der Fall $p = 1/2$ am informativsten. In diesem Kapitel leiten wir ein paar Resultate ab, die dieser Vorstellung eine handfeste mathematische Bedeutung geben. Danach behandeln wir das Hauptresultat der Informationstheorie über die Transmission von Nachrichten durch einen gestörten Kanal.

7.1 Die Entropie

Wir betrachten zunächst Laplace-Experimente. Aus einer endlichen, m Elemente umfassenden Menge S wird rein zufällig ein Element ausgewählt. Diesem Experiment ordnet man als Entropie die Zahl $\log m$ zu (Logarithmus zur Basis 2). Zur Begründung kann man anführen, daß man, wenn man es geschickt anstellt, ungefähr $\log m$ Ja-Nein-Fragen braucht, um das Resultat des Experimentes zu erfragen. Man zerlegt dazu S in zwei möglichst gleichgroße Teilmengen S_1 und S_2 und fragt, in welcher Teilmenge das Element liegt. Mit dieser Teilmenge verfährt man genauso, bis das Element gefunden ist. Ist m eine Potenz von 2, so lassen sich die Mengen immer exakt halbieren, und man braucht genau $\log m$ Fragen. Andernfalls bricht die Prozedur nach $[\log m]$ oder $[\log m] - 1$ Schritten ab.

Wir benutzen dieses Szenario, um auch für den nicht-uniformen Fall einen Ausdruck für die Entropie abzuleiten. Sei dazu $S_1, \ldots, S_k$ eine Zerlegung von

S in disjunkte Teilmengen der Mächtigkeit $m_1, \ldots, m_k$ ($m = m_1 + \cdots + m_k$). Dann läßt sich die rein zufällige Wahl eines Elements aus S in dem folgenden 2-stufigen Experiment realisieren:

1. Wähle ein zufälliges Element X aus $\{1, 2, \ldots, k\}$, und zwar x mit Wahrscheinlichkeit $p_x = m_x/m$.

2. Hat X den Wert x angenommen, so ziehe rein zufällig ein Element aus S_x.

Wir argumentieren nun, daß der Grad der Ungewißheit über den Ausgang des Gesamtexperimentes gleich der Summe der entsprechenden Größen beider Teilexperimente ist. Die Entropie des Gesamtexperiments ist $\log m$, ferner ist die Entropie des Nachfolgeexperiments im Mittel $\sum_x p_x \log m_x$. Für das Vorexperiment bleibt

$$\log m - \sum_{x=1}^{k} p_x \log m_x \;=\; -\sum_{x=1}^{k} p_x \log p_x \;.$$

Wir vereinbaren also folgende Sprechweise.

Definition. *Sei $\mu = (p_x)_{x\in S}$ eine Wahrscheinlichkeitsverteilung auf der abzählbaren Menge S. Die **Entropie** von μ ist definiert als*

$$H(\mu) \;:=\; -\sum_{x\in S} p_x \log_2 p_x$$

(mit $0 \log 0 := 0$). Ist μ die Verteilung einer S-wertigen Zufallsvariablen X, so spricht man auch von der Entropie $H(X)$ von X (bzw. von $H(X_1, \ldots, X_n)$, wenn X die Produktvariable $(X_1, \ldots, X_n)$ ist).

Die Entropie besitzt verschiedene für eine Maßzahl des Informationsgehalt wünschenswerte Eigenschaften.

1. Da $p \log p \le 0$ für $0 \le p \le 1$, ist die Entropie nichtnegativ,

$$H(\mu) \;\ge\; 0. \tag{7.1}$$

2. Seien X und Y diskrete Zufallsvariable mit Wertebereichen S und S'. Die **bedingte Entropie** von X bzgl. Y ist definiert als

$$H(X|Y) \;:=\; \sum_{y\in S'} H(X|Y=y) \cdot \mathbf{Ws}\{Y=y\},$$

mit

$$H(X|Y=y) \;:=\; -\sum_{x\in S} \mathbf{Ws}\{X=x \,\big|\, Y=y) \cdot \log \mathbf{Ws}\{X=x \,\big|\, Y=y\}\;.$$

Nach unserer Deutung der Entropie ist $H(X|Y=y)$ der Grad an Ungewißheit über den Wert von X, wenn man schon weiß, daß Y den Wert y angenommen hat. $H(X|Y)$ ist nach Art eines Erwartungswertes gebildet und daher als der mittlere Grad an Ungewißheit aufzufassen, der über den Wert von X bestehen bleibt, wenn man die Möglichkeit hat, Y zu beobachten. Dementsprechend gilt die Gleichung

$$H(X,Y) \;=\; H(Y) + H(X|Y)\;, \tag{7.2}$$

deren Beweis sich unmittelbar aus der Definition von bedingten Wahrscheinlichkeiten ergibt,

$$\begin{aligned}
&-\sum_{x,y} \mathbf{Ws}\{X=x, Y=y\} \cdot \log \mathbf{Ws}\{X=x, Y=y\} \\
&\quad= \; -\sum_{x,y} \mathbf{Ws}\{X=x, Y=y\} \cdot \log \mathbf{Ws}\{Y=y\} \\
&\qquad\quad -\sum_{x,y} \mathbf{Ws}\{Y=y\}\mathbf{Ws}\{X=x \,\big|\, Y=y\} \cdot \log \mathbf{Ws}\{X=x \,\big|\, Y=y\} \\
&\quad= \; H(Y) + H(X|Y)\;.
\end{aligned}$$

3. Gilt $X=\phi(Y)$ für Zufallsvariable X und Y und eine Abbildung ϕ, so folgt

$$H(X) \;\le\; H(Y)\;. \tag{7.3}$$

Dann gilt nämlich $\mathbf{Ws}\{X=x \,\big|\, Y=y\} = 1$ oder 0 (je nachdem, ob die Gleichung $x=\phi(y)$ erfüllt ist oder nicht) und damit $H(X|Y)=0$. Andererseits gilt nach (7.1) $H(Y|X)\ge 0$, und nach (7.2) folgt

$$H(X) \;\le\; H(X) + H(Y|X) \;=\; H(X,Y) \;=\; H(Y) + H(X|Y) \;=\; H(Y)\,.$$

4. Eine wichtige Rolle spielt die Größe

$$D(\mu\|\nu) \;:=\; \sum_x p_x \log_2 \frac{p_x}{q_x}$$

(summiert wird über alle x mit $p_x>0$) für Wahrscheinlichkeitsverteilungen $\mu=(p_x)$ und $\nu=(q_x)$ auf S. Man spricht von der **relativen Entropie** oder **Kullback-Leibler-Information** von μ bzgl ν.

Beispiel. Sind μ und ν binomial verteilt zum Parameter (n,t) bzw. (n,p), so gilt

$$\log\frac{p_x}{q_x} \;=\; x\log\frac{t}{p} + (n-x)\log\frac{1-t}{1-p}$$

und folglich

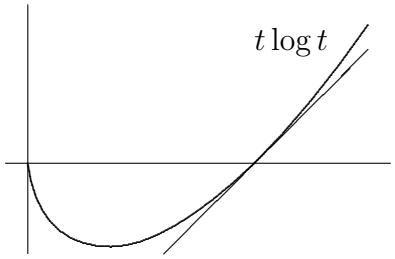
$$D(\mu\|\nu) \;=\; n\cdot\Big(t\log\frac{t}{p} + (1-t)\log\frac{1-t}{1-p}\Big)\;.$$

Die relative Entropie stimmt mit der bereits betrachteten Entropiefunktion der Binomialverteilung überein (vgl. (1.7)).

Behauptung. Es gelten die Aussagen

$$\begin{aligned} D(\mu\|\nu) &\;\geq\; 0\;, \\ D(\mu\|\nu) &\;=\; 0 \quad\Longleftrightarrow\quad \mu \;=\; \nu\;. \end{aligned} \tag{7.4}$$

Zum Beweis benutzen wir die Funktion $k(t) := t\log t$, $t \geq 0$. Sie ist strikt konvex und liegt deswegen oberhalb ihrer Tangente in 1.



Daher gilt $k(t) \geq k'(1)(t-1)$, mit Gleichheit nur im Fall $t = 1$. Es folgt

$$D(\mu\|\nu) \;=\; \sum_x q_x k\big(\tfrac{p_x}{q_x}\big) \;\geq\; \textstyle\sum_x q_x k'(1)\big(\tfrac{p_x}{q_x} - 1\big) \;=\; 0\;.$$

Gleichheit gilt nur dann, wenn $p_x/q_x = 1$ für alle x, d.h. $\mu = \nu$ gilt.

Beispiel. Ist speziell ν die uniforme Verteilung auf der Menge S mit k Elementen, also $q_x = 1/k$, so gilt $D(\mu\|\nu) = \log k - H(\mu)$, und es folgt

$$H(\mu) \;\leq\; \log k \;=\; H(\nu)\;. \tag{7.5}$$

Auf einer endlichen Menge S wird die Entropie durch die uniforme Verteilung maximiert. Dies macht Sinn: Die Ungewißheit über das Resultat einer Zufallswahl aus S ist im Fall der Gleichverteilung sicher am größten.

5. Für Zufallsvariable X und Y gilt

$$H(X,Y) \;\le\; H(X) + H(Y)\;, \tag{7.6}$$

oder äquivalent (vgl. (7.2))

$$H(X|Y) \;\le\; H(X)\;. \tag{7.7}$$

Gleichheit gilt genau dann, wenn X und Y stochastisch unabhängig sind. Zum Beweis definieren wir Verteilungen $\mu = (p_{xy})$ und $\nu = (q_{xy})$ durch

$$p_{xy} \;:=\; \mathbf{Ws}\{X = x, Y = y\}, \quad q_{xy} \;:=\; \mathbf{Ws}\{X = x\}\mathbf{Ws}\{Y = y\}\;.$$

Es gilt

$$\begin{aligned} H(X) \;&=\; -\sum_{x,y} \mathbf{Ws}\{X = x, Y = y\} \log \mathbf{Ws}\{X = x\} \\ &=\; -\sum_{x,y} \mathbf{Ws}\{X = x, Y = y\} \log \mathbf{Ws}\{X = x \,\big|\, Y = y\} \\ &\qquad + \sum_{x,y} \mathbf{Ws}\{X = x, Y = y\} \log \frac{\mathbf{Ws}\{X = x, Y = y\}}{\mathbf{Ws}\{X = x\}\mathbf{Ws}\{Y = y\}} \\ &=\; H(X|Y) + D(\mu\|\nu)\;. \end{aligned}$$

Die Behauptung folgt also aus (7.4). □

Bemerkung. Die Entropie spielt nicht nur in der von SHANNON initiierten Informationstheorie eine fundamentale Rolle. Unabhängig und schon vor Shannon haben statistische Physiker verwandte Überlegungen angestellt. BOLTZMANN hat bereits darauf hingewiesen, daß sich die Entropie bei unabhängigen Wiederholungen von Laplace-Experimenten additiv verhält: Zieht man rein zufällig eines der m' Elemente einer Menge S', anschließend unabhängig ein weiteres Element aus einer Menge S'' mit m'' Elementen, so ist das Gesamtresultat die rein zufällige Wahl eines der $m'm''$ Elemente aus dem kartesischen Produkt $S' \times S''$. Außerdem gilt $\log m'm'' = \log m' + \log m''$.

Auch Formel (**??**) läßt sich hier einordnen. Sie besagt, daß es approximativ (wir gehen vom natürlichen Logarithmus zum Logarithmus zur Basis 2 über)

$$2^{n\cdot H(\mu)}, \quad \text{mit } \mu \;:=\; (\tfrac{x_1}{n}, \ldots, \tfrac{x_k}{n})\;,$$

Möglichkeiten gibt, n unterscheidbare Kugeln auf k Schachteln zu verteilen, so daß in der i-ten Schachtel genau x_i Kugel liegen. Bei Kenntnis der Besetzungszahlen $x_1, \ldots, x_n$ sind deshalb zirka $n \cdot H(\mu)$ Ja-Nein-Fragen nötig, um festzustellen, in welcher Schachtel jede Kugel liegt. Pro Kugel ergibt das durchschnittlich $H(\mu)$ Fragen. □

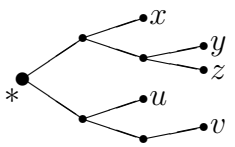
7.2 Quellenkodieren

In diesem Abschnitt geht es darum, wie man Nachrichten effizient in 01-Folgen chiffriert. Sei S eine endliche Menge. Wir fassen S als ein Alphabet auf und stellen uns die Aufgabe, seine Buchstaben (z.B. zum Zweck der Nachrichtenübertragung) durch 01-Folgen zu kodieren. Jedem $x \in S$ wird ein Kodewort $k(x)$ der Länge $|k(x)|$ zugeordnet. Wir betrachten nur **präfixfreie Kodes**, kein Kodewort soll Anfangsstück eines anderen Kodewortes sein. Damit ist garantiert, daß ein kodierter Text eindeutig entschlüsselt werden kann. Zum Beispiel ist für $S = \{x, y, z, u, v\}$ durch

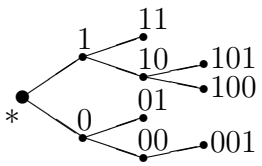
$$k(x) = 11, \quad k(y) = 101, \quad k(z) = 100, \quad k(u) = 01, \quad k(v) = 001 \tag{7.8}$$

ein präfixfreier Kode definiert. Die Liste aller Kodewörter nennt man das **Kodebuch**.

Instruktiver ist es, Kodes durch ihre **Kodebäume** zu repräsentieren. Für das Beispiel (7.8) sieht das so aus:



Allgemein hat ein Kodebaum die folgende Gestalt: Innere Knoten (inklusive die Wurzel $*$) können nach rechts (weg von der Wurzel) durch zwei Kanten verlassen werden, eine nach oben und eine nach unten (möglicherweise wird nur eine Kante gebraucht). Dann entspricht jedem Knoten eine endliche 01-Folge, sie beschreibt das Auf und Ab des Verbindungsweges durch den Baum von der Wurzel zum Knoten.



Weisen wir jedem Buchstaben aus S ein Blatt zu, so erhalten wir einen präfixfreien Kode. Es ist nicht schwer zu sehen, daß sich jeder präfixfreie Kode so darstellen läßt.

Wie konstruiert man Kodes mit kurzen Kodewörtern? Sicher ist es im allgemeinen nicht günstig, wenn alle Kodewörter gleichlang sind, besser wird man häufig verwendete Buchstaben mit kürzeren 01-Folgen kodieren. Wir

nehmen nun an, daß der Buchstabe x mit Wahrscheinlichkeit p_x auftritt ($\sum_x p_x = 1$). Die mittlere Länge des Kodes k ist dann

$$E(k) \;:=\; \mathbf{E}[|k(X)|] \;=\; \sum_x |k(x)| \cdot p_x \;,$$

dabei sei X ein zufälliger Buchstabe aus S mit Verteilung $\mu = (p_x)_x$. Wir zeigen nun, wie man Kodes von minimaler mittlerer Länge konstruiert.

Huffman-Kodes

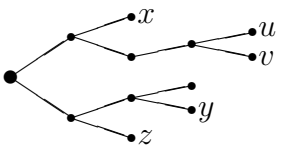
Zur Konstruktion optimaler Kodes macht man sich zwei offensichtliche Eigenschaften von Kodes minimaler mittlerer Länge zunutze.

- Gehen wir vom Kode k zum Kode $\widehat{k}$ über, indem wir die Kodewörter der Buchstaben x und y vertauschen, also $\widehat{k}(x) := k(y)$ und $\widehat{k}(y) := k(x)$ setzen, so ergibt eine einfache Rechnung die Gleichung

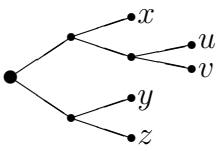
  $$E(\widehat{k}) \;=\; E(k) + (|k(x)| - |k(y)|)(p_y - p_x) \;.$$

  Gilt daher $|k(x)| > |k(y)|$ und $p_x > p_y$, so verkleinern wir durch das Vertauschen der Kodewörter die mittlere Kodelänge. Anders ausgedrückt: Hat in einem optimalen Kode x ein längeres Kodewort als y, so gilt $p_x \le p_y$.

- Besitzt der Kodebaum einen inneren Knoten, der nur durch eine einzige Kante (weg von der Wurzel) verlassen wird, so kann man diese Kante aus dem zugehörigen Kodebaum herausnehmen. Der neue Kode hat offenbar eine kleinere mittlere Länge. Genauso kann man in k jedes unbeschriftete Blatt samt zugehöriger Kante beseitigen. So läßt sich der Kodebaum

  zu dem Baum

zusammenziehen. Ein optimaler Kode hat damit die Eigenschaft, daß jeder innere Knoten (inklusive die Wurzel) nach rechts in zwei Kanten verzweigt (man sagt, der Baum ist *vollständig binär*) und jedes Blatt mit einem Buchstaben beschriftet ist.

Insbesondere treten in einem optimalen Baum die Blätter maximaler Tiefe in Paaren auf, die an einem gemeinsamen inneren Knoten hängen. Die zugehörigen Kodewörter haben maximale Länge und entsprechen deswegen den Buchstaben mit den kleinsten Wahrscheinlichkeiten. Wir können die Kodewörter dieser Buchstaben noch untereinander auswechseln, ohne daß sich die mittlere Kodelänge verändert. Daher können wir von einem optimalen Kode k ohne Einschränkung der Allgemeinheit annehmen, daß es zwei Buchstaben u und v gibt, so daß gilt:

i) $|k(u)| = |k(v)|$ *und* $|k(u)| \geq |k(x)|$ *für alle* $x \neq u, v$,

ii) u *und* v *sitzen an derselben Gabel (demselben inneren Knoten),*

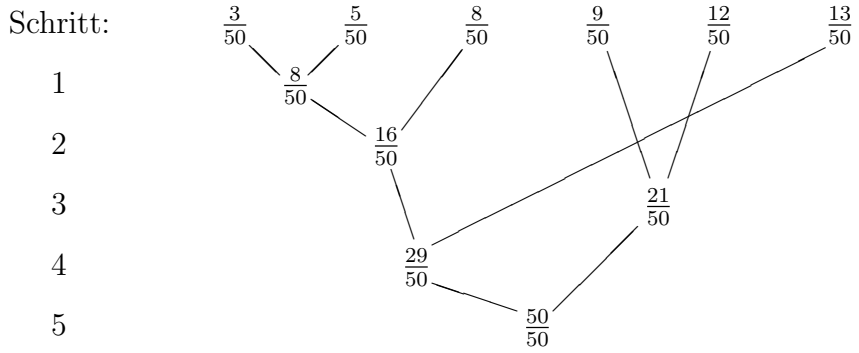
iii) $p_u \leq p_v \leq p_x$ *für alle* $x \neq u, v$.

Wir reduzieren nun das Alphabet, indem wir u und v identifizieren und dem neuen Buchstaben die Wahrscheinlichkeit $p_u + p_v$ zuordnen. Gleichzeitig beseitigen wir im Kodebaum die Gabel, an der u und v hängen. Der zugehörige innere Knoten wird zu einem Blatt, das wir mit dem neuen Buchstaben beschriften. Offenbar liefert der reduzierte Baum einen optimalen Kode für das reduzierte Alphabet. Daher dürfen wir ohne Einschränkung annehmen, daß der reduzierte Kode wieder die Bedingungen *i*)-*iii*) erfüllt.

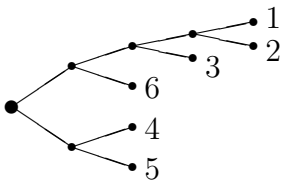
Diese Beobachtungen lassen sich als Verfahren zur schrittweisen Konstruktion optimaler Kodebäume verstehen, von der Krone hinunter zur Wurzel. Kodes, die nach dieser Vorschrift konstruiert sind, heißen **Huffman-Kodes**. Wir führen das Verfahren exemplarisch für die Verteilung μ mit den Gewichten

$$p_1 = \frac{3}{50},\ p_2 = \frac{5}{50},\ p_3 = \frac{8}{50},\ p_4 = \frac{9}{50},\ p_5 = \frac{12}{50},\ p_6 = \frac{13}{50}$$

durch. Das folgende Schema enthält die Reduktionsschritte.



Als Kodebaum erhalten wir



Der Quellenkodierungssatz

Jeden präfixfreien Kode kann man als Fragestrategie auffassen, um ein gemäß der Verteilung μ zufällig aus S ausgewähltes Element X zu erfragen: Man ermittele seine Kodierung $k(X)$, was $|k(X)|$ Ja-Nein-Fragen erfordert. Dies führt zu der Vermutung, daß bei optimaler Wahl des Kodes mit zirka $H(\mu)$ Fragen zu rechnen ist. Das vorangegangene Beispiel bekräftigt diese Erwartung: Die mittlere Länge des soeben betrachteten Kodes ist $E(k) = 2,48$, und die Entropie von $\mu = (\frac{3}{50}, \frac{5}{50}, \frac{8}{50}, \frac{9}{50}, \frac{12}{50}, \frac{13}{50})$ berechnet sich als $H(\mu) = 2,44$. Allgemeiner gilt der **Quellenkodierungssatz**.

Satz 7.1. *Für jeden präfixfreien Kode k gilt*

$$E(k) \;\geq\; H(\mu)\ ,$$

und zu jeder Verteilung μ gibt es einen präfixfreien Kode k mit

$$E(k) \;<\; H(\mu) + 1\ .$$

Der Beweis beruht auf der **Ungleichung von Fano-Kraft**.

Proposition 7.2. *Seien λ_x, $x \in S$, natürliche Zahlen. Dann gibt es genau dann einen präfixfreien Kode k mit Kodewörtern $k(x)$ der Länge λ_x für alle $x \in S$, wenn*

$$\sum_{x \in S} 2^{-\lambda_x} \;\leq\; 1$$

gilt.

Beweis. Sei k ein Kodebaum. Wir betrachten zufällige Wege durch k von der Wurzel in die Menge der Blätter, indem wir an jeder binären Verzweigung per Münzwurf entscheiden, entlang welcher Kante (weg von der Wurzel) wir den Weg fortsetzen. Wir schließen nicht aus, daß gewisse Knoten nur über eine Kante verlassen werden können, dann ist kein Münzwurf erforderlich. Trifft man bei dieser Zufallswanderung auf ein Blatt, das mit dem Buchstaben x beschriftet ist, so hat man höchstens $|k(x)|$ Münzwürfe gemacht. Daher ist die Wahrscheinlichkeit, x zu erreichen, mindestens $2^{-|k(x)|}$. Diese Wahrscheinlichkeiten können sich höchstens zu 1 aufsummieren, daher folgt

$$\sum_{x} 2^{-|k(x)|} \;\leq\; 1\;,$$

die Bedingung der Proposition ist also notwendig.

Sind umgekehrt λ_x natürliche Zahlen, die die Bedingung erfüllen, so konstruieren wir von der Wurzel aus einen vollständig verzweigenden binären Baum, der für jeden Buchstaben x ein Blatt der Tiefe λ_x vorsieht (im Kode erhält x dann ein Kodewort der Länge λ_x). Zu zeigen ist, daß man diese Konstruktion solange fortsetzen kann, bis alle Buchstaben berücksichtigt sind. Nehmen wir also an, daß die Konstruktion abgebrochen werden muß, weil alle Blätter bereits mit einem Buchstaben beschriftet sind. Wählen wir nun erneut per Münzwurf einen Zufallsweg durch den konstruierten Baum, so wird ein mit x beschriftetes Blatt mit Wahrscheinlichkeit $2^{-\lambda_x}$ erreicht. Da nach Annahme jedes Blatt einen Buchstaben trägt, summieren sich diese Wahrscheinlichkeiten zu 1 auf. Nach Voraussetzung der Proposition sind daher bereits alle Elemente von S im Baum untergebracht, und der Baum ist der gesuchte Kodebaum. □

Beweis von Satz 7.1. Nach Proposition 7.2 kann man zu jedem präfixfreien Kode k eine Wahrscheinlichkeitverteilung $\nu = (q_x)$ wählen, so daß für alle x

$$2^{-|k(x)|} \;\leq\; q_x$$

gilt. Für $\mu = (p_x)$ folgt unter Beachtung von (7.4)

$$E(k) \;=\; \sum_x |k(x)| p_x \;\geq\; -\sum_x p_x \log q_x \;=\; H(\mu) + D(\mu\|\nu) \;\geq\; H(\mu)\;,$$

also die erste Behauptung. Zum Nachweis der zweiten Behauptung wählen wir natürliche Zahlen λ_x, so daß $\lambda_x - 1 < -\log p_x \le \lambda_x$. Dann gilt

$$\sum_x 2^{-\lambda_x} \;\le\; \sum_x p_x \;=\; 1\ ,$$

nach Proposition 7.2 gibt es also einen Kode k mit $|k(x)| = \lambda_x$ für alle x. Für diesen Kode folgt wegen $|k(x)| < 1 - \log p_x$ wie behauptet

$$E(k) \;<\; \sum_x (1 - \log p_x) p_x \;=\; 1 + H(\mu)\ .$$

□

Bemerkungen.

1. Kodes, deren Kodewörter $k(x)$ eine Länge zwischen $-\log p_x$ und $-\log p_x + 1$ haben, heißen **Shannon-Kodes**. Der Beweis von Satz 7.1 zeigt, daß solche Kodes existieren und daß ihre mittlere Länge um höchstens ein bit von der mittleren Länge eines optimalen Kodes abweicht. $-\log p_x$ ist aufgerundet die Anzahl von bits, die ein solcher Kode zur Kodierung von x braucht, deswegen bezeichnet man in der Informationstheorie die Größe $-\log p_x$ auch als den **Informationsgehalt von x**.

2. Sei $\nu = (q_x)$ eine Verteilung auf S und sei k ein Kode mit Kodewörtern $k(x)$, deren Länge approximativ gleich $-\log q_x$ ist. Dann ist k also ein günstiger Kode, falls der Buchstabe x mit Wahrscheinlichkeit q_x vorkommt. Tritt x jedoch mit Wahrscheinlichkeit p_x ein, so gilt

$$E(k) \;\approx\; -\sum_x p_x \log q_x \;=\; H(\mu) + D(\mu\|\nu)\ .$$

   Die relative Entropie $D(\mu\|\nu)$ gibt also die mittlere Anzahl an zusätzlichen bits an, die erforderlich werden, wenn man den Kode an ν und nicht an vorliegende Verteilung μ angepaßt hat. □

Blockweises Kodieren von Nachrichten

Daten oder Nachrichten werden häufig wie beim Morsen Buchstabe für Buchstabe kodiert. Man kann die kodierte Nachricht aber weiter komprimieren, wenn man ganze Wörter kodiert. Wir betrachten nun zufällige Wörter $X_1X_2\ldots X_n$ der Länge n, mit S-wertigen Zufallsvariablen $X_1, \ldots, X_n$. Denkt man an die Kodierung von Texten, so wird man in das Alphabet S auch die Leerstelle und Interpunktionszeichnen aufnehmen müssen. Mit k_n bezeichnen wir den Huffman-Kode zur Kodierung von Wörtern der Länge n, d.h. von

Elementen aus S^n. Die Blätter im Kodierbaum sind nun mit ganzen Wörtern markiert. Zur Konstruktion von k_n langt es nicht mehr, die Häufigkeit der einzelnen Buchstaben zu kennen, also die Verteilung der X_i, wir benötigen die gemeinsame Verteilung von $X_1, \ldots, X_n$, d.h. die Wahrscheinlichkeiten, mit denen Wörter der Länge n auftreten.

Wir beginnen mit dem einfachsten Fall, daß $X_1, X_2, \ldots$ unabhängige, identisch verteilte Zufallsvariable mit Verteilung μ sind. Dann ist die Aussage

$$\lim_{n\to\infty} \frac{E(k_n)}{n} \;=\; H(\mu)\ . \tag{7.9}$$

gültig, der **erste Hauptsatz der Informationstheorie**. Nach Satz 7.1 gilt nämlich

$$H(X_1, \ldots, X_n) \;\le\; E(k_n) \;<\; H(X_1, \ldots, X_n) + 1\ ,$$

außerdem folgt aus der Unabhängigkeit von $X_1, \ldots, X_n$ (vgl. (7.6))

$$H(X_1, \ldots, X_n) \;=\; n \cdot H(\mu)\ .$$

Bei optimaler Kodierung von Blöcken der Länge n werden daher im Mittel pro Buchstabe zwischen $H(\mu)$ und $H(\mu) + \frac{1}{n}$ bits benötigt, und nicht mehr zwischen $H(\mu)$ und $H(\mu) + 1$ bits, wie dies nach dem Quellenkodierungssatz beim buchstabenweisen Kodieren der Fall ist.

Für einen Text in Deutsch oder einer anderen Sprache ist es allerdings völlig unangemessen, die Buchstaben als Realisationen von stochastisch unabhängigen Zufallsvariablen aufzufassen. Man betrachtet deswegen allgemeiner **stationäre Quellen**.

Definition. *Eine Folge $X_1, X_2, \ldots$ von S-wertigen Zufallsvariablen heißt **stationär**, falls $X_1, \ldots, X_n$ und $X_{m+1}, \ldots, X_{m+n}$ für alle $m, n \ge 1$ dieselbe gemeinsame Verteilung haben.*

Anschaulich gesprochen verständigen wir uns also auf die Annahme, daß die statistischen Häufigkeiten von Buchstaben und Wörtern nicht davon abhängt, auf welche Stelle des Textes man sich bezieht.

Für eine stationäre Quelle $X_1, X_2, \ldots$ gilt eine dem ersten Hauptsatz analoge Aussage. Wir beweisen die Existenz des Grenzwertes

$$\lim_{n\to\infty} n^{-1} H(X_1, \ldots, X_n)\ .$$

Zum Beweis benutzen wir die sich aus (7.2) ergebende Darstellung

$$\begin{aligned} H(X_1, \ldots, X_n) \;&=\; H(X_n | X_1, \ldots, X_{n-1}) + H(X_1, \ldots, X_{n-1}) \\ &=\; \cdots \;=\; \sum_{m=1}^{n} H(X_m | X_{m-1}, \ldots, X_1)\ . \end{aligned}$$

Die Summanden bilden eine monoton fallende Folge: In Verallgemeinerung von (7.7) gilt

$$H(X_m|X_{m-1},\dots,X_1) \;\le\; H(X_m|X_{m-1},\dots,X_2)\;,$$

und wegen der Stationarität folgt

$$H(X_m|X_{m-1},\dots,X_1) \;\le\; H(X_{m-1}|X_{m-2},\dots,X_1)\;.$$

Insbesondere konvergiert die Folge $H(X_m|X_{m-1},\dots,X_1)$. Nun ist es eine einfache Tatsache der Analysis, daß sich aus der Konvergenz einer Zahlenfolge h_m auch die Konvergenz von $n^{-1}(h_1+\cdots+h_n)$ gegen denselben Grenzwert ergibt. Damit erhalten wir die behauptete Existenz des Grenzwertes

$$H_Q \;:=\; \lim_{n\to\infty} n^{-1}H(X_1,\dots,X_n) \;=\; \lim_{m\to\infty} H(X_m|X_{m-1},\dots,X_1)\;,$$

den man als **Entropierate** der stationären Quelle bezeichnet. Kodiert man also nach der Methode von Huffman lange Blöcke, so benötigt man pro Buchstabe im Mittel H_Q bits. Je kleiner H_Q ist, desto kürzer werden die Kodewörter. Im Extremfall $X_1 = X_2 = \cdots$, dem unbeirrten Wiederholen der Nachricht X_1, gilt $H(X_1,\dots,X_n) = H(X_1)$ und $H_Q = 0$. Das andere Extrem ist der Fall von stochastisch unabhängigen Buchstaben $X_1, X_2, \dots$, dann ergibt sich $H_Q = H(X_1)$ nach (7.9). Im Allgemeinen gilt

$$0 \;\le\; H_Q \;\le\; H(X_1)\;,$$

wie aus $H(X_1,\dots,X_n) \le H(X_1)+\cdots+H(X_n) = nH(X_1)$ folgt (vgl. (7.6)). Ist die strikte Ungleichung $H_Q < H(X_1)$ erfüllt, so werden - statistisch gesehen - überflüssige Buchstaben verwendet, man sagt, die Quelle ist redundant. Eine quantitative Maßzahl dafür ist die **relative Redundanz**

$$R_Q \;:=\; 1 - \frac{H_Q}{H(X_1)}$$

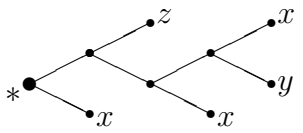
mit Werten zwischen 0 und 1.

Für die Untersuchung statistischer Eigenschaften von Sprachen hat sich das Modell einer stationären Quelle als brauchbar erwiesen. Ausführliche Untersuchungen haben ergeben, daß die Redundanz vieler europäischer Sprachen nahe bei 1/2 liegt. Salopp gesprochen könnte man jeden zweiten Buchstaben sparen. Dies ist keine Verschwendung: Ohne Redundanz würden schon kleinste Fehler in einem Text Verständigungsprobleme verursachen. - Ausführlicher berichtet F. Topsøe in seiner lesenswerten Einführung *Informationstheorie* (1975) über dieses Thema.

7.3 Simulation durch Münzwurf

Eine andere Interpretation der Entropie besagt, daß $H(\mu)$ der **Grad an Zufälligkeit** ist, der in einer Zufallsvariablen X mit Verteilung μ steckt. Um dieser Aussage eine klare Bedeutung zu geben, untersuchen wir, welchen Aufwand es erfordert, um eine solche Zufallsvariable mit Hilfe einer Folge von unabhängigen Münzwürfen zu erzeugen. Wie sich herausstellt, ist $H(\mu)$ in etwa die mittlere Anzahl von Münzwürfen, die dazu nötig ist.

Die Simulationsverfahren, um die es hier geht, lassen sich als beschriftete binäre Bäume veranschaulichen. Beispielsweise gehört der Baum



zu der Verteilung auf $S = \{x, y, z\}$ mit den Gewichten

$$p_x = \frac{1}{2} + \frac{1}{8} + \frac{1}{16} = \frac{11}{16}, \quad p_y = \frac{1}{16}, \quad p_z = \frac{1}{4} \ .$$

Ausgehend von der Wurzel $*$ wählt man per Münzwurf einen zufälligen Weg nach rechts durch den Baum, bis ein Blatt erreicht ist, an dem man dann das aus S ausgewählte Element ablesen kann. Allgemein ist ein Simulationsverfahren s durch einen vollständig binären Baum gegeben, zusammen mit einer Beschriftung seiner Blätter durch Elemente aus S. (Im Gegensatz zu den Kodierbäumen des letzten Abschnitts können nun verschiedene Blätter mit demselben Buchstaben beschriftet sein.)

Jede Verteilung $\mu = (p_x)$ auf einer abzählbaren Menge S läßt sich auf diese Weise realisieren. Das Beispiel legt nahe, wie man vorzugehen hat: Man schreibe die Gewichte von μ als Dualbruch,

$$p_x \ = \ \sum_k 2^{-\lambda_{x,k}} \ , \tag{7.10}$$

mit natürlichen Zahlen $1 \le \lambda_{x,1} < \lambda_{x,2} < \cdots$ Dann gilt

$$\sum_{x,k} 2^{-\lambda_{x,k}} \ = \ \sum_x p_x \ = \ 1 \ ,$$

nach Proposition 7.2 gibt es also einen binären Baum, der für jedes Paar (x, k) ein Blatt der Tiefe $\lambda_{x,k}$ freihält. Versieht man diese Blätter alle mit der Beschriftung x, so wird x per Münzwurf wie gewünscht mit der Wahrscheinlichkeit

$$\sum_k 2^{-\lambda_{x,k}} \ = \ p_x$$

erreicht.

Die mittlere Anzahl von Münzwürfen, die für das Simulationsverfahren s anfällt, ist durch die Formel

$$E(s) \;:=\; \sum_{b\in B} t(b)2^{-t(b)}$$

gegeben. Dabei bezeichnet B die Menge aller Blätter im zugehörigen Baum und $t(b)$ die Tiefe des Blattes b, also die Anzahl der benötigten Münzwürfe, um von der Wurzel nach b zu gelangen. Die Wahrscheinlichkeit, mit der man b erreicht, ist folglich $2^{-t(b)}$.

Der folgende Satz präzisiert, daß bei einem guten Simulationsverfahren in etwa $H(\mu)$ Münzwürfe anfallen.

Satz 7.3. *Sei μ eine Verteilung auf S. Dann gilt für jedes Verfahren s zur Simulation von μ*

$$H(\mu) \;\le\; E(s)\ .$$

Darunter gibt es eine Verfahren s mit der Eigenschaft

$$E(s) \;\le\; H(\mu)+2\ .$$

Beweis. Sei $\phi : B \to S$ die zu s gehörige Abbildung, die jedem Blatt sein Element aus S zuordnet, und sei Y das per Münzwurf aus B ausgewählte Blatt. Generiert also s die Verteilung μ, so ist $X := \phi(Y)$ nach μ verteilt. Nach (7.3) folgt

$$H(\mu) \;\le\; H(Y) \;=\; -\sum_b 2^{-t(b)}\log 2^{-t(b)} \;=\; E(s)\ ,$$

also die erste Behauptung. Weiter gilt nach (7.2)

$$\begin{aligned} E(s) \;&=\; H(Y) \;=\; H(X,Y)-H(X|Y) \;\le\; H(X,Y)\\ &=\; H(X)+H(Y|X) \;=\; H(\mu)+\sum_x H(Y|X=x)\cdot \mathbf{Ws}\{X=x\}\ . \end{aligned}$$

Zum Nachweis der zweiten Behauptung langt es also zu zeigen, daß es für μ eine Simulationsprozedur gibt, so daß $H(Y|X=x)\le 2$ für alle $x\in S$ gilt. Dies leistet das oben beschriebene, auf der Dualbruchzerlegung (7.10) beruhende Verfahren s. Für dieses Verfahren haben die mit x beschrifteten Blätter $b_{x,1}, b_{x,2},\dots$ im Baum strikt wachsende Tiefen $1\le\lambda_{x,1}<\lambda_{x,2}<\cdots$, deswegen gilt

$$\mathbf{Ws}\{Y=b_{x,k}\,\big|\,X=x\} \;=\; 2^{-\lambda_{x,k}}/p_x \;\le\; \frac{1}{2}\,\mathbf{Ws}\{Y=b_{x,k-1}\,\big|\,X=x\}\ .$$

Die zweite Behauptung des Satzes folgt nun aus der nachfolgenden Proposition. □

Proposition 7.4. *Für die Gewichte $p_1, p_2, \ldots$ einer Wahrscheinlichkeitsverteilung μ auf $\mathbb{N}$ gelte $p_{n+1} \le p_n/2$ für alle n. Dann folgt $H(\mu) \le 2$.*

Beweis. Sei ν die Verteilung mit den Gewichten $q_n := (1-p_1)^{-1}p_{n+1}$, $n \ge 1$. Eine kurze Rechnung ergibt

$$H(\mu) \;=\; -p_1 \log p_1 - (1-p_1)\log(1-p_1) + (1-p_1)H(\nu)\ .$$

Die Entropie der Verteilung mit Gewichten p_1 und $1-p_1$ ist nach (7.5) höchstens $\log 2 = 1$, außerdem gilt $p_1 \ge 1/2$ (sonst könnten sich die p_n nach Annahme nicht zu 1 aufsummieren). Daher folgt

$$H(\mu) \;\le\; 1 + H(\nu)/2\ .$$

Offenbar gilt auch $q_{n+1} \le q_n/2$ für alle n. Setzen wir also H^* als das Supremum der Entropien aller Verteilungen mit der in der Proposition angegebenen Eigenschaft, so folgt $H^* \le 1 + H^*/2$. Aus dieser Ungleichung erhalten wir $H^* \le 2$ und damit die Behauptung, vorausgesetzt, H^* ist endlich. Dies folgt aus

$$H(\mu) \;\le\; 2 + \sum_n n2^{-n}\ ,$$

denn nach Voraussetzung gilt $p_n \le 2^{-n}p_1$, und $-p\log p$ ist monoton fallend für $p \le 1/4$. □

7.4 Gestörte Nachrichtenübertragung

Mit der Nachrichtenübertragung verbinden sich verschiedene reizvolle und schwierige mathematische Fragestellungen.

Quelle $\longrightarrow$ gestörter Kanal $\longrightarrow$ Empfänger

Zunächst muß die Nachricht an der **Quelle** in eine Form gebracht werden, die ihre Versendung möglich macht. Normalerweise bedeutet dies, sie nach der Methode von Huffman in eine 01-Folge zu transformieren. Am anderen Ende der Nachrichtenstrecke muß die empfangene 01-Sequenz vom **Empfänger** entschlüsselt werden. Auf dieses Problem gehen wir hier nicht ein.

Dazwischen steht die Übertragung durch den **gestörten Kanal**. Eine unmittelbare Transmission der als 01-Folge aufbereiteten Nachricht empfiehlt sich nur dann, wenn man sich sicher sein kann, daß es keine Übertragungsfehler gibt. Sonst kann schon ein fehlerhaft empfangenes bit den Sinn der

Nachricht verfälschen oder unverständlich machen. Die Idee der **Kanalkodierung** besteht darin, daß man Redundanz in die Nachricht einführt, indem man sie künstlich verlängert, und es dem Empfänger damit möglicht macht, Fehler zu erkennen und sogar zu korrigieren. Die einfachste Idee ist es, einen **Repetitionskode** zu verwenden, der jedes bit 3-mal überträgt. Von diesen 3 bits müßten schon 2 falsch beim Empfänger ankommen, damit dieser die abgesendeten bits nicht mehr richtig erkennt. Diese Methode ist aber unökonomisch, die Theorie der **fehlerkorrigierenden Kodes**, eine wichtige, aktuelle mathematische Disziplin, hat da sehr viel bessere Vorschläge.

In diesem Abschnitt geht es um keine bestimmten Kodes sondern um die Frage, um welchen Faktor man die Länge der Nachricht mindestens strecken muß, damit eine korrekte Entschlüsselung für den Empfänger überhaupt erst durchführbar wird. Die grundlegenden Ideen gehen zurück auf SHANNON, den Begründer der Informationstheorie. Wie er betrachten wir Blockkodes. Ein **(n, m)-Blockkode** besteht aus einer Kodiervorschrift k, die 01-Folgen der Länge m in 01-Folgen der Länge n expandiert,

$$k : \{0,1\}^m \ \to \ \{0,1\}^n \ .$$

Anstelle von $u = u_1 \dots u_m$ wird das kodierte Wort $x = x_1 \dots x_n := k(u)$ gesendet. Am anderen Ende des Kanals muß dann die empfangene Nachricht mit einer Dekodierabbildung

$$d : \{0,1\}^n \ \to \ \{0,1\}^m$$

zurückübersetzt werden. Empfängt man das Wort $y = y_1 \dots y_n$, so wird es als $v = v_1 \dots v_m := d(y)$ entschlüsselt. Bezeichnen wir weiter mit Y_x das empfangene Wort aus $\{0,1\}^n$, falls $x = k(u)$ gesendet wurde, so wird die Nachricht genau dann korrekt dekodiert, falls $d(Y_x) = u$ gilt. Wir fassen Y_x als Zufallsvariable auf, denken also an den Fall, daß bei der Übertragung zufällige Fehler eintreten. Die **maximale Fehlerwahrscheinlichkeit** der Kombination eines Blockkodes k und einer Dekodierabbildung d definiert man als

$$\gamma(k,d) \ := \ \max_{u \in \{0,1\}^m} \mathbf{Ws}\big\{d(Y_{k(u)}) \neq u\big) \ ,$$

und die **durchschnittliche Fehlerwahrscheinlichkeit** als

$$\overline{\gamma}(k,d) \ := \ 2^{-m} \sum_{u \in \{0,1\}^m} \mathbf{Ws}\big\{d(Y_{k(u)}) \neq u\big) \ .$$

Der Quotient $r = m/n$ wird als die **Übertragungsrate des Blockkodes** bezeichnet, er gibt an, welchen Anteil eines bits der ursprünglichen Nachricht

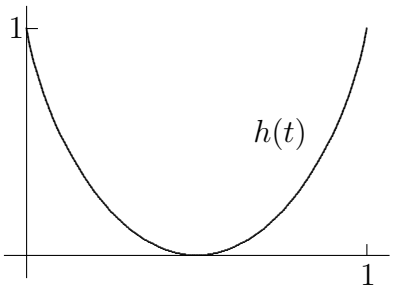
pro gesendetem bit übermittelt wird (für den Repetitionskode ist $r = 1/3$). Wir können nun unsere Fragestellung präzisieren: *Wie klein muß r mindestens sein, damit ein Blockkode der Rate r mit ausreichend kleiner Fehlerwahrscheinlichkeit existiert.* Dies erfordert eine Annahme über das Wahrscheinlichkeitsgesetz, nach dem Übertragungsfehler auftreten. Wir betrachten zunächst den einfachsten Fall eines symmetrischen gedächtnislosen Kanals.

Symmetrischer gedächtnisloser Kanal. Einzelne bits werden mit der Wahrscheinlichkeit p falsch übertragen, d.h. eine 1 anstelle einer 0, bzw. eine 0 anstelle einer 1 empfangen. Die n bits eines Wortes x werden unabhängig voneinander übertragen. Dies heißt, daß die Anzahl der Übertragungsfehler pro Wort der Länge n eine binomialverteilte Zufallsvariable zum Parameter (n, p) ist.

Für solche Kanäle gilt die folgende auf Shannon zurückgehende Aussage. Wir setzen

$$h(t) \;:=\; t\log 2t + (1-t)\log 2(1-t) \qquad 0 \le t \le 1\ .$$

Diese Funktion ist uns schon in der Entropiefunktion der $\mathrm{B}(n, 1/2)$-Verteilung begegnet (vgl. (1.7)), abgesehen davon, daß wir hier mit Logarithmen zur Basis zwei rechnen.



Satz 7.5. *Sei $p \neq 1/2$ und $\epsilon > 0$. Dann gibt es einen Blockkode, dessen Übertragungsrate mindestens $h(p)-\epsilon$ ist und der mit einer durchschnittlichen Fehlerwahrscheinlichkeit von höchstens ϵ dekodiert werden kann.*

Darüber hinaus kann man zeigen, daß die Schranke $h(p)$ nicht weiter verbessert werden kann. Es ist bemerkenswert, daß sich in diesem allgemeinen Rahmen ein solch präzises Resultat erzielen läßt. Der Beweis liefert keinen Kode mit den gewünschten Eigenschaften, dies wäre zuviel verlangt. Vielmehr konstruiert man einen für praktische Zwecke unbrauchbaren *zufälligen* Kode mit den angestrebten günstigen Eigenschaften und folgert, daß dann auch ein geeigneter deterministischer Kode existieren muß. Dabei erhält man keinen Hinweis darauf, wie so ein Kode aussehen könnte. Dies ist ein

schönes Beispiel für die **probabilistische Methode**, einer Beweismethode, die in der Kombinatorik eine wichtige Rolle spielt: Um zu beweisen, daß eine Menge ein Element mit einer bestimmten Eigenschaft enthält, zeige man, daß man bei *zufälliger Wahl* eines Elementes aus der Menge mit positiver Wahrscheinlichkeit ein solches Element erhält. Dann muß die Menge offenbar Elemente der gewünschten Eigenschaft enthalten, wo immer sie auch liegen mögen.

Beweis von Satz 7.5. Sei $p^* \in (p, 1/2)$ derart, daß $h(p^*) > h(p) - \epsilon/2$. Für vorgegebenes $n \in \mathbb{N}$ setzen wir $m := [n(h(p) - \epsilon)] + 1$. Ein (n, m)-Blockkode erreicht dann die vorgegebene Transmissionsrate.

Die Kodierabbildung wird aus der Menge aller $k : \{0,1\}^m \to \{0,1\}^n$ ausgewählt. Zum Dekodieren betrachten wir ‚Kugeln' um $y \in \{0,1\}^n$, definiert als

$$B(y) := \{z \in \{0,1\}^n : z \text{ unterscheidet sich von } y \text{ in höchstens } [np^*] \text{ Stellen } \},$$

und dekodieren k nach der folgenden Vorschrift: Gibt es zu $y \in \{0,1\}^n$ genau ein $u \in \{0,1\}^m$, so daß $k(u) \in B(y)$, so setze $d_k(y) = u$. Für andere y kann $d_k(y)$ beliebig gewählt werden.

Wir nehmen nun an, daß K ein zufälliges Element aus der Menge aller $k : \{0,1\}^m \to \{0,1\}^n$ ist, und zwar derart, daß $K(u)$, $u \in \{0,1\}^m$, unabhängige Zufallsvariable mit einer uniformen Verteilung in $\{0,1\}^n$ sind (man kann das auch so ausdrücken, daß K uniform verteilt ist in der Menge aller Kodierabbildungen). Wir fragen nach der Wahrscheinlichkeit, daß bei Benutzung dieser zufälligen Kodierung der Empfänger ein bestimmtes Wort $u = u_1 \dots u_m$ richtig dekodiert. Der Sender überträgt also $X := K(u)$. Ist das empfangene Wort $Y = Y_{K(u)}$, so wird es als Nachricht $V := d_K(Y)$ entschlüsselt. Nach Wahl der Dekodierabbildungen wird die Nachricht zumindestens dann richtig als $V = u$ entschlüsselt, falls die Ereignisse $\{K(u) \in B(Y)\}$ und $\{K(u') \notin B(Y)\}$ für alle $u' \neq u$ gemeinsam eintreten. Es langt also, die Wahrscheinlichkeit des Komplementärereignisses

$$\{K(u) \notin B(Y)\} \cup \bigcup_{u' \neq u} \{K(u') \in B(Y)\}$$

nach oben abzuschätzen. $K(u) \notin B(Y)$ impliziert, daß die Anzahl Z_p der Übertragungsfehler größer als np^* ist. Nach Annahme ist Z_p binomialverteilt zum Parameter (n, p). Für die anderen Ereignisse beachten wir, daß $K(u')$

für alle $u' \neq u$ unabhängig von $Y = Y_{K(u)}$ ist, daher folgt

$$\begin{aligned}
&\mathbf{Ws}\{K(u') \in B(Y)\} \\
&\quad = \sum_y \mathbf{Ws}\{K(u') \in B(y) \,\big|\, Y = y\} \cdot \mathbf{Ws}\{Y = y\} \\
&\quad = \sum_y \mathbf{Ws}\{K(u') \in B(y)\} \cdot \mathbf{Ws}\{Y = y\} \ .
\end{aligned}$$

Da $K(u')$ uniform in $\{0,1\}^n$ verteilt ist und $\sum_{i \le np^*} \binom{n}{i}$ Elemente in $B(y)$ enthalten sind, erhalten wir

$$\mathbf{Ws}\{K(u') \in B(Y)\} \ = \ \sum_{i \le np^*} \binom{n}{i} 2^{-n}$$

Dies ist die Wahrscheinlichkeit, daß eine binomialverteilte Zufallsvariable $Z_{1/2}$ zum Parameter $(n, 1/2)$ einen Wert kleiner oder gleich np^* annimmt. Es gibt $2^m - 1$ Worte $u' \neq u$, für die Wahrscheinlichkeit, daß u nicht richtig dekodiert wird, erhalten wir also insgesamt die Abschätzung

$$\mathbf{Ws}\{d_K(Y_{K(u)}) \neq u\} \ \le \ \mathbf{Ws}\{Z_p \ge np^*\} + 2^m \, \mathbf{Ws}\{Z_{1/2} \le np^*\} \ .$$

Diese Wahrscheinlichkeiten sind wegen $p < p^* < 1/2$ nach (3.5) und (3.6) mit wachsendem n exponentiell klein. Da wir hier in Logarithmen zur Basis 2 rechnen, erhalten wir insbesondere, wie ein Vergleich mit (3.6) zeigt,

$$2^m \, \mathbf{Ws}\{Z_{1/2} \le np^*\} \ \le 2^m 2^{-nh(p^*)} \ \le 2^{1-\epsilon n/2} \ ,$$

wobei die letzte Abschätzung aus unseren Annahmen über m und $h(p^*)$ folgt. Daher konvergiert $\mathbf{Ws}\{d_K(Y_{K(u)}) \neq u\}$ exponentiell schnell gegen 0, und zwar gleichmäßig in allen $u \in \{0,1\}^m$. Für genügend großes n folgt

$$2^{-m} \sum_u \mathbf{Ws}\{d_K(Y_{K(u)}) \neq u\} \ \le \ \epsilon \ .$$

Da $\mathbf{Ws}\{d_K(Y_{K(u)}) \neq u\}$ durch Mittelung von $\mathbf{Ws}\{d_k(Y_{k(u)}) \neq u\}$ über alle Kodierabbildungen k entsteht, muß für mindestens ein k samt Dekodierabbildung d_k

$$2^{-m} \sum_u \mathbf{Ws}\{d_k(Y_{k(u)}) \neq u\} \ \le \ \epsilon$$

gelten. Dies ist die Behauptung. □

Der Beweis zeigt, daß man durch blindes Hineingreifen in die Menge aller Kodierabbildungen einen Kode mit guten Übertragungseigenschaften findet, es

muß also viele gute Kodes geben. Es mag deswegen überraschen, daß es dennoch schwer fällt, solche Kodes explizit anzugeben. Man bedenke jedoch, daß es wesentlich darauf ankommt, daß das Kodieren und Dekodieren rechnerisch leicht zu bewerkstelligen ist und vom Computer durchgeführt werden kann. Es ist nach wie vor ein aktuelles Thema, Blockkodes zu konstruieren, die praktisch brauchbar sind, und deren Übertragungsrate nahe an die Schranke von Shannon herankommen.

Das Resultat von Shannon läßt sich wesentlich verallgemeinern. Wir definieren nun die **Kapazität** eines allgemeinen Kanals. Solch ein Kanal ist gegeben durch ein **Eingangsalphabet** S, das der Sender zum Kodieren seiner Nachricht benutzt, ein **Ausgangsalphabet** S', in dem der Empfänger die Nachrichten empfängt, und eine Matrix P von **Übergangswahrscheinlichkeiten** P_{xy}, die angeben, mit welcher Wahrscheinlichkeit der Buchstabe $y \in S'$ empfangen wird, falls der Buchstabe $x \in S$ gesendet wurde.

Um die Kapazität des Kanals zu definieren, benötigen wir den Begriff der **wechselseitigen Information** zweier Zufallsvariabler X und Y,

$$I(X\|Y) \;:=\; H(X) - H(X|Y)\ .$$

Die Bezeichnung erklärt sich aus unserer Vorstellung von $H(X)$ als Grad der Ungewißheit über den Wert von X bzw. $H(X|Y)$ als Grad der Ungewißheit über den Wert von X, wenn man Y beobachten darf. Nach (7.7) gilt

$$I(X\|Y) \;\geq\; 0\ .$$

Eine alternative Formel ist nach (7.2)

$$I(X\|Y) \;=\; H(X) + H(Y) - H(X,Y)\ ,$$

ihr entnimmt man, daß die wechselseitige Information in X und Y symmetrisch ist,

$$I(X\|Y) \;=\; I(Y\|X)\ .$$

Diese Formeln machen Sinn für beliebige Zufallsvariable X und Y. Wir stellen uns nun vor, daß X ein zufälliger Buchstabe aus S ist, der durch den Kanal mit Übergangsmatrix P gesendet wird und als zufälliger Buchstabe Y aus dem Ausgangsalphabet S' empfangen wird. Zwischen den Verteilungen μ und ν von X und Y besteht dann die Beziehung

$$\nu_y \;=\; \sum_{x\in S} \mu_x P_{xy}, \quad y \in S'\ .$$

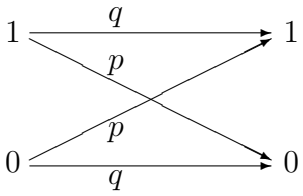
Die **Kapazität** des Kanals definiert man als

$$C \;:=\; \max_{\mu} I(X\|Y) \;.$$

Man faßt also die Verteilung μ von X ins Auge, für die die wechselseitige Information zwischen X und Y maximal ist. Die Idee ist, daß man dann aus der Beobachtung von Y die größtmögliche Information über den gesendeten Buchstaben X erhält.

Beispiele.

1. **Der symmetrische Kanal.** Beim symmetrischen Kanal ist $S = S' = \{0,1\}$ und $P_{01} = P_{10} = p$, $P_{00} = P_{11} = q = 1-p$.



   Dann gilt offenbar

$$H(Y|X=x) = -p\log p - (1-p)\log(1-p) \;=\; 1 - h(p) \;.$$

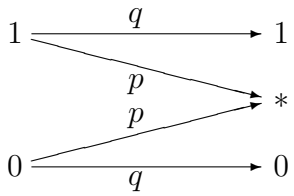
   Es folgt

$$\begin{aligned} I(X\|Y) \;&=\; H(Y) - H(Y|X) \\ &=\; H(Y) - \sum_{x=0,1} H(Y|X=x)\mathbf{Ws}\{X=x\} \\ &=\; H(Y) + h(p) - 1 \\ &\le\; h(p) \;, \end{aligned}$$

   denn es gilt $H(Y) \le \log 2 = 1$ nach (7.5). Ist X uniform auf $\{0,1\}$ verteilt, so auch Y. In diesem Fall gilt $H(Y) = 1$, für die Kapazität folgt daher

$$C \;=\; h(p) \;.$$

2. **Verlust von bits.** Wir betrachten nun einen Kanal, bei dem bits verloren gehen können, anstatt wie beim symmetrischen Kanal geflippt zu werden (**Eraser-Channel**). Das Eingangsalphabet ist wieder $S = \{0,1\}$, das Ausgangsalphabet nun $\{0,1,*\}$, und die Übergangswahrscheinlichkeiten $P_{0*} = P_{1*} = p$, $P_{00} = P_{11} = q = 1-p$.



$*$ steht für die Situation, daß bei der Übertragung das gesendete bit verloren geht. Eine kurze Rechnung ergibt

$$H(Y) \;=\; (1-p)H(X) + 1 - h(p)$$

außerdem gilt wie beim symmetrischen Kanal $H(Y|X=x) = 1-h(p)$. Daher folgt

$$I(X\|Y) \;=\; H(Y) + h(p) - 1 \;=\; (1-p)H(X)\;.$$

Der Ausdruck wird maximal, wenn X uniform verteilt ist, und es ergibt sich für die Kapazität die plausible Formel

$$C \;=\; 1-p\;.$$

□

Die für den symmetrischen binären Kanal eingeführten Begriffe übertragen sich in naheliegender Weise auf allgemeine Kanäle. Ein **(n,m)-Blockkode** besteht aus einer Kodierungsabbildung

$$k : \{0,1\}^m \;\to\; S^n\;,$$

die man mit einer Abbildung

$$d : (S')^n \;\to\; \{0,1\}^m$$

dekodiert. $r = m/n$ ist **Übertragungsrate** des Kodes, die Fehlerwahrscheinlichkeiten werden wie oben definiert. Wir betrachten wieder den Fall eines gedächtnislosen Kanals, der die einzelnen Buchstaben unabhängig voneinander gemäß der Übergangswahrscheinlichkeiten P_{xy} überträgt.

Satz 7.6. *Gegeben sei ein gedächtnisloser Kanal der Kapazität C. Dann gibt es zu jedem $\epsilon > 0$ einen Blockkode, dessen Übertragungsrate mindestens $C-\epsilon$ ist, und der mit einer maximalen Fehlerwahrscheinlichkeit von höchstens ϵ dekodiert werden kann.*

Ist umgekehrt $r \geq 0$ derart, daß für jedes $\epsilon > 0$ Nachrichten mit einer Übertragungsrate von mindestens r und einer maximalen Fehlerwahrscheinlichkeit von höchstens ϵ übertragen werden können, so folgt $r \leq C$.

Kurz zusammengefaßt ist die Kapazität die maximale Übertragungsrate, die man für einen gestörten Kanal realisieren kann. Man bemerke, daß der Satz eine Aussage über die maximale Fehlerwahrscheinlichkeit des Kodes trifft, und nicht nur über die durchschnittliche Fehlerwahrscheinlichkeit wie Satz 7.5. Ein Beweis findet sich in T.M. Cover, J.A. Thomas, *Elements of Information Theory*.