.

Wahrscheinlichkeitstheorie

U. Mortensen
.
Fachbereich Psychologie und Sportwissenschaften, Institut III
Westfälische Wilhelms-Universität
Münster
Fliednerstr. 21

Letzte Korrektur: 26. 10. 2015

Inhaltsverzeichnis

1 Elemente der Kombinatorik **7**

1.1 Die Anzahl von Paaren; Multiplets 8

1.2 Geordnete Mengen; Permutationen 10

1.2.1 Ziehen mit Zurücklegen: 10

1.2.2 Ziehen ohne Zurücklegen: 10

1.3 Teilkombinationen 13

1.3.1 Kombinationen ohne Wiederholung 13

1.3.2 Kombinationen mit Wiederholung 16

1.4 Multinomialkoeffizienten 19

2 Der Begriff der Wahrscheinlichkeit **21**

2.1 Einführung 21

2.1.1 Zufall und Wahrscheinlichkeit 21

2.1.2 Frühe Ansätze zur Definition von Wahrscheinlichkeiten . . . 23

2.1.3 Die klassische Definition (Laplace) 26

2.1.4 Die statistische Definition (v. Mises) 27

2.2 Kolmogorovs axiomatischer Ansatz 29

2.2.1 Einige Aspekte der reellen Zahlen 29

2.2.2 Maße auf Mengen 33

2.2.3 Ereignisse und Ereignisalgebren 41

2.2.4 Kolmogorovs axiomatische Definition 48

2.3 Bedingte Wahrscheinlichkeiten 52

2.4 Stochastische Unabhängigkeit 63

3 Zufällige Veränderliche und ihre Verteilungen **71**

3.1 Allgemeine Definitionen 71
3.2 Mehrdimensionale Verteilungen 80
3.3 Bedingte Verteilungen und Hazardfunktionen 82
3.4 Funktionen von zufälligen Veränderlichen 86
3.4.1 Funktionen einer zufälligen Veränderlichen 86
3.4.2 Eine Funktion zweier zufälliger Veränderlicher 90
3.5 Erwartungswerte 94
3.5.1 Der eindimensionale Fall 94
3.5.2 Bedingte Erwartungswerte 102
3.5.3 Der mehrdimensionale Fall 102
3.6 Erzeugende Funktionen 104
3.6.1 Allgemeine Definition 104
3.6.2 Erzeugende Funktionen und die Momente einer Verteilung . 105
3.7 Charakteristische Funktionen 107

4 Diskrete Wahrscheinlichkeitsverteilungen 113
4.1 Die Gleichverteilung 113
4.2 Die Bernoulli-Verteilung 115
4.3 Die Binomialverteilung 115
4.3.1 Definition 115
4.3.2 Verallgemeinerung 118
4.4 Die Multinomialverteilung 119
4.5 Die geometrische Verteilung 120
4.6 Die Poisson-Verteilung 125
4.7 Die Negative Binomialverteilung 130
4.8 Die hypergeometrische Verteilung 135
4.9 Zusammengesetzte Verteilungen 141
4.10 Produkte und Quotienten 144
4.10.1 Die Verteilung eines Quotienten von natürlichen Zahlen . . 144
4.10.2 Die Verteilung eines Produktes von natürlichen Zahlen . . . 145

5 Beispiele für stetige Verteilungen 147
5.1 Die Gleichverteilung 147

5.2 Die Exponentialverteilung 149
5.2.1 Definition 149
5.2.2 Die Exponentialverteilung als Grenzverteilung 151
5.2.3 Zum Begriff der reinen Zufälligkeit 152
5.3 Die Gamma-Verteilung 153
5.4 Die Normalverteilung 157
5.5 Die logistische Verteilung 162
5.6 Die Weibull-Verteilung 165
5.6.1 Allgemeine Definition 165
5.6.2 Die Verteilung extremer Werte 167
5.6.3 Verweildauern und Warte- bzw. Reaktionszeiten 169
5.6.4 Dosiswirkungen und Psychometrische Funktionen 170
5.7 Die log-Normalverteilung 172
5.7.1 Definition der Verteilung 172
5.7.2 Mechanismen, die auf Log-Normalverteilungen führen . . . 174
5.8 Die Cauchy-Verteilung 178
5.9 Die Beta-Verteilungen 180
5.10 Stichprobenverteilungen 181
5.10.1 Die χ^2-Verteilung 181
5.11 Die mehrdimensionale Normalverteilung 184
5.11.1 Der allgemeine Fall 184
5.11.2 Der Fall n = 2 186
5.11.3 Lineare Transformationen und Randverteilungen 189

6 Ungleichungen und Grenzwertsätze 191
6.1 Einige Ungleichungen 191
6.2 Gesetze der großen Zahlen 193
6.3 Grenzwertsätze 197
6.3.1 Der Poissonsche Grenzwertsatz 197
6.3.2 Der DeMoivre-Laplacesche Grenzwertsatz 198
6.3.3 Der Zentrale Grenzwertsatz 201

7 Die Verteilung extremer Werte 203

7.1 Einführung 203
7.2 Extremwertverteilungen als Grenzverteilungen 211
7.2.1 Grenzverteilungen für Maxima 211
7.2.2 Beispiele für die Verteilung von Maxima 216
7.2.3 Grenzverteilungen der Minima 221
7.2.4 Beispiele für die Verteilung von Minima 224
7.3 Unabhängige, nichtidentisch verteilte Variablen 226

8 Elemente der Theorie stochastischer Prozesse 227
8.1 Allgemeine Begriffe 227
8.2 Zählprozesse 238
8.2.1 Der homogene Poisson-Prozeß 240
8.2.2 Der inhomogene Poisson-Prozeß 244
8.2.3 Der zusammengesetzte Poisson-Prozeß; Shot-Noise 248
8.2.4 Geburts- und Todesprozesse 251
8.3 Markov-Ketten 259
8.3.1 Grundlegende Definitionen 259
8.3.2 Verweildauern und Semi-Markov-Prozesse 266
8.3.3 Semi-Markov-Prozesse 267
Literatur 268
Index 272

Kapitel 1

Elemente der Kombinatorik

Der Chevalier Gombaud Antoine de Méré (1607 – 1684) schätzte die guten Dinge im Leben, zu denen er auch das Würfelspiel zählte. Er fragte sich, wie die Kombinationen verschiedener Augenzahlen hinsichtlich der Gewinnchancen zu bewerten seien, und ging im Jahre 1654 seinen Freund Blaise Pascal (1623 – 1662) um Rat an.

Pascal dachte nach und diskutierte die Sache mit dem Rechtsanwalt und Mathematiker Pierre de Fermat (1607 – 1665). Sie mußten herausfinden, wie man auf systematische Weise die Anzahl der möglichen Punktzahlen zu bestimmten Stadien eines Spiels bestimmt, und de Mérés Fragen erwiesen sich als schwieriger als zunächst vermutet. Bei der Beantwortung dieser Fragen erfanden Pascal und Fermat die *Kombinatorik*. In der Kombinatorik beschäftigt man sich mit der Frage, auf wieviele Weisen Objekte angeordnet werden können, wieviele Teilmengen bestimmten Umfanges aus einer gegebenen Menge gebildet werden können, etc. Die Frage nach der Berechnung von Wahrscheinlichkeiten kann vielfach auf Fragen der Kombinatorik zurückgeführt werden; die Kombinatorik wurde zur Grundlage von so zentralen Wissenschaften wie der Statistischen Mechanik und der Festkörperphysik, und in der *Kombinatorischen Chemie* findet sie Anwendung bei der Suche nach neuen Medikamenten, bei der Millionen chemischer Komponenten kombiniert werden (Mathews (1996)). Auch bei der Planung von psychologischen Experimenten sowie bei der Formulierung psychologischer Theorien spielt die Kombinatorik oft eine wichtige Rolle, – unter anderem dann, wenn Resultate der Statistischen Mechanik zur Modellierung neuronaler Netzwerke herangezogen werden. über die Kombinatorik[1] ergeben sich also Querverbindungen zwischen so verschieden erscheinenden Wissenschaften wie Physik und Psychologie.

[1] Aber nicht *nur* über die Kombinatorik!

1.1 Die Anzahl von Paaren; Multiplets

Paare

Gegeben[2] seien zwei Mengen $A = \{a_1, \cdots, a_m\}$ und $B = \{b_1, \cdots, b_n\}$ mit je m bzw. n Elementen. Dann lassen sich die Paare (a_1, b_1), (a_1, b_2), (a_1, b_3), $\cdots$, (a_2, b_1), $(a_2, b_2), \cdots, (a_m, b_n)$ bilden; da jedes a_i mit jedem b_j gepaart werden kann, gibt es offenbar

$$N = m \cdot n \tag{1.1}$$

mögliche Paare (a_i, b_j), $i = 1, \cdots, m$; $j = 1, \cdots, n$.

Multiplets

Gegeben seien r Mengen A_1, $\cdots$, A_r; die Mengen dürfen identisch sein, d.h. $A_i = A_j$ ist zugelassen. Aus jeder Menge A_j werde eine Folge f_j von Elementen gebildet:

$$\begin{array}{lccc} f_1: & a_1, & \cdots, & a_{n_1} \\ f_2: & b_1, & \cdots, & b_{n_2} \\ & & \cdots & \\ f_r: & x_1, & \cdots, & x_{n_r} \end{array}$$

Aus jeder der Folgen f_j kann nun wieder ein Element ausgewählt werden, und aus diesen Elementen wird eine neue Folge $S_i : a_{i_1} b_{i_2} \cdots x_{i_r}$ gebildet. Es gibt genau

$$N_r = n_1 n_2 \cdots n_r \tag{1.2}$$

Folgen S_i, $i = 1, \cdots, N_r$. Denn für $r = 2$ ist N_r gerade die Anzahl möglicher Paare, $N_2 = n_1 n_2$, für $r = 3$ ist N_3 die Anzahl der möglichen Paare, die sich aus der Menge der möglichen Folgen aus 2 Elementen und einem dritten Element bilden lassen, etc.

Der auf (1.2) führende Ansatz läßt sich für viele praktische Anwendungen so in ein Paradigma über Entscheidungen mit mehreren Möglichkeiten umformulieren (Feller (1968)): es sollen r aufeinanderfolgende Wahlen getroffen werden, wobei für die k-te Wahl/Entscheidung, $1 \leq k \leq r$, gerade n_k Möglichkeiten existieren. Dann gibt es insgesamt $n_1 n_2 \cdots n_r$ mögliche Ausgünge dieser Entscheidungsfolge. Man hat dann insbesondere den

Spezialfall: Es sollen n Objekte in k Kategorien einsortiert werden. Dabei sollen mehrere Objekte in ein und dieselbe Kategorie sortiert werden dürfen. Für das erste Objekt gibt es k Möglichkeiten. Für das zweite ebenfalls, und jede Möglichkeit, das erste Objekt zu sortieren, kann mit jeder Möglichkeit, das zweite Objekt zu sortieren, kombiniert werden. Für die ersten beiden Objekte gibt es

[2] Die folgende Darstellung ist an der von Feller (1968) orientiert.

also $k \cdot k = k^2$ Möglichkeiten. Für das dritte Objekt gibt es wieder k Möglichkeiten, so dass es für drei Objekte insgesamt k^3 Möglichkeiten gibt. Es folgt, dass es für n Objekte genau

$$N_k = k^n \tag{1.3}$$

Möglichkeiten, sie in k Kategorien einzusortieren, geben muß.

Beispiel 1.1.1 100 Personen sollen (i) hinsichtlich ihres Geschlechts, (ii) ihres Zivistandes (verheiratet — unverheirated), (iii) ihres Berufs klassifiziert werden. Die Berufe werden dabei in 10 Klassen eingeteilt; angegeben werden soll die Klasse, in die der jeweilige Beruf gehört.

Die Kategorien (Geschlecht, Zivilstand, Berufsklasse) spielen hier die Rolle der Mengen A_i in der Multiplet-Betrachtung; die Subkategorien (männlich, weiblich, verheiratet, ledig etc) sind die Elemente. Man zählt alle Elemente explizit auf und haben $n_1 = 2$, $n_2 = 2$, $n_3 = 10$. Dann gibt es insgesamt $N_r = N_3 = n_1 n_2 n_3 = 2 \cdot 2 \cdot 10 = 40$ Klassen. Hierbei sind also (männlich, verheiratet, Tischler), (männlich, verheiratet, Buchhalter), (weiblich, unverheiratet, Akademikerin) etc die möglichen Klassen, in die eine Person einsortiert werden kann. Es gibt 40^{100} mögliche Zuordnungen der Personen zu diesen 40 Klassen. □

Beispiel 1.1.2 (Bälle in Zellen) Es sollen n Bälle in k Zellen einsortiert werden. Für jeden der n Bälle gibt es k Möglichkeiten, also gibt es k^n Möglichkeiten, die Bälle auf die Zellen zu verteilen.

Diese Situation ist ein Paradigma, auf das viele Fragestellungen zurückgeführt werden können. Man kann z.B. nach der Menge der möglichen Resultate fragen, wenn ein Würfel mit sechs Seiten n-mal geworfen wird. Bei jedem Wurf liegt genau eine Seite oben. Man kann nun den Würfel als "Ball" auffassen und die Seiten des Würfels als "Zelle". Zeigt der Würfel nach einem i-ten Wurf eine bestimmte, etwa die j-te Seite, so entspricht dies dem Fallen des i-ten Balles in die j-te Zelle. Bei jedem Wurf stehen alle Zellen zur Verfügung. Bei n Würfen gibt es also 6^n mögliche Folgen.

Ein Hochhaus habe 10 Stockwerke, zwischen denen ein Aufzug verkehrt, der an jedem Stockwerk hält. Der Aufzug beginnt seine Reise mit 8 Passagieren. Auf wieviele Weisen können sich die 8 Passagiere auf die Stockwerke verteilen? Diese Frage ist offenbar äquivalent der Frage, auf wieviele Weisen man 8 Bälle auf 10 Zellen verteilen kann: denn für jede Person gibt es 10 Möglichkeiten. den Aufzug zu verlassen, also auf $k^n = 10^8$ verschiedene Weisen.

Man wählt eine Stichprobe von 10 Personen. Auf wieviele Weisen können die Geburtstage dieser Personen über die Tage eines Jahres verteilt sein? Man setze als Annäherung die Anzahl der Tage im Jahr mit 365 fest. Die $n = 10$ Geburtstage sind jetzt die Bälle, die $k = 365$ Tage des Jahres sind die Zellen. Also gibt es 365^{10} mögliche Anordnungen der 10 Geburtstage in einem Jahr. □

1.2 Geordnete Mengen; Permutationen

Es sei wieder eine Menge $A = \{a_1, \cdots, a_n\}$ von n Elementen gegeben. Ein geordnetes Arrangement $a_{j1}, \cdots, a_{jr}$ von $r \leq n$ Elementen aus A heißt geordnete Stichprobe von r Elementen, oder auch r-Permutation. Damit ist eine r-Permutation ein Element des r-fachen Cartesischen Produktes von A mit sich selbst, d.h.

$$A^r = \underbrace{A \times A \times \cdots \times A}_{r\text{-mal}}$$

$$A^r = \{(a_{j1}, \cdots, a_{jr}); \quad a_{js} \in A, \quad s = 1, \cdots, r\}$$

Man hat nun zwei Möglichkeiten: man kann die Stichprobe, d.h. die r-Permutation, durch Ziehen mit Zurücklegen oder durch Ziehen ohne Zurücklegen bilden; der Ausdruck Ziehen erinnert an das Ziehen einer Kugel aus einer Urne.

1.2.1 Ziehen mit Zurücklegen:

Die Elemente werden nach jeder "Wahl" — nach jedem Ziehen aus der Urne — wieder zurückgelegt und stehen deshalb für die nächste Wahl wieder zur Verfügung. Nach (1.2) gibt es dann

$$N_r = |A^r| = \underbrace{n \cdot n \cdots n}_{r\text{-mal}} = n^r \tag{1.4}$$

mögliche Folgen $(a_{j1}, \cdots, a_{jr})$, wobei mit $|A^r|$ wie üblich die Anzahl der Elemente in A^r bezeichnet wird. In den Folgen können Elemente wiederholt auftreten; die Folgen heißen deswegen auch r-Permutation mit Wiederholung.

> **Anmerkung:** Das Paradigma der Bälle in Zellen kann auch hier angewendet werden: Die Menge A ist die Menge der Zellen (Kategorien), und man wirft einen Ball mehrfach in diese Menge, wobei er jeweils in genau eine Zelle trifft. Nach jedem Wurf kann eine getroffene Zelle wieder getroffen werden, die Zellen werden also "mit Zurücklegen gezogen". Hier ist n die Anzahl der Zellen und r die Anzahl der Würfe; (1.3) und (1.4) haben dann also die gleiche Bedeutung.

1.2.2 Ziehen ohne Zurücklegen:

Der Fall $n > r$: Für das erste Element hat man n mögliche Wahlen. Für das zweite gibt es noch $n - 1$ Möglichkeiten, so dass es nach (1.1) $n(n-1)$ Möglichkeiten für die ersten beiden Wahlen gibt. Für die dritte Wahl bleiben noch $n - 2$ Möglichkeiten, von denen jede mit jedem aus den ersten beiden Wahlen resultierenden Paaren kombiniert werden kann. Nach (1.1) gibt es dann $n(n-1)(n-2)$

mögliche Folgen für die ersten drei Elemente. Bezeichnet man mit $(n)_r$ die Anzahl der möglichen Folgen von r Elementen, so hat man

$$(n)_r = n(n-1)(n-2)\cdots(n-r+1). \tag{1.5}$$

Bei diesen Folgen sind keine zwei Elemente identisch, und sie heißen deshalb auch *r-Permutationen ohne Wiederholung.*

Beispiel 1.2.1 Wieviele Verteilungen der 10 Geburtstage aus Beispiel 1.1.2 in einem Jahr mit 365 Tagen gibt es derart, dass keine zwei oder mehr Personen am gleichen Tag Geburtstag haben?

Eine solche Verteilung ist eine 10-Permutation ohne Wiederholung. Denn aus den $n = 365$ Tagen "ziehen" wir den ersten Geburtstag (durch Wahl einer Person) und haben dafür n Möglichkeiten. Für den zweiten Geburtstag haben wir noch $n-1$ Möglichkeiten, etc, so dass sich insgesamt

$$(365)_{10} = 365(365-1)\cdots(365-10+1)$$

mögliche Verteilungen ergeben. □

Der Spezialfall n = r:

Es sei insbesondere $r = n$. Aus (1.5) erhält man sofort

$$n! := (n)_n = n(n-1)(n-2)\cdots 2\cdot 1; \tag{1.6}$$

$n!$ wird dabei als "n-Fakultät" gelesen. $n!$ ist die Anzahl der möglichen Anordnungen von n Elementen auf n Plätzen, oder die Anzahl der möglichen Reihenfolgen von n Elementen. Eine solche Anordnung wird auch *Permutation* genannt. Für $n = 0$ setzt man

$$0! = 1 \tag{1.7}$$

Dieser Ausdruck ist nicht aus (1.6) ableitbar; die Setzung (1.7) ist aber so gewählt, dass alle Formeln, in denen $n!$ auftritt, für den Fall $n = 0$ sinnvoll bleiben.

Für größere Werte von n ist $n!$ langwierig zu berechen. Man macht dann von der *Stirlingschen Formel*

$$n! \approx n^{n+1/2}e^{-n}\sqrt{2\pi} \tag{1.8}$$

Gebrauch. Den Beweis für diese Näherungsformel findet man u.a. in Feller (1968).

Die vorangegangenen Ergebnisse können zur besseren übersicht wie folgt zusammengefaßt werden:

Satz 1.2.1 *In einer Population von n Elementen gibt es bei der Stichprobenbildung mit Zurücklegen gerade n^r Stichproben mit Zurücklegen vom Umfang r, und bei Stichprobenbildung ohne Zurücklegen gibt es $(n)_r$ mögliche Stichproben.*

Beispiel 1.2.2 In einem Experiment zur Struktur des Gedächtnisses wird eine Versuchsperson (Vp) gebeten, die Namen von fünf amerikanischen Filmschauspielern zu nennen, so, wie sie ihr gerade einfallen. Der Zufall will es, dass die Vp gerade 100 amerikanische Filmschauspieler in ihrem Gedächtnis gespeichert hat. Es soll ein reines Zufallsmodell des Gedächtnisses angenommen werden; die Voraussagen aufgrund eines solchen Modelles liefern eine Art Maßstab, anhand dessen die Voraussagen von Modellen, in denen bestimmte Annahmen über die Struktur des Gedächtnisses formuliert werden, beurteilt werden können. Dieses Zufallsmodell wird durch die folgenden Annahmen charakterisiert: (i) das Gedächtnis der Vp kann mit einer Urne verglichen werden, in der die Namen gut durchmischt liegen; (ii) die Vp greift aufs Geratewohl in diese "Urne", um einen Namen zu wählen; (iii) hat sie ihn genannt, vergißt sie sofort, dass sie ihn genannt hat, und der Name steht für die folgende Wahl eines Namens wieder zur Verfügung. Nach Versuchsanweisung besteht die Aufgabe der Vp darin, 5 Namen zu nennen. Dem Zufallsmodell zufolge besteht die Nennung der 5 Namen in der Bildung einer Stichprobe vom Umfang 5 mit Zurücklegen. Nach Satz 1.1 kann sie dann $100^5 = (10^2)^5 = 10^{10}$ Stichproben von Namen bilden. Modifiziert man die Annahme (iii) derart, dass die Vp einen bereits genannten Namen als genannt erinnert und dementsprechend für eine weitere Wahl nicht mehr in Betracht zieht, so bildet sie eine Stichprobe ohne Zurücklegen und nach Satz 1.1 kann sie nurmehr

$$(100)_5 = 100 \cdot 99 \cdot 98 \cdot 97 \cdot 96 = 9034502400$$

Stichproben von fünf Namen erzeugen. □

Beispiel 1.2.3 Beim TAT (Thematic Apperception Test) werden der zu testenden Person nacheinander 10 Karten vorgelegt, auf denen je ein Bild zu sehen ist, das die Person interpretieren soll. Der Vermutung nach drückt sich der seelische Zustand der Person in ihren Bildauslegungen aus. Es besteht nun Grund zu der Annahme, dass die Reihenfolge, in der die Bilder der Person vorgelegt werden, einen Effekt auf die von der Person gegebenen Interpretationen hat. Ein Psychologe kommt auf die Idee, alle möglichen Reihenfolgen der Karten auf ihren Effekt auf die gegebenen Interpretationen zu überprüfen. Nach (1.5) hat er $n! = 3628800$ mögliche Reihenfolgen in seiner Untersuchung zu erproben. Wird er dies tun? □

Beispiel 1.2.4 Bei einer Familienfeier gilt es, 10 Personen an einem runden Tisch zu plazieren. Dabei sollen aber, so denkt sich der Gastgeber, die Base I. und der Vetter J. nicht nebeneinander sitzen, da sonst über den Ausgang der Feier keine gesicherte Aussage mehr möglich ist. Zwischen wievielen möglichen Anordnungen der Gäste, bei denen Base I. und Vetter J. nicht nebeneinander sitzen, kann der Gastgeber wählen?

Zur Beantwortung dieser Frage betrachtet man zunächst die möglichen Anordnungen, bei denen Base I. und Vetter J. nebeneinander sitzen. Dazu numeriert man zunächst die Plätze durch. Base I. möge nun auf Platz 1 und Vetter J. auf

Platz 2 sitzen. Die beiden sitzen jetzt benachbart, und die restlichen 8 Gäste können irgendwie angeordnet werden; dafür gibt es nach (1.5) 8! Anordnungen. Ebenso kann Base I auf Platz 2 und Vetter J. auf Platz 1 sitzen, und wieder gibt es 8! mögliche Anordnungen für die restlichen Gäste. Also gibt es insgesamt $2 \cdot 8!$ Anordnungen, bei denen Base I. und Vetter J. benachbart auf den Plätzen 1 und 2 sitzen. Natürlich können die beiden auch auf irgendwelchen anderen nebeneinander positionierten Plätzen sitzen, um benachbart zu sitzen. Benachbarte Plätze sind nun also (1,2). (2,3), (3,4), (4,5), (5,6), (6,7), (7,8), (8,9), (9,10), (10,1), also gibt es insgesamt $10 \cdot 2 \cdot 8!$ Anordnungen, bei denen Base I. und Vetter J. nebeneinander sitzen. Insgesamt gibt es aber 10! überhaupt mögliche Anordnungen der Gäste; also gibt es $10! - 10 \cdot 2 \cdot 8! = 2822400$ mögliche Anordnungen, bei denen der Ausgang der Feier vorhersagbar bleibt. □

1.3 Teilkombinationen

1.3.1 Kombinationen ohne Wiederholung

Es sei wieder eine Menge $A = \{a_1, \cdots, a_n\}$ gegeben; A stelle eine Population dar. Gesucht ist nun die Anzahl $C(n, r)$ der möglichen Teilpopulationen bzw. Stichproben aus A mit $r \leq n$ verschiedenen Elementen. Dabei soll es auf die Anordnung der Elemente nicht ankommen. Die Elemente der Stichprobe können also beliebig auf r Plätze verteilt werden. Eine solche Teilmenge von A heißt *r-Kombination der n Elemente (von A) ohne Wiederholung* .

Es sei nun eine Stichprobe von $r \leq n$ verschiedenen Elementen aus A gegeben. Die Elemente der Stichprobe können von 1 bis r durchnummeriert werden. Nach (1.6) gibt es gerade $r!$ mögliche Reihenfolgen für diese Durchnummerierung. Werden die Elemente der Stichprobe nacheinander aus der Population A gezogen, so gibt es $r!$ mögliche Reihenfolgen, in der diese Stichprobe gebildet werden kann. Betrachten wir für den Moment Stichproben, die aus den gleichen Elementen bestehen, die aber in verschiedener Reihenfolge gezogen wurden, als verschieden an, so gibt es $C(n, r) \cdot r!$ verschiedene Stichproben vom Umfang r. Die Stichproben sind ohne Zurücklegen gebildet worden, und nach (1.5) gibt es dann $(n)_r$ die Anzahl der überhaupt möglichen Stichproben vom Umfang r. Dann gilt aber $(n)_r = C(n, r) \cdot r!$, und hieraus folgt sofort

$$C(n, r) = \frac{(n)_r}{r!}. \tag{1.9}$$

Die Anzahl $C(n, r)$ wird auch mit

$$\binom{n}{r}$$

bezeichnet; man spricht dies "n über r".

Eine Teilmenge oder Stichprobe von r Elementen aus einer Population von n Elementen kann also auf $C(n,r) = \binom{n}{r}$ verschiedene Weisen gebildet werden, wenn es auf die Reihenfolge nicht ankommt. Solche eine Teilmenge ist eindeutig durch die Menge der $n-r$ Elemente von A, die nicht zu der in Frage stehenden Stichprobe gehören, bestimmt. Daraus folgt sofort, dass es so viele Teilmengen/Stichproben vom Umfang $n-r$ geben muß, wie es Teilmengen/Stichproben vom Umfang r gibt. Damit hat man

$$C(n,r) = C(n,n-r), \tag{1.10}$$

bzw.

$$\binom{n}{r} = \binom{n}{n-r} \tag{1.11}$$

Multipliziert man nun den Zähler und den Nenner von $(n)_r/r!$ mit $(n-r)!$, so erhält man

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}, \quad r \geq 0. \tag{1.12}$$

wegen (1.6) folgt dann

$$\binom{n}{0} = \frac{n!}{0!n!} = 1. \tag{1.13}$$

Die Größen $C(n,r) = \binom{n}{r}$ sind als *Binomialkoeffizienten* bekannt. Es gilt ja

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^n b^{n-k}.$$

Setzt man nun $a = b = 1$, so erhält man sofort

$$2^n = \sum_{k=0}^{n} \binom{n}{k} \tag{1.14}$$

d.h. die Summe der Binomialkoeffizienten ist gleich 2^n. Da aber $\binom{n}{k}$ gerade die Anzahl der möglichen Teilmengen/Stichproben vom Umfang r aus einer Menge vom Umfang n ist, so liefert (1.14) die Anzahl der überhaupt möglichen Teilmengen, die aus A gebildet werden können. Die Menge der überhaupt möglichen Teilmengen einer Menge A heißt die *Potenzmenge* $\mathcal{A}$ *von* A; es ist also $|\mathcal{A}| = 2^n$. Hierbei wird die leere Menge, d.h. die Menge, die gar kein Element enthält, mitgezählt; in diesem Fall ist $r = 0$. Die Anzahl der Teilmengen mit mindestens einem Element ist also gleich $2^n - 1$.

Es sei noch darauf hingewiesen, dass (1.9) das Bildungsgesetz für das Pascalsche Dreieck liefert:

$$\begin{array}{c} 1 \\ 1\ 1 \\ 1\ 2\ 2\ 1 \end{array}$$

1 3 4 3 1
1 4 7 7 4 1
etc.

Die Zahl $\binom{n}{r}$ steht an der r-ten Stelle der n-ten Zeile.

Beispiel 1.3.1 Ein Fragebogen enthalte 10 Fragen, die entweder mit "ja" oder mit "nein" beantwortet werden sollen. Die Antworten einer Person lassen sich dann als eine Folge von "ja"- und "nein"-Antworten codieren. Wieviele mögliche Folgen mit (i) 3 "ja"-Antworten, (ii) 5 "ja"-Antworten und (iii) 7 "ja"-Antworten gibt es?

Man kann eine Folge von Antworten als eine Teilmenge der 10 Fragen auffassen, nämlich als eine Teilmenge, die aus den Fragen besteht, die mit "ja" beantwortet wurden. Die Anzahl der Teilmengen mit r Elementen ist aber $\binom{n}{r}$; für (i) erhalten man also

$$\binom{n}{r} = \binom{10}{3} = \frac{10 \cdot 9 \cdot 8}{1 \cdot 2 \cdot 3} = 120.$$

Für (ii) folgt

$$\binom{n}{r} = \binom{10}{5} = \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5} = 252$$

und für (iii) hat man

$$\binom{n}{r} = \binom{10}{7} = \binom{10}{3} = 120;$$

hierbei ist von (1.10) Gebrauch gemacht worden. □

Beispiel 1.3.2 In einem Gremium soll der Vorsitzende gewählt werden. Es herrscht anscheinend übereinstimmung, dass ein bestimmter Kandidat das Amt übernehmen soll, und die vorgeschriebene geheime Wahl scheint nur eine Formsache zu sein. Gleichwohl ergibt die Auszählung der Stimmen, dass dem Kandidaten drei Stimmen fehlen. Alle Mitglieder des Gremiums bis auf fünf bekennen freimütig, für den Kandidaten gestimmt zu haben, und man vermutet, dass drei von diesen fünf Mitgliedern nicht für den Kandidaten gestimmt haben. Wieviele mögliche Dreiergruppen lassen sich aus dieser Menge mit fünf Elementen bilden? Offenbar

$$C(5,3) = \binom{5}{3} = \frac{5 \cdot 4 \cdot 3}{1 \cdot 2 \cdot 3} = 10$$

Es gibt also 10 mögliche Koalitionen gegen den Kandidaten. □

Beispiel 1.3.3 Ein Versuch wird n-mal wiederholt. Jeder Versuch endet mit einem "Erfolg" oder mit einem "Misserfolg". Ein Erfolg wird mit einer "1", ein Misserfolg wird mit einer "0" kodiert. Wieviele mögliche Folgen mit genau $k \leq n$ Einsen und $n - k$ Nullen gibt es?

Die Frage kann auf die nach der Anzahl der Teilmengen mit genau k Elementen zurückgeführt werden: die Elemente einer Teilmenge können mit der "1" markiert werden, die der Restmenge mit der "0". Es gibt aber gerade $\binom{n}{k}$ mögliche Teilmengen mit k von n möglichen Elementen, also gibt es $\binom{n}{k}$ mögliche Folgen mit genau k Einsen und $n-k$ Nullen. □

Beispiel 1.3.4 In einer Untersuchung zur Struktur des Langzeitgedächtnisses wird einer Versuchsperson (Vp) eine Geschichte vorgelesen. Im Anschluß daran wird die Vp gebeten, die fünf Blumennamen zu nennen, die insgesamt in der Geschichte vorgekommen sind; die Vp zu Beginn des Experimentes nicht instruiert worden, insbesondere auf Blumennamen zu achten. Angenommen, die Vp sucht nun gemäß dem in Beispiel 1.2.2 angedeuteten Gedächtnismodell fünf Blumennamen nach den Zufall aus dem Gedächtnis, wobei sie insgesamt n Blumennamen im Gedächtnis gespeichert haben möge. Wieviele Möglichkeiten hat sie, fünf Namen auszuwählen, von denen gerade k, $0 < k \leq 5$, zu den tatsächlich in der Geschichte genannten gehören?

Der Fall $k = 5$ ist offenbar der einfachste Fall, denn es gibt nur eine solche Möglichkeit.

Nun sei $0 < k < 5$. Dies bedeutet, dass sie von den 5 Namen gerade k wählt. Es kommt nicht darauf an, welche k dieser 5 Wörter sie zieht, und demnach gibt es $\binom{5}{k}$ Möglichkeiten, eine Teilmenge von k Elementen aus der Menge von 5 möglichen Namen zu ziehen. Da die Vp aber insgesamt 5 Namen aus dem Gedächtnis wählt, zieht sie noch $5-k$ "falsche" (d.h. nicht zur Menge der fünf in der Geschichte vorkommenden Namen gehörenden) Namen. Die gezogene Menge von falschen Namen ist ein Element der Menge der Teilmenge mit $5-k$ falschen Namen. Wir benötigen die Anzahl der Elemente in dieser Teilmenge, denn jede der $\binom{5}{k}$ Möglichkeiten, k "richtige" Namen zu nennen, kann mit jeder Möglichkeit, $5-k$ falsche zu nennen, auftreten. Insgesamt sind $n-5$ Namen "falsch". Dann gibt es gerade $\binom{n-5}{5-k}$ Möglichkeiten, $5-k$ falsche Namen aus dieser Menge auszuwählen. Mithin gibt es gerade $N_k = \binom{5}{k}\binom{n-3}{5-k}$ Möglichkeiten, k richtige und $5-k$ falsche Namen zu nennen. Es sei angemerkt, dass dieser Ausdruck für N_k die Spezialfälle $k = 0$ und $k = 5$ enthält. □

1.3.2 Kombinationen mit Wiederholung

Es sei A wieder eine n-elementige Menge, aus der r-mal jeweils ein Element gezogen wird; das Element kann beliebig gewählt werden. Ein bestimmtes Element von A kann also mehrfach gezogen werden. Dabei soll es auf die Anordnung der Elemente in der Stichprobe nicht ankommen. Solche Stichproben heißen *Kombinationen mit Wiederholung vom Umfang* r. Dieser Fall liegt z.B. dann vor, wenn die Elemente von A wieder Kategorien (Zellen) sind, in die r Objekte (Bälle) sortiert werden sollen. Es seien r_i, $i = 1, \cdots, n$ die Häufigkeiten, mit denen das

erste, zweite, etc Element von A gewählt wurde. Dann muß

$$r_1 + r_2 \cdots + r_n = r \tag{1.15}$$

gelten. Die Kombinationen vom Umfang r sind die verschiedenen n-Tupel $(r_1, \cdots, r_n)$, die der Gleichung (1.15) genügen. Es soll also nicht zwischen den Objekten, die klassifiziert werden, unterschieden werden, sondern zwischen den verschiedenen $(r_1, \cdots, r_n)$.

Ein Beispiel ist die Menge $A = \{1, 2, 3, 4\}$. Die Kombinationen mit Wiederholung vom Umfang $r = 2$ sind durch die Paare (1,1), (2,2), (3,3), (4,4), (1,2), (1,3), (1,4), (2,3), (2,4), (3,4) gegeben; zwischen den Paaren (1,2) und (2,1) etc wird also nicht unterschieden. In der Sprache der Zellen und Bälle heißt dies z. B., dass zwei Bälle in 4 Zellen sortiert werden sollen. Im Falle (1,1) sind beide Bälle in die Zelle "1" gelangt. Im Falle (1,2) ist der erste Ball in die erste Zelle, der zweite in die zweite sortiert worden. Da die Bälle nicht unterschieden werden (oder: da es auf die Unterschiede zwischen den Bällen nicht ankommt), ist dieses Ergebnis äquivalent dem Ergebnis (2,1), bei dem der erste Ball in die zweite und der zweite in die erste Zelle sortiert wurde. Die r_i können nur die Werte 0, 1 oder 2 annehmen. Man findet $K(n, r) = K(4, 2) = 10$ Kombinationen mit Wiederholung vom Umfang $r = 2$.

Allgemein gilt

$$K(n, r) = \binom{n + r - 1}{r}. \tag{1.16}$$

Beweis: Der Beweis dieser Aussage ist dem Buch von Plachky et al. (1983) entnommen.

Offenbar kommt es bei der Bildung einer Kombination mit Wiederholung vom Umfang r nicht auf die Werte a_i der in Frage stehenden Menge (Population) an, sondern nur auf ihre Indices i, $1 \leq i \leq n$. Deshalb genügt es, wie im vorangegangenen Beispiel die Menge $A = \{1, 2, \cdots, n\}$ zu betrachten.

Es sei nun A^r das r-fache Cartesische Produkt von A mit sich selbst. A^r ist also die Menge der möglichen r-Tupel von Elementen aus A. Es seien nun $(a_1, \cdots a_r) \in A^r$ und $(b_1, \cdots, b_r) \in A^r$ zwei Elemente (r-Tupel) aus A^r, und es werde die Klasse $[a_1, \cdots, a_r]$ wie folgt definiert: es gelte

$$[a_1, \cdots, a_r] = [b_1, \cdots, b_r]$$

genau dann, wenn sich die Tupel $(a_1, \cdots, a_r)$ und $(b_1, \cdots, b_r)$ nicht durch die Elemente, sondern nur durch deren Anordnung unterscheiden. Die beiden r-Tupel können also durch eine Permutation der Elemente ineinander überführt werden. Die Reihenfolge der Elemente in $[a_1, \cdots, a_r]$ soll also irrelevant sein. Im obigen Beispiel ist dann also $[1, 2] = \{(1, 2), (2, 1)\}$. Dann bedeutet es keine Einschränkung, wenn man $a_1 \leq a_2 \leq \cdots \leq a_r$ annimmt. Wir betrachten nun die Menge

$$A_n^r = \{(a_1, \cdots, a_r) | (a_1, \cdots, a_r) \in A^r\}$$

Dies ist die Menge der möglichen Klassen für r Elemente aus A. Für die Menge $A = \{1, 2, 3, 4\}$ ist die Menge A_n^r oben vollständig angegeben worden; sie enthält 10 Elemente.

Weiter werden die Mengen

$$M := \{1, 2, \cdots, n + r - 1\}, \quad \mathfrak{P}(M) := \{B | B \subset M, |B| = r\}$$

eingeführt. $\mathfrak{P}(M)$ ist die Menge der Teilmengen von M, die den Umfang r haben. Jeder Klasse $[a_1, \cdots, a_r]$ werde nun die Folge $a_1, a_2 + 1, a_3 + 3, \cdots, a_r + r - 1$ zugeordnet. Diese Folge besteht aus Elementen, die eine Teilmenge von M bilden, d.h. sie entspricht einem $B \in \mathfrak{P}(M)$. Diese Zuordnung von Klassen zu Teilmengen aus $\mathfrak{P}(M)$ ist offenbar umkehrbar eindeutig: jeder Klasse entspricht aufgrund der Zuordnungsregel genau eine Teilmenge, und jeder Teilmenge $B \in \mathfrak{P}(M)$ entspricht genau eine Klasse. Damit ist die Anzahl $|A_n^r|$ dieser Klassen gerade gleich der Anzahl der Teilmengen in $\mathfrak{P}(M)$. Diese Menge hat aber genau $n + r - 1$ Elemente, und nach (1.9) gibt es dann gerade

$$|\mathfrak{P}(M)| = \binom{n + r - 1}{r}$$

Teilmengen von $\mathfrak{P}(M)$, und damit ist die Behauptung (1.16) bewiesen. □

Beispiel 1.3.5 Es werden drei nicht unterscheidbare Würfel gleichzeitig geworfen. Wieviele unterscheidbare Resultate gibt es? Zeigen die Würfel W_1, W_2 und W_3 die Augenzahlen "1", "2" und "4", so ist dieses Ergebnis nicht unterscheidbar von dem, bei dem der Würfel W_2 die Augenzahl "1" und W_1 die Augenzahl "2" zeigt, denn die Würfel sind als nicht unterscheidbar vorausgesetzt worden. Als "Population" kann man hier die Menge $A = \{1, \cdots, 6\}$ wählen; A repräsentiert die möglichen Augenzahlen für einen Würfel. Das Werfen mit drei Würfeln kann als Bildung einer Stichprobe von drei Elementen aus A aufgefaßt werden, wobei die gezogenen Elemente jedesmal wieder zurückgelegt werden. Der Wurf dreier Würfel liefert also eine Kombination mit Wiederholung vom Umfang 3. Es gibt

$$A_n^r = A_6^3 = \binom{6 + 3 - 1}{3} = \frac{8 \cdot 7 \cdot 6}{1 \cdot 2 \cdot 3} = 56$$

mögliche unterscheidbare Resultate. □

Beispiel 1.3.6 Ein Psychotherapeut hat einem Klienten die Aufgabe gestellt, 30 Tage lang jeden Abend seine (des Klienten) durchschnittliche Befindlichkeit für diesen Tag auf einer Skala mit 5 Befindlichkeitskategorien zu beurteilen. Am Ende bestimmt der Therapeut die Häufigkeiten $r_1, \cdots, r_5$ und versucht, daraus etwas über die durchschnittliche Befindlichkeit des Klienten zu erfahren. Wieviele verschiedene Verteilungen $(r_1, \cdots r_5)$ gibt es? Offenbar

$$\binom{5 + 30 - 1}{30} = 1496$$

□

Zur Klärung des Begriffes der Kombination mit Wiederholung vom Umfang r mag es sinnvoll sein, noch einmal die in (1.3) gegebene Größe zu betrachten. Dort sollten n Objekte in k Kategorien sortiert werden, und nach (1.3) ist die Anzahl der Möglichkeiten, dies zu tun, durch $N_k = k^n$ gegeben. Bei den Kombinationen mit Wiederholung vom Umfang r wird eine ähnliche Situation betrachtet: r Objekte werden auf n Kategorien verteilt. Die Gesamtzahl möglicher Anordnungen der r Objekte in den n Kategorien ist, gemäß (1.3), durch n^r gegeben. Zwei Anordnungen werden dabei als verschieden betrachtet, wenn sie sich auch nur um die Zuordnung zweier Objekte unterscheiden, auch wenn die Häufigkeiten $r_1, \cdots, r_n$ sich nicht unterscheiden. Bei den Kombinationen mit Wiederholung vom Umfang r dagegen ist man nicht an der unterschiedlichen Zuordnung der Objekte interessiert, sondern an der Anzahl der verschiedenen n-Tupel $(r_1, \cdots, r_n)$.

1.4 Multinomialkoeffizienten

Gegeben sei eine Menge $A = \{a_1, \cdots, a_n\}$ mit n Elementen, die nach k Kategorien sortiert werden sollen. Dabei mögen r_1 Elemente in die erste Kategorie, r_2 in die zweite und schließlich r_k in die k-te Kategorie fallen; dabei gilt natürlich $r_1 + r_2 + \cdots + r_k = n$, denn es sollen ja alle Elemente von A sortiert werden. Es darf dabei vorkommen, dass bestimmte Kategorien gar nicht besetzt werden, dass also $r_j = 0$ für einige j, $0 \leq j < n$. Die Frage ist nun,auf wieviel verschiedene Weisen die n Elemente auf die k Kategorien aufgeteilt werden können, so dass die Besetzungshäufigkeiten stets r_1, $r_2, \cdots, r_k$ sind.

Es sei $N(r_1, \cdots, r_k)$ die gesuchte Anzahl; die $N(r_1, \cdots, r_k)$ heißen *Multinomialkoeffizienten.*

Satz 1.4.1 *Es gilt*

$$N(r_1, \cdots, r_k) = \frac{n!}{r_1! r_2! \cdots r_k!} \tag{1.17}$$

Beweis: Um (1.17) einzusehen, beginnt man mit der Auswahl der Elemente für die erste Kategorie. Eine Stichprobe von r_1 Elementen aus n kann auf $C(r_1, n) = \binom{n}{r_1}$ verschiedene Weisen erfolgen. Für die Auswahl der Elemente für die zweite Kategorie sind dann noch $n - r_1$ Elemente übrig, und eine Stichprobe von r_2 Elementen kann aus ihnen auf $C(n - r_1, r_2) = \binom{n-r_1}{r_2}$ verschiedene Weisen erfolgen. Die ersten beiden Stichproben können insgesamt auf $C(r, r_1)C(n - r_1, r_2)$ verschiedene Weisen gebildet werden. Nachdem man schließlich die $(k-1) - te$ Stichprobe gebildet hat, verbleiben noch $r_k = n - r_1 - \cdots - r_{k-1}$ Elemente von A, die in die k-te Kategorie fallen. Also muß

$$N(r_1, \cdots, r_k) = C(n, r_1)C(n - r_1, r_2) \cdots C(n - r_1 - \cdots - r_{k-2}, r_{k-1})$$

sein, – bei der Wahl der Elemente für die k-te Kategorie ist man ja bereits festgelegt. Setzt man die entsprechenden Ausdrücke für die Kombinationen ohne Wiederholung ein, so ergibt sich (1.17). □

Beispiel 1.4.1 (Q-sort Methode nach Stephenson (1953)) Gelegentlich kann es interessant sein, eine Person sich selbst oder andere beurteilen zu lassen. So kann man eine Person sich selbst vor und nach einer Therapie beurteilen lassen, oder um die Struktur von Stereotypen zu erfassen kann man eine Person Mitglieder bestimmter Gruppen beurteilen lassen. Bei dieser Beurteilung sollen 100 Aussagen, die jeweils auf einer Karte stehen, über die zu beurteilende Person auf einer Skala von 1 bis 9 eingeordnet werden, wobei die Kategorie "1" für die Aussage "Trifft auf die Person nicht zu" und die Kategorie "9" für "Trifft für die Person 100 %-ig zu" steht. Dabei wird nun vorgeschrieben, wieviele Karten in die einzelnen Kategorien eingeordnet werden sollen, aber nicht, welche. Man spricht deshalb auch von *forced choice distributions*.

Die Häufigketen r_1, r_2, $\cdots$, r_9 sind also vorgegeben, und es muß $r_1 + r_2 + \cdots + r_9 = 100$ gelten. Bei einer solchen Vorgabe ist dann die Anzahl der möglichen Verteilungen der Karten durch (1.17) gegeben, wobei $n = 100$ ist. Man macht sich leicht klar, dass die Berechnung von $N(r_1, \cdots r_9)$ mit dem Taschenrechner auf direkte Weise kaum möglich ist. Man kann die Schwierigkeiten durch Anwendung von Stirlings Formel (1.8) umgehen, denn $n! \approx \sqrt{2\pi} \cdot n^{n+1/2} e^{-n}$ für $n = 100$ läßt sich z.B. durch Zerlegung der Faktoren n^n und e^{-n} berechnen: $n^{100} = n^{25} n^{25} n^{25} n^{25}$, und e^{-n} läßt sich auf ähnliche Weise berechnen. Die Details sollen hier nicht vertieft werden.

Zur Illustration soll nun der vereinfachte Fall von $n = 12$ Karten und 3 Kategorien betrachtet werden. Kategorie 1 stehe für "Trifft nicht zu", Kategorie 2 für "unentschieden" und Kategorie 3 für "Trifft zu". Es sei der beurteilenden Person vorgeschrieben, in jede Kategorie je 4 Karten zu legen. Dann ist $r_1 = r_2 = r_3 = 4$. Nach (1.17) gibt es dann

$$N(r_1, r_2, r_3) = \frac{12!}{4!4!4!} = 34650$$

mögliche Verteilungen der Karten. Der die Untersuchung duchführende Psychologe kann nun der Ansicht sein, dass die mittlere Kategorie (unentschieden) möglichst wenig gewählt werden sollte, und setzt $r_1 = r_3 = 5$ und $r_2 = 2$ fest. Dann ergibt sich für die Gesamtzahl möglicher Verteilungen der Karten

$$N(r_1, r_2, r_3) = \frac{12!}{5!2!5!} = 16632.$$

Es gibt hier also weniger Möglichkeiten als im Fall gleicher Häufigkeiten. Umgekehrt kann man auch den Fall betrachten, bei dem die mittlere Kategorie die Häufigste sein soll, es sei etwa $r_1 = r_3 = 3$, $r_2 = 6$. Dann ist

$$N(r_1, r_2, r_3) = \frac{12!}{3!6!3!} = 55440$$

die Anzahl möglicher Verteilungen. □

Kapitel 2

Der Begriff der Wahrscheinlichkeit

Probability is the most important concept in modern science, especially as nobody has the slightest notion what it means. (Bertrand Russell, 1929)

2.1 Einführung

2.1.1 Zufall und Wahrscheinlichkeit

Zum Begriff des Zufalls Umgangssprachlich werden Ereignisse dann als 'zufällig' bezeichnet, wenn ihr Eintreten nicht mit Sicherheit vorausgesagt werden kann oder konnte. Ob man sich bei einem Besuch eines Supermarkts mit einem Schnupfenvirus angesteckt wird, ob man bei einer Autofahrt in einen Unfall verwickelt wird, ob einem die nächste Ziehung der Lottozahlen einen Millionengewinn bescheren wird läßt, – all diese Ereignisse lassen sich nicht mit Sicherheit voraussagen und sind deshalb zufällig. Eine Teilmenge der Menschen vertritt die Ansicht, dass es keinen Zufall gibt, alles Geschehen also einem großen Plan folgt, den wir leider nur nicht kennen. Es wird damit angenommen, dass es in der Welt eine alle durchdringende Kausalität gibt und jedes Ereignis also vollständig determiniert ist. Eine andere Teilmenge von Menschen geht davon aus, dass es in bestimmten Bereichen der Physik keine Kausalität gibt. Selbst wenn man annimmt, dass grundsätzlich der Satz vom zureichenden Grund gilt, demzufolge jedes Ereignis eine Ursache hat, erweist sich der Begriff der Kausalität als äußerst evasiv: da auch die Ursache für ein Ereignis ein Ereignis ist, muß auch die Ursache eine Ursache haben, so dass man letztlich zunm Anbeginn der Zeiten zurückgehen muß, wenn man die "wirkliche" Ursache eines Ereignisses feststellen will – aber das ist offenbar eine unkösbare Aufgabe. Eine lesenswerte Abhandlung

über den Kausalitätsbegriff hat bereits Bertrand Russel verfasst (On the notion of cause[1]), es ist unmöglich, das Resultat des Wurfs eines Würfels zu berechnen, ebenso wie es unmöglich ist, den Verlauf einer Billardkugel *exakt* zu berechnen, wenn sie mehrfach auf andere Kugeln oder an die Bande stößt, – diese Unmöglichkeit ergibt sich nicht nur aus der Unmöglichkeit, wirklich alle Faktoren, die den Lauf einer Kugel bestimmen, mit in die Rechnung einzubeziehen, sondern auch aus der Notwendigkeit, mit endlich vielen Nachkommastellen rechnen zu müssen, so dass sich notwendig numerische Fehler aufaddieren, die das Endergebnis dann als 'zufällig' weil nicht genau vorhersehbar bzw. berechenbar erscheinen lassen. Eine sehr gründliche Abhandlung neueren Datums zum Problem der Kausalität findet man bei Earman (1986).

Für die Wahrscheinlichkeitstheorie sind philosophische Betrachungen über das "Wesen" des Zufall und Notwendigkeit (z.B. Monod (1970)) (man denke an den berühmten Satz vom Zureichenden Grund) allerdings unerheblich. Selbst wenn alle Ereignisse vollständig kausal determiniert sind, haben wir oft nur unvollständige Information über das kausale Netz, dass letzlich zum Eintreten eines Ereignisses führt. Überdies redet man nicht nur von zufälligen Ereignissen, wenn es um Autounfälle, Klausurnoten, erfolgreichen Bewerbungen etc geht. Nicht berechenbare Effekte gehen in so gut wie jede Untersuchung ein, und wenn man Glück hat, sind sie vernachlässigbar, etwa wenn man mit den heute zur Verfügung stehenden Uhren Untersuchungen zum Fallgesetz in einem fast perfekten Vakuum macht. In anderen Untersuchungen ist das Verhältnis von systematischen und zufälligen Effekten weniger günstig und man benötigt statistische Tests, um zu entscheiden, ob die Ergebnisse einer Untersuchung mit einer bestimmten Theorie kompatibel sind oder nicht. Hier sind nicht nur die Ergebnisse einer Untersuchung zufällige Ereignisse in dem Sinne, dass sie nicht exakt vorhersagbar sind – wären sie exakt vorhersehbar, müßte keine Untersuchung mehr durchgeführt werden – , sondern auch Einschätzungen der "Wahrheit" von Theorien. Der Ansatz, auch Theorien oder Hypothesen Wahrscheinlichkeiten zuzuordnen, ist unter Philosophen durchaus umstritten, im täglichen Wissenschaftsbetrieb geht man aber mit diesem Ansatz ganz entspannt um. Man muß ja entscheiden, welcher Hypothese von mehreren man mit den nächsten Untersuchungen nachgehen soll, und man wird sich dabei oft für diejenige Hypothese entscheiden, die vor dem Hintergrund von Wissen und bereits vorliegenden Daten als die plausibelste erscheint, und das ist dann oft diejenige Hypothese, die den Wissenschaftlern mit größter Wahrscheinlichkeit als wahr erscheint. Die Feststellung der Wahrheit einer Hypothese ist in diesem Sinne ebenfalls ein zufälliges Ereignis. Die Beziehung zwischen Daten und Theorien oder Hypothesen war und ist Gegenstand ungezählter Diskussionen und Publikationen und die verschiedenen Positionen können hier allenfalls ansatzweise vorgestellt werden.

[1] In *Mysticism and Logic*, London 1918

2.1.2 Frühe Ansätze zur Definition von Wahrscheinlichkeiten

Die Versuche, den Begriff der Wahrscheinlichkeit formal zu fassen, beginnen zumindest in Europa erst um die Mitte des 16-ten Jahrhunderts mit Versuchen, bei Würfel- und Kartenspielen die Wahrscheinlichkeit des Gewinnens zu erhöhen. Einer der ersten, die demnach den Wahrscheinlichkeitsbegriff mathematisch zu fassen suchten, war der seinerzeit berühmte italienische Arzt und Erfinder[2] Gerolamo Cardano (1501 - 1576), ein Zeitgenosse Galileo Galileis, der sich seinen Lebensunterhalt teilweise durch Würfel- und Kartenspielen verdienen mußte (Penrose (1994), andere Quellen sagen, Cardano sei schlicht spielsüchtig gewesen) und deshalb versuchte, durch systematische Betrachtung der jeweiligen Spielsituation einen objektiven numerischen Wert für Wahrscheinlichkeiten zu finden um seine seine Gewinnchancen zu maximieren. Man vermutet, dass Cardano das erste Buch über Wahrscheinlichkeitsrechnung (Liber de ludo aleae - Buch über Würfelspiele) geschrieben hat. Auch Galileo Galilei versuchte, den Begriff der Wahrscheinlichkeit zu definieren. Später wandte sich Antoine Gombaud, Chevalier de Méré, an Blaise Pascal (1623 – 1662) mit der Bitte, ihm bei der Lösung eines Problems beim Glücksspiel zu helfen, worauf es zu einem Briefwechsel zwischen Pascal und Pierre de Fermat (1607 – 1665) kam, in dem die Frage nach der Wahrscheinlichkeit mit einiger Grundsätzlichkeit diskutiert wurde. Der niederländische Mathematiker und Physiker Christiaan Huyghens (1629 – 1695) hörte bei einem Aufenthalt in Paris davon und erweiterte die Wahrscheinlichkeitstheorie, schrieb ein Buch darüber, wodurch die Frage nach der Charakterisierung von Wahrscheinlichkeiten in die Niederlande und nach England gebracht wurde. Huyghens orientierte sich allerdings immer noch am Modell des Glücksspiels.

Eine erste Abhandlung über mögliche Anwendungen in der Wissenschaft wurde von dem Schweizer Mathematiker Jakob Bernoulli (1654 – 1705) verfasst, die posthum 1713 in Basel erschien. Sie enthielt bereits Elemente der Kombinatorik. U.a. wurde hier zum ersten Mal der Begriff der Permutation eingeführt. Der Franzose Abraham de Moivre (1667 – 1754) publizierte 1718 in England sein Werk *The doctrine of chances* mit Resultaten, die auch heute noch zum Standard der Wahrscheinlichkeitstheorie gehören (die Normalverteilung als Grenzverteilung der Binomialverteilung). Die Arbeiten Bernoullis und de Moivres waren der Beginn der Anwendung der Statistik und Wahrscheinlichkeitsrechung in den Wissenschaften. Der englische Pfarrer und Mathematiker Thomas Bayes (1701 – 1761) leitete ein Theorem her, dass 1764 posthum veröffentlicht wurde, und das heute nach ihm benannt wird: der Satz von Bayes, der für die Statistik von zentraler Bedeutung werden sollte. Bayes führte hier zum ersten Male den Begriff der bedingten Wahrscheinlichkeit ein: die Wahrscheinlichkeit eines Ereignisses A kann sich verändern, wenn ein anderes Ereignis B eingetreten ist. Bis dahin hatte man nur

[2]Für die Kutsche Kaiser Karl des Fünften erfand er eine Achskonstruktion, die es beim Durchfahren einer Kurve dem inneren Rad erlaubt, sich langsamer zu drehen als das äußere Rad. Heutigen Autofahrern ist diese Konstruktion als *Kardanwelle* bekannt.

unbedingte Wahrscheinlichkeiten betrachtet, die Information, die durch andere Ereignisse geliefert wird, wurde nicht berücksichtigt. An einem Würfelbeispiel macht man sich unmittelbar klar, was mit einer bedingten Wahrscheinlichkeit gemeint ist: man stelle sich vor, man solle die Zahl raten, die nach dem Wurf eines Würfels oben liegt. Alle Zahlen sind gleich wahrscheinlich, also kann man erwarten, dass man in einem Sechstel der Fälle korrekt rät. Nun wird dem Ratenden die Information gegeben, die gewürfelte Zahl sei eine gerade Zahl. Dann kann nur eine der Zahlen 2, 4 oder 6 oben liegen, und man kann jetzt erwarten, in einem Drittel der Fälle korrekt zu raten. Es ist eine bis heute diskutierte Frage, ob man das Bayessche Theorem zur Bewertung der Wahrscheinlichkeit von Hypothesen verwenden kann oder nicht. Philosophen wie Karl Popper verneinen dies, andere, mehr aus der Wissenschaft kommende Wahrscheinlichkeitstheoretiker wie etwa der Physiker E. T. Jaynes (2003) bejahen die Frage und nennen den Kritischen Rationalisten Popper einen Irrationalisten.

In der Tat gibt es bis heute miteinander konkurrierende Interpretationen des Begriffs der Wahrscheinlichkeit. Eine verhältnismäßig grobe Unterscheidung ist die zwischen subjektiven oder epistemischen und objektiven Wahrscheinlichkeiten. Einige Theoretiker oder empirischen Wissenschaftler sind der Ansicht, dass subjektive Komponenten in der Wissenschaft ausgeschlossen werden müssen, so dass Wahrscheinlichkeiten nur als objektive Größen definiert werden sollten. Andere wiederum weisen auf die Tatsache hin, dass insbsondere die empirischen Wissenschaften induktive Komponenten enthalten, deren Beschreibung am besten über das Bayessche Theorem gelinge, in dessen Anwendung dann subjektive Einschätzungen der Wahrscheinlichkeit, dass bestimmte Hypothesen wahr sind, eingehen. Manche Vertreter der Auffassung, dass Wahrscheinlichkeiten objektive Grössen seien, lassen sich eine Art Hintertür zur Interpretation von Wahrscheinlichkeiten als subjektiver Größe offen: sie sprechen von *statistischer Wahrscheinlichkeit* (Feller (1968), p. 4) und argumentieren, dass sich der Begriff der Wahrscheinlichkeit nicht auf Urteile beziehe, sondern auf Gedankenexperimente (conceptual experiments). Bevor man von Wahrscheinlichkeiten reden könne, müsse man sich auf ein idealisiertes Modell eines bestimmten Gedankenexperimentes (Münzwurf, das Zählen von Unfällen auf einem Autobahnabschnitt, Bestimmung von Reaktionszeiten) einigen. Man müsse sich (i) über die Menge der möglichen Ergebnisse des Experimentes und (ii) über die zugeordneten Wahrscheinlichkeiten einigen. Weiter müsse man sich vorstellen, dass das Experiment viele Male durchgeführt wird - auch wenn diese Durchführung praktisch unmöglich ist. Die Zuordnung eines Wahrscheinlichkeitswertes, etwa $p(A) = .7$ für ein zufälliges Ereignis A, soll dann bedeuten, dass A zumindest bei der vorgestellten mehrfachen Wiederholung in 70% der Fälle auftritt. In einigen Klassikern der Wahrscheinlichkeitstheorie (Keynes, 1920/1980), Jeffreys, 1961/2003) wird die Beziehung zwischen den subjektiven und den objektiven Aspekten der Wahrscheinlichkeit ausführlich diskutiert. Sowohl die "Subjektivisten" wie die "Objektivisten" stehen vor der Frage, wie man Ereignissen numerische Werte, die die

Wahrscheinlichkeiten repräsentieren, zuordnen kann. Ein Ansatz der Subjektivisten, Wahrscheinlichkeiten numerisch zu fassen, besteht darin, wie über einen Wettquotienten q zu definieren: die Aussage, die Chance stehen heute 4 zu 3, dass es noch ein Gewitter gibt, ist irgendwie verständlich, und da $q = p/(1-p)$, p die Wahrscheinlichkeit, mit der es ein Gewitter geben wird, läßt sich p nun ausrechnen, denn es ist ja $p = q/(1+q)$, hier also $p = 4/7$. Objektivisten sind mit einem solchen Ansatz unzufrieden, sie verweisen darauf, dass man Wahrscheinlichkeiten in wissenschaftlichen Untersuchungen durch die relativen Häufigkeiten, mit denen die interessierenden Ereignisse eintreten, abschätzen könne, und die relativen Häufigkeiten keineswegs immer den subjektiven Einschätzungen folgen. So bleibt die Frage, wie man objektive Wahrscheinlichkeiten bestimmen soll. Der Versuch, den Begriff der Wahrscheinlichkeit formal zu fassen und der Zuordnung von numerischen Werten einen objektiven Charakter zu geben hat relativ spät und, wie oben schon angedeutet, nicht im Zusammenhang mit wissenschaftlichen Fragen begonnen. Dies mag eine Implikation der Auffassung sein, dass Gesetzmäßigkeiten in der Natur notwendig deterministisch sind; Einsteins Argument, Gott würfele nicht, zeigt, dass diese Auffassung bis in die Neuzeit geteilt wurde und wird, und die kritischen Anmerkungen zur Anwendung statistischer Methoden in den Sozialwissenschaften und in der Psychologie sind oft hoch emotional und polemisch. Diesen Anmerkungen liegen oft philosophische Annahmen zugrunde, die im Deutschen Idealismus, letzlich im Absoluten Idealismus Hegels verankert sind, wonach der 'lebendige Geist' nur irgendwie ganzheitlich ('Das Ganze ist das Wahre') und jedenfalls nicht mathematisch oder statistisch fassbar sei. Ein anderer Grund mag darin liegen, dass auch in den Naturwissenschaften mit Ausnahme der Astronomie zunächst relativ wenig gemessen wurde. Hier verwundert es ein wenig, dass Galilei sich nicht ausführlich auf das Problem fehlerbehafteter Messungen eingelassen hat, denn ihm standen bei seinen Versuchen keine Uhren im heutigen Sinne zur Verfügung,d.h., seine Messungen enthielten einen hohen Fehleranteil. Dass er trotzdem die richtigen Schlüsse aus seinen Daten zog spricht für seine wissenschaftliche Begabung. Erst Jean-Marie Legendre (1752 – 1833) und Carl Friedrich Gauß (1777 – 1855) entwickelten unabhängig voneinander im Zusammenhang mit astronomischen Messungen eine Methode zum Fehlerausgleich (Methode der Kleinsten Quadrate), bei der der Wahrscheinlichkeitsbegriff benötigt wird, um die Abweichung von Modell und Daten zu bewerten. Dabei wird klar, dass man nicht davon ausgehen kann, dass die möglichen Meßergebnisse nicht alle die gleiche Wahrscheinlichkeit haben können, – es ist offenkundig, dass im Allgemeinen sehr große und sehr kleine (relativ zum Mittelwert aller Messungen) Messungen weniger häufig auftreten als Messungen in der Nähe des Mittelwerts.

Auf diese Entwicklungen kann hier nicht weiter eingegangen werden. Zwei Ansätze sollen im folgenden Abschnitt etwas ausführlicher vorgestellt werden, weil sie zum Teil immer noch als ein Art erste Einführung des Wahrscheinlichkeitsbegriffs verwendet werden: der Ansatz von Pierre Simon de Laplace (1749 – 1827)

und der von Richard von Mises (1883 – 1953). Laplace geht davon aus, dass zufällige (Elementar-)Ereignisse gleichwahrscheinlich sind, auf ihn geht die noch oft angetroffene Definition von Wahrscheinlichkeit als dem Verhältnis von "günstigen zu möglichen" Fällen zurück. von Mises definierte Wahrscheinlichkeiten als Grenzwert relativer Häufigkeiten. Dies ist der *frequentistische Ansatz*. Beide Ansätze definieren Wahrscheinlichkeiten als objektive Größen, lassen sich aber nicht aufrechterhalten. Die heute verwendete Definition von Wahrscheinlichkeiten geht auf den russischen Mathematiker Andrei Nikolajewitsch Kolmogorov (1903 – 1987) zurück. Er lieferte eine axiomatische Definition, in der weder von günstigen und möglichen Fällen noch von relativen Häufigkeiten die Rede ist und die als Basis für eine subjektive wie für eine objektive Interpretation von Wahrscheinlichkeiten dienen kann.

Einen kurzen Überblick über die frühe Entwicklung des Wahrscheinlichkeitsbegriffes findet man unter*http://www.secondmoment.org/articles/probability.php*, etwas ausführlicher sind Gnedenko (1978), Dinges und Rost (1982) und Wirth (1999); eine ausführlichere Diskussion verschiedenen Ansätze wird im Skriptumn Wissenschaftstheorie III (2)[3] gegeben.

2.1.3 Die klassische Definition (Laplace)

Laplace (1831) definierte die Wahrscheinlichkeit eines zufälligen Ereignisses A als das Verhältnis der Anzahl der für das Ereignis "günstigen" Fälle zur Gesamtzahl n der "möglichen" Fälle, wobei ein "Fall" das Ergebnis eines Versuchs ist, – man spricht von solchen Ergebnissen auch als von den Elementarereignissen eines Versuchs.

$$p(A) := \frac{n(A)}{n}. \tag{2.1}$$

Dabei ist $n = |\Omega|$ die Anzahl der Elementarereignisse (mit $|M|$ wird stets die Anzahl der Elemente einer Menge M bezeichnet), und $n(A) = |A|$, $A \subset \Omega$. Aus (2.1) folgt sofort, dass alle Elementarereignisse $\omega \in \Omega$ die gleiche Wahrscheinlichkeit haben. Denn (2.1) soll ja für alle zufälligen Ereignisse gelten, also insbesondere auch für die Elementarereignisse. Für $A = \omega$, ω ein Elementarereignis, folgt aber sofort $n(A) = |A| = |\omega| = 1$, so dass $p(\omega) = 1/n$ für alle $\omega \in \Omega$. In der Tat schreibt Laplace (zitiert nach Dinges et al. (1982)): "Die Wahrscheinlichkeitstheorie besteht in einer Zurückführung aller Ereignisse derselben Art auf eine gewisse Anzahl von gleich möglichen Fällen, d.h. von solchen Fällen, über deren Eintreten wir gleich wenig wissen, und in der Bestimmung derjenigen Anzahl von Fällen, die für das Ereignis günstig sind, desen Wahrscheinlichkeit wir suchen". Wahrscheinlichkeiten sind für Laplace nur Ausdruck mangelnden Wissens; er glaubte an einen vollständigen Determinismus allen Geschehens. Hätte man nur alle Informationen über die Vorgänge in der Natur und genügend Intelligenz,

[3] http://www.uwe-mortensen.de/SkriptenWissTheorie.html

diese Informationen auch zu verarbeiten, würde man die Zukunft bis in alle Zeiten vorhersagen können.

Die Laplacesche Definition bezieht sich

(i) nur auf endliche Mengen Ω von Elementarereignissen und nimmt

(ii) die Elementarereignisse als gleichwahrscheinlich an.

Eine solche Definition ist schon deswegen problematisch, weil man oft vor der Aufgabe, Wahrscheinlichkeiten bezüglich einer unendlichen Menge Ω erklären zu müssen steht: Man denke nur an die Messung von Reaktionszeiten, bei der man alle Zeitpunkte in einem Intervall $(0, T]$ betrachtet, wobei T hinreichend groß ist, so dass praktischg alle Reaktinoszeiten kleiner als T sind. Dieses Intervall enthält nicht nur unendlich viele, sondern überabzählbar unendlich viele Zeitpunkte (Abschnitt 2.2.2 enthält weitere Ausführungen zu den Begriffen abzählbar und überabzählbar) (ii) ist die Laplacesche Definition zirkulär: denn zwar wird die Wahrscheinlichkeit eines Ereignisses A durch die rechte Seite von (2.1) erklärt, aber die Definition basiert bereits auf dem Begriff der gleichen Wahrscheinlichkeit der Elementarereignisse. Nur unter dieser Annahme macht es Sinn, den Quotienten $n(A)/n$ für $p(A)$ zu bilden. Es ist unklar, was "Anzahl der günstigen Fälle" etwa bei der Messung von Reaktionszeiten ist, und warum soll es stets eine Menge Ω von gleichwahrscheinlichen Reaktionszeiten geben? Die Forderung nach der Existenz einer solchen Menge mag für bestimmte Situationen - Glücksspiele etwa - sinnvoll sein, muß aber nicht für beliebige Situationen in der Welt, in denen zufällige Ereignisse beobachtet werden können, erfüllbar sein. Die Neurowissenschaftlerin, die Ströme durch Ionenkanäle in der Membran eines Neurons mißt, mißt jeweils Verläufe $i(\cdot)$ aus einer überabzählbaren Menge möglicher Verläufe, und warum alle diese möglichen Verläufe gleichwahrscheinlich sein sollen, ist völlig unklar.

2.1.4 Die statistische Definition (v. Mises)

Es liegt nahe, die Wahrscheinlichkeit eines Ereignisses A mit der Häufigkeit, mit der A bei einer hinreichend großen Zahl von Wiederholungen eines Versuches unter konstanten Bedingungen in Beziehung zu setzen. Schon im Jahre 2238 v. Chr. beobachteten die Chinesen die Häufigkeiten von Mädchen- und Knabengeburten und setzen den Anteil auf 1/2 fest; daraus läßt sich die Vermutung ableiten, dass ein Neugeborenes mit der Wahrscheinlichkeit 1/2 ein Junge bzw. ein Mädchen ist. Laplace interessierte sich ebenfalls für diese Häufigkeiten und fand aufgrund einer Untersuchung der Geburtsregister, dass der Anteil der Jungen an der Gesamtzahl von Geburten ziemlich konstant bei $22/43 \approx .5163$ liege. Nach den Laplaceschen Zählungen ist also die Wahrscheinlichkeit, dass eine zufällig gewählte schwangere Frau ein Mädchen gebiert, $\approx .48$. Der schottische Arzt und Mathematiker Dr. John Arbuthnot (1667 – 1735) wertete 82 Geburtsstatistiken aus und fand, dass

aus allen Statistiken hervorging, dass *signifikant* mehr Jungen als Mädchen geboren wurden. In den *Philosophical Transactions of the Royal Society of London 27 (1710), 186–190* veröffentlichte er seine Schlußfolgerung aus diesem Befund:

> AMONG innumerable Footsteps of Divine Providence to be found in the Works of Nature, there is a very remarkable one in the exact Ballance that is maintained between the Numbers of Men and Women; for by this means it is provided, that the Species may never fail, nor perish, since every Male may have its Female, and of a proportional Age. This Equality of Males and Females is not the Effect of chance but Divine Providence, working for a good End, which I thus demonstrate: ...

Das vollständige Argument findet man in
http://www.york.ac.uk/depts/maths/histstat/arbuthnot.pdf.
Arbuthnots Artikel enthält übrigens den ersten Versuch, einen Signifikanztest zu konstruieren.

Der österreichische Mathematiker Richard Edler von Mises (1883 – 1953) schlug in seinem 1919 erschienenen Werk *Grundlagen der Wahrscheinlichkeitsrechnung*[4] vor, die Wahrscheinlichkeit eines Ereignisses A durch den Limes

$$p(A) := \lim_{n\to\infty} \frac{N(A)}{N} \tag{2.2}$$

zu definieren. Hierbei ist N die Gesamtzahl der Versuche und $N(A)$ ist die Gesamtzahl der Versuche, bei denen A beobachtet wurde. Wird N größer, so wird auch $N(A)$ entsprechend größer. Der Quotient $N(A)/N$ soll gegen einen Grenzwert $p(A)$ streben.

Bei dieser Definition wird also nicht, wie bei Laplace, eine Beziehung zwischen der Wahrscheinlichkeit eines Ereignisses und der Anzahl der Elementarereignisse, die einerseits A und andererseits Ω definieren, hergestellt. Nach v. Mises besteht der Vorteil der Definition (2.2) darin, dass keine "metaphysischen" Annahmen über die Wahrscheinlichkeiten von Ereignissen gemacht werden (wie etwa die Annahme gleicher Wahrscheinlichkeiten für die 'Elementarereignisse', also die Ergebnisse eines Versuchs). Man müsse eben nur eine hinreichend große Anzahl von Beobachtungen unter gleichen Bedingungen machen und wird dann eine Zahl $N(A)/N$ finden, die der gewünschten Zahl $p(A)$ bis auf ein paar Stellen nach dem Komma entspricht. Auf diese Weise könne die Wahrscheinlichkeit empirisch bestimmt werden.

Das Problem bei diesem Ansatz ist, dass vorausgesetzt muß, dass der Limes in (2.2) tatsächlich existiert. Die Frage ist, wie man wissen kann, dass er existiert.

[4] v. Mises Betrachtungen findet man in v. Mises, R.: *Probability, Statistics and Truth*, Dover Publications New York 1981, ein Übersetzung des 1928 beim Springer-Verlag Berlin erschienenen Buchs *Wahrscheinlichkeit, Statistik und Wahrheit*, in dem er seinen Wahrscheinlichkeitsbegriff noch einmal elaboriert.

Kann man die Existenz des Limes nicht von etwas unmittelbar Gegebenem ableiten, so enthält das Postulat seiner Existenz genau so viel "Metaphysik" – die v. Mises ja gerade vermeiden wollte – wie das Postulat der Gleichwahrscheinlichkeit bei Laplace.

Neben der Existenz des Limes muß v. Mises "Zufälligkeit" fordern. Damit meint er, dass der Limes in (2.2) unabhängig ist von der Art und Weise, in der die Beobachtungen erhoben werden. Hier wird also der Begriff der Zufälligkeit zur Erklärung des Begriffes der Wahrscheinlichkeit herangezogen, was zumindest eine spezielle Definition der "Zufälligkeit" erfordert. Es ist ebenso denkbar, dass man die Zufälligkeit über die Wahrscheinlichkeit von Ereignissen definiert. Die sogenannte statistische Definition der Wahrscheinlichkeit stößt also auf eine Reihe logischer Schwierigkeiten, die den v. Misesschen Ansatz als unbefriedigend erscheinen lassen.

2.2 Kolmogorovs axiomatischer Ansatz

Wahrscheinlichkeiten lassen sich ohne Bezug auf bestimmte inhaltliche Interpretationen des Wahrscheinlichgkeitsbegriffs definieren, nämlich als Maße auf Mengen. Wer sich für die mathematischen Hintergründe des Maßbegriffs nicht interessiert, kann gleich Abschnitt 2.2.3 springen. Wer mit den Eigenschaften der Menge der rellen Zahlen nicht besonders vertraut ist und den allgemeinen Maßbegriff kennen lernen möchte, kann durch die folgenden Abschnitte gehen. Der Maßbegriff wird zwar sehr allgemein vorgestellt, aber die Details werden nicht weiter elaboriert; dies würde den Rahmen dieser Darstellung völlig sprengen.

Die folgenden Betrachtungen sind sehr grundsätzlich; wer direkter auf die Definition von Wahrscheinlichkeiten geführt werden will, kann gleich zu Abschnitt 2.2.3 gehen.

2.2.1 Einige Aspekte der reellen Zahlen

Mit $\mathbb{R}$ wird die Menge aller reeller Zahlen bezeichnet; mit $\mathbb{R}_+$ wird oft die Menge der positiven Zahlen gekennzeichnet. Sie enthält die Menge der natürlichen Zahlen $\mathbb{N} = \{0, 1, 2, 3, \ldots\}$ als Teilmenge, ebenso die Menge der ganzen Zahlen, $\mathbb{Z} = \{\ldots, 3, -2, -1, 0, 1, 2, 3, \ldots\}$. Eine weitere Teilmenge ist die Menge der rationalen Zahlen $r = p/q$, $p, q \in \mathbb{N}$ bzw. $p, q \in \mathbb{Z}$. Der Ausdruck 'rational' leitet sich vom lateinischen *ratio* für Bruch, Quotient her. Außer den rationalen Zahlen gibt es dann noch die Menge $\mathbb{I}$ irrationalen Zahlen, d.h. der Zahlen, die sich *nicht* als ratio, d.h. Quotient zweier natürlicher Zahlen darstellen lassen. Ein Beispiel für eine irrationale Zahl ist die Länge der Diagonale eines Quadrats mit der Seitenlänge 1. Denn nach dem Satz des Pythagoras ist das Quadrat der Länge der Diagonalen durch $1^2 + 1^2 = 2$ gegeben, und die Länge ist dann $\sqrt{2}$. Es gibt aber keine zwei natürlichen Zahlen p und q derart, dass $\sqrt{2} = p/q$. Um dies zu

sehen, werde angenommen, dass es sie gäbe. Man kann annehmen, dass p und q teilerfremd sind, d.h. dass sie kein gemeinsames Vielfaches zweier natürlichen Zahlen a und b sind. Denn dann hätte man $p/q = ka/(kb)$ und k würde sich herauskürzen, so dass $p/q = a/b$. Dan ist $2 = p^2/q^2$ und damit $p^2 = 2q^2$, woraus folgt, dass p^2 eine gerade Zahl ist, das sie das Produkt des Faktors 2 mit q^2 ist. Dann muß auch p eine gerade Zahl sein. Denn es sei $p = 2k + 1$ ungerade, $k \in \mathbb{N}$; dann folgt $(2k + 1)^2 = (2k)^2 + 1^2 + 2 \cdot (2k \cdot 1) = 4k^2 + 4k + 1$, d.h. für ungerades p ist p^2 ebenfalls ungerade, d.h. es muß $p = 2k$ gelten. Dann folgt aber $p^2 = 4k^2 = 2q^2$, also $2k^2 = q^2$, mithin ist auch q eine gerade Zahl. Dann aber sind p und q nicht mehr teilerfremd, weil sie beide den Faktor 2 enthalten, was der Forderung widerspricht, das p/q teilerfremd ist. Also kann $\sqrt{2}$ keine rationale Zahl sein. Es läßt sich zeigen, dass rationale Zahlen entweder nur endlich viele Dezimalstellen haben, oder dass im Falle unendlich vieler Dezimalstellen diese periodisch sind. Dagegen haben irrationale Zahlen unendlich viele Dezimalstellen ohne Periodizität.

Die Anzahl der Elemente in einer Menge M ist die *Mächtigkeit* von M; sie wird oft mit $|M|$ bezeichnet. Hat man zwei Mengen M_1 und M_2 und kann man jedem Element von M_1 genau ein Element von M_2 zuordnen, und umgkehrt jedem Element von M_2 genau ein Element von M_1, so haben die beiden Mengen dieselbe Mächtigkeit, $|M_1| = |M_2|$. Die Menge $\mathbb{N}$ der natürlichen Zahlen hat unendlich viele Elemente; das Symbol für diese Mächtigkeit ist $\aleph_0$, ($\aleph$ ist der erste Buchstabe des hebräischen Alphabets), also $|\mathbb{N}| = \aleph_0$. Für die rationalen Zahlen gilt der folgende Satz:

Satz 2.2.1 *Die Menge $\mathbb{Q}$ der rationalen Zahlen ist abzählbar.*

Beweis: Eine rationale Zahl ist als Quotient p/q, $p, q \in \mathbb{N}$ definiert. Die Mächtigkeit von $\mathbb{Q}$ ist demnach gleich der Anzahl der Paare (p, q). Es sei $p + q$ die *Höhe* des Paares (p, q). Die Anzahl der Paare mit vorgegebener Höhe n, $n > 1$, ist gleich $n - 1$:

$$P_n = \{(1, n-1), (2, n-2), \ldots, (n-1, 1)\}, \quad |P_n| = n - 1 \tag{2.3}$$

Offenbhar ist P_n abzählbar, denn P_n ist eine natürliche Zahl, und der natürlichen Zahl n wird die natürliche Zahl P_n zugeordnet, so dass man eine Abbildung $\mathbb{N} \to \mathbb{N}$ hat. Das gleiche gilt auch für $|P_n|(n-1)$, die eine natürliche Zahl ist, d.h. $|P_n|(n-1)$ entspricht wiederum eine Abbildung $\mathbb{N} \to \mathbb{N}$. Also ist die Menge der rationalen Zahlen wiederum abzählbar, so dass $|\mathbb{Q}| = |\mathbb{N}|| = \aleph_0$. □

Definition 2.2.1 *Es sei $M \subseteq \mathbb{R}$ und $\omega_1, \omega_2 \in M$ mit $\omega_1 < \omega_2$. Gibt es zwischen ω_1 und ω_2 unendlich viele Elemente $\omega \in M$, so heißt M* überall dicht.

Satz 2.2.2 *Die Menge $\mathbb{Q}$ der rationalen Zahlen ist überall dicht.*

Beweis: Es seien $p, r \in \mathbb{Q}$. Dann existiert stets eine rationale Zahl $q \in \mathbb{Q}$ derart, dass $p < q < r$. Es genügt zu zeigen, dass irgendeine rationale Zahl zwischen $p, r \in \mathbb{Q}, p < r$, liegt, etwa $q := (p+r)/2$. Dann folgt $2p = p+p < p+r < r+r = 2r$ und q erfüllt offenbar die Bedingung $p < q < r$. In der gleichen Weise kann man fortfahren und Zahlen zwischen p und q sowie q und r finden, etc., und das heißt, dass Q überall dicht ist. □

Anmerkung: Satz 2.2.2 gilt für beliebige reelle Zahlen $a, c \in \mathbb{R}$, die Nebenbedingung $b = (a + c)/2 \in \mathbb{Q}$ entfällt, um $a < b < c$ mit $b = (a + c)/2$ zu zeigen.

Verglichen mit der Mächtigkeit der rellen Zahlen ($\mathbb{R} = \mathbb{Q} \cap \mathbb{I}$) ist dies eine "kleine" Zahl, – die Mächtigkeit von $\mathbb{I}$ ist *sehr* viel größer als die von $\mathbb{Q}$. Dies ist der Grund, warum es bei der Abbildung $f(x)$, (2.6), so viel Werte gleich 1 mehr gibt als Werte gleich Null.

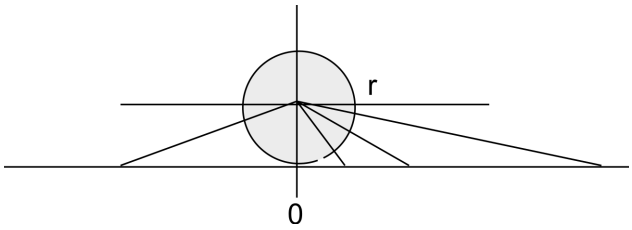
Die Frage ist nun, ob sich die Mächtigkeit von $\mathbb{R}$ abschätzen läßt. Die irrationalen Zahlen müssen zwischen den rationalen Zahlen liegen, und da die rationalen Zahlen überall dicht liegen, müssen die offenbar beliebig kleinen Differenzen zwischen zwei rationalen Zahlen noch mit den irrationalen Zahlen aufgefüllt werden, d.h. die rationalen Zahlen müssen in die irrationalen Zahlen eingebettet sein. Um nun $|\mathbb{R}|$ abzuschätzen, betrachtet man zunächst die Menge aller Untermengen einer Menge M. Es sei zunächst $|M| < \infty$. Weiter sei $A \subseteq M$. Man kann nun eine Funktion definieren, die jedem $\omega \in A$ einen von zwei möglichen Werten zuordnet, je nachdem, ob $\omega \in M$ zur Teilmenge A gehört oder nicht:

$$\phi(\omega) = \begin{cases} 0, & \omega \notin A, \\ 1, & \omega \in A \end{cases} \tag{2.4}$$

Der Teilmenge A entspricht genau eine Folge von Nullen und Einsen, wobei die Länge der Folge durch die Anzahl $|M|$ der Elemente von M gegeben ist; entspricht einem Element von M eine 1, so gehört dieses Element zu A, und wird dem Element eine 0 zugeordnet, so gehört das Element nicht zu A. Die Anzahl verschiedener Folgen ist dann gleich der Anzahl verschiedener Teilmengen von M. Diese Anzahl ist leicht zu bestimmen. Die Elemente können durchnummeriert werden, wobei die Reihenfolge der Nummerierung unwichtig ist. Für das erste Element hat man zwei Möglichkeiten: ihm wird entweder die 0 oder die 1 zugeordnet. Für das zweite Element gilt dasselbe, und für die beiden ersten Elemente gibt es dann insgesamt $2 \cdot 2 = 2^2 = 4$ Möglichkeiten: 00, 01, 10 und 11. Für das dritte Element gibt es wieder zwei Möglichkeiten, und jede dieser Möglichkeiten kann mit jeder der 2^2 Verteilungen von 0 und 1 für die ersten beiden Elemente kombiniert werden, so dass es $2 \cdot 2^2 = 2^3$ Möglichkeiten für die ersten drei Elemente gibt. So fährt man fort und erhält schließlich $2^{|M|}$ Möglichkeiten für die Aufteilungen von Nullen und Einsen, und dies ist die Anzahl der möglichen Teilmengen von M. Man schreibt dafür auch $\mathfrak{P}(P)$ oder auch einfach 2^M. Die Menge aller Teilmengen von M ist die Potenzmenge) von M.

Man betrachtet nun die Abbildung $\phi\colon \mathbb{N} \to \{0, 1\}$, d.h. jeder Zahl $n \in \mathbb{N}$ wird

Abbildung 2.1: Umkehrbar eindeutige Abbbildung eines Intervalles auf $\mathbb{R}$



eine Zahl i_n zugeordnet, und i_n ist entweder gleich 0 oder gleich 1. Damit wird eine unendliche Folge

$$i_1, i_2, i_3, \ldots, i_n, \ldots \tag{2.5}$$

erzeugt. Umgekehrt entspricht jeder derartigen Folge eine Abbildung φ, $\varphi(n) = 0$ oder $\varphi(n) = 1$. Dies bedeutet, dass die Menge der Folgen der Form (2.5) gleich der Menge aller Teilmengen von $\mathbb{N}$ ist, und deren Mächtigkeit ist dann gleich $2^{\aleph_0}$. Setzt man vor jede Folge (2.5) einen Dezimalpunkt, so erhält man jeweils einen Dualbruch, und der stellt eine reelle Zahl dar. Daraus folgt, dass die Mächtigkeit von $\mathbb{R}$ durch $|\mathbb{R}| = 2^{\aleph_0}$ gegeben ist. Offensichtlich gilt $2^{\aleph_0} > \aleph_0$, so dass die Menge *aller* reellen Zahlen eine Machtigkeit größer als die der rationalen Zahlen hat:

Definition 2.2.2 *Eine Menge habe unendlich viele Elemente, die aber durchnummeriert werden können, so dass jedem Element von M genau eine natürliche Zahl zugeordnet werden kann. Dann heißt M* abzählbar. *Die Mächtigkeit von $\mathbb{R}$ ist $\boldsymbol{c} = 2^{\aleph_0}$ und heißt die* Mächtigkeit des Kontinuums*; $\mathbb{R}$ ist* überabzählbar.

Satz 2.2.3 *Es sei $I_{a,b} = [a, b] \subset \mathbb{R}$, $a < b$. Dann ist $|I_{a,b}| = 2^{\aleph_0}$, d.h. die Mächtigkeit von $I_{a,b}$ ist gleich der der reellen Zahlen.*

Beweis: Gegeben seien zwei Mengen A und B und ϕ sei eine umkehrbar eindeutige Abbildung[5] $\phi\colon A \to B$, d.h. $\phi(a_1) = b_1$, $\phi(a_2) = b_2$ und $b_1 = b_2 \Rightarrow a_1 = a_2$. Dann gilt $|A| = |B|$. Man muß also nur zeigen, dass eine Abbildung $\phi\colon I_{a,b} \to \mathbb{R}$ existiert. Dazu betrachte man die Geraden, die den Mittelpunkt des Kreises mit der x-Achse, die $\mathbb{R}$ repräsentiert, verbinden (vergl. Abb. 2.1). Sie schneiden den Viertelkreis jeweils an genau einer Stelle x', und treffen auf die x-Achse an genau einer Stelle x. Auf diese Weise wird eine umkehrbar eindeutige Abbildung f zwischen den Punkten in $[0, T]$ und den Punkten auf der x-Achse definiert, $f(x') = x$. Dies bedeutet, dass $[0, T]$ und $\mathbb{R}$ dieselbe Mächtigkeit haben. $\square$

Teilmengen aus $M \subset \mathbb{R}$ können bestimmte Eigenschaften haben. So heißt M *nach unten beschränkt*, wenn es eine Zahl $a \in \mathbb{R}$ gibt derart, dass $a \leq x$ für alle

[5] bijektive Abbildung.

$x \in M$, und nach *oben beschränkt*, wenn es ein $b \in \mathbb{R}$ gibt derart, dass $x \leq b$ für alle $x \in M$. Eine Menge heißt *offen*, wenn $M = \{x|a < x < b, x \in M, a, b \in M^c\}$, wobei $M^c \subset \mathbb{R}$ die Komplementärmenge von M ist. Eine Menge, z.B. ein Intervall aus $M \subset \mathbb{R}$, ist *abgeschlossen*, wenn das Komplement M^c von M offen ist. Ist eine Menge $M \subset \mathbb{R}$ sowohl beschränkt wie auch abgeschlossen, so heißt M *kompakt*. So sei $[a, b] = \{x|a \leq x \leq b,\ a, b \in \mathbb{R}\}$ ein Intervall, bei dem die Randpunkte a und b zum Intervall gehören; das Intervall heißt demnach *abgeschlossen*. Nun sei das Intervall $(a, b) = \{x|a < x < b,\ a.b \in \mathbb{R}\}$ gegeben. Bei diesem Intervall gehören die Punkte a und b nicht zum Intervall; es heißt dementsprechend *offenes* Intervall. Die Intervalle $(a, b] = \{x|a < x \leq b,\ a, b \in \mathbb{R}\}$ und $[a, b) = \{x|a \leq x < b,\ a, b \in \mathbb{R}\}$ sind dementsprechend *halboffene* Intervalle.

2.2.2 Maße auf Mengen

In Abschnitt 2.2.4 werden Wahrscheinlichkeiten als Maße auf Mengen bzw. auf Teilmengen von Ω definiert werden. Die Grundlage dafür liefert die allgemeine Maßtheorie. Man kann sagen, dass die Wahrscheinlichkeitstheorie im Wesentlichen Maßtheorie ist. Gleichwohl würde eine vollständige maßtheoretische Darstellung der Wahrscheinlichkeitstheorie den Rahmen dieses Textes sprengen. Es werden nur einige grundlegende Sachverhalte vorgestellt.

Es gibt Funktionen, die sich anhand des Riemannschen Integralbegriffs nicht integrieren lassen, bei denen es aber als sinnvoll erscheint, ein Integral zu berechnen. Man betrachte z.B. die Funktion

$$f(x) = \begin{cases} 0, & x \in \mathbb{Q} \\ 1, & x \in \mathbb{I} \end{cases}, \quad x \in \mathbb{R} \tag{2.6}$$

wobei $\mathbb{Q}$ die Menge der rationalen Zahlen und $\mathbb{I}$ die Menge der irrationalen Zahlen ist. Die Funktion läßt sich nicht exakt zeichnen, da sie in nachgerade bizarrer Weise zwischen den Werten 0 und 1 hin und her springt. Aber wie im vorangegangenen Abschnitt gezeigt wurde ist die Mächtigkeit von $\mathbb{Q}$ "einfach" $\aleph_0$, während die von $\mathbb{I}$ deutlich größer als $\aleph_0$ ist. Daraus folgt, dass die Punkte mit $f(x) = 1$ sehr viel dichter liegen als die Punkte mit $f(x) = 0$. Dieser Sachverhalt impliziert, dass sich f doch integrieren läßt, wozu allerdings ein anderer Integralbegriff (Lebesgue-Integral) nötig ist, der auf einem allgemeinen Maßbegriff beruht

In den meisten praktischen Anwendungen der Wahrscheinlichkeitstheorie muß allerdings nicht mit dem Lebesgue-Integral operiert werden, das Riemann-Integral ist dafür hinreichend. Dementsprechend soll hier keine allgemeine Einführung in die Maßtheorie gegeben werden. Gleichwohl sollen einige wenige Begriffe aus der Maßtheorie vorgestellt werden, die deutlich machen, warum man spezielle Mengensysteme betrachten muß, um Wahrscheinlichkeiten überhaupt definieren zu können. So betrachte man ein einfaches Würfelspiel. Die Menge der möglichen Ergebnisse eines Wurfs ist durch $\Omega = \{1, 2, 3, 4, 5, 6\}$ gegeben. Spieler A gewinnt,

wenn nach einem Wurf eine gerade Zahl oben liegt, und Spieler B gewinnt, wenn eine ungerade Zahl oben liegt. Die für A günstige Menge von Ergebissen besteht aus $\{2, 4, 6\}$, die für B günstige Menge ist $\{1, 3, 5\}$. Die interessierenden Ereignisse sind also durch Teilmengen Ω definiert. Man kann sich fragen, wieviele verschiedene Teilmengen überhaupt definiert werden können. Wie im vorangegangenen Abschnitt gezeigt wurde, gibt es bei einer endlichen Menge insgesamt 2^n Teilmengen. In einer bestimmten experimentellen Situation werden nicht alle Teilmengen von Interesse sein, aber zumindest im Prinzip wird es möglich sein, die Wahrscheinlichkeit für ein Ereignis zu berechnen, das einer dieser Teilmengen entspricht. Die Situation ist aber grundsätzlich anders, wenn man Ereignisse betrachtet, die durch Teilmengen der reellen Zahlen definiert sind, – etwa wenn die Wahrscheinlichkeit von Zeiten bestimmte werden soll. Zeiten sind im Allgemeinen auf einem Kontinuum definiert. Es liegt nahe, bestimmte Intervalle aus dem Kontinuum zur Definition von Ereignissen zu wählen, und in der Tat zeigt es sich, dass für Menge *aller* Teilmengen von $\mathbb{R}$ gar kein Wahrscheinlichkeitsmaß existiert. Es sollen Ereignisse betrachtet werden, die gewissermaßen numerisch kodiert werden können, – was stets möglich ist, wie noch gezeigt werden wird. Im einfachsten Fall hat man $A \to 1$ und $\neg A \to 0$. Eine andere Möglichkeit ist $A = A_n \to n \in \mathbb{N}$: dies ist das Ereignis, dass ein Ereignis n-mal aufgetreten ist. Schließlich hat man bei vielen Messungen den Fall, das $A \to x \in \mathbb{R}$ gilt: A ist eine Messung, die den Wert x aus $\mathbb{R}$ geliefert hat. In jedem Fall hat man es mit numerischen Repräsentationen von Ereignissen zu tun, und diese Repräsentationen sind Teilmengen von $\mathbb{R}$. Es zeigt sich, dass man mit einen sehr allgemeinen Typ von Systemen von Teilmengen auskommt:

Definition 2.2.3 *Gegeben sei eine Menge Ω und ein System τ von* offenen *Teilmengen $A \in \Omega$ derart, dass die Bedingungen*
1. Ω und die leere Menge $\emptyset$ gehören zu τ,
2. Beliebig viele (endliche oder unendliche) Vereinigungen $\cap_k A_k$ und Durchschnitte $\cap_k A_k$ offener Mengen gehören zu τ.
Dann heißt das System τ eine Topologie *auf Ω, und das Paar (Ω, τ) heißt* topologischer Raum.

Man sagt auch, (Ω, τ) definiert eine *Topologie* auf Ω. Natürlich muß erklärt werden, was mit "offen" gemeint ist. In Bezug auf $\mathbb{R}$ ist eine Menge M "offen", wenn es eine Umgebung $O_\varepsilon(x)$ für alle $x \in M$ gibt, in der nur Punkte aus M sind. Dies sind z.B. die am Ende des vorangegangenen Abschnitts erwähnten offenen Intervalle (a, b) mit $a < b$. Es gilt

Satz 2.2.4 *Eine Teilmenge $M \subset \mathbb{R}$ ist offen genau dann, wenn die Komplementärmenge $M^c = \mathbb{R} \setminus M$ abgeschlossen ist.*

Beweis: Ein Punkt $x \in \mathbb{R}$ heißt *Kontaktpunkt* von M, wenn jede Nachbarschaft von x einen Punkt enthält, der zu M gehört. Wenn M offen ist, so existiert für

jedes $x \in M$ eine Nachbarschaft, die nur Punkte aus M enthält. Demnach ist kein Punkt $x \in M$ ein Kontaktpunkt des Komplements $M^c = \mathbb{R} \setminus M$. Ist also x ein Kontaktpunkt von M^c, so muß $x \in M^c$ sein, also ist M^c abgeschlossen. Ist umgekehrt M^c abgeschlossen, so folgt, dass für alle $x \in M$ eine Umgebung mit Punkten nur aus M existiert, und das heißt, dass M offen ist. □

Satz 2.2.5 *Die Vereinigung beliebig vieler offener Mengen ist offen, und der Durchschnitt einer endlichen Menge offener Mengen ist offen.*

Beweis: Der Beweis ergibt sich unmittelbar aus dem Satz 2.2.4. □

Es wird noch einmal zu der Frage nach den Mengen, die Ereignisse repräsentieren, zurückgekehrt. Die folgend Defintion ist für dies Frage von zentraler Bedeutung:

Definition 2.2.4 *Es sei Ω eine nichtleere Menge und $\mathcal{A}$ sein ein System von Teilmengen von Ω. $\mathcal{A}$ heißt* Algebra*, wenn $\mathcal{A}$ die Bedingungen*
1. $\emptyset \in \mathcal{A}$,
2. $A, B \in \mathcal{A} \Rightarrow A \cup B \in \mathcal{A}$,
3. $A \in \mathcal{A} \Rightarrow A^c = \Omega \setminus A \in \mathcal{A}$. Gilt zusätzlich
*4. Für $A_1, A_2, A_3, \ldots \in \mathcal{A} \Rightarrow \bigcap_n A_n \in \mathcal{A}$, so heißt $\mathcal{A}$ σ-*Algebra*.*

Satz 2.2.6 *Es seien $\mathcal{A}$ und $\mathcal{B}$ zwei σ-Algebren. Dann ist der Durchschnitt $\mathcal{A} \cap \mathcal{B}$ ebenfalls eine σ-Algebra.*

Beweis: Es seien $\mathcal{A}_j$, $j = 1, 2, \ldots$ Sigma-Algebren, und es sei $\mathcal{A} = \bigcap_j \mathcal{A}_j$ der Durchschnitt dieser Sigma-Algebren. Sei $A \in \mathcal{A}$. Dann ist $A \in \mathcal{A}_j$ für alle j und folglich ist $A^c \in \mathcal{A}_j$ für alle j. Ist $A_k \in \mathcal{A}$ füßr alle $k = 1, 2, \ldots$, so ist $A_k \in \mathcal{A}_j$ für alle k und j, also folgt $\bigcup_k A_k \in \mathcal{A}_j$ für alle j und also auch $\bigcup_k A_k \in \mathcal{A}$. □

Definition 2.2.5 *Es sei $A \subset \Omega$ eine Teilmenge vn Ω. Dann ist $\sigma(A) = \{\emptyset, \Omega, A, A^c\}$ die kleinste σ-Algebra, die A enthält. Dann heißt $\sigma(A)$ die von A* erzeugte σ-Algebra*, und A heißt ihr* Erzeuger*. Ist $\mathcal{E}$ ein Mengensystem aus 2^Ω und ist Σ die Menge aller σ-Algebren, die $\mathcal{E}$ enthalten, so ist*

$$\sigma(\mathcal{E}) = \bigcap_{\mathcal{F} \in \Sigma} \mathcal{F}$$

die kleinste σ-Algebra, die $\mathcal{E}$ enthält und $\mathcal{E}$ heißt die von $\mathcal{E}$ erzeugte σ-Algebra*.*

Definition 2.2.6 *Es sei Ω ein topologischer Raum und $\mathcal{O}$ sei das System der offenen Teilmengen von Ω. Dann heißt $\mathcal{B}(\Omega) = \sigma(\mathcal{O})$ die* Borelsche σ-Algebra über Ω*, und die Elemente von $\mathcal{B}(\mathcal{O})$ werden* Borelmengen *genannt.*

Anmerkung: Mit Satz 2.2.4 zeigt sich, dass die Borelmengen von Ω so gut wie alle Mengen sind, die sich mit $\cup$, $\cap$, $\setminus$, $[\cdot,\cdot)$, $(\cdot,\cdot]$ etc anschreiben lassen. $\square$

Mengen können Maße zugeordnet werden. Insbesondere können bestimmten Systemen von Teilmengen von $\mathbb{R}$ (und damit auch von $\mathbb{Q}$) Maße zugeordnet werden. Es läßt sich aber zeigen, dass nicht allen Teilmengen von $\mathbb{R}$ Maße zugeordnet werden können, man muß sich auf σ-Algebren beschränken; dementsprechend heißen die Elemente einer σ-Algebra *messbar*. Um dies einzusehen zu können, muß man sich mit der allgemeinen Definition von Maßen vertraut machen.

Definition 2.2.7 *Es sei Ω eine Menge und $\mathcal{A}$ eine σ-Algebra auf Ω. Ein* Maß μ *auf $\mathcal{A}$ ist eine Abbildung $\mu : \mathcal{A} \to [0,\infty]$ mit den Eigenschaften*
1. $\mu(\emptyset) = 0$,
2. Für $\{A_n\}_{n\in\mathbb{N}} \subset \mathcal{A}$, $A_n \cap A_m = \emptyset$ *für* $n \neq m$ *ist*

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n). \tag{2.7}$$

$(\Omega, \mathcal{A}, \mu)$ *steht für* Maßraum.

Anmerkung: Man beachte, dass μ auf $[0,\infty]$ definiert ist, d.h. der Fall $\mu(\Omega) = \infty$ wird in der allgemeinen Definition eines Maßes nicht ausgeschlossen. Wahrscheinlichkeiten werden jedoch als *normierte Maße* definiert, d.h. es wird $\mu(\Omega) = 1$ gesetzt (vergl. die folgende Definition). $\square$

Definition 2.2.8 *Es sei $(\Omega, \mathcal{A}, \mu)$ ein Maßraum. Das Maß μ heißt*
1. μ heißt endliches Maß*, wenn $\mu(\Omega) < \infty$.*
2. Wahrscheinlichkeitsmaß*, wenn $\mu(\Omega) = 1$,*
3. σ-endliches Maß*, falls eine Folge $\{A_n\}_{n\in\mathbb{N}} \subset \mathcal{A}$ existiert, so dass*

$$\bigcup_{n=1}^{\infty} A_n = \Omega, \quad \text{mit } \mu(A_n) < \infty, \ \forall n \in \mathbb{N}. \tag{2.8}$$

Der folgende Satz charakterisiert grundlegende Eigenschaften von Maßen:

Satz 2.2.7 *Es sei $(\Omega, \mathcal{A}, \mu)$ ein Maßraum. Dann gelten die folgenden Aussagen:*

1. **Endliche Additivität***: Für ein gegebenes $n \in \mathbb{N}$ seien $A_1, A_2, \ldots, A_n \in \mathcal{A}$, mit $A_I \cap A_j = \emptyset$ für $i \neq j$. Dann gilt*

$$\mu(A_1 \cup \cdots \cup A_n) = \mu(A_1) + \cdots + \mu(A_n). \tag{2.9}$$

2. **Isotonie** *Für $A \in \mathcal{A}, B \in \mathcal{A}$ und $A \subset B$ gilt $\mu(A) \leq \mu(B)$.*
3. **Subtraktivität** *Für $A \in \mathcal{A}, B \in \mathcal{A}$, $A \subset B$ und $\mu(A) < \infty$ folgt*

$$\mu(B - A) = \mu(B) - \mu(A)$$

4. **Sub-Additivität** *Es gelte* $\{A_n\}_{n\in\mathbb{N}} \subset \mathcal{A}$. *Dann folgt*

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \mu(A_n). \tag{2.10}$$

5. **Stetigkeit von unten** *Es sei* $\{A_n\}_{n\in\mathbb{N}} \subset \mathcal{A}$, *und*

$$A_1 \subset A_2 \subset \cdots \subset \bigcup_{n=1}^{\infty} A_n = A \in \mathcal{A}$$

sei eine isotone Folge von messbaren Mengen mit einem messbaren Grenzwert. Dann gilt

$$\lim_{n\to\infty} \mu(A_n) = \mu(A). \tag{2.11}$$

6. **Stetigkeit von oben***: Es sei* $\{A_n\}_{n\in\mathbb{N}} \subset \mathcal{A}$, *und sei*

$$A_1 \supset A_2 \supset \cdots \supset \bigcap_{n=1}^{\infty} A_n = A \in \mathcal{A}$$

eine isotone Folge messbarer Mengen, mit $\mu(A_1) < \infty$. *Dann gilt*

$$\lim_{n\to\infty} \mu(A_n) = \mu(A). \tag{2.12}$$

Beweis: Billingsley. □

Zum Abschluß dieses Abschnitts soll noch der wichtige Satz von Vitali[6] vorgestellt werden, wenn auch ohne Beweis. Für endliche Mengen können alle möglichen Teilmengen, also die Potenzmenge der Ausgangsmenge, betrachtet werden. Aber in vielen Anwendungen hat man es nicht mit endlichen Mengen (ürfel etc) zu tun, sondern mit abzählbaren Mengen bzw. mit $\mathbb{R}^n$. Die Potenzmenge von $\mathbb{R}^n$ – oft mit $2^{\mathbb{R}^n}$ bezeichnet – ist allerdings so kompliziert, dass man klaum eine Intuition von ihr entwickeln kann, und man kann davon ausgehen, dass kaum jemals in einem Experiment die Menge *aller* möglichen Teilmengen eine Menge ist, die von Interesse ist. Abgesehen davon läßt sich der folgende Satz beweisen:

Satz 2.2.8 *(Satz von Vitali (1905)) Es existiert kein Maß* μ, *dass auf allen Teilmengen von* $\mathbb{R}^n$ *erklärt werden könnte.*

Beweis: s. Elstrodt (2005), Satz III.3.3. □

Es sei nun $G_x \subset \mathbb{R}$ irgendeine offene Teilmenge von $\mathbb{R}$ derart, dass für ein gegebenes Element $x \in M$ auch $x \in G_x$ gilt. Dann ist intuitiv klar, dass

$$M \subset \bigcup_{\forall x\in M} G_x \tag{2.13}$$

[6] Giuseppe Vitali (1875 – 1932), italienischer Mathematiker

gilt. Es sei $\mathcal{G} = \{G_x | x \in M\}$ ein Mengensystem; $\mathcal{G}$ heißt dann *offene Überdeckung* von M. Gilt insbesondere $\mathcal{G} = \{G_{x_1}, G_{x_2}, \ldots, G_{x_n}\}$, $n < \infty$ und überdeckt $\mathcal{G}$ die Menge M, also $M \subset \mathcal{G}$, so wird gesagt, dass $\mathcal{G}$ eine *endliche Überdeckung* von M enthält. Der Begriff der Überdeckung soll an einigen einfachen Beispielen illustriert werden:

1. Es sei $M = \{1, 2, \ldots, n\}$, $G_k = \{k-1, k+1\}$ für $k = 1, \ldots, n$, $\mathcal{G} = \{G_1, G_2, \ldots, G_n\}$ ist dann eine offene Überdeckung von M, die für $n < \infty$ auch eine endliche Überdeckung ist.
2. Nun sei $M = \mathbb{N}$ und $G_k = \{k-1, k+1\}$. $\mathcal{G} = \{G_1, G_2, \ldots\}$ ist eine offene Überdeckung, – enthält aber keine endliche Überdeckung.
3. Es sei $M = (0, 1)$. Für $x \in M$ sei $G_x = \{x/2, 3x/2\}$. Eine endliche Überdeckung hätte die Form $\mathcal{G} = G_{x_1}, G_{x_2} \cup \ldots \cup G_{x_m}$, $m < \infty$. Für $x_1 < x_2 < \cdots < x_m$ gehört aber z.B. $x_1/3 \notin \mathcal{G}$, d.h. es existiert keine endliche Überdeckung von M.
4. Nun sei $M = [0, 1]$, d.h. die Randpunkte 0 und 1 gehören nun zu M. Das unter 3. definierte System $\mathcal{G}$ werde durch $G_0 = \{-1/10, 1/19\}$ und $G_1 = \{-9/10, 9/10\}$ ergänzt. Dann ist $\mathcal{G}$ eine offene und endliche Überdeckung von M.

Es gibt Mengen, die für jede Überdeckung eine endliche Überdeckung enthält; diese Mengen werden im folgenden Satz charakterisiert:

Satz 2.2.9 *(Satz von Heine-Borel:) Es sei $M \subset \mathbb{R}$. M ist kompakt genau dann, wenn jede ihrer Überdeckungen eine endliche Überdeckung enthält.*

Beweis: z.B. Heuser (2001), p. 229. □

Es sei $\Omega \subseteq \mathbb{R}$. Es gibt Teilmengen von Ω, für die zwar ein Maß existiert, das aber stets gleich Null ist. Solche Mengen heißen *Nullmengen* oder *vernachlässigbare Mengen*. Insbesndere zeigt sich, dass abzählbare Teilmengen und damit auch die Menge $\mathbb{Q}$ der rationalen Zahlen selbst das Maß Null haben. Für die praktische Answendung der Wahrscheinlihkeitstheorie spielen diese Mengen kaum eine Rolle, für das theoretische Verständnis des Wahrscheinlichkeitsbegriff aber schon, zumal sie die Definition bestimmter Maße motivieren.

Satz 2.2.10 *Die Menge $\mathbb{Q}$ kann von abzählbar vielen Intervallen beliebig kleiner Länge überdeckt werden.*

Beweis: Da $\mathbb{Q}$ abzählbar ist, kann $\mathbb{Q}$ in der Form $\mathbb{Q} = \{r_1, r_2, r_3, \ldots\}$ dargestellt werden, wobei $r_k \in \mathbb{Q}$ ist, $k \in \mathbb{N}$. Sei $\varepsilon > 0$ beliebig (klein); r_k liegt dann in einem Intervall

$$I_k = (r_k - \varepsilon/2^{k+1}, r_k + \varepsilon/2^{k+1}). \tag{2.14}$$

I_k hat die Länge

$$|I_k| = r_k + \varepsilon/2^k - (r_k + \varepsilon/2^k) = \frac{\varepsilon}{2^k}.$$

Dann ist

$$\sum_{k=1}^{n} |I_k| = \frac{\varepsilon}{2}\left(1 + \frac{1}{2} + \frac{1}{2^2} + \cdots + \frac{1}{2^{n-1}}\right).$$

Die Reihe in der Klammer ist von der Form $S_n = 1+q+q^2+\cdots+q^n$ mit $q = 1/2$. Bekanntlich gilt

$$S_n = \frac{1-q^{n+1}}{1-q},$$

und mit $q = 1/2$ ergibt sich $\lim_{n\to\infty} S_n = 2$. Dann folgt

$$\sum_{k=1}^{\infty} |I_k| = \varepsilon. \tag{2.15}$$

Dies heißt, dass $\mathbb{Q}$ durch eine abzählbare Anzahl von Intervallen I_k überdeckt werden kann. □

Folgerung: Da ε beliebig klein gewählt werden kann, folgt für $\varepsilon \to 0$ auch $|I_k| \to 0$ und damit auch $\sum_k |I_k| \to 0$. □

Definition 2.2.9 *Es sei M eine Menge, die durch abzählbar viele Intervalle überdeckt werden kann. Dann heißt M* Nullmenge.

Da $\mathbb{Q}$ durch abzählbar viele Intervalle überdeckt werden kann, ist $\mathbb{Q}$ offenbar eine Nullmenge. Dieser Sachverhalt hat auf den ersten Blick gegenintuitiv erscheinende Implikationenen in Bezug auf die Wahrscheinlichkeit, mit der Werte aus $\mathbb{Q}$ auftreten können, wenn zufällige Ereignisse durch Teilmengen reeller Zahlen repräsentiert werden.

Dies ist bemerkenswert, denn die Implikation von Satz 2.2.2 ist ja, dass die rationalen Zahlen in $\mathbb{R}$ überall dicht liegen, d.h. zwischen zwei rationalen Zahlen liegen immer unendlich viele weitere rationale Zahlen, aber jede rationale Zahl ist wegen der Überdeckung durch die I_k stets in irrationale Zahlen eingebettet.

Satz 2.2.11 *Es sei M eine kompakte Menge. Dann ist M genau dann eine Nullmenge, wenn für jedes $\varepsilon > 0$ eine endliche Überdeckung offener oder abgeschlossener Intervalle existiert, deren Längensumme $\leq \varepsilon$ ist.*

Beweis: Wegen der vorausgesetzten Kompaktheit von M existiert nach dem Satz von Heine-Borel eine endliche Überdeckung von M. □

Anmerkung: Satz 2.2.10 impliziert offenbar, dass $\mathbb{Q}$ eine Nullmenge ist. Dieser Sachverhalt hat eine intuitiv nicht sehr naheliegende Implikation: in einem Experiment werden (Mess-)Werte beobachtet, die aus einem Kontinuum stammen, zum Beispiel Zeiten, d.h. es wird gemessen, wie lange ein Prozess dauert, und die dafür beobachtete Zeit variiert unter sonst konstanten Bedingungen zufällig von einem experimentellen Durchgang zum anderen. Dann ist die Wahrscheinlichkeit,

dass die Zeit einen Wert aus $\mathbb{Q}$ annimmt, gleich Null, – obwohl es nicht unmöglich ist, dass ein Wert aus $\mathbb{Q}$ auftritt. Für die praktische Forschung hat dieser Sachverhalt kaum Konsequenzen, zumal Zeiten nur mit endlicher Genauigkeit gemessen werden können, aber er illustriert eine Eigentümlichkeit des Wahrscheinlichkeitsbegriffs. Vergl. etwa die Exponentialverteilung in Abschnitt 5.2.2. □

Im Folgenden werden noch einige Maße vorgestellt. die für die Definition des Wahrscheinlichkeitsbegriffs von Belang sind.

Definition 2.2.10 *Es sei*

$$\lambda^n(I^n_{a,b}) = \prod_{i=1}^{n}(b_i - a_i) \tag{2.16}$$

ein Maß, wobei

$$I^n_{a,b} = (a_1, b_1] \times (a_2 - b_2] \times \cdots \times (a_n, b_n] \in \mathbb{R}^n.$$

Dann heißt $\lambda^n(I^n_{a,b})$ Lebesgue-Maß.

Die Intervalle $(a_i, b_i]$ sind also halboffene Intervalle. Die halb-offenen Mengen sind *Borel-Mengen.* Für $n = 1$ ist das Lebesgue-Maß gerade gleich der Länge eines Intervalls, für $n = 2$ erhält man die Fläche eines Rechtecks, etc. $\lambda^n(I^n_{a,b})$ heißt *Lebesgue-Maß* oder auch *Lebesgue-Borel-Maß.* Über Lebesgue-Borel-Maße lassen sich Wahrscheinlichkeitsmaße über *maßerzeugende Funktionen* erklären, worauf in Abschnitt 2.2 näher eingegangen werden wird. Hier sei angemerkt, dass wegen (2.14) Nullmengen und damit abzählbare Mengen nicht in das Maß $\lambda^n_{a,b}$ eingehen; so tragen z.B., die rationalen Zahlen nichts zum Wert von $\lambda^n_{a,b}$ bei. Diese Eigenschaft vererbt sich auf allgemeine Wahrscheinlichkeitsmaße.

Definition 2.2.11 *Es sei* μ *ein Maß auf* $(\Omega, \mathcal{A})$*,* $\mathcal{A}$ *eine* σ*-Algebra, und es gelte insbesondere für* $\omega \in \Omega$

$$\mu\colon \mathcal{A} \to \mathbb{N}, \quad \mu(A) = \begin{cases} |A|, & |A| < \infty, \\ \infty, & \textit{sonst} \end{cases} \tag{2.17}$$

Dann heißt μ Zählmaß.

Anmerkungen: Die Null gehört in diesem Text stets zu $\mathbb{N}$, also $\mathbb{N} = \{0, 1, 2, \ldots\}$. Ein Zählmaß gibt also einfach die Anzahl der Elemente in einer Menge an. □

Definition 2.2.12 *Es sei* μ *ein Maß auf* $(\Omega, \mathcal{A})$*,* $\mathcal{A}$ *eine* σ*-Algebra und* μ *ein Maß. Es gelte*

$$\delta(A) = \begin{cases} 0, & \omega \notin A \in \mathcal{A}, \\ 1, & \omega \in A \in \mathcal{A} \end{cases} \tag{2.18}$$

Dann heißt δ Dirac-Maß.

Bemerkung: Paul Dirac (1902– 1984), britische Physiker, führte die nach ihm benannte Dirac-Funktion ein: $\delta(x) = 0$ für $x \neq 0$, $\delta(x) = \infty$ für $x = 0$ und $\int_{\mathbb{R}} \delta(x)dx = 1$; es handelt sich dabei um eine *verallgemeinerte Funktion.* Es gilt $\int(\delta(x - x_0)f(x)dx = f(x_0)$, wobei das Integral von $-\infty$ bis $+\infty$ läuft. Mit der Dirac-Funktion lassen sich also Funktionswerte für bestimmte der unabhängigen Variablen bestimmen. Das Dirac-Maß leistet genau dies, weil es erlaubt, Maße für singuläre x-Werte zu definieren. Man spricht auch von einem *Punktmaß* oder von einer *Punktmasse.*

2.2.3 Ereignisse und Ereignisalgebren

Ergebnisse und Ereignisse Um den Begriff des zufälligen Ereignisses zu präzisieren, wird zunächst der Begriff des Versuchs eingeführt. Ein Versuch ist eine durch eine Reihe von Bedingungen definierte Situation oder Anordnung, in der *bestimmte* Ergebnisse beobachtet werden können. Man wird also nicht *alle* überhaupt möglichen Ereignisse betrachten: ein plötzlicher Stromausfall, der Absturz eines Sportflugzeugs auf das Laborgebäude, der Stich einer durch das geöffnete Fenster eingedrungenen Wespe etc sind unvorhergesehehene und deshalb zufällige Ereignisse, auf die allerdings beim einem bestimmten Versuch nicht fokussiert wird. Dies sind Ereignisse, die mit dem Versuch *verbunden* sind:

Definition 2.2.13 *Gegeben sei ein Versuch mit möglichen Versuchsergebnissen $\omega \in \Omega$, wobei Ω sei die Menge der als Resultat des Versuchs möglichen Ergebnisse sei. Ω heißt* Stichprobenraum. *A sei ein zufälliges Ereignis, das dann als eingetreten gilt, wenn ein Ergebnis $\omega \in \Omega_A \subseteq \Omega$ eingetreten ist. A heißt mit dem Versuch* verbunden *genau dann, wenn nach Durchführung des Versuchs entschieden werden kann, ob A eingetreten ist oder nicht.*

Den Ausdruck 'Stichprobenraum' (engl: *sample space*) hat man eingeführt, weil ja jede Menge von Messungen oder allgemein Beobachtungen ω zu einer Stichprobe aus der Menge der überhaupt möglichen Beobachtungen führt.

Beispiel 2.2.1 Beim Münzwurf besteht der Stichprobenraum aus 'Kopf', 'Zahl' und 'Rand', $\Omega = \{Kopf, Zahl, Rand\}$, beim Würfel ist er durch $\Omega = \{1, 2, 3, 4, 5, 6\}$ gegeben. □

Beispiel 2.2.2 Bei einem Versuch zur Verkehrssicherheit sind Verkehrsunfälle an einer bestimmten Kreuzung an bestimmten Wochentagen zu bestimmten Tageszeiten von Interesse. $\Omega = \{1, 2, 3, \ldots\}$ enthält die möglichen Anzahlen k von Verkehrunfällen in den genannten Zeitabschnitten. A sei das zufällige Ereignis, dass $k > 3$, und die zugehörige Teilmenge Ω_A aus Ω ist $\Omega_A = \{4, 5, 6, \ldots\}$. Einer der im Zusammenhang mit dem Versuch den Verkehr beobachtenden Polizisten erleidet einen Herzinfarkt. Der Herzinfarkt war nicht vorhergesehen worden und

ist insofern ein zufälliges Ereignis, das allerdings nicht mit dem Versuch verbunden ist, da es sich nicht um einen Verkehrsunfall handelt. □

Beispiel 2.2.3 Bei einem Versuch in einem neurowissenschaftlichen Labor sollen elektrische Ströme durch Ionenkanäle in der Zellmembran eines bestimmten Typs von Neuronen untersucht werden. Der Stichprobenraum Ω ist dann die Menge der Verläufe $i(\cdot)$ während einer Messung; $i(\cdot)$ ist eine Funktion der Zeit, $i(t)$ ist die Stromstärke zur Zeit t, d.g. $\Omega = \{i(\cdot)|t_0 \leq t \leq t_1\}$, und $[t_0, t_1]$ ist das Zeitintervall, innerhalb dessen ein Strom gemessen wird. Es wird angenommen, dass die $i(\cdot)$ innerhalb $[t_0, t_1]$ stetig und differenzierbar sind. A sei das zufällige Ereignis, dass das Maximum des gemessenen Stroms, also des gemessenen Verlaufs $i(\cdot)$, kleiner als ein bestimmter Wert i_{max} ist. A ist eingetreten, wenn ein Verlauf aus einer Teilmenge Ω_A der Verläufe mit einem Maximum kleiner als i_{max} aus der Menge Ω der überhaupt möglichen Verläufe innerhalb $[t_0, t_1]$ gemessen wurde. Während der Messungen kommt es zu einem Stromausfall im gesamten Laborgebäude. Der Stromausfall war nicht vorhersehbar, da durch Blitzeinschlag verursacht, und ist deshalb ein zufälliges Ereignis, das allerdings nicht mit dem Versuch der Messung von Ionenströmen verbunden ist. □

Man kann drei Arten von zufälligen Ereignissen unterscheiden:

1. $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ mit $n < \infty$. Ω ist also endlich; einfache Beispiele sind die Versuche des Münzwurfs mit $\Omega = \{$Kopf, Zahl, Rand$\}$, oder des Würfelwurfs mit $\Omega = \{1, 2, 3, 4, 5, 6\}$, aber auch die Klausur ist ein Beispiel, wobei $\Omega = \{0, 1, 2, \ldots, j, \ldots, n\}$ die Menge der möglichen Ergebnisse ist; – j ist die Anzahl der Punkte, die ein Teilnehmer der Klausur insgesamt erhält.

2. $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n, \ldots\}$. Die möglichen Ergebnisse $\omega = \omega_j$ können durchnummeriert werden, aber ihre Anzahl ist nicht endlich, sondern unendlich; man spricht von einer *abzählbar unendlichen* Menge Ω. Ein einfaches Beispiel für eine solche Menge liefert das Roulette-Spiel: $\omega_j = j$ ist ist die Anzahl der Versuche, die benötigt werden, bis zum ersten Mal die Kugel auf ein rotes Feld fällt, $j \leq 1$. Da sich die Wahrscheinlichkeit für "rot" oder "schwarz" von einem Wurf zum nächsten nicht ändert kann es *im Prinzip* unendlich viele Würfe benötigen, bis die Kugel auf einem roten Feld landet. Dass die Wahrscheinlichkeit für unendlich viele Würfe gleich Null ist, ist dabei unerheblich, denn eine Wahrscheinlichkeit von Null bedeutet noch nicht die Unmöglichkeit eines solchen Ergebnisses, wie sich noch zeigen wird.

3. $\Omega = \{x|x \in I\}$, wobei $I = \{x|a \leq x < b, a, b, x \in \mathbb{R}$, d.h. I ist eine Teilmenge der Menge $\mathbb{R}$ reellen Zahlen, oder gleich der Menge der rellen Zahlen, in Abhängigkeit von der Definition von $a, b \in \mathbb{R}$. In diesem Fall ist jedes Element des Kontinuums ein mögliches Ergebnis. Ein einfaches Beispiel ist die

Messung einer Reaktionszeit t: hier ist $\Omega = \{t|t \in (0, \infty)\}$, oder die Stunde von 15 h bis 16 h, innerhalb der Kurt seine Freundin Lisa anrufen will, – so hat er es ihr versprochen. Der Anruf kann zu jedem Zeitpunkt innerhalb des Intervalls $I = [15, 16]$ erfolgen. Die Menge der Zeitpunkte in Ω bzw. I kann nicht durchnummeriert werden, I bzw. Ω enthält *überabzählbar unendlich* viele Elemente. Auch die Menge $\Omega = \{i(\cdot)|t_0 \leq t \leq t_1\}$ aus Beispiel 2.2.3 gehört in die Klasse der Stichprobenräume mit überabzählbar vielen Elementen.

Anmerkung: Ist Ω endlich oder abzählbar, so heißt Ω *diskret*.

Ergebnisse (Elementarereignisse) eines "Experiments" sind zufällige Ereignisse. Wie bereits ausgeführt lassen sich weitere zufällige Ereignisse als Teilmengen von Ω definieren. So können zwei Spieler vereinbaren, dass derjenige von ihnen gewinnt, der beim Würfeln eine gerade Zahl wirft, wenn also das zufällige Ereignis $A = \{2, 4, 6\} \subset \Omega = \{1, 2, 3, 4, 5, 6\}$ eintritt. Ein Roulette-Spieler mit begrenztem Kapital ist an dem zufälligen Ereignis interessiert, dass die Zahl k der Würfe, die ausgeführt werden, bis zum ersten Mal die Kugel auf einem roten Feld landet, kleiner als 6 ist, $A = \{k|k < 6\} \subset \Omega = \{1, 2, 3, \ldots\}$, denn er möchte durch Doublieren gewinnen; bei dieser Strategie wird nach jedem Wurf der Kugel der Einsatz verdoppelt, wenn die Kugel auf einem schwarzen Feld gelandet ist.

Die Menge *aller* möglichen zufälligen Ereignisse ist dann durch die Menge der Teilmengen $\mathfrak{P}(\Omega)$ (auch: 2^{Ω}) von Ω definiert. Der Punkt dabei ist aber, dass man kaum jemals an allen möglichen Teilmengen interessiert ist. Es sei noch einmal an Kurt erinnert, der versprochen hat, seine Freundin Lisa zwischen 15 h und 16 h anzurufen; jeder Zeitpunkt t mit $15 \leq t \leq 16$ ist ein für den Anruf möglicher Zeitpunkt. Zur Menge aller möglichen Teilmengen des Intervalls [15, 16] gehört die Menge der Zeitpunkte, die durch eine rationale Zahl repräsentiert werden, also durch eine Zahl, die durch einen Quotienten p/q mit $p, q \in \mathbb{N}$, $\mathbb{N}$ die Menge der natürlichen Zahlen, definiert ist: $A = \{t|15 \leq t \leq 16, t = p/q, p, q \in \mathbb{N}\}$. Lisa wird sich für dieses Ereignis kaum interessieren. Interessanter ist es, das Intervall [15, 16] in Subintervalle $I_1, I_2, \ldots, I_n$ aufzuteilen, die zufällige Ereignisse $t \in I_1, \ldots, t \in I_n$ definieren.

Die Frage ist nun, wie man überhaupt Ereignissen, also Teilmengen von Ω, Wahrscheinlichkeiten zuordnen soll. Einen völlig neuen Weg beschritt dabei der russische Mathematiker Andrei Nikolajewitsch Kolmogorov (1903 – 1987). Kolmogorov ging von Resultaten einer relativ neuen Theorie der Integration von Funktionen aus. Der von Bernhard Riemann (1826 – 1866) entwickelte Integralbegriff hatte sich als nicht hinreichend allgemein erwiesen, um bestimmte Volumen in höherdimensinalen Räumen berechnen zu können, so dass die Entwicklung eines allgemeineren Integralbegriffs nötig wurde. Wie schon angtemerkt, wurde dieses Problem von Émile Borel (1871 – 1956) und Henri Léon Lebesgue (1875 – 1956) gelöst. Der von ihnen eingeführte Begriff des Maßes einer Menge erlaubte eine neue, axiomatische Definition des Wahrscheinlichkeitsbegriffs. Nach

dieser Definition ist eine Wahrscheinlichkeit einfach, d.h. ohne weiteren Bezug auf inhaltliche Aspekte der Wahrscheinlichkeit, ein Maß auf einer Menge. Eine inhaltliche Charakterisierung als Maß subjektiver Sicherheit oder Ungewißheit bezüglich des Eintretens eines Ereignisses, oder als 'Propensität' (engl. propensity – Neigung zu etwas) natürlicher Prozesse, bestimmte Ereignisse zu begünstigen (eder Philosoph Karl Popper) muß gar nicht gegeben werden. Bei diskreten Ergebnismengen Ω kann man dabei im Prinzip die Menge aller Teilmengen von Ω betrachten, – auch man bei Anwendungen der Wahrscheinlichkeitstheorie normalerweise gar nicht an *allen* Teilmengen interessiert ist. Die Situation ist anders, wenn Ω durch ein Kontinuum, etwa durch ein Intervall $I = [a, b] \subset \mathbb{R}$ definiert ist. Nach Satz 2.2.8 existiert für bestimmte Teilmengen gar kein Maß und damit keine Wahrscheinlichkei. Man muß sich demnach auf bestimmte Klassen von Teilmengen beschränken. Dies sind die nach Borel benannten *Borel-Mengen.* Die Borel-Mengen sind – in $\mathbb{R}$ – alle Intervalle

$$(-\infty, a) = \{x \in \mathbb{R} | -\infty < x < a\},$$

zusammen mit abzählbaren Vereinigungen und Durchschnitten solcher Intervalle, und damit die Menge aller der offenen und abgeschlossenen Intervalle in $\mathbb{R}$ bzw $I \subset \mathbb{R}$. Dazu muß kurz erklärt werden, was unter einer offenen, er abgeschlossenen und einer halboffenen Menge verstanden wird. Es seien $a, b \in \mathbb{R}$ mit $a < b$. Das Intervall $(a, b) = \{x | a < x < b\}$ heißt *offen*, weil nur die rellen Zahlen zwischen a und b, nicht aber die Zahlen a und b selbst zum Intervall gehören. Dementsprechend heißt das Intervall $[a, b] = \{x | a \leq x \leq b\}$ *abgeschlossen*, – a und b gehören zum Intervall. Die Intervalle $(a, b] = \{x | a < x \leq b\}$ und $[a, b) = \{x | a \leq x < b\}$ heißen *halboffen*, weil entweder nur a oder nur b zum Intervall gehören. Ist z.B. $\Omega = \{x | a \leq x < b, a, b \in \mathbb{R}\}$ und will man dieses Intervall in Teilintervalle aufteilen, so werden diese Intervalle als halboffene Intervalle (= Teilmengen) definiert sein: $I_1 = \{x | a \leq x < t_1\}$, $I_2 = \{x | t_1 \leq x < t_2\} \ldots$, $I_j = \{x | t_{j-1} \leq x < t_j\}$, etc., denn die t_j können nicht zu zwei Teilintervallen gleichzeitig gehören. Die Borel-Mengen bilden eine σ-Algebra:

Definition 2.2.14 *Es sei $\mathcal{A}$ eine Menge von Teilmengen aus einer Menge Ω, für die die Bedingungen*

1. Sowohl Ω wie auch die leere Menge $\emptyset$ sind Element von $\mathcal{A}$,
2. Für jedes Element $A \in \mathcal{A}$ ist auch $A^c \in \mathcal{A}$,
3. Für jedes $n \geq 2$ und alle $A_1, A_2, \ldots, A_n \in \mathcal{A}$ ist $\bigcup_{k=1}^{n} A_k \in \mathcal{A}$.
Dann heißt $\mathcal{A}$ eine Algebra. *Gilt außerdem für alle $A_1, A_2, A_3, \ldots$ die Bedingung*
4. $\bigcup_{k=1}^{\infty} A_k \in \mathcal{A}$, so heißt $\mathcal{A}$ eine σ-Algebra.

Allen Teilmengen, die eine σ-Algebra bilden, läßt sich nach Lebesgue ein Maß zuordnen, so dass statt von Borel-Mengen auch von (Lebesgue-) messbaren Mengen gesprochen wird.

Generell wird man in Bezug auf einen Versuch eine Menge $\mathcal{A}$ von als Teilmengen von Ω definierten zufälligen Ereignissen erklären. Der Versuch korrespondiert dann zu einem Paar $(\Omega, \mathcal{A})$; etwas lax wird gelegentlich auch gesagt, das Paar $(\Omega, \mathcal{A})$ sei ein 'zufälliger Versuch'.

Es sei also ein durch ein Paar $(\Omega, \mathcal{A})$ beschriebener Versuch gegeben, wobei Ω die Menge der möglichen Ergebnisse und $\mathcal{A}$ die Menge der mit dem Versuch verbundenen zufälligen Ereignisse sei. Für zufällige Ereignisse A, B C aus $\mathcal{A}$ gelten dann die folgenden Aussagen:

1. *Komplementärereignisse:* Ist A ein zufälliges Ereignis aus $\mathcal{A}$, so ist das Ereignis, dass A nicht eintritt, ebenfalls ein zufälliges Ereignis aus $\mathcal{A}$; es wird oft mit A^c bezeichnet und heißt das zu A *komplementäre Ereignis* (das c in A^c steht für das englische 'complementary'). Alternative Notationen sind $\bar{A}$, $\sim A$ oder $\neg A$, – man spricht auch von "nicht-A". In diesem Text werden hauptsächlich die Bezeichungen $\bar{A}$ oder $\neg A$ verwendet werden. Natürlich ist A auch das Komplementärereignis von A^c, denn $(A^c)^c = A$, ebenso wie $\neg(\neg A) = A$, etc.

2. *Konjunktion von Ereignissen:* Es seien A und B zwei beliebige zufällige Ereignisse aus $\mathcal{A}$. Dann ist $C := A$ und B ebenfalls ein zufälliges Ereignis aus $\mathcal{A}$. C ist eingetreten, wenn sowohl die Ereignisse A wie auch B eingetreten sind. Dieses Ereignis heißt die *Konjunktion* der Ereignisse A und B; man schreibt $\cap$ oder $\wedge$ oder gelegentlich auch & für "und", d.h. $C = A \cap B$, $C = A \wedge B$ oder $C = A\&B$. In manchen Texten wird auch von der *polnischen Notation* $C = AB$ Gebrauch gemacht. In diesem Text werden nur die Zeichen $\cap$ oder $\wedge$ verwendet werden.

   Es seien A_k, $k = 1, \ldots, n$ zufällige Ereignisse aus $\mathcal{A}$. Dann ist auch

   $$A_1 \cap \cdots \cap A_n = \bigcap_{k=1}^{n} A_k$$

   ein zufälliges Ereignis aus $\mathcal{A}$.

3. *Vereinigung von Ereignissen:* Es seien A und B irgend zwei zufällige Ereignisse aus $\mathcal{A}$. Dann ist $D := A$ oder B wieder ein zufälliges Ereignis; D ist eingetreten, wenn entweder A, oder B oder beide Ereignisse A, B eingetreten sind. Bei diesem Oder handelt es sich also um das einschließende Oder. D heißt *Vereinigung* der Ereignisse A und B. Man schreibt $\cup$ oder $\vee$ für das einschließende Oder, also $D = A \cup B$ oder $D = A \vee B$.

   Es seien A_k, $k = 1, \ldots, n$ zufällige Ereignisse aus $\mathcal{A}$. Dann ist auch

   $$A_1 \cup \cdots \cup A_n = \bigcup_{k=1}^{n} A_k$$

   ein zufälliges Ereignis aus $\mathcal{A}$.

4. *Das sichere und das unmögliche Ereignis:* Die Ereignisse A und A^c aus Ω können offenbar nicht zugleich eintreten. Demnach heißt die Konjunktion $A \cap A^c$ das *unmögliche Ereignis.* Für dieses Ereignis hat man das Symbol $\emptyset$, so dass $A \cap A^c = \emptyset$.

   Andererseits muß eines der Ereignisse aus Ω eintreten: $A \cup A^c$ ist demnach das *sichere Ereignis.* Offenbar gilt $A \cup A^c = \Omega$, so dass Ω auch das sichere Ereignis bezeichnet.

   Es sei A ein beliebiges zufälliges Ereignis. Für das sichere Ereignis Ω und für das unmögliche Ereignis $\emptyset$ gelten die Aussagen

   (i) $\Omega \cup A = \Omega$, $\Omega \cap A = A$,

   (ii) $\emptyset \cup A = A$, $\emptyset \cap A = \emptyset$.

   Denn Ω ist das sichere Ereignis, und $\Omega \cup A$ tritt ein, wenn Ω oder A oder beide eintreten. Aber Ω tritt stets ein, und damit ist dann auch stets $\Omega \cup A$ eingetreten. $\Omega \cap A$ tritt ein, wenn Ω und A eintreten; also kann $\Omega \cap A$ nur eintreten, wenn A eingetreten ist. Die Aussagen (ii) folgen ebenfalls sofort aus den Definitionen von $\cap$ und $\cup$.

5. *Kombinationen von zufälligen Ereignissen:* Da mit den zufälligen Ereignissen A und B aus $\mathcal{A}$ auch die Komplementärereignisse $\bar{A}$ und $\bar{B}(\neg B)$ zufällige Ereignisse aus $\mathcal{A}$ sind folgt sofort, dass mit $A \cup B$ und $A \cap B$ auch $\bar{A} \cup B$, $A \cup \bar{B}$, $\bar{A} \cup \bar{B}$, $\bar{A} \cap B$, $A \cap \bar{B}$ und $\bar{A} \cap \bar{B}$ zufällige Ereignisse aus $\mathcal{A}$ sein müssen.

6. *Implikation von Ereignissen:* Das Ereignis $D = \bar{A} \cup B$ mit $A, B \in \mathcal{A}$ tritt genau dann ein, wenn mindestens eines der Ereignisse $\bar{A}$, B eingetreten ist. Ist nun A eingetreten, so ist nicht $\bar{A}$ eingetreten; damit D als eingetreten gelten kann, muß nun B eingetreten sein. In diesem Sinne impliziert das Ereignis A das Ereignis B; man schreibt $A \to B$.

7. *Äquivalente Ereignisse:* Gilt einerseits $A \to B$ und andererseits $B \to A$, so heißen A und B auch *äquivalent* und man schreibt $A \equiv B$. Alle unmöglichen Ereignisse sind äquivalent, und ebenso sind alle sicheren Ereignisse äquivalent.

8. *Differenz von Ereignissen:* Es seien $A, B \in \Omega$ zufällige Ereignisse. Dann ist $A \cap \bar{B}$ wieder ein zufälliges Ereignis, das eintritt, wenn A, aber nicht B eintritt. $A \cap \bar{B}$ ist die *Differenz der Ereignisse* A und B; andere Schreibweisen dafür sind $A \setminus B$ oder $A - B$. Da $A \cap \bar{B} \neq \bar{A} \cap B$, folgt $A - B \neq B - A$, d.h. die Differenz von Ereignissen ist nicht symmetrisch. Vergl Abbildung 2.2 (b), Seite 55.

9. *Symmetrische Differenz von Ereignissen:* Es gelte insbesondere $C := A \cap \bar{B}$ und $D := \neg A \cap B$. Das Ereignis $C \cup D$ tritt ein wenn entweder nur A

oder nur B eingetreten ist, nicht aber eine Kombination von A mit B. Man schreibt für C auch $A\Delta B$; Δ definiert das *ausschließende Oder* oder die *symmetrische Differenz.*

10. *Disjunktive und konträre Ereignisse:* Gilt $A \cap B = \emptyset$ so heißen die zufälligen Ereignisse $A, B \in \Omega$ *disjunkt* oder *unvereinbar* miteinander. A und $\neg A = A^c$ sind notwendig disjunkt.

    Gilt für $A, B \in \Omega$ die Aussage $A \cup B = \Omega$, $A \cap B = \emptyset$, so heißen A und B *konträr* und es gilt $B = \bar{A} = A^c$.

    Tritt $A \cup B$ mit Sicherheit ein, so schreibt man $A \cup B = \Omega$. Die Ereignisse $\emptyset$ und Ω sind komplementär.

11. *DeMorgansche Regeln:* Es seien $A, B \in \Omega$; dann gelten für die Disjunktion und für die Konjunktion die beiden Regeln

$$A \cup B = \overline{(\bar{A} \cap \bar{B})} = \neg(\bar{A} \cap \bar{B}) \tag{2.19}$$

$$A \cap B = \overline{(\bar{A} \cup \bar{B})} = \neg(\bar{A} \cup \bar{B}). \tag{2.20}$$

    Denn es sei $C = A \cup B$. C tritt ein, wenn A oder B oder beide Ereignisse eintreten. Dann tritt $\neg C$ ein, wenn weder A noch B eintreten, d.h. wenn $\bar{A} \cap \bar{B}$ eintritt; d.h. aber $C = \neg(\bar{A} \cap \bar{B})$. $D = A \cap B$ ist eingetreten, wenn sowohl A als auch B eingetreten sind. D ist nicht eingetreten, wenn A und $\bar{B}$ oder $\bar{A}$ und B oder $\bar{A}$ und $\bar{B}$ eintreten; dies ist aber äquivalent $\bar{A} \cup \bar{B}$, so dass $D = \overline{(\bar{A} \cup \bar{B})} = \neg(\bar{A} \cup \bar{B})$.

    Die Beziehungen (2.19) und (2.20) heißen *DeMorgansche Regeln.* Sie lassen sich auf mehr als zwei Ereignisse verallgemeinern:

$$A_1 \cup \cdots \cup A_n = \neg(\bar{A}_1 \cap \cdots \cap \bar{A}_n) \tag{2.21}$$

$$A_1 \cap \cdots \cap A_n = \neg(\bar{A}_1 \cup \cdots \cup \bar{A}_n) \tag{2.22}$$

Über die Negation $\neg$ bzw. die Komplementbildung sowie die Vereinigung $\cup$ lassen sich also alle möglichen Kombinationen von zufälligen Ereignissen bilden, die selber wieder zufällige Ereignisse sind. Man kann sagen, dass die Menge der zufälligen Ereignisse gegenüber der Negation $\neg$, der Vereinigung $\cup$ und der Konjunktion $\cap$ *abgeschlossen* sind. Es sei $\mathcal{A}$ eine Menge zufälliger Ereignisse (Teilmengen aus Ω). Im Grunde genügt es, festzustellen, dass für jedes Ereignis A auch $A^c = \neg A \in \mathcal{A}$, also das zugehörige Komplementärereignis in $\mathfrak{P}a$ ist, denn wie auch immer A aus anderen Ereignissen zusammengesetzt ist, es ist am Ende stets als eine Teilmenge von Elementen aus Ω definiert. In der Mathematik haben derartige abgeschlossene Mengen , wegen ihrer Rolle bei bestimmten mathematischen Betrachtungen, einen speziellen Namen:

Bei endlichen Mengen Ω, etwa für einen Würfel, macht man sich leicht klar, dass die Menge aller Teilmengen von Ω sicherlich eine σ-Algeba bildet. Da Ω eine

endliche Menge ist, könnte man den Punkt 4. als problematisch empfinden. Aber nach 1. ist ja $\emptyset$ stets in $\mathcal{A}$ enthalten, so dass man die Folge A_1, $A_2 = A_1 \cup \emptyset$, $A_3 = A_1 \cup \emptyset \cup \emptyset$ etc bilden kann, wobei A_1 durch eine Teilmenge von $\{1, 2, 3, 4, 5\}$ definiert ist, und da $A_2 = A_1 \cup \emptyset = A_1$ ist und eine analoge Feststellung für alle weiteren A_k gilt, ist 4. trivialerweise erfüllt.

2.2.4 Kolmogorovs axiomatische Definition

Kolmogorovs Definition vermeidet die Nennung aller inhaltlichen Aspekte des Wahrscheinlichkeitsbegriffs und fokussiert auf die formale Struktur der Zuordnung von Wahrscheinlichkeiten zu Ereignissen. Es genügt, Wahrscheinlichkeit als normiertes (das Maß ist maximal gleich 1), additives Maß zu spezifizieren, wobei hier nicht ausführlich auf maßtheoretische Details eingegangen wird – eine vorzügliche moderne Darstellung findet man in Billingsley (1979)[7].

Wie man den Ereignissen konkrete, numerische Werte zuweist, geht nicht in die Definition ein; dazu muß man Annahmen formulieren, die nicht Teil der Definition des Begriffs der Wahrscheinlichkeit sind. In bestimmten Fällen *kann* man annehmen, dass gewisse Ereignisse - die von Kolmogorov so genannten Elementarereignisse, sie entsprechen dem im englischen gebräuchlichen Ausdruck *outcome* bzw. dem deutschen Ergebnis (eines Münzwurfs, einer Messung, etc) – gleichwahrscheinlich sind, muß es aber nicht. In der mathematischen Statistik wird diskutiert, unter welchen Bedingungen Wahrscheinlichkeiten durch relative Häufigkeiten geschätzt werden können, – die Art der Schätzung von Wahrscheinlichkeiten ist ebenfalls nicht Bestandteil der Definition von Wahrscheinlichkeiten.

Definition 2.2.15 *(Kolmogorov)*[8]*: Es sei Ω ein Stichprobenraum und $\mathcal{A}$ sei eine auf Ω erklärte σ-Algebra. Bezüglich ($\mathcal{A}$, A) wird ein Wahrscheinlichkeitsmaß p erklärt mit den folgenden Eigenschaften:*

(K1) $p(\Omega) = 1$

(K2) Für alle $A \in \mathcal{A}$, $p(A) \geq 0$

(K3) Sind A_1, A_2, $\cdots$ disjunkte Teilmengen aus $\mathcal{A}$, so gilt

$$p(A_1 \cup A_2 \cup \cdots) = p(A_1) + p(A_2) + \cdots$$

Die Anzahl der Teilmengen A_i, $i = 1, 2, \cdots$ ist nicht notwendig endlich. Man sagt, das Tripel (Ω, $\mathcal{A}$, p) definiere einen Wahrscheinlichkeitsraum*, auch kurz W-Raum genannt.*

[7] Billilngsley, P.: *Probability and Measure*, John Wiley & Sons, New York 1979

[8] Kolmogorov, A.N.: Grundbegriffe der Wahrscheinlichkeitsrechnung, Springer-Verlag, Berlin 1933

Es wird keinerlei Einschränkung in Bezug auf Ω gemacht: Ω kann endlich, abzählbar oder überabzählbar unendlich sein. Bevor auf Fragen der Bestimmung numerischer Werte für $p(A)$ eingegangen wird, sollen einige unmittelbare Implikationen der Kolmogovschen Definition hergeleitet werden. Sie sind nützlich für die Bestimmung numerischer Werte.

Satz 2.2.12 *Es sei Ω eine Menge von Elementarereignissen und $\mathcal{A}$ eine auf Ω erklärte σ-Algebra. p sei ein Wahrscheinlichkeitsmaß, dass der Kolmogorovschen Definition genügt. Für ein beliebiges $A \in \mathcal{A}$ folgt*

(a) $p(\emptyset) = 0$

(b) $p(\bar{A}) = 1 - p(A)$

(c) Für $A \subset B \subset \Omega$ folgt $p(\bar{A} \cap B) = p(B) - p(A)$

(d) Für $A \subset B \subset \Omega$ folgt $p(A) \leq p(B) \leq 1$

(e) Für $A \in \mathcal{A}$, $B \subset A$ gilt $p(A \cup B) = p(A) + p(B) - p(A \cap B)$.

Beweis:

(a) Es ist $\Omega = \Omega \cup \emptyset\}$, und Ω und $\emptyset$ sind disjunkt, d.h. $\Omega \cap \emptyset$. Nach K3 gilt dann aber $p(\Omega) = p(\Omega \cup \emptyset) = p(\Omega) + p(\emptyset)$. Wegen K1 ist aber $p(\Omega) = 1$, so dass $1 = 1 + p(\emptyset)$, woraus $p(\emptyset) = 0$ folgt.

(b) Es ist $A \cup \bar{A} = \Omega$ und $A \cap \bar{A} = \emptyset$, also nach K1 und K3 $p(A \cup \bar{A}) = p(\Omega) = p(A) + p(\bar{A}) = 1$, mithin $p(\bar{A}) = 1 - p(A)$.

(c) $A \subset B$ impliziert $B = A \cup (\bar{A} \cap B)$, wobei $A \cap (\bar{A} \cap B) = \emptyset$, d.h. A und $\bar{A} \cap B$ sind disjunkt. Nach K3 gilt dann $p(B) = p(A) + p(\bar{A} \cap B)$, und daraus folgt sofort $p(\bar{A} \cap B) = p(B) - p(A)$.

(d) Aus dem Beweis von (c) folgt wegen K2 $p(\bar{A} \cap B) \geq 0$ schon $p(B) \geq p(A)$. Nach (a) ist aber $p(B) = 1 - p(\bar{B})$ und nach K2 ist $p(\bar{B}) \geq 0$; deshalb ist $p(B) \leq 1$. Dies gilt für alle $B \subset \Omega$.

(e) $A = (A \cap B) \cup (A \cap \bar{B})$, $B = (B \cap A) \cup (B \cap \bar{A})$, und $A \cup B = (A \cap \bar{B}) \cup (B \cap \bar{A})(A \cap B)$, so dass $p(A \cup B) = p(A \cap \bar{B}) + p(\bar{A} \cap B) + p(A \cap B)$. Andererseits ist $p(A) + p(B) = 2p(A \cap B) + p(A \cap \bar{B}) + p(\bar{A} \cap B)$. Der Vergleich dieser beiden Gleichungen liefert (e).

Abb. 2.2 (b), Seite 55, verdeutlicht den Satz. Offenbar ist $A = (A - B) \cup (A \cap B)$, $B = (B - A) \cup (A \cap B)$, und $A - B)$, $B - A$ und $A \cap B$ sind paarweise diskjunkt. Also ist einerseits $P(A \cup B) = P(A - B) + P(B - A) + P(A \cap B)$, andererseits ist $P(A) + P(B) = P(A - B) + P(B - A) + 2P(A \cap B)$, so dass $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. $\square$

Abzählbares Ω: Die Menge Ω der Elementarereignisse sei abzählbar. Dann können die Elementarereignisse durchnummeriert werden; es sei ω_i das i-te Elementarereignis, $i = 1, ...$ Zwei beliebige Elementarereignisse ω_i und ω_j, $i \neq j$, können nicht zugleich auftreten (sonst wären sie entgegen der Annahme identisch, oder sie hätten gemeinsame Elemente, d.h. sie wären zerlegbar, aber dann wären sie keine Elementarereignisse). Da $\Omega \in \mathcal{A}$ für jede σ-Algebra auf Ω gilt $\omega_i \in \mathcal{A}$ für alle i. Es sei $A = \{\omega_i | r \leq i \leq s,\ 1 \leq r < s \leq n\}$ ein beliebiges Ereignis. Wegen K3 folgt dann sofort

$$p(A) = p(\omega_r) + \cdots + p(\omega_s) = \sum_{i=1}^{n} p(\omega_i) \tag{2.23}$$

$p(A)$ läßt sich also aus den Wahrscheinlichkeiten der Elementarereignisse berechnen. Die Argumentation gilt natürlich auch für $|\Omega| < \infty$. .

Überabzählbares Ω: Ω sei überabzählbar. Diese Situation ist bei Messungen von Werten aus einem Kontinuum von Werten gegeben. Beispiele sind Messungen von Zeiten, Positionen, etc. Um von Wahrscheinlichkeiten ungleich Null sprechen zu können, müssen Ereignisse durch Intervalle definiert werden.

Definition 2.2.16 *Es sei $F\colon \mathbb{R} \to \mathbb{R}$ eine Funktion mit den Eigenschaften*
1. $\lim_{x\to-\infty} F(x) = 0, \quad \lim_{x\to\infty} F(x) = 1,$
2. $F(a) \leq F(b)$ *für* $a < b$ *(Isotonie, d.h. F ist ordnungserhaltend).*
Dann heißt F maßerzeugende Funktion.

Mittels einer maßerzeugenden Funktion läßt sich ein Wahrscheinlichkeitsmaß durch

$$P(I_{a,b}) = F(b) - F(a) \tag{2.24}$$

definieren. Allgemeiner soll gelten

$$P\left(\bigcup_{n=1}^{m} I_{a_n,b_n}\right) = \sum_{n=1}^{m} (F(b_n) - F(a_n)), \quad m < \infty. \tag{2.25}$$

Definition 2.2.17 *Die maßerzeugende Funktion F heißt* Verteilungsfunktion.

Beispiel 2.2.4 Es sei $\Omega \subset \mathbb{R}$ und $\lambda(\Omega) < \infty$ sei ein Lebesgue-Maß. Weiter sei $A \in \mathcal{B}(\Omega)$. Dann ist

$$\lambda(A) = \frac{\lambda(A)}{\lambda(\Omega)}. \tag{2.26}$$

ist insbesondere $\Omega = [c, de]$ mit $c < d$, so hat man

$$P(I_{a,b}) = F(b) - F(a), \quad I_{a,b} \subset [c, d] \tag{2.27}$$

und

$$F(x) = \begin{cases} 0, & x \leq c \\ \frac{x-c}{c-d}, & c < x < d \\ 1, & x \geq d \end{cases} \tag{2.28}$$

F definiert die *Gleichverteilung* auf Ω. □

Definition 2.2.18 *Es sei F eine Verteilungsfunktion; dann gilt*

$$\frac{dF(x)}{dx} = f(x) \geq 0; \tag{2.29}$$

f heißt die zu F gehörige Dichtefunktion.

Anmerkung: In der Maßtheorie entspricht f die *Radon-Nikodym-Derivierte.* □

Es sei $(0, x_1] = \{x|0 < x \leq x_1\}$ ein halboffenes Intervall. f ist so definiert, dass

$$P(x \in (0, x_1])) = \int_0^{x_1} f(x)dx \tag{2.30}$$

die Wahrscheinlichkeit ist, mit der ein x aus $(0, x_1]$ gemessen wird. Aus diesem Ausdruck kann ein Ausdruck für die Wahrscheinlichkeit eines beliebigen Ereignisses – also für ein beliebiges halboffenes Intervall – gefunden werden: wegen

$$(0, x_2] = (0, x_1] \cup (x_1, x_2]$$

hat man

$$P(x \in (x_1, x_2]) = \int_0^{x_2} f(x)dx - \int_0^{x_1} f(x)dx = \int_{x_1}^{x_2} f(x)dx. \tag{2.31}$$

(2.31) kann als Verallgemeinerung von (2.23) auf den überabzählbaren Fall gesehen werden.

Beispiel 2.2.5 Es werden Zeiten auf dem Intervall (t_0, ∞) gemessen, $0 < t_0$. Die Dichtefunktion sei durch

$$f(t) = \begin{cases} 0, & t \leq t_0 \\ \lambda \exp(-\lambda t), & t > t_0. \end{cases}, \quad \lambda \in \mathbb{R}_+ = \{x|x \in R, x > 0\}. \tag{2.32}$$

gegeben. Die Wahrscheinlichkeit, dass $t \leq t_1$ ist dann

$$P(t \leq t_1) = \lambda \int_{t_0}^{t_1} e^{-\lambda\tau} d\tau = e^{-\lambda t_0} - e^{-\lambda t_1}. \tag{2.33}$$

□

Aus (2.31) folgt sofort die obige Bemerkung, dass die Wahrscheinlichkeit für eine spezielle Messung x_1 gleich Null ist:

$$P(x = x_1) = \int_{x_1}^{x_1} f(x)dx = 0. \tag{2.34}$$

Wahrscheinlichkeiten ungleich Null ergeben sich also nur für Intervalle mit von Null verschiedener Länge. Der Ausdruck (2.34) bedeutet *nicht*, dass das Ereignis $\{x_1\}$ unmöglich ist, – bei jedem Versuch tritt ja genau ein Element aus Ω auf! Im Übrigen sei noch auf die Anmerkung 2.2.2 zu Satz 2.2.10 auf Seite 39 hingewiesen. $P(t \leq t_1)$ ist die Wahrscheinlichkeit, dass t einen Wert im Intervall $(0, t_1]$ (hier also $t_0 = 0$) annimmt. Dann ist die Wahrscheinlichkeit, dass $t \in \mathbb{Q}$, dass also t eine rationale Zahl ist, gleich Null. Diese Aussage ist nicht äquivalent zu (2.34, denn (2.34) gilt für *alle* x_1 aus einem Kontinuum, während sich Anmerkung 2.2.2 auf endliche Teilintervalle eines Kontinuums bezieht.

Das Ereignis A ist vernachlässigbar, wenn endlich oder abzählbar viele Intervalle $I_1, I_2, \ldots$ existieren derart, dass $A \subset \cup_k I_k$ und $\sum_k P(I_k) < \varepsilon$,. d.hj. A hat das Lebesgue-Maß 0. Dann gilt

Satz 2.2.13 *Es sei $A_1, A_2, \ldots$ eine endlich oder abzählbar unendliche Folge vernachlässigbarer Ereignisse. Dann ist ihre Vereinigung $\cup_k A_k$ vernachlässigbar.*

Beweis: Für jedes $n \in \mathbb{N}$ können Intervalle $I_{n1}, I_{n2}, \ldots$ ausgewählt werden derart, dass

$$A_n \subset \cup_k I_{nk}, \quad \sum_k P(I_{nk}) < \frac{\varepsilon}{2^n}.$$

Nun gilt aber $\cup_n A_n \subset \cup_k I_{nk}$ und

$$\sum_n \sum_k P(I_{nk}) < \sum_n \frac{\varepsilon}{2^n} < \varepsilon,$$

(vergl. den Beweis von Satz 2.2.10, Seite 38). Dies bedeutet, dass $\cup_n A_n$ vernachlässigbar ist. □

Da die Menge $\mathbb{Q}$ der rationalen Zahlen abzählbar ist, bedeutet Satz 2.2.13, dass die Wahrscheinlichkeit, dass ein ω mit aus $\mathbb{Q}$ stammender Repräsentation stets gleich Null ist, sofern $\mathcal{A}$ auf $\Omega \subseteq \mathbb{R}$ definiert ist; dies gilt für jedes Intervall aus $\mathbb{R}$ (vergl. Abbildung 2.1, Seite 32). Dies heißt nicht, dass es unmöglich ist, dass ein derartiges $\omega \in \Omega$ auftritt. Das Ergebnis kann so ausgedrückt werden: es ist *fast sicher* (f.s.) (*almost sure*, a.s.), dass ein $\omega \in \mathbb{Q}^c$ eintritt.

2.3 Bedingte Wahrscheinlichkeiten

Definition 2.3.1 *Es sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum, und $A, B \in \mathcal{A}$ seien irgend zwei zufällige Ereignisse. Dann sei $P(A|B)$ die Wahrscheinlichkeit*

von A, wenn B bereits eingetreten ist. $P(A|B)$ heißt bedingte Wahrscheinlichkeit *von A, gegeben B, im Unterschied zu $P(A)$, der* unbedingten *Wahrscheinlichkeit von A.*

Zur Vereinfachung seien A und B nicht nur die Namen für zwei zufällige Ereignisse, sondern auch für die zu ihnen korrespondierenden Teilmengen aus Ω.

Anmerkung: Es sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum und $A \in \mathcal{A}$. Dann ist

$$P(A) = P(A|\Omega). \tag{2.35}$$

Der Ausdruck "wenn B bereits eingetreten ist" in Definition 2.3.1 kann durch "wenn $\Omega = B$ gesetzt wird" ersetzt werden. Dem entspricht der Sachverhalt, dass Wahrscheinlichkeiten von Ereignissen stets in Bezug auf allgemeine Randbedingungen definiert sind, die wiederum die Menge Ω definieren. In diesem allgemeinen Sinn sind Wahrscheinlichkeiten stets bedingte Wahrscheinlichkeiten, und dies wird durch (2.35) ausgedrückt. □

Die Aussage 'wenn B bereits eingetreten ist' ist also gleichbedeutend mit der Aussage, dass nur noch Ereignisse aus B beobachtet werden können. Dieser Sachverhalt gleichbedeutend mit der Reduktion des Stichprobenraums Ω auf $\Omega_B = B$. Ein einfaches Beispiel ist die unbedingte Wahrscheinlichkeit, dass eine "2" gewürfelt wurde, und die bedingte Wahrscheinlichkeit, dass eine "2" gewürfelt wurde, wenn bekannt ist (= unter der Bedingung), dass eine gerade bzw. eine ungerade Zahl gewürfelt wurde; im "unbedingten" Fall – gemeint ist, dass alle sechs Seiten des Würfels gleichwahrscheinlich sind und nichts weiter bekannt ist – ist die Wahrscheinlichkeit der "2" gleich 1/6, im "bedingten" Fall (= es wurde eine gerade Zahl gewürfelt) ist $\Omega_B = B = \{2, 4, 6\}$ und die Wahrscheinlichkeit einer "2" ist 1/3.

Gesucht ist ein zur Definition 2.3.1 korrespondierender Ausdruck für $P(A|B)$, mit dem die bedingte Wahrscheinlichkeit berechnet werden kann. Offenbar kann unter der Bedingung, dass B der Stichprobenraum ist, das zufällige Ereignis A nur eintreten, wenn $A \cap B \neq \emptyset$ ist, so dass $P(A|B) = 0$, wenn $A \cap B = \emptyset$, und $P(A|B) = 1$ wenn $A \cap B = B$ ist. Es sei $\tilde{P}$ ein Wahrscheinlichkeitsmaß, dass aus P hervorgeht, wenn Ω durch B ersetzt wird, d.h. $\tilde{P}(B) = 1$. Mit "hervorgehen" ist gemeint, dass die Relationen zwischen den Wahrscheinlichkeiten von Ereignissen, wie sie durch das Maß P bestimmt werden, sich auf das Maß $\tilde{P}$ übertragen sollen, $\tilde{P}$ ist gewissermaßen nur eine Skalierung von P. Demnach soll gelten, dass eine Zahl $\mu \in (0, 1)$ existiert derart, dass

$$\tilde{P}(B) = \mu P(B) = 1 \Rightarrow \mu = \frac{1}{P(B)}. \tag{2.36}$$

Wenn der Stichprobenraum von Ω auf B reduziert wird, kann A nur eintreten, wenn $A \cap B \neq \emptyset$. Demnach kann man

$$P(A|B) = \tilde{P}(A) = \mu P(A \cap B)$$

setzen. Wegen (2.36) hat man dann

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{2.37}$$

Folgerungen:

1. Es ist $P(A|B) = 0$ genau dann, wenn $A \cap B = \emptyset$. $P(B) > 0$ muß vorausgesetzt werden. Sicherlich gilt

   $$B = B \cap \bar{A} \cup B \cap A,$$

   so dass

   $$P(B) = P(B \cap \bar{A}) + P(B \cap B),$$

   denn sicherlich ist $(B \cap \bar{A}) \cap (B \cap A) = \emptyset$, und wegen $P(B \cap \bar{A}) \geq 0$ folgt $P(B) \geq P(A \cap B)$. Diese Relation garantiert, dass $0 \leq P(A|B) \leq 1$ gilt.

   Andererseits gilt $B \subseteq A \Rightarrow P(A|B) = 1$, denn für jedes $\omega \in B$ ist auch $\omega \in A$. Vergl. Abb. 2.2 (a) und (c).

2. Es gilt

   $$P(A^c|B) = 1 - P(A|B), \tag{2.38}$$

   bzw.

   $$P(A^c|B) + P(A|B) = 1. \tag{2.39}$$

   Denn

   $$P(A \cup A^c|B) = \frac{P(A \cup A^c \cap B}{P(B)} = \frac{P(\Omega \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1,$$

   und andererseits ist

   $$P(A \cup A^c|B) = P(A|B) + P(A^c|B),$$

   da $A \cap A^c = \emptyset$, so dass (2.38) folgt.

3. Es ist

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

   so dass

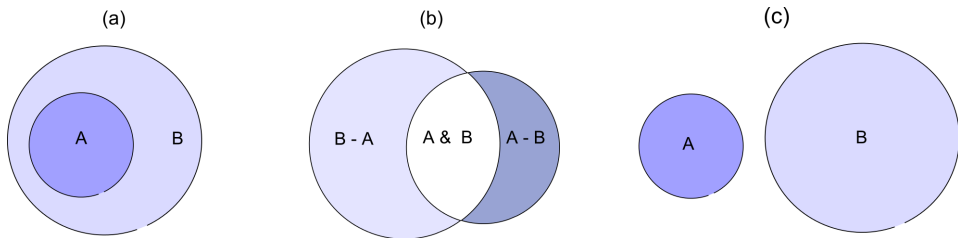
   $$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A). \tag{2.40}$$

   Daraus folgt sofort

   $$P(A|B) = P(B|A)\frac{P(A)}{P(B)}. \tag{2.41}$$

   $P(A|B)$ und $P(B|A)$ werden gelegentlich *zueinander inverse* Wahrscheinlichkeiten genannt. (2.41) bedeutet, dass $P(A|B) = P(B|A)$ nur dann, wenn $P(A)/P(B) = 1$, wenn also $P(A) = P(B)$ gilt.

Abbildung 2.2: Bedingte Wahrscheinlichkeiten. Die Fälle (a), (b) und (c) sind jeweils in einen Stichprobenraum Ω eingebettet, der hier nicht explizit gezeigt wird.



4. Es sei $A \subset B$ (vergl Abb. 2.2). Dann ist

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

Umgekehrt ist

$$P(B|A) = \frac{P(A)}{P(A)} = 1.$$

5. **Konjunktionen von Ereignissen:** Es seien A_1, A_2, A_3 zufällige Ereignisse. Gesucht ist ein Ausdruck für $P(A_1 \cap A_2 \cap A_3)$. Zunächst sei festgestellt, dass

$$P(A_3|A_1 \cap A_2) = \frac{P(A_1 \cap A_2 \cap A_3)}{P(A_1 \cap A_2)}$$

gilt. Aber aus der Definition der bedingten Wahrscheinlichkeit folgt

$$P(A_1 \cap A_2) = p(A_1)P(A_2|A_),$$

woraus sich

$$P(A_1 \cap A_2 \cap A_3) = p(A_1)P(A_2|A_1)(A_3|A_1 \cap A_2) \tag{2.42}$$

ergibt. Man zeigt leicht auf analoge Weise, dass

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdots P(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1}) \tag{2.43}$$

gilt.

Eines der einfachsten Beispiele für eine bedingte Wahrscheinlichkeit liefert der Würfel:

Beispiel 2.3.1 Ein fairer Würfel wird von einem Spieler geworfen. Der Spieler verrät nun, dass die geworfene Augenzahl $i \geq 3$ ist. Wie groß ist die Wahrscheinlichkeit, dass die Augenzahl gerade ist?

Man kann $A = \{i \text{ ist gerade}\}$, $B = \{i \geq 3\}$ setzen. Gesucht ist die Wahrscheinlichkeit $p(A|B)$. Es ist $\Omega = \{1, 2, 3, 4, 5, 6\}$, und offenbar ist $p(A \cap B) = p(\{4, 6\})$, $p(B) = p(\{3, 4, 5, 6\})$. Die Annahme, dass der Würfel "fair" ist, soll bedeuten, dass die $\omega = i \in \Omega$, $i = 1, ..., 6$ alle gleichwahrscheinlich sind, so dass $p(i) = 1/6$ für alle i, $1 \leq i \leq 6$. Nach K3 ist dann $p(A \cap B) = p(i = 4 \cup i = 6) = p(4) + p(6) = 2/6$, $p(B) = p(i = 3 \cup \cdots \cup i = 6) = 4/6$. Dann ergibt sich

$$p(A|B) = \frac{2/6}{4/6} = \frac{1}{2}$$

□

Interessantere Beispiele ergeben sich nach einigen Vorbetrachtungen. Zunächt wird der folgende Satz bewiesen:

Satz 2.3.1 *(Satz von der totalen Wahrscheinlichkeit): Es sei Ω eine beliebige Menge von Elementarereignissen und A_i, $i = 1, ..., n$ seien paarweise disjunkte Teilmengen von Ω, die Ω ausschöpfen, d.h. $A_i \cap A_j = \emptyset$ für alle $i \neq j$, $A_1 \cup A_2 \cup \cdots \cup A_n = \Omega$. Weiter sei $B \subseteq \Omega$ eine beliebige Teilmenge von Ω. Dann ist*

$$p(B) = p(B|A_1)p(A_1) + p(B|A_2)p(A_2) + \cdots + p(B|A_n)p(A_n). \tag{2.44}$$

Beweis: Wegen $\bigcup_{i=1}^{n} A_i = \Omega$ kann der Durchschnitt zumindest eines Ereignisses A_i mit B nicht leer sein. Also kann man $B = (B \cap A_1) \cup \cdots \cup (B \cap A_n)$ schreiben, und da die A_i disjunkt sind, folgt $(B \cap A_i) \cap (B \cap A_j) = \emptyset$ für alle $i \neq j$. Daraus folgt nach Axiom K3 $p(B) = p(B \cap A_1) + \cdots + p(B \cap A_n)$, und substituiert man nach (2.40) $p(B|A_i)p(A_i)$ für $p(B \cap A_i)$, so erhält man (2.44). □

Mithilfe des Satzes von der totalen Wahrscheinlichkeit läßt sich $p(A|B)$ auch in anderer Form schreiben:

Satz 2.3.2 *(Satz von Bayes): Für die zufälligen Ereignisse A_i und B mögen die Voraussetzungen des Satzes 2.3.1 gelten. Dann folgt*

$$p(A_i|B) = \frac{p(B|A_i)p(A_i)}{\sum_{i=1}^{n} p(B|A_i)p(A_i)}, \quad i = 1, \ldots, n \tag{2.45}$$

Beweis: Der Satz folgt unmittelbar aus der Definition von $P(A|B)$ und dem Satz von der totalen Wahrscheinlichkeit. □

Beispiel 2.3.2 (Röntgen-Reihenuntersuchung auf Tuberkulose) Es gelte:

(1) 90% der TBC-Kranken werden durch die Röntgen-Reihenuntersuchung entdeckt, 10% bleiben unentdeckt.
(2) 99% der TBC-Freien werden als TBC-frei diagnostiziert,

(3) 1% werden als TBC-verdächtig eingestuft, obwohl sie TBC-frei sind,
(4) Nur .1% der Bevölkerung haben TBC.
Eine Person wird zufällig gewählt, untersucht und als verdächtig eingestuft. Wie groß ist die Wahrscheinlichkeit, dass die Person wirklich an TBC leidet?

Es sei A das zufällige Ereignis, dass die untersuchte Person an TBC leidet, und B sei das zufällige Ereignis, dass die Röntgendiagnose positiv ist, d.h. dass die Person als "verdächtig" eingestuft wird. Gegeben sind

(a) $P(A) = .001$
(b) $P(\bar{B}|\bar{A}) = .99$, $P(B|\bar{A}) = .01$
(c) $P(B|A) = .9$.

Gesucht ist $P(A|B)$. Es ist

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\bar{A})P(\bar{A})} \\ &= \frac{.9 \times .001}{.9 \times .001 + .01 \times .99} \approx .083 \end{aligned}$$

Dieses Ergebnis ist vielleicht überraschend: zwar ist der Röntgen-Test sehr zuverlässig, da 90% der Kranken gefunden und 99% der Gesunden als solche klassifiziert werden, doch haben nur ca. 8% der Personen mit positivem Befund tatsächlich TBC. Der Vorteil der Reihenuntersuchung liegt darin, dass eine gründliche und teure Detailuntersuchung nur auf einen kleinen Teil der Bevölkerung angewandt werden muß. Eine ähnliche Situation hat man, wenn man z.B. mit psychologischen Tests Unfall- oder Schlaganfallopfer auf Hirnschäden untersucht. Eine genaue medizinische Untersuchung ist aufwendig, muß aber nur auf einen kleinen Teil der wirklich Verdächtigen angewendet werden. □

Beispiel 2.3.3 Brustkrebsrisiko: Dawes (2001) berichtet von dem kalifornischen Arzt C. S. Rogers, der 1977 - 1979 einen "pionering approach" zur Verhinderung von Brustkrebs praktiziert hat. Der Radiologe John N. Wolfe aus Detroit hatte nämlich anhand von Mammographien die These aufgestellt, dass ca 57 % der Frauen ein Brustgewebe haben, das er Hochrisikogewebe (HR) nannte, während die übrigen Frauen ein Niedrigrisikogewebe (NR) haben. Nach Einschätzung des Radiologen würden 92 % der Frauen mit HR-Gewebe im Alter zwischen 40 und 59 Jahren Brustkrebs bekommen. Von den Frauen mit NR-Gewebe würden nur 7.5 % Brustkrebs bekommen. Rogers empfahl nun den Frauen mit HR, sich beide Brüste amputieren zu lassen (*prophylaktische Masectomie*), bevor der Krebs ausbrechen würde. Nach einer Rekonstruktion der Brüste auf Siliconbasis würden die Frauen dann wieder "wie Frauen" aussehen. Innerhalb von 2 Jahren haben sich 90 Frauen dieser Operation unterzogen.

Dawes konnte aus den publizierten Prozentzahlen die Häufigkeiten ausrechnen, mit denen in einer Stichprobe von 1000 Fällen Frauen mit HR- und NR-

Gewebe enthalten sein würden und wieviele aus diesen beiden Gruppen Krebs bekommen würden: Die Daten in der Tabelle erlauben eine Bewertung der intui-

Tabelle 2.1: Risikogewebe und Krebshäufigkeit

	Krebs		
	nein	ja	$\sum$
HR	499	71	570
NR	424	6	430
$\sum$	923	77	1000

tiven Abschätzungen des Radiologen und, daraus resultierend, eine Evaluation der Aktivitäten des Dr. Rogers. Man muß nur von der Formel für bedingte Wahrscheinlichkeiten ausgehen:

$$p(A|B) = \frac{p(A \cap B)}{p(B)}.$$

Die Wahrscheinlichkeit, dass eine Frau ein HR-Gewebe *und* Krebs hat ist demnach

$$p(\text{HR} \cap \text{Krebs}) = \frac{71}{1000} = .071,$$

während der Anteil der Frauen mit Krebs durch

$$p(\text{Krebs}) = \frac{77}{1000} = .077$$

ist. Demnach folgt

$$p(\text{HR}|\text{Krebs}) = \frac{p(\text{HR} \cap \text{Krebs})}{p(\text{Krebs})} = \frac{.071}{.077} = .92.$$

Aber $p(\text{HR}|\text{Krebs})$ ist nicht gleich $p(\text{Krebs}|\text{HR})$:

$$p(\text{Krebs}|\text{HR}) = \frac{p(\text{HR} \cap \text{Krebs})}{p(\text{HR})} = \frac{71}{570} = .12;$$

also"nur" ca 12 % der Frauen mit HR-Gewebe bekommen tatsächlich Krebs, - das ist substantiell weniger als 92 %! Der Unterschied zwischen diesen beiden bedingten Wahrscheinlichkeiten ergibt sich aus den unterschiedlichen Grundquoten $p(\text{HR}|\text{Krebs})$ und $p(\text{HR}) = 77/570 = .135$ von HR- und NR-Gewebe. Denn es ist ja

$$p(\text{Krebs}|\text{HR}) = p(\text{HR}|\text{Krebs})\frac{p(\text{Krebs})}{p(\text{HR})}.$$

Da aber $p(\text{Krebs})/p(\text{HR}) = 77/570 = .135$, wird der hohe Wert $p(\text{HR}|\text{Krebs}) = .92$ um eben diesen Faktor deutlich auf eben 12 % reduziert! Zwar ist diese bedingte Wahrscheinlichkeit fast doppelt so hoch wie die bedingte Wahrscheinlichkeit, unter der NR-Bedingung an Krebs zu erkranken, aber längst nicht hoch genug, um den drastischen Empfehlungen Dr. Rogers zu folgen. □

Beispiel 2.3.4 In einem Experiment zur Psychologie der Mustererkennung haben die Versuchspersonen (Vpn) die Aufgabe, "verrauscht" dargebotene Muster zu identifizieren: die Muster werden kurzfristig (50 ms) mit zufällig verteilten Punkten überlagert dargeboten. Ziel der Untersuchung ist es, Informationen über die Art und Weise, in der das sensorische System Musteraspekte herausfiltert, zu erfahren sowie die Entscheidungsprozesse der Vpn zu charakterisieren. Insbesondere wird für die letzteren angenommen, dass die Vpn nach dem Bayes-Theorem entscheiden.

Es seien A_1, A_2 und A_3 die im Experiment verwendeten Muster. Sie werden in zufälliger Folge dargeboten; die Wahrscheinlichkeit, dass ein bestimmtes Muster A_i in einem bestimmten Versuchsdurchgang (Trial) dargeboten wird, ist $P(A_i) = 1/3$. Es sei B eine bestimmte Aspektmenge, d.h. ein Teilmuster, dass in allen drei Mustern enthalten ist. Gesucht ist die bedingte Wahrscheinlichkeit einer Antwort $P(A_i|B)$, d.h. die Wahrscheinlichkeit, dass die Vp mit "A_i" antwortet, wenn sie nur das Teilmuster B wahrgenommen hat. Im Laufe des Experiments hat sie gelernt, dass $P(B|A_1) = .25$, $P(B|A_2) = .35$, $P(\bar{B}|A_3) = .65$. Gesucht ist $P(A_1|B)$. Es ist

$$\begin{aligned} P(A_1|B) &= \frac{P(B|A_1)P(A_1)}{P(B)} = \\ &= \frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3)} \end{aligned}$$

und $P(B|A_3) = 1 - P(\bar{B}|A_3) = .45$. Mithin ist

$$P(A_1|B) = \frac{.25(1/3)}{(1/3)(.25 + .35 + .45)} = \frac{.25}{.25 + .35 + .45} = .238 \approx .24.$$

Auf analoge Weise findet man $P(A_2|B) = .333$, $P(A_3|B) = .429$. Eine Entscheidungsregel der Vp könnte lauten: Nehme dasjenige Muster A_i als das Dargebotene an, für das die bedingte Wahrscheinlichkeit $P(A_i|B)$ maximal ist. Dann würde die Vp mit Wahrscheinlichkeit 1 das Muster A_3 als gezeigt bezeichnen. Eine andere Regel wäre, die Muster A_i probabilistisch zu wählen, und zwar jeweils mit Wahrscheinlichkeiten, die den bedingten Wahrscheinlichkeiten $P(B|A_i)$ entsprechen. □

Urteilsfehler, die bei der Vorhersage von Ereignissen gemacht werden, beruhen zum Teil ebenfalls auf Fehleinschätzungen von bedingten Wahrscheinlichkeiten. Dazu muß man den Begriff der Hazardrate oder der bedingten Fehlerrate (*conditional failure rate*) einführen.

Beispiel 2.3.5 Hazardfunktion Erfolg einer Therapie läßt sich gelegentlich dadurch beurteilen, dass man die Zeit τ bis zum Wiederauftreten der Symptome der Erkrankung misst. Für $\tau = \infty$ treten die Symptome nie mehr ein,der Patient

ist endgültig geheilt. Generell wird man sagen, je größer τ, desto besser die Therapie. Man bestimmt nun eine zeitliche Einheit - Stunde, oder Tag, oder Woche, etc. und kann nun nach der Wahrscheinlichkeit fragen, dass τ in einer bestimmten Einheit liegt. Die in Frage stehende zeitliche Einheit beginne zum Zeitpunkt t und habe die Dauer Δt, d.h. die Einheit ist durch das Intervall $(t, t + \Delta t]$ bestimmt. Man kann nun die absolute Wahrscheinlichkeit $p(\tau \in (t, t+\Delta t])$ betrachten, oder aber die bedingte Wahrscheinlichkeit

$$\phi_{\Delta t}(t) = P(\tau \in (t, t + \Delta t] | \tau > t), \tag{2.46}$$

also die bedingte Wahrscheinlichkeit, dass $\tau \in (t, t+\Delta t]$ fällt unter der Bedingung, dass $\tau > t$ ist. Nun gilt für zwei zufällige Ereignisse A und B mit $p(B) > 0$, dass $p(A|B) = p(A \cap B)/p(B)$. Setzt man $A = \tau \in (t, t + \Delta t]$ und $B = \tau > t)$, so erhält man

$$\tau \in (t, t + \Delta t] \cap \tau > t = \tau \in (t, t + \Delta t],$$

denn das Ereignis $\tau \in (t, t + \Delta t]$ ist ja in $\tau > t$ enthalten. Damit ist

$$\phi_{\Delta t}(t) = \frac{p(\tau \in (t, t + \Delta t])}{p(\tau > t)} \tag{2.47}$$

Einen interessanten Spezialfall erhält man, wenn diese bedingte Wahrscheinlichkeit konstant für alle t ist. Dies bedeutet ja, dass das in Frage stehende zufällige Ereignis - also etwa das Auftreten der Symptome - mit größer werdendem t nicht mehr oder weniger wahrscheinlich wird: man kann sagen, dass das Ereignis dann *rein zufällig* auftritt. Betrachtet man etwa die Intervalle $(t_1, t_1 + \Delta t]$, $(t_2, t_2 + \Delta t]$, $(t_3, t_3 + \Delta t]$ etc., wobei $t_2 = t_1 + \Delta t$, $t_3 = t_2 + \Delta t = t_1 + 2\Delta t$ etc ist, so ist der Fall

$$\phi_{\Delta t}(t_k) = \text{konstant für alle } k$$

nicht gleichbedeutend mit der Aussage, dass

$$P(\tau \in (t_k, t_k + \Delta t]) = \text{konstant für alle } k,$$

woraus folgt, dass diese Wahrscheinlichkeiten nicht reine Zufälligkeit widerspiegeln. Diese wiederum etwas gegenintuitive Aussage wird später noch genauer diskutiert werden. □

Das folgende Beispiel ist zwar nur eine Denksportaufgabe, es illustriert aber die Anwendung der bedingten Wahrscheinlichkeit auf erwartete Handlungen:

Beispiel 2.3.6 Das Gefangenenproblem Drei Verurteilte A, B und C warten auf ihre Hinrichtung, die am kommenden Tag stattfinden soll. Da kommt ein Wärter und sagt ihnen, dass einer von ihnen begnadigt worden sei, aber er dürfe nicht sagen, wer es ist. Jeder der drei Gefangenen hat somit eine Überlebenswahrscheinlichkeit von $1/3$. A schafft es, den Wärter zu bestechen, ihm einen

Mitgefangenen zu nennen, der *nicht* begnadigt worden ist. Der Wärter teilt ihm mit, B sei nicht begnadigt worden. A freut sich, denn er folgert, dass entweder er oder C begnadigt wurde, seine Überlebenschance also von 1/3 auf 1/2 gestiegen sei.

A irrt sich, denn die Entscheidung, ob er oder ein anderer begnadigt wurde, ist ja schon vor der Befragung gefallen, und der Wärter kann ihm stets einen Mitgefangenen nennen, der nicht begnadigt wurde: wurde A begnadigt, kann der Wärter zufällig zwischen zwei der Gefangenen, die nicht begnadigt wurden, wählen, und wenn A nicht begnadigt wurde, so kann der Wärter immer noch denjenigen Gefangenen nennen, der ebenfalls nicht begnadigt wurde. Mit Wahrscheinlichkeit 2/3 wurde A nicht begnadigt, und mit eben dieser Wahrscheinlichkeit wird vom Wärter der Mitgefangene genannt, der ebenfalls nicht begnadigt wurde, und das heißt auch, dass mit Wahrscheinlichkeit 2/3 der Gefangene C begnadigt wurde. An As Wahrscheinlichkeit, begnadigt worden zu sein (1/3) hat sich nichts geändert.

Zur Übung kann man den Apparat der bedingten Wahrscheinlichkeiten auf As Problem anwenden. Es seien a, b und c die zufälligen Ereignisse, dass A, B oder C begnadigt worden sind, und es gelte

$$p(a) = p(b) = p(c) = 1/3. \tag{2.48}$$

W_A, W_B und W_C seien die zufälligen Ereignisse, dass der Wärter A, B oder C als den nicht Begnadigten nennt. Da A fragt und der Wärter ihm nicht sagen darf, ob er begnadigt wurde oder nicht, ist $p(W_A) = 0$. A ist nun an den bedingten Wahrscheinlichkeiten $p(a|W_B)$ und $p(c|W_B)$ interessiert; $p(b|W_B) = 0$, da der Wärter ja schon gesagt hat, dass B nicht begnadigt wurde, $p(c|W_B) = p(W_B|c)p(c)/p(W_B) = (1/3)/P(W_B)$, denn $P(W_B|c) = 1$, denn $p(W_A|c) = 0$, da A als Fragendem nichts ihn selbst Betreffendes mitgeteilt werden darf, und $p(W_C|c) = 0$, da C nicht als nicht begnadigt genannt werden kann, da C ja begnadigt wurde, mithin $p(W_B|c) = 1$; nach (2.39), Seite 54, muß ja

$$P(W_a|c) + P(W_b|c) + P(W_c|c) = 1$$

gelten. Dann hat man

$$\begin{aligned} p(a|W_B) &= \frac{p(W_B|a)p(a)}{p(W_B)} = \frac{(1/2)(1/3)}{P(W_B)} \\ p(c|W_B) &= \frac{p(W_b|c)p(c)}{p(W_B)} = \frac{1/3}{p(W_B)} \end{aligned}$$

und

$$p(W_B) = p(W_B|a)p(a) + P(W_B|b)p(b) + P(W_B|c)p(c) = \frac{1}{2}\frac{1}{3} + 0\frac{1}{3} + 1\frac{1}{3} = \frac{1}{2}.$$

Dann ist

$$p(a|W_B) = \frac{p(W_B|a)p(a)}{p(W_B)} = \frac{(1/2)(1/3)}{1/2} = \frac{1}{3}$$

und

$$p(c|W_B) = \frac{p(W_b|c)p(c)}{p(W_B)} = \frac{1/3}{1/2} = \frac{2}{3}.$$

Der intuitive Schluß des Gefangenen A, seine Überlebenswahrscheinlichkeit sei auf $1/2$ gestiegen, ist nicht berechtigt.

Das Gefangenenproblem ist analog zum berühmten Ziegenproblem konstruiert. Es gibt drei Türen, und hinter einer von ihnen ist ein Auto verborgen, dass man gewinnen kann. Hinter den anderen beiden Türen steht je eine Ziege. Eine Person soll raten, hinter welcher Tür das Gefährt steht. Die Wahrscheinlichkeit, dass sie korrekt rät, ist $1/3$. Der Spielleiter will der Person behilflich sein und öffnet, nachdem eine Tür geraten aber noch nicht geöffnet worden ist, eine Tür, hinter der eine Ziege steht. Die ratende Person kann nun, wenn sie möchte, die Tür wechseln in der Hoffnung, ihre Chance auf den Gewinn zu erhöhen. Die Frage ist, ob sie ihre Chance durch den Wechsel erhöht.

Um die Frage zu beantworten, muß man allerdings nicht den langen Weg der Diskussion bedingter Wahrscheinlichkeiten gehen. Denn mit der Wahrscheinlichkeit $2/3$ wählt die ratende Person eine der beiden "falschen" Türen und der Spielleiter hat nur eine Möglichkeit, eine Tür zu öffnen, hinter der eine Ziege steht, hinter der anderen Tür steht dann notwendig das Auto, und das heißt, dass man mit der Wahrscheinlichkeit $2/3$ das Auto gewinnt, wenn man die Tür wechselt. □

Inkrementelle Konfirmation und Diskonfimation von Hypothesen: Hypothesen sind Aussagen, für die es entweder noch keinen Beweis gibt oder für die es auf der Basis bereits vorhandenen Wissens keinen deduktiven Beweis geben kann. Man versucht dann, empirische Evidenz für oder gegen die Hypothese zu finden. Man habe aus einem Experiment *Evidenz* E, also Daten, gewonnen, die mit einer Hypothese H verträglich sind oder nicht; im ersteren Fall hat E einen bestätigenden oder konfirmatorischen Effekt, im zweiten Fall wird der bisher existierende Grad von Bestätigung reduziert. Die Begriffe der Bestätigung oder der Reduktion der Bestätigung sind vage, zumal die Relation von E zu H häufig nicht alternativ als entweder bestätigend oder nicht bestätigend zu sehen ist. Zur exakteren Fassung der Beziehung zwischen E und H kann das Byessche Theorem dienen.

Dazu werde noch ein allgemeines Wissen K (von 'knowledge') eingeführt. Gesucht ist ein Ausdruck für die bedingte Wahrscheinlichkeit $P(H|K,E)$, wobei K,E vereinfachend für $K \cap E$ steht. Eine direkte Anwendung der Formel für die bedingte Wahrscheinlichkeit führt auf

$$P(H|K,E) = \frac{P(H,K,E)}{P(K,E)}.$$

Die Beziehung (2.42), Seite 55, liefert dann

$$P(H,K,E) = P(K)P(H|K)P(E|H,K)$$

und weiter hat man $P(E, K) = P(E|K)P(K)$, so dass

$$P(H|K, E) = \frac{P(H|K)P(E|H, K)}{P(E|K)}. \tag{2.49}$$

Hempel (1945a, b) führte den *hypothetico-deduktiven Ansatz* (HD-Ansatz) ein. Die Idee dabei ist, dass aus einer Hypothese H in Kombination mit dem vorhandenen Wissen K eine Folgerung über einen im Prinzip beobachtbaren Sachverhalt E deduziert wird. Dann wird ein Experiment oder eine Untersuchung durchgeführt, in dem oder in der überprüft wird, ob E tatsächlich der Fall (E) ist oder nicht ($\neg E$). Dieser Ansatz ist eingebettet in den Versuch der logischen Empiristen (Wiener Kreis), eine induktive Logik zu begründen, die die deduktive Logik vervollständigen sollte, damit gesichertes Wissen entstehen kann. Auf die extensive Diskussion dieses Ansatzes kann hier nicht eingegangen werden, aber es kann gezeigt werden, wie der HD-Ansatz über Bayes Theorem implementiert werden kann. Die Darstellung ist an Earman (1992)[9] orientiert.

Zunächst wird der HD-Ansatz ein wenig formalisiert. Dazu wird angenommen, dass (a) die Kombination $\{H, K\}$ die Aussage E semantisch impliziert; damit ist eben gemeint, dass E aus H und K deduziert wird. Man schreibt dafür auch $\{H, K\} \models E$ ($\models$ steht für 'impliziert semantisch'). Weiter wird (b) $0 < P(H|K) < 1$ angenommen, und schließlich wird (c) $0 < P(E|K) < 1$ postuliert. (b) und (c) bedeuten, dass H und K zwar aus dem Wissen K abgeleitet werden können, es aber nicht sicher ist, ob H und E auch wirklich gelten. (a) bedeutet, dass $P(E|H, K) = 1$, denn E ist ja eine deduzierte Implikation, so dass aus (2.49)

$$P(H|K, E) = \frac{P(H|K)}{P(E|K)}. \tag{2.50}$$

folgt. Wegen (c) folgt $P(H|K)/P(E|K) > P(H|K)$, also $P(H|K, E) > P(H|K)$. Die Wahrscheinlichkeit für H unter der Bedingung K *und* E ist demnach größer als die Wahrscheinlichkeit von H unter der Bedingung K, so dass E die Wahrscheinlichkeit von H *inkrementell erhöht*. Der Satz von Bayes erlaubt es also, zu spezifizieren, was es heißen soll, dass die "Evidenz" E zur Bestätigung von H beiträgt. Weiter folgt aus (2.50), dass die inkrementelle Konfirmation von H um so größer ist, je kleiner $P(E|K)$ ist. Ein kleiner Wert von $P(E|K)$ bedeutet, dass auf der Basis des vorhandenen Wissens die Beobachtung von E wenig wahrscheinlich ist; die tatsächliche Beobachtung von E bedeutet dementsprechend einen großen Informationsgewinn relativ zu K.

2.4 Stochastische Unabhängigkeit

Zwei zufällige Ereignisse sind *abhängig*, wenn $p(A|B) \neq p(A)$, denn das Eintreten von B erfordert dann eine andere Beurteilung der Wahrscheinlichkeit von A.

[9] Earman, J.: Bayes or bust? A critical examination of Bayesian confirmation theor. MIT Press, Cambridge, London 1992

Es gibt dann irgendeine Art von Zusammenhang zwischen den Ereignissen. So sind etwa in der psychologischen Diagnostik gewisse Abhängigkeiten durchaus erwünscht: ist B ein Symptom und A ein durch das Symptom vorhersagbares Merkmal, so hätte man gern eine Abhängigkeit der Art $p(A|B) \approx 1$.

Hier soll insbesondere die stochastische Unabhängigkeit von zufälligen Ereignisse charakterisiert werden.

Definition 2.4.1 *Es sei $(\Omega, \mathcal{A})$ der betrachtete Ereignisraum und A, B seien zwei beliebige Ereignisse aus $\mathcal{A}$. Gilt*

$$p(A|B) = p(A), \tag{2.51}$$

so heißen A und B stochastisch unabhängig.

Gilt (2.51), so ist klar, dass B die Vorhersage von A nicht erleichtert, denn die Wahrscheinlichkeit für das Eintreten von A wird durch die Kenntnis von B nicht verändert. Wegen $p(A|B) = p(B|A)p(A)/p(B)$ folgt mit (2.51) sofort, dass auch $p(B|A) = p(B)$, so dass auch B von A stochastisch ist.

Nun gilt allgemein $p(A|B) = p(A \cap B)/p(B)$. Setzt man $p(A)$ für $p(A|B)$ ein, so erhält man $p(A) = p(A \cap B)/p(B)$, woraus sofort

$$p(A \cap B) = p(A)p(B) \tag{2.52}$$

folgt; dies ist die *Multiplikationsregel*; sie gilt nur im Falle stochastischer Unabhängigkeit der Ereignisse A und B. Die Multiplikationsregel wird, wie eben gezeigt wurde, von der stochastischen Unabhängigkeit impliziert. Umgekehrt folgt aus der Gültigkeit der Multiplikationsregel auch die stochastische Unabhängigkeit der Ereignisse A und B. Denn angenommen, wir schreiben (2.52). Andererseits gilt allgemein nach (4.1) $p(A \cap B) = p(B|A)p(A)$, so dass wir $p(A)p(B) = p(B|A)p(A)$ und damit $p(B|A) = p(B)$ folgern können; dies heißt aber, dass A und B stochastisch unabhängig sind.

Der Fall der stochastischen Unabhängigkeit ist ein Spezialfall von Abb. 2.2 (b), wenn nämlich

$$(P(B - A) + P(A \cap B))(P(A - B) + P(A \cap B)) = P(A \cap B) \tag{2.53}$$

ist: sei $P(B - A) = .3$; $P(A - B) = .2$, $P(A \cap B) = .2$. Dann ist $(.3 + .2)(.2 + .2) = .5 \times .4 = .2$. Stochatische Unabhängigkeit meint also nicht den Fall in Abb. 2.2, (c), wo A und B disjunkt sind. Der disjunkte Fall bedeutet Abhängigkeit: $P(A|B) = P(B|A) = 0$, dann $P(A \cap B) = 0$.

Spezialfälle:

1. Es seien A und B zufällige Ereignisse und es gelte $p(A) = 0$. Dann ist $p(A \cap B) = p(B|A)p(A) = 0 = p(A)p(B)$. Also sind A und B stochastisch unabhängig.

2. Es gelte $A = \Omega$ und mithin $p(A) = 1$. Dann folgt $p(B|A) = p(B)$ und mithin $p(A \cap B) = p(B|A)p(A) = p(B)p(A)$, und A und B sind wiederum stochastisch unabhängig.

Sind A und B stochastisch unabhängig, so sind auch A und $\bar{B}$ stochastisch unabhängig.

Es ist $p(\bar{A}|B) = p(\bar{A} \cap B)/p(B) = p(B - A \cap B)/p(B)$. Es ist

$$p(B) = p(B - A \cap B \cup A \cap B) = p(B - A \cap B) + p(A \cap B),$$

da $(B - A \cap B) \cap (A \cap B)) = \emptyset$. Nach Voraussetzung gilt $p(A \cap B) = p(A)p(B)$, so dass $p(B) = p(B - A \cap B) + p(A)p(B)$, also $p(B - A \cap B) = p(B)(1 - p(A))$, mithin

$$p(\bar{A}|B) = \frac{p(B)(1 - p(A))}{p(B)} = p(\bar{A}),$$

d.h. aber, dass $\bar{A}$ und B stochastisch unabhängig sind. □

Praktische und theoretische Fragen legen es nun nahe, die Definition der stochastischen Unabhängigkeit von zwei auf mehr Ereignisse zu verallgemeinern. Bei der Analyse von Tests und Fragebögen ist es z. B. gelegentlich wichtig, annehmen zu können, dass alle Aufgaben oder Fragen unabhängig voneinander beantwortet werden. Bei experimentellen Untersuchungen möchte man die Annahme machen können, dass die verschiedenen Versuchspersonen (Vpn) unabhängig voneinander auf die experimentelle Aufgabe reagieren, oder dass eine Vp in jedem Versuchsdurchgang unabhängig von den vorausgegangenen reagiert. Das folgende Beispiel zeigt aber, dass man von der stochastischen Unabhängigkeit zweier Ereignisse nicht auf die Unabhängigkeit aller Ereignisse schließen darf:

Beispiel 2.4.1 Es werden zwei Würfel, W_1 und W_2, geworfen. Dann ist $\Omega = \{(i, j)|1 \leq i, j \leq 6\}$ die Menge der Elementarereignisse. Ein Elementarereignis ω ist also ein Paar (i, j), i die gewürfelte Augenzahl von W_1 und j die entsprechende Augenzahl von W_2. Es werde angenommen, dass $p(\omega) = 1/|\Omega| = 1/36$ für alle $\omega \in \Omega$ gilt, die Elementarereignisse also gleichwahrscheinlich sind. Weiter sei $A_k = \{W_k \quad \text{zeigt eine gerade Augenzahl}\}$ für $k = 1, 2$. Nun ist

$$p(A_1 \cap A_2) = \frac{|A_1 \cap A_2|}{|\Omega|} = \frac{9}{36} = \frac{1}{4}.$$

Andererseits ist

$$p(A_1) = \sum_{j=1}^{6} p(2, j) + \sum_{j=1}^{6} p(4, j) + \sum_{j=1}^{6} p(6, j) = \frac{3 \cdot 6}{36} = \frac{1}{2}.$$

Auf analoge Weise findet man, dass $p(A_2) = 1/2$, so dass sich schließlich $p(A_1)p(A_2) = p(A_1 \cap A_2) = 1/4$ ergibt. Also sind A_1 und A_2 stochastisch unabhängig. Nun

werde das weitere Ereignis $A_3 = \{i + j \quad \text{ist gerade}\}$ eingeführt. Man findet $p(A_1) = p(A_3) = 1/2$. Es ist

$$p(A_3|A_1) = \frac{p(A_1 \cap A_3)}{P(A_1)} = \frac{|A_1 \cap A_3|}{|A_1|} = \frac{1}{2} = P(A_3),$$

so dass A_1 und A_3 stochastisch unabhängig sind. Analog findet man $P(A_3|A_2) = 1/2$, so dass auch A_2 und A_3 stochastisch unabhängig sind. Nun ist aber $\{A_1 \cap A_2 \cap A_3\} = \{A_1 \cap A_2\}$, denn sind zwei Zahlen gerade, so ist auch ihre Summe gerade, so dass $p(A_1 \cap A_2 \cap A_3) = p(A_1 \cap A_2) = 1/2$. Andererseits ist $p(A_1)p(A_2)p(A_3) = (1/2)^3 = 1/8$, die Ereignisse A_1, A_2 und A_3 sind also insgesamt *nicht* stochastisch unabhängig. □

Definition 2.4.2 *Es sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum. Die die zufälligen Ereignisse $A_i \in \mathcal{A}$, $i = 1, \cdots, n$ heißen*

1. paarweise stochastisch unabhängig *(bezüglich des Wahrscheinlichkeitsmaßes P), wenn*
$$p(A_i \cap A_j) = p(A_i)p(A_j)$$
für alle i, j mit $i \neq j$ gilt;

2. stochastisch unabhängig *(bezüglich P), wenn*
$$p(A_{i_1} \cap \cdots \cap A_{i_m}) = p(A_{i_1}) \cdots p(A_{i_m})$$
für je endlich viele paarweise verschiedene Indices
$$i_1, \ldots, i_m \ \in \{1, \ldots, m\}.$$

Gilt z.B. die Produktregel für alle Ereignisse A_i, d.h.

$$p(A_1 \cap \cdots \cap A_n) = p(A_1) \cdots p(A_n),$$

so folgt daraus noch nicht die paarweise stochastische Unabhängigkeit Es sei etwa $\Omega = \{1, 2, 3\}$, $p(1) = 1/2$, $p(2) = 1/2$, $p(3) = 0$. Weiter sei $A_1 = \{1\}$, $A_2 = \{2\}$, $A_3 = \{3\}$. Dann ist $p(A_1 \cap A_2 \cap A_3) = p(\emptyset) = 0$, weiter ist $p(A_1)p(A_2)p(A_3) = 0$, $p(A_1 \cap A_2) = p(\emptyset) = 0$, aber $p(A_1)p(A_2) = 1/4$. Sind dagegen zufällige Ereignisse $A_i \ \epsilon \ A$ stochastisch unabhängig im Sinne von Definition 2.4.2, so sind sie auch paarweise stochastisch unabhängig; dies folgt unmittelbar aus der Definition.

Beispiel 2.4.2 In einem westlichen Industrieland gibt es 21 arbeitende AKWs. Die Wahrscheinlichkeit, dass während eines Jahres in keinem AKW ein GAU, geschieht, werde von den AKW-Befürwortern mit .999 angegeben. Wie groß ist n die durchschnittliche Wahrscheinlichkeit für ein AKW, keinen GAU zu liefern?

Es sei A_i das Ereignis, dass im i-ten AKW ein GAU innerhalb eines Jahres geschieht. Es werde angenommen, dass die A_i, $1 \leq i \leq 21$ stochastisch unabhängig im Sinne Von Def. 2.4.2 (2) sind. Dann sind auch die $\bar{A}_i$ in diesem Sinne stochastisch unabhängig, und die Wahrscheinlichkeit eines GAUs ist durch

$$p(\text{kein GAU innerhalb eines Jahres}) = \prod_{i=1}^{n} p(\overline{A}_i)$$

$n = 21$, gegeben. Die durchschnittliche Wahrscheinlichkeit, $\bar{q}$, *keinen* GAU zu liefern, ist durch das geometrische Mittel der $p(\bar{A}_i)$ gegeben. Dann ist $\bar{q}^{21} = .999$, so dass $\bar{q} = .999^{21} = .9995$ ist.

Wir können jetzt noch berechnen, wie groß die Wahrscheinlichkeit für das Ereignis A, dass mindestens ein GAU eintritt. Dies ist die Wahrscheinlichkeit, dass mindestens eines der Ereignisse A_i eintritt. Offenbar ist dies Ereignis komplementär zu dem Ereignis, dass kein GAU eintritt, so dass

$$p(A) = 1 - \overline{A}_i$$

gelten muß. Es ist $p(A) = 1 - .999 = .001$. □

Anmerkung: Die Wahrscheinlichkeit, dass mindestens eines der Ereignisse A_i eintritt, kann im Prinzip auch über die Formel $p(A_1 \cup \cdots \cup A_n)$ berechnet werden; in der Tat ist dieser Ausdruck die direkte übersetzung der Aussage "mindestens ein Ereignis" in einen Formelausdruck. Tatsächlich ist dieser Ausdruck aber sehr unhandlich, denn

$$p(A_1 \cup \cdots \cup A_n) = p(A_1) + \cdots + p(A_n) - R,$$

wobei $R = -(p(A_1 \cap A_2) + p(A_1 \cap A_3) + \cdots + p(A_1 \cap A_2 \cap A_3) + \cdots)$ ist. Man benötigt also die Wahrscheinlichkeiten für alle möglichen Konjunktionen der Ereignisse. Abgewesen davon, dass diese Größen gar nicht immer bekannt sind ist es rechnerisch sehr unökonomisch, auf diese Weise vorzugehen, denn der Ausdruck

$$p(A) = 1 - \prod_{i=1}^{n} p(\overline{A}_i) = 1 - \prod_{i=1}^{n} (1 - p(A_i)) \tag{2.54}$$

benötigt nur die Wahrscheinlichkeiten $p(A_i)$.

Beispiel 2.4.3 In der psychologischen Diagnostik ist man u. a. darum bemüht, bestimmte Persönlichkeitsmerkmale durch andere "vorherzusagen". Wir betrachten hier der Einfachheit wegen nur Merkmale, die entweder "vorhanden" oder "nicht vorhanden" sind. Sicherlich ist es so, dass stochastisch unabhängige Merkmale nicht zur wechselseitigen Vorhersage taugen. Eine Schwierigkeit für die Diagnostik entsteht nun daraus, dass aus der paarweisen stochastischen Unabhängigkeit noch keine allgemeine stochastische Unabhängigkeit folgt. Dazu betrachten

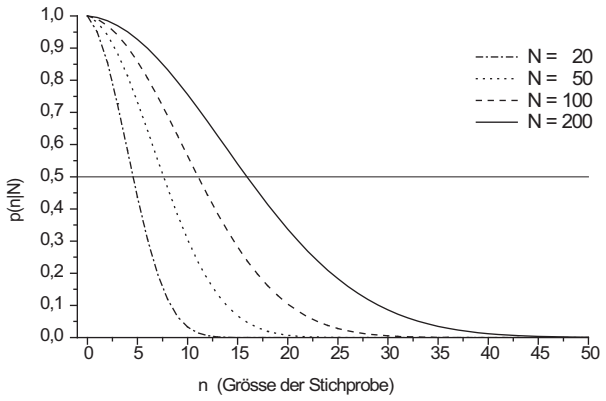
wir drei Merkmale M_1, M_2 und M_3. Ist eines der Merkmale M_i bei einer (zufällig gewählten) Person vorhanden, so werde gesagt, dass zufällige Ereignis A_i sei eingetreten, $i = 1, 2, 3$. Längere Beobachtungen mögen nun ergeben haben, dass $p(A_1) = .9$, $p(A_2) = .7$ und $p(A_3) = .5$ gilt. Weiter hat man $p(A_1 \cap A_2) = .63$, $p(A_1 \cap A_3) = .45$, $p(A_2 \cap A_3) = .35$ sowie $p(A_1 \cap A_2 \cap A_3) = .14$ gefunden.

Man rechnet sofort nach, dass für alle Paare (i, j) mit $i \neq j$ $p(A_i \cap A_j) = p(A_i)p(A_j)$ gilt, dass aber nicht $p(A_1 \cap A_2 \cap A_3) = p(A_1)p(A_2)p(A_3)$ gilt. Wegen der paarweisen stochastischen Unabhängigkeit kann man nicht von einem Merkmal auf ein anderes schließen. Kennt man aber bei einer Person die Ausprägung zweier Merkmale M_i und M_j, so kann man aber nicht folgern, dass das dritte Merkmal M_k ebenfalls stochastisch unabhängig von M_i und M_j auftritt. Man hat es hier mit *konfiguralen Abhängigkeiten* zu tun, die sich in den bedingten Wahrscheinlichkeiten $p(A_1|A_2 \cap A_3)$, $p(A_2|A_1 \cap A_3)$, $p(A_3|A_1 \cap A_2)$ äußern. Anhand der gegebenen Werte berechnet man leicht $p(A_1|A_2 \cap A_3) = .4 \neq p(A_1)$, $p(A_2|A_1 \cap A_3) = .3111 \neq p(A_2)$ und $p(A_3|A_1 \cap A_2) = .2222 \neq p(A_3)$.

Bedenkt man, dass der hier beschriebene Fall wegen der Beschränkung auf drei Merkmale eine starke Vereinfachung darstellt, so wird deutlich, dass die Diagnostik vor einer außerordentlich komplexen Aufgabe steht. □

Beispiel 2.4.4 Gegeben sei eine Population der Größe N, aus der eine Stichprobe vom Umfang n gezogen wird. Die Elemente der Stichprobe werden stochastisch unabhängig voneinander, aber *mit Zurücklegen* gezogen. Wie groß ist die Wahrscheinlichkeit, dass die Elemente der Stichprobe alle verschieden sind, also ein Element der Population höchstens einmal gezogen wird? Es seien bereits $j < n$

Abbildung 2.3: Wahrscheinlichkeit von Stichproben ohne Mehrfachziehungen beim unabhängigem Ziehen mit Zurücklegen



Elemente aus der Population gezogen worden, und keines dieser Elemente sei mehr als einmal gezogen worden. Das Ziehen eines Elements aus der Population

ist ein zufälliges Ereignis, für den $j+1$-ten Zug etwa E_{j+1} $(j+1 \leq n)$. Die Wahrscheinlichkeit, beim $j+1$-ten Zug ein Element zu ziehen, dass noch nicht in der bisher gezogenen Stichprobe enthalten ist, ist

$$p_{j+1} = p(E_{j+1}) = \frac{N-j}{N} = 1 - \frac{j}{N},$$

Für $j = 0$ ist noch kein Element gezogen worden, so dass für den ersten Zug $p_{0+1} = p_1 = 1 - 0/N = 1$ gilt, für den zweiten $p_{1+1} = p_2 = 1 - 1/N$, etc.. Die Ereignisse E_{j+1} sollen stochastisch unabhängig sein; also folgt

$$\begin{aligned} qP(E_1 \cap \cdots \cap E_n) &= p(E_1) \cdots p(E_n) \\ &= \left(1 - \frac{0}{N}\right)\left(1 - -\frac{1}{N}\right) \cdots \left(1 - \frac{n}{N}\right) \\ &= \prod_{j=0}^{n-1} \left(1 - \frac{j}{N}\right) \end{aligned} \tag{2.55}$$

In Abbildung 2.3 werden die Wahrscheinlichkeiten für zwei Populationen, eine vom Umfang $N = 50$ und eine vom Umfang $N = 100$, dargestellt; die Zahlen in Klammern an den x-Achsen stellen jeweils den Stichprobenumfang dar. Man sieht, dass für größer werdendes n die Wahrscheinlichkeit, jedes Element der Stichprobe nur einmal zu ziehen, sehr schnell gegen Null geht. □

Kapitel 3

Zufällige Veränderliche und ihre Verteilungen

3.1 Allgemeine Definitionen

Wird ein Merkmal gemessen, so ist die Messung im allgemeinen mit zufälligen Fehlern behaftet. In diesem Sinne sind die Werte der Messungen zufällige Variable: die gemessenen Werte treten in Abhängigkeit von den Ausprägungen der Merkmale, aber auch von zufälligen Ereignissen auf. Um zu einer allgemeinen Charakterisierung von zufälligen Variablen (ZV; auch Zufallsvariablen) zu kommen, muß die Beziehung zwischen gemessenen Werten und zufälligen Ereignissen explizit gemacht werden. Nun reflektiert die Zufälligkeit einer Variablen die Zufälligkeit von Ereignissen, und deshalb liegt es nahe, eine zufällige Variable als Funktion dieser Ereignisse zu definieren. Eine Funktion ist aber eine Abbildung einer Menge in eine andere. Diese Abbildung muß aber gewissen Bedingungen genügen: (i) tritt das zufällige Ereignis $A \in \mathcal{A}$ ein, so wird man fordern, dass X Werte aus einer bestimmten Menge $B(A) \in \mathbb{R}$ annimmt, wobei $B(A) = B$ der Wertebereich von X ist und $\mathbb{R}$ ist die Menge der reellen Zahlen, und (ii) umgekehrt soll jede Teilmenge $B_0 \subset B$ einem zufälligen Ereignis $A_0 \subset A \in \mathcal{A}$ entsprechen. Die Forderung (i) ist sicherlich selbstverständlich; durch (ii) soll sichergestellt werden, dass jedem Wert von X auch tatsächlich ein beobachtbares Ereignis entspricht. Sicherlich entspricht damit dem Mengensystem A ein bestimmtes System B von Teilmengen reeller Zahlen. Weiter soll natürlich die Wahrscheinlichkeit $P_x(B)$, mit der ein Wert aus B gemessen wird, der Wahrscheinlichkeit $P(A)$ entsprechen, mit der das Ereignis $A \in \mathcal{A}$ auftritt. Das Wahrscheinlichkeitsmaß P *induziert* gewissermaßen ein Wahrscheinlichkeitsmaß P_x im Mengensystem B.

Definition 3.1.1 *Es sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum. Eine zufällige Veränderliche (Zufallsvariable; ZV) ist eine Abbildung $X : \Omega \to \mathbb{R}$, die den Bedingungen*

(i) $P_x(\{X \in B \subset \mathbb{R}\}) = P(X^{-1}(B))$, $X^{-1}(B) = A = \{\omega \in \Omega | X(\omega) \in B\}$,

(ii) $P_x(\{X = -\infty\}) = P_x((X = \infty\}) = 0$

genügt. P_x *ist das durch* P *auf* B *induzierte Wahrscheinlichkeitsmaß, wobei* B *durch die Definition von* X *festgelegt wird.*

Jedem Elementarereignis $\omega \epsilon \Omega$ wird also genau eine reelle Zahl zugeordnet; dabei kann aber verschiedenen $\omega \epsilon \Omega$ dieselbe Zahl zugeordnet werden. Nimmt X Werte aus einer bestimmten Menge B aus $\mathbb{R}$ an, so sollen den Elementen aus B die Elemente einer Teilmenge A aus Ω entsprechen. Dabei ist es zulässig, dass A aus nur einem Element besteht. Die Wahrscheinlichkeit, dass X Werte aus der Menge B annimmt, soll gleich der Wahrscheinlichkeit sein, dass das zufällige Ereignis A eintritt. Der Begriff der induzierten Wahrscheinlichkeit ist insofern wichtig, als ja die Definition einer zufälligen Veränderlichen im allgemeinen nicht eindeutig ist. So sei A das Ereignis, dass eine zufällig gewählte Person eine Körperlänge zwischen 1.70 $[m]$ und 1.75 $[m]$ hat, und X seien Längenmessungen in Metern. Dem zufälligen Ereignis A entspricht dann das zufällige Ereignis, eine Messung aus dem Bereich $B(A) = (1.70,\ 1.75)$ zu erhalten, d.h. A entspricht dem Ereignis $\{1.70 < X < 1.75\}$. Wir können aber auch vereinbaren, die Körperlänge in Zentimetern zu messen. Bedeuten nun 1.70 und 1.75 Zentimeter und ist weiterhin A die eben betrachtete Teilmenge von Personen, so ist sicherlich $P(A) \neq P_x(\{1.70 < X < 1.75\})$, vielmehr ist $P(A) = P_x(\{170 < X < 175\})$. Zur Definition von X gehört eben auch die Wahl der Maßeinheit, und damit wird eine der Maßeinheit entsprechende Menge B von reellen Zahlen festgelegt. Bei der Betrachtung von Funktionen von zufälligen Veränderlichen wird der Begriff der durch eine zufällige Veränderliche induzierten Wahrscheinlichkeit weiter erläutert.

Der Index $_x$ kann im allgemeinen weggelassen werden; es kann aber nützlich sein, ihn bei der Betrachtung von Funktionen oder Transformationen von z. V. explizit zu benützen.

Eine Menge M heißt bekanntlich *abzählbar*, wenn sie entweder nur endlich viele Elemente hat oder wenn sie zwar unendlich viele Elemente hat, die aber mithilfe der natürlichen Zahlen durchnummeriert werden können. Ist M nicht abzählbar, so heißt sie auch *überabzählbar*. Die Menge $M = \{x | x \in [0, 1]\}$ ist überabzählbar.

Definition 3.1.2 *Ist der Wertebereich einer ZV* X *abzählbar, so heißt die ZV* diskret. *Ist der Wertebereich überabzählbar, so heißt die ZV* stetig.

Beispiel 3.1.1 Ein einfaches Beispiel ist wieder der Münzwurf. Die zufällige Veränderliche X werde wie folgt definiert:

$$X(\omega) = \begin{cases} 1, & \text{wenn } \omega = K \\ 0, & \text{wenn } \omega = W \end{cases}$$

X ist offenbar diskret. Es werde die Annahme $P(\omega = K) = P(\omega = W) = 1/2$ gemacht. Dann ist $P_x(X = 1) = P_x(X = 0) = 1/2$; diese beiden Werte definieren hier die Wahrscheinlichkeitsverteilung von X. Man sieht, warum P_x durch P induziert wird: hätte man $X(\omega) = 5$ für $\omega = K$ und $X(\omega) = -5$ für $\omega = W$ definiert, so erhielten wir $P_x(X = 1) = P_x(X = 0) = 0$, denn die Werte 1 und 0 können nun nicht mehr auftreten. □

Beispiel 3.1.2 Beim Wurf eines Würfels kann die Menge der Elementarereignisse durch die sechs möglichen Seiten, die bei einem Wurf oben liegen können, repräsentiert werden. Es sei $\Omega = \{\omega_1, \cdots, \omega_6\}$. Dann definiert $X : \Omega \to \mathbb{R}$, $\omega_i \mapsto i$, $i = 1, \cdots, 6$ eine zufällige Veränderliche; X ist offenkundig diskret. Die möglichen Werte von X sind die Augenzahlen, die jeder Seite zugeordnet werden.

Es sei $A = \{2, 4, 6\}$. Die Abbildung $Y : \Omega \to \mathbb{R}$, die durch

$$Y(\omega) = \begin{cases} 1, & \text{wenn } \omega \in A \\ 0, & \text{wenn } \omega \notin A \end{cases}$$

definiert ist, ist ebenfalls eine ZV. Y nimmt den Wert 1 an, wenn eine gerade Zahl gewürfelt wurde, und eine 0 für den Fall einer ungeraden Zahl. Den möglichen Werten von Y entsprechen jeweils drei Elementarereignisse. Dem Ereignis $\{Y = 1\}$ entspricht das Ereignis A, mithin ist $P(Y = 1) = P(A)$, und dementsprechend ist $P(Y = 0) = P(\bar{A}) = 1 - P(A)$. Nimmt man weiter an, dass die Elementarereignisse ω gleichwahrscheinlich sind, so ist $P(Y = 1) = P(Y = 0) = 1/2$. □

Zufällige Veränderliche wie Y, die nur zwei Werte, 0 oder 1, annehmen können, heißen auch *Bernoulli-Variable*.

Beispiel 3.1.3 Eine Münze werde dreimal geworfen. Bei jedem Wurf liegt entweder "Kopf" (K) oder "Wappen" (W) oben. Es sei $\Omega_o = \{K, W\}$. Die Menge der Elementarereignisse für das aus einem dreimaligen Wurf bestehende "Experiment" seien die möglichen Folgen aus Elementen aus Ω_0, also $\Omega_0^3 = \Omega_0 \times \Omega_0 \times \Omega_0$. Dann werde ein ZV $X : \Omega \to \mathbb{N} \subset \mathbb{R}$ definiert durch $\omega \mapsto i \in \mathbb{N}$, genau dann, wenn $\omega \in \Omega$ genau i-mal das W enthält, wenn also eine Folge von Würfen gerade i-mal W enthält. Die $|\Omega| = 2^3 = 8$ Elemente aus Ω werden also den vier natürlichen Zahlen 0, 1, 2 und 3 zugeordnet. Es ist $X(KKK) = 0$, $X(WKK) = X(KWK) = X(KKW) = 1$, $X(WWK) = X(WKW) = X(KWW) = 2$, $X(WWW) = 3$. X ist diskret. Man hat $p(X = 0) = P(KKK)$, $P(X = 1) = p(WKK \cup KWK \cup KKW) = p(WKK) + p(KWK) + p(KKW)$, $P(X = 2) = P(WWK \cup WKW \cup KWW) = P(WWK) + P(WKW) + P(KWW)$ denn die Ereignisse WKK, KWK und KKW können nicht zugleich auftreten (sind durchschnittsfremd), und $p(X = 3) = p(WWW)$. Nimmt man an, dass jede Folge von drei Würfen gleichwahrscheinlich ist, so ist insbesondere $P(X = 0) = 1/8$, $P(X = 1) = 3/8$, $P(X = 2) = 3/8$, $P(X = 3) = 1/8$. □

Beispiel 3.1.4 In einem Experiment zu Lernen von Begriffen hat die Vp die Aufgabe, die Reizmuster aus einer bestimmten Menge von Mustern verschiedenen Klassen zuzuordnen. Der Einfachheit halber falle ein Muster entweder in die Klasse A oder in die Klasse B. Dabei muß sie die Regel, nach der die Klassifikation zu erfolgen hat, aufrund der "richtig"- und "falsch"-Meldungen des Versuchsleiters, die er nach jeder Darbietung und erfolgter Klassifikation eines Reizmusters gibt, erschließen. Es wird angenommen, dass die Vp Hypothesen über die Zuordnungsregel bildet und anwendet, wobei sie eine Regel so lange anwendet, wie sie zu "richtig"-Antworten des Versuchleiters führt. Bei einer "falsch"-Antwort verwirft sie die Hypothese und bildet eine neue. Es sei $\Omega = \{1, 2, 3, \cdots\} = \mathbb{N}$, $X : \Omega \to \mathbb{N} \subset \mathbb{R}$ sei die ZV, die die Anzahl der Hypothesen repräsentiert, die die Vp bilden muß, bis sie eine gefunden hat, die für alle Muster zu "richtig"-Antworten führt. X ist diskret. Die Anzahl der möglichen Hypothesen wird im allgemeinen endlich sein. Man könnte aber die Hypothese aufstellen, dass die Vp Hypothesen wählt wie man Kugeln aus einer Urne "zufällig" wählt, wobei eine schon einmal gewählte Hypothese wieder gezogen werden darf. Dann kann X unendlich viele Werte annehmen. □

Beispiel 3.1.5 Im Rahmen eines psychiatrischen Forschungsprogramms sollen die sozialen Akte schizophrener Patienten erfaßt werden. Dazu wird die Anzahl dieser Akte innerhalb einer Woche bestimmt. Für jeden Patienten wird eine ZV X definiert, die die Anzahl sozialer Akte repräsentiert ($X = k$ bedeutet also, dass das zufällige Ereignis $A_k = \{$die Menge der sozialen Akte umfaßt k Elemente$\}$ eingetreten ist. Die Definition von A_k fällt gewissermaßen mit der von X zusammen). Dann ist $P_x(X = k) = p_k = P(A_k)$ die durch P induzierte W-Verteilung von X. □

Stetige zufällige Veränderliche treten in den folgenden Beispielen auf:

Beispiel 3.1.6 Ein Autofahrer fährt auf einer Straße, auf der unvermittelt ein Hindernis auftritt. Von dem Moment t_o an, in dem er das Hindernis wahrnimmt, bis zu dem Zeitpunkt, in dem er auf die Bremse tritt, vergeht eine gewisse Zeit $t > 0$.

Ohne Einschränkung der Allgemeinheit kann man $t_0 = 0$ setzen. Es sei $\Omega = \{\omega | \omega$ Zeiten größer/gleich $t_0\}$. Die Abbildung $X : \Omega \to \mathbb{R}_+$, $\mathbb{R}_+ = \{x | x \in \mathbb{R}, x > 0\}$ definiert eine ZV, die die Zeit bis zur Reaktion (d.h. bis zur Betätigung des Bremspedals) repräsentiert. Dann wird jedem $\omega \in \Omega$ (durch eine Uhr) eine reelle, positive Zahl $X(\omega)$ zugeordnet, und diese Zuordnung wird durch die Wahl der Zeiteinheit mitbestimmt. Die Beziehung zwischen dem Wert einer stetigen zufälligen Veränderlichen und der Wahrscheinlichkeit, mit der er auftritt, ist etwas komplizierter als bei diskreten Variablen und muß etwas ausführlicher behandelt werden. □

Beispiel 3.1.7 Es werde noch einmal das Beispiel 3.1.5 betrachtet; X sei wieder die Zeit bis zur Reaktion, gemessen in einer bestimmten Zeiteinheit. X ist eine stetige ZV. Es sei $A \in \mathcal{A}$, $\mathcal{A}$ die zugrundeliegende σ-Algebra, ein bestimmtes Zeitintervall und $\tilde{A}$ das zugehörige Intervall auf der reellen Achse; die Breite des Intervalles $\tilde{A}$ hängt von der gewählten Zeiteinheit ab. $P_x(\tilde{A}) = P(A)$ ist die durch das Wahrscheinlichkeitsmaß P induzierte W-Verteilung von X. Es sei $\omega \epsilon \Omega$ eine bestimmte Reaktionsdauer, die als $X[\cdot]$ ausgedrückt werden kann. $[\cdot]$ ist die gewählte Zeiteinheit, etwa Millisekunden. □

Die folgende Definition gilt für diskrete wie für stetige zufällige Veränderliche gleichermaßen; sie wird aber benötigt, um für stetige ZV Wahrscheinlichkeitßuordnungen zu treffen:

Definition 3.1.3 *Es sei* $(\Omega, \mathcal{A}, P)$ *ein Wahrscheinlichkeitsraum. Die Abbildung* $F : B \to [0,1]$, $F(x) = P(X \leq x) = P(X \ \epsilon(-\infty, x])$ *heißt* Verteilungsfunktion *der ZV* X.

Aus dieser Definition folgen sofort einige wichtige Eigenschaften der Verteilungsfunktion; zur Erinnerung werde vorher noch der Begriff des Limes von rechts (bzw, links) erläutert. Es sei $\varphi(x)$ eine Funktion von x. Es sei $a < x$. Man kann nun "von oben" x gegen a streben lassen. Existiert der Grenzwert $\lim_{x \downarrow a} \varphi(x) = \varphi_+(a)$, so heißt $\varphi_+(a)$ *rechtsseitiger Grenzwert.* Analog kann man für $x < a$ den Grenzwert von $\varphi(x)$ betrachten; dieser sei $\varphi_-(a) = \lim_{x \uparrow a} \varphi(x)$; $\varphi_-(a)$ heißt *linksseitiger Grenzwert.* Die Grenzwerte $\varphi_-(a)$ und $\varphi_+(a)$ müssen nicht übereinstimmen. Ist $\varphi_+(a) > \varphi_-(a)$, so macht die Funktion φ an der Stelle a einen Sprung der Größe $\varphi_+(a) - \varphi_-(a)$.

Satz 3.1.1 *Es sei* X *eine zufällige Veränderliche und* $F : B \to [0,1]$ *ihre Verteilungsfunktion. Dann hat* F *die folgenden Eigenschaften*

(a) $0 \leq F(x) \leq 1$

(b) für $x < y$ *folgt* $F(x) \leq F(y)$

(c) $F(-\infty) = 0$, $F(\infty) = 1$

(d) $\lim_{h \to 0} F(x-h) = F(x)$, $h > 0$, *d.h.* F *ist stetig von rechts.*

(e) für $a < b$ *gilt* $p(X \ \epsilon\, (a,b]) = p(X \leq b) - p(X \leq a) = F(b) - F(a)$

Beweis: (a) Da $P(X \leq x)$ eine Wahrscheinlichkeit ist, muß $0 \leq F(x) \leq 1$ sein. (b) Es seien x und y zwei beliebige Werte von X, die die Bedingung $x \leq y$ erfüllen. Da $(-\infty, x] \subseteq (-\infty, y]$ muß nach Satz 4.3.1 (d) $p((-\infty, x]) \leq p((-\infty, y])$ und damit $F(x) \leq F(y)$ sein, d.h. F muß eine monoton wachsende Funktion von x sein. (c) Da $X < \infty$ folgt wegen (b) sofort, dass $F(\infty) = 1$ sein muß; analog folgt

$F(-\infty) = 0$. (d) Eine Funktion heißt *stetig von rechts*, wenn für jede Folge y_1, y_2, ... , die einen Punkt x von rechts konvergiert, der Grenzwert $F(x)$ existiert. Es sei y_1, $y_2, ...$ eine solche Folge. Dann ist das Intervall $(-\infty,\ x]$ der Durchschnitt der Intervalle $(-\infty,\ y_n]$, und es ist $p((-\infty, x]) = p(\cap(-\infty, y_n]) = \lim_{n\to\infty} F(y_n)$. (e) Es sei A das zufällige Ereignis, dass X im Intervall $(-\infty, a]$ liegt, und B sei das zufällige Ereignis, dass X im Intervall $(-\infty, b]$ liegt und es sei $a < b$. Dann ist $\bar{A}$ das Ereignis, dass X in (a, ∞) liegt. Dann ist $\bar{A} \cap B$ das Ereignis, dass X ϵ (a, b] und nach Satz 3.1, (c) ist dann $p(\bar{A} \cap B) = p(X\ \epsilon(a, b]) = p(B) - p(A) = F(b) - F(a)$.

□

In vorangegangenen Satz wurde die Wahrscheinlichkeit, dass X in einem halboffenenem Intervall liegt, angegeben. Natürlich läßt sich auch die Wahrscheinlichkeit, dass X einen Wert im abgeschlossenen Intervall $[a, b]$ annimmt, ausdrücken. In Satz 3.1.1, (e) fehlt nur der Wert der Wahrscheinlichkeit, dass X den Wert a annimmt. Aus Satz 3.1.1 (e) ergibt sich dann

$$p(X\ \epsilon[a, b]) = F(b) - F(a) + p(X = a), \tag{3.1}$$

und für die Wahrscheinlichkeit, dass X im offenen Intervall (a, b) liegt, ergibt sich

$$p(X\ \epsilon(a, b)) = F(b) - F(a) - p(X = b). \tag{3.2}$$

Satz 3.1.1 (e) zeigt, dass die Wahrscheinlichkeit des Ereignisses $X\ \epsilon(a, b]$ über die Verteilungsfunktion F ausgedrückt werden kann. (3.1) und (3.2) erweitern dieses Resultat für offene und geschlossene Intervalle. Läßt man nun b gegen a streben, so strebt nach (3.1) $P(X\ \epsilon(a, b))$ gegen $P(X = a)$, sofern F in a einen Sprung (der Höhe $P(X = a)$) hat. Ist aber F stetig in a, so ist die Höhe des Sprunges und damit $P(X = a)$ gleich Null.

In Satz 3.1.1 (e) wird die Wahrscheinlichkeit, dass X einen Wert in einem (halb-offenem) Intervall $(a, b]$ annimmt, durch die Differenz $F(b) - F(a)$ ausgedrückt. Für $a = -\infty$, $b = x$ folgt dann sofort

$$F(x) = p(-\infty < X \leq x). \tag{3.3}$$

Wir betrachten zunächst abzählbare Mengen Ω. Die Werte der zufälligen Veränderlichen können dann durchnummeriert werden, d.h. X nimmt die Werte $x_1, x_2, x_3, \ldots$ an Es sei P_x die W-Verteilung von X; für jeden X-Wert x_i existiert eine Wahrscheinlichkeit $P_x(x_i)$. Allgemein gilt $P(X\ \epsilon(a, b]) = F(b) - F(a)$, und $F(x) = P(X \leq x)$ ist eine monoton steigende Funktion von x. Nun sei $x = x_j + \varepsilon < x_{j+1}$. Es haben aber Werte von x zwischen zwei benachbarten Werte, x_j und x_{j+1} die Wahrscheinlichkeit Null, und da andererseits $F(x)$ eine monoton steigende Funktion von x ist, kann F für diese Werte von x nicht kleiner werden. F kann aber auch nicht wachsen, also muß $F(x)$ für $x_j < x < x_{j+1}$ konstant sein. Erst, wenn x den Wert x_{j+1} annimmt, macht F einen Sprung, und

zwar um den Betrag $p(X = x_{j+1}) = F(x_{j+1}) - F(x_j)$. Wir können die $P(X = x_j)$ in einem Stabdiagramm darstellen. Offenbar ist

$$F(x_k) = \sum_{i=1}^{k} P(X = x_i) \tag{3.4}$$

Man beachte, dass X nicht notwendig eine Anzahl von Elementen oder Ereignissen bedeuten muß. Ω kann z.B. eine Gruppe von Personen sein, und X ist eine nach irgendwelchen Kriterien gewählte Durchnummerierung der Personen. $F(x_k) = P(X \leq x_k)$ mit $x_k = k$ ist dann die Wahrscheinlichkeit, dass man bei zufälliger Wahl eine der Personen wählt, deren Nummer zwischen 1 und k liegt.

Beispiel 3.1.8 Münzwurf Für den Münzwurf ist die zufällige Veränderliche X gemäß

$$X(\omega) = \begin{cases} 1, & \omega = K, \\ 0, & \omega = W, \end{cases}$$

und es ist $p(X = 1) = p(X = 0) = 1/2$. Die Verteilungsfunktion ist dann durch

$$F(x) = \begin{cases} 0, & x < 0 \\ \frac{1}{2}, & 0 \leq x \\ 1, & x \geq 1 \end{cases}$$

gegeben. In $x = 0$ springt die Funktion F um den Betrag 1/2, d.h. um die Wert der Wahrscheinlichkeit für das Ereignis $\{X = 0\}$. In $x = 1$ springt F wieder um den Betrag 1/2, d.h. um den Wert der Wahrscheinlichkeit des Ereignisses $\{X = 1\}$. Die Intervalle $(x_i - \Delta x, x_i]$ sind hier die Intervalle zwischen den natürlichen Zahlen -1, 0, 1, 2, ... □

Beispiel 3.1.9 Für den Würfel (vergl. Beispiel 3.1.1) ist $\Omega = \{\omega | \omega$ ist eine der sechs Seiten des Würfels$\}$, und mit Laplace kann man anehmen, dass $P(\omega) = 1/6$ für alle $\omega \epsilon \Omega$. üblicherweise (und trivialerweise) wählt man die ZV $X : \Omega \to \{1, 2, 3, 4, 5, 6\}$, $X(\omega) \mapsto i$, $1 \leq i \leq 6$. Dann ist $P_x(i) = P(\omega) = 1/6$ für alle ω bzw. i die durch P induzierte Wahrscheinlichkeitsverteilung. Die Verteilungsfunktion ist durch

$$F(k) = P(X \leq k) = \begin{cases} 0, & X < 1 \\ \sum_k P_k k, & 1 \leq k \leq 6 \\ 1, & X > 6 \end{cases} \tag{3.5}$$

wobei $P_k = P(X = k) = 1/6$. Die Wahrscheinlichkeitsverteilung $p(X = k)$ ist eine diskrete Gleichverteilung, da (i) X nur diskrete Werte annimmt, und (ii) alle $p(x = k)$ gleich groß sind.

□

Nach Satz 3.1.1, Gl. (3.1) ist $p(X \ \epsilon(a, b]) = F(b) - F(a)$. Betrachten wir nun die Reaktionszeiten aus Beispiel 3.1.8. Gesucht ist Repräsentation für $F(x)$, die

zu (3.4) analog ist. Es wird nun vorausgesetzt, dass F eine stetige und auf dem Definitionsbereich differenzierbare Funktion ist[1]. Es werde $b = a + \Delta x$ gesetzt. Wegen der Differenzierbarkeit existiert dann, mit $x = a$, jedenfalls der Grenzwert

$$\frac{dF}{dx} := f(x) = \lim_{\Delta x \to 0} \frac{F(x + \Delta x) - F(x)}{\Delta x} \tag{3.6}$$

aus $\Delta x > 0$ und $F(x + \Delta x) > F(x)$ folgt, dass stets auch $f(x) \geq 0$ für alle x aus dem Definitionsbereich von F.

Ersetzt man Δx durch dx, so hat man

$$F(x + dx) - F(x) = p(x \leq X \leq x + dx) = f(x)dx. \tag{3.7}$$

Die Wahrscheinlichkeit, dass X zwischen x und $x+dx$ liegt, ist also durch $f(x)dx$ gegeben. Man muß hier bedenken, dass es sich nur um eine symbolische Schreibweise handelt, denn dx entspricht ja keinem bestimmten numerischen Wert; in der Tat gilt aber für hinreichend kleines $\Delta x \, \varepsilon \, \mathbb{R}$, dass $p(x \leq X \leq x+\Delta x) \approx f(x)\Delta x$. Die Ableitung $f(x) = dF(x)/dx$ ist keine Wahrscheinlichkeit.

Definition 3.1.4 *Es sei B der Wertebereich von X, F: $B \to [0, 1]$ sei eine Verteilungsfunktion Dann heißt $f = dF/dx \geq 0$ die zu F gehörige* Dichtefunktion.

Wegen $f(x) = dF(x)/dx$ erhält man nun sofort eine Darstellung für F :

$$F(x) = \int_{-\infty}^{x} f(\xi)\, d\xi \tag{3.8}$$

d.h. $F(x)$ ist das Integral der Dichtefunktion f. Im Falle einer stetigen Verteilungsfunktion wird also die Summe in (3.7) für $\Delta x \to 0$ durch ein Integral ersetzt. Wegen (3.2) erhält man aus (4.6.11) sofort

$$\int_{-\infty}^{\infty} f(\xi)\, d\xi = 1, \tag{3.9}$$

d.h. die Dichtefunktion f muß stets so normiert sein, dass das Integral von f über dem Definitionsbereich gleich 1 wird. Weiter folgt für $p(X \, \epsilon (a, b])$ wegen (3.1) aus (4.6.12) sofort

$$p(X \ \in (a, b]) = F(b) - F(a) = \int_{-\infty}^{b} f(\xi)d\xi - \int_{-\infty}^{a} f(\xi)d\xi. \tag{3.10}$$

Beispiel 3.1.10 Der Autofahrer in Beispiel 3.1.6 pralle zur Zeit T auf das Hindernis, falls er es nicht rechtzeitig bemerkt und abbremst. Es sei wieder t_0 der Zeitpunkt, zu dem er das Hindernis bemerke, $t_0 \leq T$. Es sei X die Reaktionszeit, und es gelte

$$p(X \leq t) = \int_{t_0}^{t} f(\tau)d\tau, \tag{3.11}$$

[1]Es genügt, wenn F bis auf abzählbar viele Punkte differenzierbar ist.

wobei die zusätzliche Annahme $t \leq T$ sinnvoll ist. Die Wahrscheinlichkeitdichte f ist dann auf dem Intervall $[t_0, T]$ definiert, und es gilt $g(\tau) = 0$ für $X < t_0$ und $t > T$. Insbesondere gelte $f(X) = 1/(T - t_0)$ für alle $X \epsilon [t_0, T]$; $f(X)$ ist damit konstant für alle Werte von t aus $[t_0, T]$. Die Dichte f heißt deshalb auch stetige Gleichverteilung. Man findet dann, dass $P(X \leq t) = (t - t_0)/(T - t_0)$ für $t \epsilon [t_0, T]$. Für $t = T$ folgt $P(X \leq T) = 1$. □

Der Vollständigkeit halber soll noch der Begriff des Stieltjes-Integrals eingeführt werden, der bei der Betrachtung von Verteilungsfunktionen und damit zusammenhängenden Größen wie Erwartungswerten (vergl. Abschn. 3.5) gelegentlich auftaucht.

Definition 3.1.5 *Es seien $\varphi(x)$ und $\psi(x)$ auf dem Intervall $[a, b]$ der reellen Achse erklärte Funktionen. $[a, b]$ werde in Teilintervalle $a < x_1 < x_2 < \cdots x_m < b$ zerlegt. Es sei $\xi_k \in [x_k, x_{k+1}]$. Dann heißt*

$$\sigma = \sum_{k=0}^{m-1} \varphi(\xi_k)[\psi(x_{k+1}) - \psi(x_k)]$$

Stieltjes-Summe. *Es sei $I \in \mathbb{R}$. Gibt es zu jedem $\epsilon > 0$ ein $\delta > 0$ derart, dass für jedes $\lambda = \max(x_{k+1} - x_k) < \delta$ die Bedingung $|\sigma - I| < \epsilon$ für beliebige Zerlegung von $[a, b]$ und beliebiger Wahl von $\xi_k \in [x_k,\ x_{k+1}]$ erfüllt ist, so heißt I* Stieltjes-Integral von φ nach ψ*; man schreibt*

$$I = \int_a^b \varphi(x) d\psi(x).$$

Für $g(x) = x$ geht das Stieljes-Integral in das gewöhnliche Riemann-Integral über. Für $\varphi(x) \equiv 1$ erhält man

$$I = \int_a^b d\psi(x)$$

Es sei $\eta = d\psi/dx$, so dass $\psi = \int \eta dx$. Dann ist $d\psi = \eta dx$ und $\int_a^b d\psi(x) = \int_a^b \eta dx = \psi$. Für die Verteilungsfunktion F und die Dichte f gilt aber gerade die Beziehung $f = dF/dx$, $F = \int f dx$; wir können also F als Stieltjes-Integral $F = \int dF$ schreiben. Die Wahrscheinlichkeit $P(X \leq t)$ in (3.1) etwa kann dann in der Form

$$P(X \leq t) = \int_{t_0}^t dF(\tau)$$

geschrieben werden.

Es sei F eine stetige Wahrscheinlichkeitsverteilung. Es werde noch die Wahrscheinlichkeit betrachtet, dass X einen speziellen Wert x annimmt. Es sei $a = x$, $b = x + \Delta x$. Nach (3.1) ist dann

$$p(X\ \epsilon(x, x + \Delta x]) = F(x + \Delta x) - F(x).$$

Dann ist

$$p(X = x) = \lim_{\Delta x \to 0} p(X \ \epsilon (x, x + \Delta x])$$

und wegen (3.10) folgt sofort $p(X = x) = 0$. Man mache sich hier noch einmal klar, dass zwar die Wahrscheinlichkeit, dass X einen bestimmten Wert x annimmt, stets gleich Null ist, das Ereigis $\{X = x\}$ aber keinesweges unmöglich ist; nur für diskrete Verteilungen kann $p(X = x) \neq 0$ sein. Es sei F wieder eine stetige Verteilungsfunktion, und es seien $x_1, x_2, x_3, ...$ verschiedene Werte von X, die die Bedingung

$$x_j - x_i > \varepsilon > 0$$

für $i \neq j$ erfüllen. Für jeden der Punkte x_i gilt jetzt $p(X = x_i) = 0$. Weiter sei $M = \{x_1, x_2, x_3, ...\}$; M kann als zufälliges Ereignis interpretiert werden, dass dann eingetreten ist, wenn X einen der Werte x_i angenommen hat. Dann ist nach dem Axiom K3 (s. Definition 2.2.15):

$$p(X \ \epsilon \ M) = p(X = x_1) + p(X = x_2) + p(X = x_3) + ...$$

und da die $p(X = x_i) = 0$ für alle i folgt, dass $p(X \ \epsilon \ M) = 0$, unabhängig davon, wieviele Elemente M enthält, und unabhängig davon, wie groß die Differenz $\varepsilon = x_i - x_j$ ist, so lange nur $\epsilon > 0$ ist. Man spricht in diesem Falle auch von singulären Punkten. Man wird bei stetigen zufälligen Veränderlichen also nicht alle möglichen Teilmengen von Ω als mögliche zufälligen Ereignisse betrachten, sondern sich auf σ-Algebren beschränken, die weniger als die Menge aller Teilmengen enthalten. Alle Teilmengen, die aus singulären Punkten bestehen, haben ja die Wahrscheinlichkeit Null. Die einzig "interessanten" Teilmengen aus $\mathbb{R}$ sind die offenen, halboffenen und geschlossenen Intervalle, denn nur für sie kann P_x Werte ungleich Null annehmen (dies sind die Borel-Mengen) Teilmengen aus singulären Punkten heißen dann auch Mengen mit (Wahrscheinlichkeits-) Maß Null. Ereignisse, die solchen Mengen entsprechen, sind dennoch nicht unmöglich!

3.2 Mehrdimensionale Verteilungen

Eine zufällige Veränderliche repräsentiert im allgemeinen nur ein Merkmal. Häufig ist man aber an der gleichzeitigen Betrachtung mehrerer Merkmale interessiert. Dieses Interesse wird z.B. in der Regressionsrechnung illustriert.

Dies führt zu der folgenden

Definition 3.2.1 *Es seien $X_1, \cdots, X_n$ zufällige Veränderliche. Dann heißt*

$$X = (X_1, \cdots, X_n)$$

auch mehrdimensionale (insbesondere: n-dimensionale) zufällige Veränderliche *oder auch* zufälliger Vektor*, und*

$$F(x_1, \cdots, x_n) = P(X_1 \leq x_1 \cap \cdots \cap X_n \leq x_n) \tag{3.12}$$

heißt die gemeinsame Verteilung *der* $X_1, \cdots, X_n$.

Statt $P(X_1 \leq x_1 \cap \cdots \cap X_n \leq x_n)$ wird auch einfach $P(X_1 \leq x_1, \cdots, X_n \leq x_n)$ geschrieben.

Die mehrdimensionale ZV X heißt *diskret*, wenn die X_i nur abzählbar viele Werte annehmen können. Dann gilt wieder

$$F(x_1, \cdots, x_n) = \sum^{x_1} \cdots \sum^{x_n} P(X_1 = x_1, \cdots, X_n = x_n) \tag{3.13}$$

wobei die gemeinsame Wahrscheinlichkeitsverteilung ist; die Summation erstreckt sich jeweils bis zum Wert x_i.

Die mehrdimensionale ZV heißt *stetig*, wenn es eine nichtnegative Dichtefunktion $f(x_1, \cdots, x_n)$ gibt derart, dass

$$F(x_1, \cdots, x_n) = \iint_{-\infty}^{\infty} f(x_1, \cdots, x_n)\, dx_1 \cdots dx_n \tag{3.14}$$

Setzt man in (3.13) und (3.14) die maximal möglichen Werte für $x_1, \cdots, x_n$ ein, so ergibt sich $F(x_1, \cdots, x_n) = 1$.

Erstreckt an den Integrationsbereich in (3.14) aller Variablen bis auf eine, x_i, auf den Gesamtbereich $(-\infty, \infty)$, so erhält man die *Randverteilung* F_i von x_i, d.h.

$$F_i(x_i) = \int_{-\infty}^{\infty} \int_{-\infty}^{x_i} \int_{-\infty}^{\infty} f(x_1, \cdots, x_n)\, dx_1 \cdots dx_n \tag{3.15}$$

Eine analoge Aussage gilt für (3.13).

Zwischen den $x_1, \cdots, X_n$ können beliebige Abhängigkeiten bestehen. Der Fall, dass die X_i völlig unabhängig voneinander variieren, ist ein Spezialfall:

Definition 3.2.2 *Die zufälligen Veränderlichen* $X_1, \cdots, X_n$ *heißen* stochastisch unabhängig *, wenn*

$$F(x_1, \cdots, x_n) = F(x_1) \cdots F(x_n) \tag{3.16}$$

gilt.

Erstreckt man den Integrationsbereich in (3.16) aller Variablen bis auf eine, x_i, auf den Gesamtbereich $(-\infty, \infty)$, so erhält man die Randverteilung F_i von x_i, d.h.

$$F_i(x_i) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{x_i} \cdots \int_{-\infty}^{\infty} f(x_1, \cdots, x_n) dx_1 \cdots dx_n \tag{3.17}$$

Eine analoge Aussage gilt für (3.13).

Definition 3.2.3 *Es sei* $X = (X_1, \cdots, X_n)'$ *ein Vektor. Sind die* X_i, $1 \leq i \leq n$, *zufällige Veränderliche, so heißt* X zufälliger Vektor.

3.3 Bedingte Verteilungen und Hazardfunktionen

Es seien A und B zwei zufällige Ereignisse. Man erinnere sich daran, dass die bedingte Wahrscheinlichkeit $P(A|B)$ definiert ist durch

$$P(A|B) = \frac{p(A \cap B)}{p(B)}. \tag{3.18}$$

Bedingte Wahrscheinlichkeiten treten auch in Zusammenhang mit Fragen nach der Verteilungsfunktion zufälliger Veränderlicher auf. Man hat die folgende

Definition 3.3.1 *Es sei X eine zufällige Veränderliche mit der Verteilungsfunktion F_x und A ein zufälliges Ereignis. Die* bedingte Verteilung *von X ist die bedingte Wahrscheinlichkeit*

$$F_x(x|A) = \frac{P(X \leq x|A)}{P(A)}, \tag{3.19}$$

wobei $\{X \leq x|A\} = \{\omega|\omega \in A \subseteq \Omega\} \cap \{\omega|\omega \in \Omega,\ X(\omega) \leq x\}$. Die Funktion

$$f(x|A) = \frac{dF(x|A)}{dx} \tag{3.20}$$

heißt bedingte Dichte *bzw.* bedingte Wahrscheinlichkeitsfunktion.

Es folgt sofort wieder

$$F(\infty|A) = 1, \quad F(-\infty|A) = 0. \tag{3.21}$$

Beispiel 3.3.1 Es sei Ω die Menge der Studenten einer Universität, und es sei A das Ereignis, dass ein(e) zufällig ausgewählte(r) Student(in) $\alpha \in \Omega$ Psychologie im Hauptfach studiert. Weiter sei X der Intelligenzquotient eines Studenten. Werden Personen aus der Population der Studenten zufällig gewählt, so ist X eine zufällige Veränderliche. In bezug auf die *Gesamtpopulation* gelte $E(X) = 100$, $Var(X) = 225$. Gesucht ist die Wahrscheinlichkeit $P(X \leq 110|A)$.

Aus den Rektoratsunterlagen weiß man, dass der Anteil der Studierenden mit dem Hauptfach Psychologie gleich .022 ist. Aus den ausgedehnten Untersuchungen des Fachbereichs Psychologie weiß man, dass $P(X \leq 110 \cap A) = .00556$ ist. Aus 3.19) folgt dann sofort $P(X \leq 110|A) = .00556/.022 \approx .25$. □

Es sollen jetzt noch zufällige Ereignisse betrachtet werden, die in bezug auf die zufällige Veränderliche X definiert sind.

Satz 3.3.1 *Es sei X eine zufällige Veränderliche mit der Verteilungsfunktion F_x, und es seien $A = \{X \leq a\}$, $\bar{A} = \{X > a\}$ komplementäre zufällige Ereignisse. Dann gilt*

$$F_x(x|A) = \begin{cases} F_x(x)/F_x(a), & x \leq a \\ 1, & x > a \end{cases} \tag{3.22}$$

und

$$F_x(x|\bar{A}) = \begin{cases} (F_x(x) - F_x(a))/(1 - F(a)), & x > a \\ \\ 0, & x \leq a \end{cases} \tag{3.23}$$

Beweis: Der Beweis ergibt sich sofort aus (3.19): für (3.22) hat man $P(A) = F_x(a)$ und $P(X \leq x \cap X \leq a) = P(X \leq x)$ für $x \leq a$, $P(X \leq x \cap X \leq a) = F_x(a)$ für $x > a$. Für 3.23) hat man $P(\bar{A}) = 1 - F_x(a)$ und $P(X \leq x \cap X > a) = F_x(x) - F_x(a)$ für $x \geq a$, und $P(X \leq x \cap X > a) = 0$ für $X \leq a$. □

Aus (3.22) ergibt sich sofort die Dichtefunktion

$$f_x(x|A) = \begin{cases} \frac{dF_x(x)}{dx}/F_x(a) = f_x(x)/\int_{-\infty}^{\infty} f_x(x)dx, & x < a \\ 0, & x \geq a \end{cases} \tag{3.24}$$

Aus (3.23) findet man sofort

$$f_x(x|A) = \frac{f_x(x)}{1 - F_x(a)} = \frac{f_x(x)}{\int_a^{\infty} f_x(x)dx}, \quad x \geq a \tag{3.25}$$

Die Dichtefunktion $f_x(x|A) = f_x(x|X \geq a)$ ist von besonderem Interesse:

(i) In Studien zur Reliabilität von Materialien etwa sei X diejenige Belastung eines Materials, bei dem das Material bricht: dann ist $f_x(x|X \geq a)dx$ die Wahrscheinlichkeit, dass das Material bei einer Belastung zwischen a und $a + dx$ bricht unter der Bedingung, dass es bis zur Belastung a nicht gebrochen ist.

(ii) In der Psychophysik untersucht man u.a. die Entdeckbarkeit eines bestimmten Reizes. Es sei X diejenige Intensität des Reizes, bei der der Reiz mit einer bestimmten Wahrscheinlichkeit entdeckt wird; $f_x(x|X \geq a)dx$ ist die Wahrscheinlichkeit, dass er mit einer Intensität zwischen a und $a + dx$ entdeckt wird unter der Bedingung, dass er bei der Intensiät a noch nicht entdeckt worden ist.

(iii) Es sei t das Alter eines Menschen, in dem eine bestimmte Krankheit ausbricht; t wird hier als zufällige Veränderliche betrachtet. Dann ist $f_t(t|t \geq t_0)dt$ die Wahrscheinlichkeit des Ausbruchs der Krankheit in der Zeit zwischen t und $t + dt$ unter der Bedingung, dass die Krankheit bis zum Zeitpunkt t_0 noch nicht ausgebrochen ist.

(iv) Es sei t die Zeit, die eine Person bis zu einer bestimmten Reaktion benötigt (Auffinden eines Wortes, Lösen einer Aufgabe, Betätigung des Bremspedals). t ist sicherlich eine zufällige Veränderliche. Dann ist $f_t(t|t \geq t_0)dt$ wieder die Wahrscheinlichkeit der Reaktion zwischen t_0 und $t_0 + dt$ unter der Bedingung, dass die Reaktion nicht bis zum Zeitpunkt t_0 erfolgte.

Bayesscher Satz Der Satz von Bayes kann für Wahrscheinlichkeitsfunktionen bzw. - dichten angeschrieben werden. Sind sowohl die zufälligen Veränderlichen X und Y diskret, ht man sofort

$$P(X = x|Y = y) = \frac{P(Y = y|X = x)P(X = x)}{P(Y = y)}. \tag{3.26}$$

Sind sowohl X wie Y stetig, so folgt aus dem Begriff der bedingten Dichte

$$f_X(Y = y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}, \quad f_Y(y|X = x) = \frac{f_{X,Y}(x, y)}{f_X(x)}$$

$$f_X(x|Y = y) = \frac{f_Y(y|X = x)f_X(x)}{f_Y(y)} \tag{3.27}$$

$f_Y(y)$ kann wieder aus dem Satz der totalen Wahrscheinlichkeit bestimmt werden:

$$f_Y(y) = \int_{-\infty}^{\infty} f_Y(y|X = \xi)f_X(\xi)d\xi. \tag{3.28}$$

Natürlich können auch Kombinationen von diskreten und stetigen Verteilungen betrachtet werden; so erhält man für den Fall, dass X stetig und Y diskret ist,

$$P(X = x|Y = y) = \frac{F_Y(y|X = x)P(X = x)}{f_Y(y)}. \tag{3.29}$$

Der Fall, dass X diskret und Y stetig ist wird analog erklärt.

Definition 3.3.2 *Es sei X eine zufällige Veränderliche mit der Verteilungsfunktion F_x. Dann heißt die Funktion*

$$\lambda(x) = f_x(x|X \geq x) = \frac{f_x(x)}{1 - F_x(x)} \geq 0 \tag{3.30}$$

Hazardfunktion *(engl.:* failure rate *oder* hazard rate function*).*

Hazardfunktionen haben eine große praktische Bedeutung, da aus der Form der Hazardfunktion Aussagen über die Zufälligkeit von Ereignissen abgeleitet werden können. Denn die Hazardfunktion läßt sich über den Begriff der bedingten Wahrscheinlichkeit herleiten: man kann etwa nach der Wahrscheinlichkeit fragen, mit der die zufällige Veränderliche X einen Wert zwischen x und $x + \Delta x$ annimmt unter der Bedingung, dass $X > x$ gilt. Es ist

$$\begin{aligned} P(X \in (x, x + \Delta x)|X > x) &= \frac{P(X \in (x, x + \Delta x) \cap X > x)}{P(X > x)} \\ &= \frac{P(X \in (x, x + \Delta x))}{1 - F(x)} \end{aligned} \tag{3.31}$$

Läßt man hier Δx beliebig klein werden,so erhält man hieraus wegen

$$\lim_{\Delta x \to 0} P(X \in (x, x + \Delta x)) = f(x)dx$$

gerade den Ausdruck (3.30). Bevor ein Beispiel gegeben wird, soll noch die Verteilungsfunktion F über λ ausgedrückt werden:

Satz 3.3.2 *Es sei $X \geq 0$ eine zufällige Veränderliche mit der Hazardfunktion $\lambda(x)$. Dann ist*

$$F(x) = 1 - \exp\left(-\int_0^x \lambda(\xi)\, d\xi\right). \tag{3.32}$$

Beweis: Es ist $f(x) = dF(x)/dx = F'(x)$, d.h. $\lambda(x) = F'(x)/(1 - F(x))$. Dann ist

$$\int_0^\infty \lambda(\xi)\, d\xi = -\log(1 - F(x)),$$

wie man durch Differentiation der rechten Seite sofort bestätigt. Multipliziert man beide Seiten mit -1 und bildet $\exp(\log(1 - F(x)))$, so wird man auf (3.32) geführt. □

Beispiel 3.3.2 Es sei $\lambda(\xi) =$ *konstant* für alle ξ. Wegen der Beziehung 3.31 kann diese Konstanz so gedeutet werden, dass die bedingte Wahrscheinlichkeit $P(X \in (x + \Delta x)|X > x)$ für alle x konstant, d.h. aber unabhängig von x ist. *In diesem Sinne* ist also die Wahrscheinlichkeit für das Auftreten von X gleichverteilt über die möglichen x-Werte und insofern *rein zufällig*: die Wahrscheinlichkeit, einen Wert in einem Intervall $(x, x + \Delta x)$ zu beobachten, wird nicht größer oder kleiner, je größer oder kleiner der Wert von x ist! Aus (3.32) folgt überdies, dass die Verteilung von X dann durch

$$F(x) = 1 - e^{-\lambda x}$$

gegeben ist; dies ist die Exponentialverteilung. Wegen der Bedeutung von $\lambda =$ *konstant* spielt diese Verteilung in vielen Anwendungen eine zentrale Rolle. □

Es werde nun eine 2-dimensionale Dichte $f(x, y)$ mit der Verteilung $F(x, y)$ betrachet. Es sei etwa $A = \{X \leq x\}$; man kann dann die bedingte Verteilung

$$f(y|A) = \frac{P(X \leq x \cap Y \leq y)}{P(X \leq x}) \tag{3.33}$$

betrachten. Es werde $F_{xy}(x, y) = P(X \leq x \cap Y \leq y)$ gesetzt. Dann folgt die *bedingte Dichte*

$$f(y|X \leq x) = \frac{F_{xy}(x, y)}{F_x(x)} = \frac{\int_{-\infty}^x f_{xy}(\xi, y) d\xi}{\int_{-\infty}^\infty \int_{-\infty}^x f_{xy}(\xi, y)\, d\xi dy} \tag{3.34}$$

Es sei weiter $A = \{x_1 \leq X \leq x_2\}$. Dann ist

$$F_y(y|A) = \frac{P(x_1 \leq X \leq x_2 \cap Y \leq y)}{P(x_1 \leq X \leq x_2)} \tag{3.35}$$

und die entsprechende Dichte ist

$$f_y(y|A) = \frac{\int_{x_1}^{x_2} f_{xy}(x,y)dx}{F_x(x_2) - F_x(x_1)} \tag{3.36}$$

Schließlich ist noch die bedingte Verteilung $F(y|X = x)$ von Interesse. Diese läßt sich aus (3.35) herleiten, indem man $x_1 = x$ und $x_2 = x + \Delta x$ setzt und dann Δx gegen Null gehen läßt. Man erhält

$$F_y(y|X = x) = \lim_{\Delta x \to 0} F_y(y|A) = \frac{\int_{-\infty}^{y} f_{xy}(x,\eta)\, d\eta}{f_x(x)} \tag{3.37}$$

Differenziert man bezüglich y, so erhält man die bedingte Dichte für y:

$$f_y(y|x) = \frac{f_{xy}(x,y)}{\int_{-\infty}^{\infty} f_{xy}(x,y)\, dy} \tag{3.38}$$

3.4 Funktionen von zufälligen Veränderlichen

3.4.1 Funktionen einer zufälligen Veränderlichen

Die Verteilungsfunktion einer zufälligen Variablen X ordnet den Werten von X Wahrscheinlichkeiten zu. Nun können die Werte von X Meßwerte sein, und die sind häufig nur bis auf eine lineare Transformation eindeutig, und es zeigt sich, dass eine Skalentransformation auch eine Transformation des Wahrscheinlichkeitsmaßes notwendig macht. Statistische Fragestellungen führen aber ebenfalls dazu, das Wahrscheinlichkeitsmaß von Funktionen von zufälligen Veränderlichen zu betrachten. Die Durchführung solcher Transformationen soll kurz beschrieben werden.

Es sei X eine ZV und h sei eine Funktion, so dass $Y = h(X)$; Y ist wieder eine ZV. Die Wahrscheinlichkeit, dass Y einen Wert aus der Menge B annimmt, ist dann gleich der Wahrscheinlichkeit, dass X einen Wert aus einer bestimmten Menge A annimmt, so dass $P(Y \epsilon B) = P(X \epsilon A)$. Dieser Sachverhalt soll an einem einfachen Sachverhalt inllustriert werden:

Beispiel 3.4.1 Die ZV X nehme die Werte -1 und +1 mit den Wahrscheinlichkeiten $P(X = -1) = P(X = +1) = 1/2$ an. Andere Werte als -1 und +1 haben die Wahrscheinlichkeit Null. Nun sei $h(X) = Y = 2X$, d.h. $Y = -2$ oder $Y = +2$, andere Werte nimmt Y nicht an. Sicherlich ist jetzt $P(Y = -2) = P(X = -1) = 1/2$ und $P(Y = +2) = P(X = +1) = 1/2$. Für die Verteilungsfunktion F_x von X

erhält man $F_x(x) = 0$ für $x > -1$, $F_x(-1) = 1/2$, $F_x(+1) = P(X = -1 \cup X = +1) = 1$. Dann ist $F_y(y) = P(Y \leq y) = P(2X \leq y) = P(X \leq y/2) = F_x(y/2)$. Die Verteilungsfunktion F_y von Y kann also durch die Verteilungsfunktion F_x von X ausgedrückt werden. □

Beispiel 3.4.2 Es sei nun $Y = h(X) = -X$, und X habe die Verteilungsfunktion F_x. Gesucht die Verteilungsfunktion F_y von Y sowie die zugehörige Wahrscheinlichkeitsfunktion bzw. Wahrscheinlichkeitsdichte.

$$F_y(y) = P(Y \leq y) = P(-X > y) = P(X > -y) = 1 - P(X \leq -y).$$

a. X diskret:

Ist X diskret und ist y eine Sprungstelle, so ist

$$F_y(y) = 1 - F_x(-y) - P(X = -y)$$

(vergl. (4.6.2)). $P(Y = -X)$ ist die Wahrscheinlichkeitsfunktion.

b. X stetig:

Ist X stetig, so ist $F_y(y) = 1 - F_x(-y)$. Gesucht ist die Dichtefunktion für Y. Es sei f die Dichte für die ZV X; dann gilt $f(x) = dF_x(x)/dx$. Die Dichte von Y sei durch $g(Y)$ gegeben, und es muß $g(y) = dF_y(y)/dy$ gelten. Aber dann ist

$$g(y) = \frac{dF_y(y)}{dy} = \frac{d(1 - F_x(-y))}{dy} = -\frac{dF_x(-y)}{dy}.$$

Nun sei $\eta = -y$; dann ist der Kettenregel entsprechend $dF_x(\eta)/dy = (dF_x(\eta)/d\eta)(d\eta/dy) = f_x(\eta)$ü$-1 = -f_x(-y)$, und mithin ist die Dichte $g(y)$ durch

$$g(y) = -(-f_x(-y)) = f_x(-y)$$

gegeben. □

Beispiel 3.4.3 Es sei X eine stetige zufällige Veränderliche und es gelte $Y = aX + b$. F_x sei die Verteilungsfunktion und f_x sei die Dichtefunktion von X. Gesucht ist die Dichtefunktion f_y und die Verteilungsfunktion F_y. Wir müssen zwei Fälle, $a < 0$ und $a > 0$ betrachten: **(i)** $a > 0$: Es ist $h(X) = Y = aX + b$,

also ist $h^{-1}(h(X)) = X = h^{-1}(aX + b) = (Y - b)/a$. Dann ist

$$\begin{aligned} F_y(y) &= P(Y \leq y) = P(aX + b \leq y) = P\left(X \leq \frac{y-b}{a}\right) \\ &= P(X \leq h^{-1}(y)) = F_x(h^{-1}(y)), \end{aligned}$$

d.h.

$$F_y(y) = F_x(h^{-1}(y)) = F_x\left(\frac{y-b}{a}\right). \tag{3.39}$$

Daraus ergibt sich die Dichte f_y durch Differentiation:

$$f_y(y) = \frac{dF_y(y)}{dy} = \frac{dF_x(h^{-1}(y))}{dh^{-1}}\frac{dh^{-1}}{dy} = f_x\left(\frac{y-b}{a}\right)\frac{1}{a}. \tag{3.40}$$

Die Dichtefunktion f_y ist also gleich der Dichte f_x an der Stelle $(y-b)/a$, multipliziert mit $1/a$.

(ii) $a < 0$: Wir setzen $\alpha = -a$, $\alpha > 0$. Dann ist $Y = aX + b = -\alpha X + b$ und $h^{-1}(y) = X = -(y-b)/\alpha$

$$\begin{aligned} F_y(y) &= P(-\alpha X + b) = P\left(X \leq -\frac{y-b}{\alpha}\right) = \\ P\left(\frac{y-b}{\alpha} < X\right) &= 1 - P\left(X \leq \frac{y-b}{\alpha}\right) = \\ &= 1 - P(X \leq -h^{-1}(y)) \end{aligned}$$

so dass

$$F_y(y) = 1 - F_x\left(\frac{y-b}{\alpha}\right). \tag{3.41}$$

Die Dichte f_y folgt daraus wieder durch Differentiation:

$$f_y(y) = \frac{dF_y(y)}{dy} = -\frac{dF_x(h^{-1})}{dh^{-1}}\frac{dh^{-1}}{dy} = -f_x\left(-\frac{y-b}{\alpha}\right)\cdot -\frac{1}{\alpha}, \tag{3.42}$$

Ist X diskret, so erhält man

$$F_y(y) = 1 - F_x\left(\frac{y-b}{a}\right) - P\left(X = \frac{y-b}{a}\right), \tag{3.43}$$

wenn $(y-b)/a$ eine Sprungstelle von X ist. In allen anderen Punkten ist $P(X = (y-b)/a) = 0$.

□

Das Vorgehen in den beiden vorangegangenen Beispielen kann verallgemeinert werden. Es sei h eine streng monotone Funktion der zufälligen Veränderlichen X auf einem Intervall $[a, b]$, d.h. für $x_1 < x_2$ gelte entweder $h(x_1) < h(x_2) - h$ ist streng monoton steigend — oder $h(x_1) > h(x_2)$, dann ist h streng monoton fallend. Dann existiert für alle $X \in [a, b]$ die inverse Funktion h^{-1}. Wir suchen die Verteilung bzw. die Dichte für $Y = h(X)$. Dabei müssen die Fälle streng monoton steigender und streng monoton fallender Funktionen zunächst getrennt betrachtet werden.

(i) Streng monoton steigende Funktionen. Es sei $y = g(x)$. Wenn g streng monoton steigend auf [a, b] ist, so folgt aus $g(x) \leq y$ auch $x \leq g^{-1}(y)$ und es ist $dg/dx = g' > 0$ auf $[a, b]$. Dann hat man

$$P(Y \leq y) = P(h(X) \leq y) = P(X \leq h^{-1}(y)) = F_x(h^{-1}(Y)), \tag{3.44}$$

woraus. sich die Dichte

$$g(y) = \frac{dF_x(h^{-1}(y))}{dy} = f_x(h^{-1}(y))\frac{dh^{-1}(y)}{dy} \tag{3.45}$$

ergibt.

(ii) Streng monoton fallende Funktionen. Hier gilt, für alle x_1, $x_2 \in [a, b]$, $g(x_1) > g(x_2)$ für $x_1 < x_2$. Setzt man $x = x_1$ und $y = x_2$, so sieht man, dass aus $g(x) \leq y$ die Aussage $x > g^{-1}(y)$ folgt. Weiter muß $dg/dx < 0$ sein, und wegen $dg^{-1}/dy = 1/g'(x)$ folgt $dg^{-1}/dy < 0$. Demnach hat man nun

$$F_y(y) = P(g(x) \leq y)) = P(x > g^{-1}(y)) = 1 - P(X \leq g^{-1}(y)), \tag{3.46}$$

und man erhält die Dichte

$$f_y(y) = -f_x(g^{-1}(y))\frac{dg^{-1}}{dy}. \tag{3.47}$$

Da aber $dg^{-1}/dy < 0$, ist $f_y(y) \geq 0$, wie es sein soll. Der Ausdruck für die Dichte von $y = g(x)$ läßt sich dementsprechend gemäß

$$f_y(y) = f_x(g^{-1}(y))\left|\frac{dg^{-1}}{dy}\right| \tag{3.48}$$

zusammenfassen.

Natürlich hat man es im allgemeinen Funktionen mit Funktionen zu tun, die nicht überall auf dem interessierenden Intervall von X-Werten streng monoton sind. In diesem Fall muß man (3.48) auf die streng monotonen Teilbereiche anwenden.

Beispiel 3.4.4 Ein für spätere Anwendungen interessantes Beispiel für eine nicht überall streng monotone Funktion ist $Y = h(X) = X^2$. Dann ist $Y \geq 0$ unabhängig vom Vorzeichen von X. Es soll nun wieder die Verteilung F_y von Y bestimmt werden. Offenbar muß $F_y(y) = 0$ für $y < 0$ sein, und

$$F_y(y) = P(Y \leq y) = P(X^2 \leq y)$$

für $y > 0$. Es sei $y = \sqrt{x_0}$; dann gilt offenbar $X^2 \geq y$ für $-x_0 \leq X \leq x_0$, d.h. für $-\sqrt{y} \leq X \leq \sqrt{y}$. Damit hat man

$$F_y(y) = \begin{cases} 0, & y < 0 \\ P(Y < y) = P(X^2 < y) = P(-\sqrt{y} < X < \sqrt{y}), & y \geq 0 \end{cases}$$

d.h. aber

$$F_y(y) = F(\sqrt{y}) - F(-\sqrt{y}) \quad \text{für} \quad y > 0, \quad F_y(y) = 0 \quad \text{für} \quad y \leq 0$$

Daraus ergibt sich durch Differentiation die Dichte

$$f_y(y) = \frac{f_x(\sqrt{y}) + f_x(-\sqrt{y})}{2\sqrt{y}} \tag{3.49}$$

Ist X aber diskret, so ist

$$F_y(y) = 0 \text{ für } \; y \leq 0, F(\sqrt{y}) - F(-\sqrt{y}) - P(X = -\sqrt{y}), \, y > 0. \tag{3.50}$$

Für die Wahrscheinlichkeitsfunktion ergibt sich

$$P(Y = y_i) = P(X^2 = y_i) = P(X = -\sqrt{y_i}) + P(X = \sqrt{y_i}). \tag{3.51}$$

□

Es soll kurz die Rolle des Faktors $|dg^{-1}/dy|$ in (3.48) betrachtet werden, um die Diskussion der Transformation mehrdimensionaler Veränderlicher vorzubereiten. Die Verteilungsfunktion von Y ist durch

$$F_y(a) = \int_{-\infty}^{a} f_x(g^{-1}(y)) \left| \frac{dg^{-1}}{dy} \right| dy$$

gegeben, also durch das Integral der Dichte von X, wobei aber nicht einfach das Differential dy, sondern eben $|dg^{-1}/dy|dy$ statt des Differentials dx auftritt, d.h. $dx = |dg^{-1}/dy|dy$. Der Faktor $|dg^{-1}/dy|$ kann also als eine der Transformation g entsprechende Transformation der Länge des Differentials dx angesehen werden.

3.4.2 Eine Funktion zweier zufälliger Veränderlicher

Vielfach ist eine zufällige Z als Funktion $h(x, y)$ zweier anderer zufälliger Veränderlicher definiert, z.B. ist $Z = X + Y$. Gesucht ist die Verteilungsfunktion $G(z) = P(Z \leq z)$, mit der ist dann auch die Dichte f_z gegeben.

Es ist möglich, dass der gleiche Z-Wert für verschiedene Werte von X und Y resultieren. Das Ereignis $\{Z \leq z\}$ tritt also dann ein, wenn die Punkte (X, Y) aus einer bestimmten Menge B_z stammen. Die Menge B_z ist die Menge der Punkte $(X.Y)$, für die $h(X, Y) \leq z$ gilt. Also ist

$$\{Z \leq z\} = \{(X, Y)|(X, Y)\epsilon \; B_z\}.$$

Die (X, Y) sind Realisierungen der ZV X und Y und stellen Elementarereignisse aus $\Omega = \mathbb{R} \times \mathbb{R}$ dar. Für stetige X und Y müssen wir also über den Bereich B_z integrieren, um $P(Z \leq z)$ zu bestimmen:

$$G(z) = P(Z \leq z) = \iint_{B_z} f(x, y) \, dydz, \tag{3.52}$$

wobei $f(x, y)$ die gemeinsame Dichte der zufälligen Veränderlichen X und Y ist. Die Dichte $g(z)$ der ZV Z ist dann durch $g(z) = dG(z)/dz$ gegeben. Das Vorgehen wird für die folgenden Funktionen erläutert.

Die Funktion $Z = X + Y$

Wir wollen $G(z) = P(X + Y) \leq z)$ bestimmen. Es ist $X + Y \leq z$ genau dann, wenn $X \leq z - Y$ ist. Wir setzen $\zeta = z - Y$ und erhalten eine Definition der Menge B_z :

$$B_z = \{(X, Y)|X \leq \zeta = z - Y\}.$$

Daraus ergibt sich die Integration über B_z : für festes Y ist

$$P(X \leq z - Y) = \int_{-\infty}^{z-Y} f(x, y)\mathrm{dxdy} \ = P(X + Y \leq z). \tag{3.53}$$

Dies gilt für jeden Y-Wert. Wir müssen also noch über Y integrieren, um $G(z) f$ür alle Y zu bestimmen, die der Bedingung $X + Y \leq z$ genügen. Damit ist

$$G(z) = \int_{-\infty}^{\infty} \int_{-\infty}^{z-Y} f(x, y)\mathrm{dxdy}, \tag{3.54}$$

und für die Dichte f_z folgt durch Differentiation nach z

$$g(z) = \int_{-\infty}^{\infty} f(z - y,\, y)dy. \tag{3.55}$$

Bisher sind keine Enschränkungen bezüglich der Dichte $f(x, y)$ gemacht worden, d.h. insbesondere, dass die Variablen X und Y in beliebiger Weise voneinander abhängen dürfen. Es werde nun insbesondere angenommen, dass X und Y stochastisch unabhängig seien. Dann ist $f(x, y) = f_x(x) f_y(y)$ und aus (3.55) folgt sofort

$$g(z) = \int_{-\infty}^{\infty} f_x(z - y) f_y(y)dy. \tag{3.56}$$

Das Integral (3.56) heißt auch *Faltungsintegral*, und man sagt $g(z)$ sei durch *Faltung* der Dichten f_x und f_y gegeben.

Beispiel 3.4.5 Es seien X und Y zwei exponentialverteilte Variablen mit den Dichten $f_x(x) = \lambda \exp(-\lambda x)$, $x \geq 0$, $\lambda > 0$, $f_x(x) = 0$ für $x < 0$, $f_y(y) = \mu \exp(-\mu y)$, $y \geq 0$, $\mu > 0$, $f_y(y) = 0$ für $y < 0$. Gesucht ist die Verteilung und Dichte von $z = x + y$. Aus (3.56) folgt dann für die Dichte

$$g(z) = \lambda\mu e^{-\lambda z} \int_0^z e^{(\lambda-\mu)y} dy = \begin{cases} \lambda\mu\,(e^{-\lambda z} - e^{-\mu z})/(\mu - \lambda), & \lambda \neq \mu \\ \lambda^2 z e^{-\lambda z}, & \lambda = \mu \end{cases} \tag{3.57}$$

Diese Verteilung von Z heißt auch *Gamma-Verteilung.* □

Die Funktion $Z = X/Y$

Gegeben seien die zufälligen Veränderlichen mit der gemeinsamen Dichte $f(x, y)$, gesucht ist die Dichte der zufälligen Veränderlichen $Z = X/Y$.

Das zufällige Ereignis $\{Z \leq z\}$ tritt ein genau dann, wenn $\{X/Y \leq z\}$ und damit $\{X \leq zY\}$ eintritt. Um die Menge B_z zu bestimmen, sind zwei Fälle zu unterscheiden: (i) $Y > 0$: für feste Werte von $Y > 0$ und z hat man $B_z^+ = \{(X, Y)|X \leq zY\}$; (ii) $Y < 0$, $z > 0$: Dann existiert ein $x = -x_0(Y)$ derart, dass $-x_0(Y) = zY$, d.h. $-x_0/Y = z$, und für alle $X < -x_0(Y)$ ist $X/Y > z$, also ist $B\bar{z} = \{(X, Y)|Y < 0,\ X > -x_0(Y)\} = \{(X, Y)|X > zY, Y < 0\}$. Der gesuchte Bereich B_z ist durch die Vereinigung $B_z^+ \cup B_z^-$ gegeben. Da $B_z^- \cap B_z^+ = \emptyset$ folgt $P(Z \leq z) = P(B_z^-) + P(B_z^+)$, d.h.

$$G(z) = P(Z \leq z) = \int_0^\infty \int_{-\infty}^{Yz} f(x, y)\, dxdy + \int_{-\infty}^0 \int_{Yz}^\infty f(x, y)\, dxdy, \tag{3.58}$$

und für die Dichte folgt

$$g(z) = \frac{dG(z)}{dz} = \int_0^\infty y\, f(x, y)dy - \int_{-\infty}^0 y\, f(x, y)dy. \tag{3.59}$$

Da im zweiten Integral $y < 0$, kann man die beiden Integrale zusammenfassen und erhält

$$g(z) = \int_{-\infty}^\infty |y| f(zy, y)dy. \tag{3.60}$$

Beispiel 3.4.6 Es seien X und Y zwei stochastisch unabhängige, exponentialverteilte zufällige Veränderliche mit den Dichten $f_x(x) = \lambda exp(-\lambda x)$ für $x \geq 0$, $\lambda > 0$, und $f_y(y) = \mu \exp(-\mu y)$, $y \geq 0$, $\mu > 0$. Gesucht ist die Dichte des Quotienten $Z = X/Y$. Für die gemeinsame Dichte gilt wegen der stochastischen Unabhängigkeit $f(x, y) = d_x(x) f_y(y)$. Nach 3.60) hat man dann

$$g(z) = \lambda\mu \int_0^\infty y\, \exp(-\lambda zy - \mu y)dy = \frac{\lambda\mu}{(\lambda z + \mu)^2}. \tag{3.61}$$

□

Die Funktion $Z = XY$

Es sei wieder die $f(x, y)$ die gemeinsame Dichte der zufälligen Veränderlichen X und Y. Gesucht ist die Dichte des Produkts $Z = X\,Y$. Um B_z zu bestimmen, betrachtet man wieder das Ereignis $\{Z \leq y\}$. Dieses Ereignis tritt genau dann ein, wenn $\{XY \leq z\}$, d.h. wenn $\{X \leq z/Y\}$ eintritt. Für festes $Y < 0$ hat man dann

$$P(Z \leq z/Y) = \int_{-\infty}^{z/Y} f(x, y)\, dxdy$$

und durch Integration über Y erhält man die Wahrscheinlichkeit von $Z \leq z/Y$ für alle $Y > 0$. Für $Y < 0$ gilt dann analog

$$P(Z \leq z/Y) = \int_{z/Y}^{\infty} f(x,y)\,dxdy$$

und durch Integration über alle $Y > 0$ erhält man $P(Z \leq z/Y)$ für alle $Y < 0$. Zusammenfassend hat man

$$G(z) = P(Z \leq z) = \int_{-\infty}^{0}\int_{z/Y}^{\infty} f(x,y)\,dxdy + \int_{0}^{\infty}\int_{-\infty}^{z/Y} f(x,y)\,dxdy. \tag{3.62}$$

Daraus folgt die Dichte durch Differentiation nach z :

$$g(z) = \int_{-\infty}^{\infty} f_y(z/y, y)\left|\frac{1}{y}\right| dy \tag{3.63}$$

b. Zwei Funktionen zweier zufälliger Veränderlicher

Es seien $U = g(X,Y)$, $V = h(X,Y)$ zwei Funktionen der zufälligen Veränderlichen X und Y, und gesucht ist die gemeinsame Verteilung F_{uv} von U und V sowie die zugehörige Dichtefunktion f_{uv}.

Es sei B_{uv} die Menge der Punkte (x,y), für die

$$\{g(X,Y) \leq u, \quad h(X,Y) \leq v\} = \{(X,Y) \in B_{uv}\}$$

gilt. Dann ist

$$F_{uv}(u,v) = \iint_{B_{uv}} f_{xy}(x,y)\,dx\,dy$$

Es sei weiter $\partial(g,h)/\partial(x,y)$ die Jacobi-Determinante

$$\frac{\partial(g,h)}{\partial(x,y)} = \begin{vmatrix} g_x & g_y \\ h_x & h_y \end{vmatrix}$$

wobei $\quad g_x = \partial g(x,y)/\partial x$, etc. Aus der Analysis ist nun bekannt, dass

$$f_{xy}(g(x,y), h(x,y)) = f_{uv}(u,v)\frac{\partial(g,h)}{\partial(x,y)}$$

(vergl. etwa Courant (1963), p. 224). Dann folgt

$$f_{uv}(u,v) = f_{xy}(g(x,y), \quad h(x,y)\frac{\partial(x,y)}{\partial(g,h)}, \quad \frac{\partial(x,y)}{\partial(g,h)} = \left[\frac{\partial(g,h)}{\partial(x,s)}\right]^{-1} \tag{3.64}$$

und natürlich

$$P(U,V \text{ in } S) = \iint_S f_{xy}(G(u,v),\, H(u,v))\frac{\partial(x,y)}{\partial(g,h)}\,dudv, \tag{3.65}$$

wobei G und H die zu g und h inversen Funktionen sind.

3.5 Erwartungswerte

3.5.1 Der eindimensionale Fall

In empirischen Untersuchungen wird häufig eine Stichprobe von Meßwerten $x_1, \cdots, x_n$ erhoben, um bestimmte Parameter der untersuchten Population abzuschätzen. So wird z.B. das arithmetische Mittel $\bar{x}$ und die Stichprobenvarianz s^2 berechnet, $\bar{x}$ kann dann etwa als Schätzung für den Parameter μ und s^2 kann als Schätzung der Varianz σ^2 der "Fehler" e_i im Modell $x_i = \mu + e_i$ interpretiert werden. Nun sind aber die Werte von $\bar{x}$ und s^2 aber stichprobenabhängig, d.h. sie variieren von einer Stichprobe zur nächsten. Die Frage, um welche Werte sie variieren, führt zum Begriff des Erwartungswertes.

Um die Definition des Begriffes zu motivieren, sei an die Definition des arithmetischen Mittels erinnert. Kommt der Meßwert x_i gerade n_i-mal unter den Meßwerten vor, so ist ja

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \sum_{i=1}^{k} \frac{n_i}{n} x_i = \sum_{i=1}^{k} \hat{p}_i x_i, \quad n_1 + \cdots + n_k = n$$

wobei $\hat{p}_i$ die relative Häufigkeit des Meßwertes x_i in der Stichprobe ist; es gibt also k Gruppen *verschiedener* Messwerte. Für die Varianz s^2 erhält man analog

$$\hat{s}^2 = \sum_{i=1}^{k} \hat{p}_i (x_i - \overline{x})^2$$

die höheren Stichprobenmomente lassen sich auf ähnliche Weise einführen. Ersetzt man nun die relativen Häufigkeiten $\hat{p}_i$ durch die "wahren" Wahrscheinlichkeiten p_i, so wird man Konstante erhalten, die für die Verteilung der X-Werte in der Population charakteristisch sind. Diese überlegung führt zu der folgenden

Definition 3.5.1 *Es sei X eine zufällige Variable; X kann diskret oder stetig sein. Dann heißt*

$$E(X) = \sum_i p_i x_i \tag{3.66}$$

falls X diskret ist, bzw.

$$E(X) = \int_{-\infty}^{\infty} x f(x)\, dx \tag{3.67}$$

falls X stetig ist und falls $\int_{-\infty}^{\infty} |x| f(x) dx < \infty$, der Erwartungswert *von X. Dabei ist p_i, $i \in \mathbb{N}$, $\mathbb{N}$ die Menge der natürlichen Zahlen, die Wahrscheinlichkeitsfunktion von X im diskreten Fall, $\sum_i$ ist die Summe über alle $p_i x_i$, und $f(x)$ ist die Wahrscheinlichkeitsdichte im stetigen Fall.*

Im stetigen Fall existiert $E(X)$ also nur, wenn $\int |x| f(x) dx < \infty$, andernfalls hat X keinen Erwartungswert.

Kommentar: Den Übergang von der Summe (3.66) zum Integral (3.67) macht man sich wei folgt klar: Bei einer stetigen zufälligen Veränderlichen X ist die Wahrscheinlichkeit, dass X einen Wert zwischen a und $a+\Delta x$ annimmt, durch das Integral (= der Fläche unter der Dichtefunktion zwischen a und $a+\Delta x$)

$$F(a+\Delta x) - F(a) = \int_a^b f(x)dx$$

gegeben. Teilt man nun den Wertebereich von X in k Intervalle der Breite Δx auf und bezeichnet man mit x_j den Wert in der Mitte des j-ten Intervalles, und bezeichnet man mit p_j den Wert von $\int_{a_j}^{b_j} xf(x)dx$, wobei a_j der untere Wert des j-ten Intervalles und b_j der obere Wert dieses Intervalles ist, so erhält man

$$\sum_{j=1}^{k} p_j x_j = \sum_{j=1}^{k} x_j \int_{a_j}^{b_j} f(x)dx = \sum_{j=1}^{k} \int_{a_j}^{b_j} x_j f(x)dx \to \int_{-\infty}^{\infty} xf(x)dx = E(X),$$

Man kann unter Verwendung des Begriffs des Stieltjes-Integrales (vergl. Definition 3.1.5) auch

$$E(X) = \int_{-\infty}^{\infty} xdF(x) \tag{3.68}$$

schreiben.

Der Erwartungswert unterscheidet sich also vom Stichprobenmittelwert dadurch, dass über alle möglichen Realisationen der zufälligen Veränderlichen X gemittelt wird. Die explizite Berechnung des Erwartungsertes setzt deshalb i. a. die Kenntnis der Wahrscheinlichkeitsverteilung von X voraus. Die Redeweise vom Erwartungswert einer zufälligen Veränderlichen ist insofern irreführend, als die explizite Berechnung eines Erwartungswertes eben an die Annahme einer bestimmten Wahrscheinlichkeitsverteilung bzw. einer Dichte gebunden ist. Die Erwartungswerte für bestimmte Wahrscheinlichkeitsverteilungen werden im folgenden Kapitel bestimmt. Hier sollen einige allgemeine Aussagen über Erwartungswerte und damit zusammenhängende Größen hergeleitet werden; die Annahme spezieller Wahrscheinlichkeitsverteilungen bzw. -dichten wird dabei nicht notwendig sein. Die Aussagen erweisen sich aber als nützlich bei der Bestimmung von Erwartungswerten für bestimmte Verteilungen.

Satz 3.5.1 *Für die zufälligen Veränderlichen X, $X_1, \cdots, X_n$ existiere jeweils der Erwartungswert. Dann gelten die folgenden Aussagen:*

$$E(X) = b, \quad \textit{für } X = b \textit{ eine Konstante}; \tag{3.69}$$

$$E(X_1 + \cdots + X_n) = E(X_1) + \cdots + E(X_n) \tag{3.70}$$

$$E(aX + b) = aE(X) + b \tag{3.71}$$

$$E(X_1 X_2) = E(X_1)E(X_2) \Leftrightarrow X_1, X_2 \textit{stochastisch unabhängig} \tag{3.72}$$

Beweis: Es genügt, den Beweis für diskrete Variable durchzuführen, für stetige ist er analog.

(3.69) Für $X = b$ ist $P(X = b) = 1$, $P(X \neq b) = 0$. Dann ist $E(X) = 1 \cdot b = b$.

(3.70) Es sei $n = 2$. Dann ist $E(X) = \sum_i \sum_j (x_i + x_j) P(x_i \cap x_j) = \sum_i \sum_j x_i P(x_i \cap x_j) \sum_i \sum_j x_j P(x_i \cap x_j) = \sum_i x_i P(x_i) + \sum_j x_j P(x_j) = E(X_i) + E(X_j)$. Es sei $Z = X_i + X_j$; man kann nun die Summe $Z + X_k$ analog betrachten, etc.

(3.71) $E(aX+b) = \sum_i (ax_i+b)p(x_i) = \sum_i ax_i p(x_i) + \sum_i bP(x_i) = a \sum_i x_i P(x_i) + b \sum_i P(x_i) = aE(X) + b$, da $\sum_i P(x_i) = 1$.

(3.72) $E(X_1 X_2) = \sum_i \sum_j x_{1i} x_{2j} P(x_{1i} \cap x_{2j})$ und $P(x_{1i} \cap x_{2j}) = P(x_{1i}) P(x_{2j})$ genau dann, wenn X_1 und X_2 stochastisch unabhängig; dann ist aber

$$\sum_i \sum_j x_{1i} x_{2j} P(x_{1i} \cap x_{2j}) = \sum_i x_i P(x_{1i}) \sum_j x_{2j} P(x_{2j}).$$

Für stetige Variable ergibt sich der Beweis, indem man von den Wahrscheinlichkeitsfunktionen zu den Dichten übergeht und statt der Summen- Integralzeichen schreibt. □

Beispiel 3.5.1 Es sei X eine zufällige Veränderliche und es gelte $X = \mu + e$, wobei μ eine Konstante ist und e der Meßfehler bei den Realisierungen von X ist. Dann gilt $E(X) = E(\mu + e) = E(\mu) + E(e) = \mu + E(e)$. Der Erwartungswert von X hängt dann also nicht nur von μ sondern auch von $E(e)$ ab, und $E(e)$ wiederum hängt ab von den Bedingungen, unter denen die Realisierungen von X gewonnen werden. Es gilt nicht notwendig $E(e) = 0$. □

Definition 3.5.2 *Es seien X und Y zwei zufällige Veränderliche, wobei $Y = h(X)$ eine Funktion von X ist. Dann heißt*

$$E(Y) = \sum_i h(x_i) P(X = x_i) \tag{3.73}$$

bzw.

$$E(Y) = \int_{-\infty}^{\infty} h(x) f(x) dx \tag{3.74}$$

der Erwartungswert der Funktion $h(x)$. *Ist insbesondere $Y = h(X) = X^k$, $0 < k \in \mathbb{N}$, k eine Konstante, so heißt $E(Y) = E(X^k)$ k-tes* Moment *der zufälligen Veränderlichen X.*

Die Definition des k-ten Momentes enthält die des Erwartungswertes, denn für $k = 1$ wird $E(Y) = E(X)$. Der Erwartungswert heißt deshalb auch *erstes Moment.*

Definition 3.5.3 *Es seien X und Y zwei zufällige Veränderliche mit den Erwartungswerten $E(X)$ und $E(Y)$. Dann heißt*

$$Kov(X, Y) = E[(X - E(X))(Y - E(Y))] \tag{3.75}$$

die Kovarianz *von X und Y. Für $X = Y$ heißt*

$$Kov(X, X) = Var(X) = \sigma^2 = E[(X - E(X))^2] \tag{3.76}$$

die Varianz *der zufälligen Veränderlichen X; σ ist dann die Streuung von X. Sind σ_x und σ_y die Streuungen von X bzw. von Y, so heißt*

$$\rho = \frac{Kov(X, Y)}{\sigma_x \sigma_y} \tag{3.77}$$

die (Produkt-Moment-) Korrelation *der zufälligen Veränderlichen X und Y.*

Offenbar ist $h(X) = (X - E(X))^2$ eine Funktion von X, und $Var(X) = E(h(X))$. Durch Ausmultiplizieren und Vereinfachen folgt sofort

$$\begin{aligned} Kov(X, Y) &= E(XY) - E(X)E(Y) && (3.78) \\ Var(X) &= E(X^2) - E^2(X). && (3.79) \end{aligned}$$

Aus (3.76) erhält man sofort die

Folgerung: *Die zufälligen Veränderlichen X und Y seien stochastisch unabhängig. Dann gilt $Kov(X, Y) = 0$.*

Denn nach Satz 3.5.1 gilt bei stochastischer Unabhängigkeit von X und Y $E(XY) = E(X)E(Y)$; eingesetzt in (3.76) erhält man die Folgerung. Die Umkehrung, dass aus $\mathrm{Kov}(X, Y) = 0$ auch die stochastische Unabhängigkeit folgt, gilt nicht.

Satz 3.5.2 *Es seien X und Y zwei zufällige Veränderliche mit den Erwartungswerten $E(X)$ und $E(Y)$. Dann gilt*

$$\begin{aligned} Kov(aX + b, cY + d) &= ac\, Kov(X, Y) && (3.80) \\ Var(aX + b) &= a^2 Var(X) && (3.81) \\ Var(X_1 + \cdots + X_n) &= Var(X_1) + \cdots + Var(X_n) \\ &\quad + 2 \sum_{i<j} Kov(X_i, X_j). && (3.82) \end{aligned}$$

Beweis: Zu (3.80) Aus der Definition der Kovarianz folgt

$$\begin{aligned} Kov(aX + b, cY + d) &= E[(aX + b - aE(X) - b)(cY + d - cEY - d)] \\ &= E[(aX - aE(X))(cY - cE(Y))] = acKov(X, Y) \end{aligned}$$

gemäß Satz 3.5.1.

Zu (3.81): Die Gleichung folgt aus (3.80), wenn man $X = Y$ und $a = c$ setzt.

Zu (3.82): Man setze $X = X_1 + \cdots + X_n$. Dann ist

$$E(X^2) = \sum_{i=1}^{n} E(X_i^2) + 2\sum_{i<j} E(X_i X_j)$$

und

$$(E(X))^2 = E^2(X) = \sum_{i=1}^{n} E^2(X) + 2\sum_{i<j} E(X_i)E(X_j).$$

Nun ist $Var(X) = E(X^2) - E^2(X)$, und durch Rearrangieren der rechten Seite erhält man (3.82). □

Beispiel 3.5.2 Es sei X eine Bernoulli-Variable, d.h. X nehme den Wert $X = 1$ an, wenn das zufällige Ereignis $A =$ "Erfolg") eingetreten ist, und den Wert $X = 0$, wenn $\bar{A}$ (kein "Erfolg") eingetreten ist. Es sei $P(X = 1) = p$, $P(X = 0) = q = 1 - p$. Nach (3.66) ist dann der Erwartungswert von X durch

$$E(X) = p \cdot 1 + q \cdot 0 = p \tag{3.83}$$

gegeben. Um die Varianz von X zu finden, muß man zunächst $E(X^2)$ bestimmen. Es ist $E(X^2) = p \cdot 1^2 + q \cdot 0^2 = p$. Somit ist

$$Var(X) = E(X^2) - E^2(X) = p - p^2 = p(1-p) = pq. \tag{3.84}$$

Weiter seien $X_1, \cdots, X_n$ stochastisch unabhängige Bernoulli-Variable. Gesucht ist der Erwartungswert und die Varianz von $X = X_1 + \cdots + X_n$. Es folgt sofort

$$E(X) = E(X_1 + \cdots + X_n) = \sum_i E(X_i) = np, \tag{3.85}$$

und

$$Var(X) = \sum_{i=1}^{n} Var(X_i) = np(1-p) = npq, \tag{3.86}$$

da wegen der stochastischen Unabhängigkeit der X_i die Kovarianzen zwischen ihnen verschwinden.

Die relative Häufigkeit des Ereignisses $X = 1$ ist durch

$$\hat{p} = \sum_{i=1}^{n} \frac{x_i}{n}$$

gegeben. Wir können noch den Erwartungswert und die Varianz von $\hat{p}$ berechnen. Es ist

$$E(\hat{p}) = E\left(\frac{1}{n}X\right) = \frac{1}{n}E(X) = p, \tag{3.87}$$

denn $E(X) = np$ nach (3.85), und

$$Var(\hat{p}) = Var\left(\frac{1}{n}X\right) = \frac{1}{n^2}Var(X) = \frac{p(1-p)}{n}, \tag{3.88}$$

da ja $Var(X) = np(1-p)$ nach (3.86). Faßt man also $\hat{p}$ als Schätzung für p auf, so sieht man, dass man im Mittel gerade den Wert p erhält; für eine gegebene Stichprobe wird $\hat{p}$ zwar von p abweichen, aber nicht systematisch. Man sagt, die Schätzung $\hat{p}$ für p sei *erwartungstreu*. $Var(\hat{p}) = p(1-p)/n$ gibt die Varianz der Schätzungen $\hat{p}$ um p an. Für $n \to \infty$ geht sie offenbar gegen Null, d.h. je größer der Wert von n, desto näher liegt im allgemeinen $\hat{p}$ an p (dies bedeutet nicht, dass für eine vorliegende Stichprobe $\hat{p}$ nicht auch einmal erheblich von p abweichen könnte!)

Im übrigen wird hier sehr deutlich, dass der Erwartungswert und damit zusammenhängende Größen wie die Varianz von der Wahrscheinlichkeitsverteilung der X-Werte abhängen, $E(X)$ ist hier ja gerade durch den Wert p definiert. □

Die Aussagen im vorangegangenen Beispiel sind Spezialfälle einer allgemeineren Aussage:

Satz 3.5.3 *Es seien $x_1, \ldots, x_n$ stochastisch unabhängige Realisierungen einer zufälligen Veränderlichen X mit dem Erwartungswert $\mu = E(X)$ und der Varianz $\sigma^2 = Var(X)$. Dann gilt*

$$E(\bar{x}) = E(X) = \mu \tag{3.89}$$

$$Var(\bar{x}) = \frac{1}{n}Var(X) = \frac{1}{n}\sigma^2 \tag{3.90}$$

Beweis: Es ist

$$E(\bar{x}) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \frac{1}{n}\sum_{i=1}^{n} E(X) = \frac{1}{n}nE(X) = E(X),$$

denn die x_i kommen nach Voraussetzung aus der gleichen Population (sind Realisierungen derselben ZV) und haben deshalb alle den gleichen Erwartungswert $\mu = E(X)$, und $1/n$ spielt die Rolle des Faktors a in Satz 3.5.1, (3.71). Weiter ist

$$\mathrm{Var}(\bar{x}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(x_i) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n},$$

nach Satz 3.5.1 (3.70) mit $a = 1/n$, und (3.71) unter der Berücksichtigung der Tatsache, dass wegen der geforderten stochastischen Unabhängigkeit $Kov(x_i, x_j) = 0$ für alle i, j mit $i \neq j$, und da $Var(x_i) = \sigma^2$ für alle i. □

Die Abhängigkeit von $E(\bar{x})$ und $Var(\bar{x})$ von der speziellen Wahrscheinlichkeitsverteilung wird hier nicht explizit deutlich, sie steckt aber in der Definition des Erwartungswertes und der Varianz. Man sieht, dass allgemein $\bar{x}$ ein erwartungstreuer Schätzer für $\mu = E(X)$ ist, und dass die Varianz der zufälligen Veränderlichen $\bar{x}$($\bar{x}$ hängt ja von der Stichprobe ab) mit wachsendem n gegen Null geht, $\bar{x}$ bei großem n also eher näher an $E(X)$ liegt als bei kleinem n.

In Satz 3.5.3 werden Aussagen über den Erwartungswert und die Varianz des arithmetischen Mittels $\bar{x}$ gemacht. Es liegt nahe, analoge Aussagen über den Erwartungswert der Stichprobenvarianz s^2 herzuleiten:

Satz 3.5.4 *Es sei $(x_1, \cdots, x_n)$ eine Stichprobe von n Realisierungen einer zufälligen Veränderlichen X mit dem Erwartungswert $\mu = E(X)$ und der Varianz $Var(X) = \sigma^2$. Es sei $\bar{x}$ das arithmetische Mittel und $s^2 = \sum_i (x_i - \bar{x})^2/n$ die Varianz der Stichprobe. Dann gilt*

$$E(s^2) = \frac{n-1}{n}\sigma^2. \tag{3.91}$$

Beweis: Es ist

$$\begin{aligned} \sigma^2 &= E[(x-\mu)^2] = E[x^2] - \mu^2 && (3.92)\\ \sigma_{\bar{x}}^2 &= E[(\bar{x}-\mu)^2] = E[\bar{x}^2] - \mu^2, && (3.93) \end{aligned}$$

so dass

$$\begin{aligned} E[x^2] &= \sigma^2 + \mu^2 && (3.94)\\ E[\bar{x}^2] &= \sigma_{\bar{x}}^2 + \mu^2 && (3.95) \end{aligned}$$

Es ist

$$E(s^2) = E\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right) = E[x^2] - E[\bar{x}^2],$$

d.h. aber

$$E(s^2) = \sigma^2 + \mu^2 - \sigma_{\bar{x}}^2 - \mu^2 = \sigma^2\left(1 - \frac{1}{n}\right) = \frac{n-1}{n}\sigma^2,$$

denn nach (3.90) ist $\sigma_{\bar{x}}^2 = \sigma^2/n$. □

Biasfreie Schätzung von σ^2 Dieses Ergebnis bedeutet, dass s^2 keine erwartungstreue Schätzung für σ^2 ist, im Durchschnitt weicht s^2 von σ^2 um den Faktor $(n-1)/n$ ab. σ^2 wird also im Durchschnitt systematisch unterschätzt. Multipliziert man dementsprechend s^2 mit dem Faktor $n/(n-1)$, so bekommt man im Durchschnitt eine Schätzung, die gerade gleich σ^2 ist; es ist

$$\hat{s}^2 = \frac{n}{n-1}s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{3.96}$$

also eine Schätzung für σ^2, die frei von einem systematischen Fehler ist. Man sagt auch, $\hat{s}^2$ sei eine bias-freie (erwartungstreue) Schätzung für σ^2. Die Gleichung (3.96) liefert die Begründung, weshalb man bei der Berechnung der Stichprobenvarianz durch $n-1$ statt durch n teilt.

Es sei Z eine zufällige Veränderliche, mit dem Erwartungswert $E(Z) = 0$ und der Varianz $Var(Z) = 1$; Z heißt dann Standardvariable oder standardisierte Variable. Bekanntlich kann man eine zufällige Veränderliche X mit dem Erwartungswert $\mu = E(X)$ und der Varianz $\sigma^2 = Var(X)$ standardisieren, indem man von X zu $(x-\mu)/\sigma$ übergeht; tatsächlich findet man sofort, dass $E[(x-\mu)/\sigma) = E(X)/\sigma - \mu/\sigma = \mu/\sigma - \mu/\sigma = 0$ und $Var[(x-\mu)/\sigma] = Var(X)/\sigma^2 = \sigma^2/\sigma^2 = 1$. Man kann nun irgendwelche zufällige Veränderliche als durch stochastisch unabhängige Standardvariablen zusammengesetzt auffassen. Um dies zu sehen, werde die Aussage des folgenden Satzes bewiesen:

Satz 3.5.5 *Es seien Z_1 und Z_2 zwei stochastisch unabhängige Standardvariable, d.h. es sei $E(Z_1) = E(Z_2) = 0$ und $Var(Z_1) = Var(Z_2) = 1$. Weiter seien a_1, a_2, a_3, b_1, b_2 b_3 reelle Zahlen. Die zufälligen Veränderlichen X und Y seien durch*

$$X = a_1 Z_1 + a_2 Z_2 + a_3, \quad Y = b_1 Z_1 + b_2 Z_2 + b_3 \tag{3.97}$$

definiert. Dann gilt

$$\begin{aligned} E(X) &= a_3, \quad E(Y) = b_3 && (3.98)\\ Var(X) &= a_1^2 + a_2^2, \quad Var(Y) = b_1^2 + b_2^2 && (3.99)\\ Kov(X,Y) &= a_1 b_1 + a_2 b_2 && (3.100) \end{aligned}$$

Beweis: (3.98) folgt sofort aus der Tatsache, dass $E(Z_1) = E(Z_2) = 0$. (3.99) folgt sofort aus Satz 3.5.2 (ii) und (iii) unter Berücksichtigung der Tatsache, dass die Kovarianz von stochastisch unabhängigen Variablen stets gleich Null ist. Um (3.100) einzusehen, wendet man die Definition der Kovarianz an:

$$\begin{aligned} Kov(X,Y) &= E[(a_1 Z_1 + a_2 Z_2)(b_1 Z_1 + b_2 Z_2)] \\ &= E[a_1 b_1 {Z_1}^2 + a_2 b_2 {Z_2}^2 + a_1 b_2 Z_1 Z_2 + a_2 b_1 Z_2 Z_1) \\ &= a_1 b_1 E({Z_1}^2) + a_2 b_2 E({Z_2}^2) + a_1 b_2 E(Z_1 Z_2) + a_2 b_1 E(Z_2 Z_1) \end{aligned}$$

Aber $E({Z_1}^2) = E({Z_2}^2) = 1$ nach Voraussetzung, und

$$E(Z_1 Z_2) = E(Z_1)E(Z_2),$$

wegen der stochastischen Unabhängigkeit der Z_i. Da aber $E(Z_1) = E(Z_2) = 0$, folgt (3.100). □

Die Erwartungswerte, Varianzen und die Kovarianz der Variablen X und Y können also durch geeignete Wahl der Konstanten a_i und b_j, über die die stochastisch

unabhängigen zufälligen Veränderlichen in die Definition der Variablen X und Y eingehen, repräsentiert werden; die Kovarianz von X und Y ist trotz der stochastischen Unabhängigkeit der Z_1 und Z_2 nicht Null, wenn nur die Konstanten die Bedingung $a_1b_1 + a_2b_2 = 0$ nicht genügen. Allgemein läßt sich sagen, dass sich die Verteilungen von X und Y als Faltungen der Variablen a_1Z_1 und a_2Z_2 *bzw.* b_1Z_1 und b_2Z_2 ergeben; diese Aussagen gelten unabhängig von der speziellen Form der Verteilungen von Z_1 und Z_2. Der eigentlich interessante Aspekt des Satzes ist aber, dass X und Y als lineare Kombinationen unabhängiger Variabler repräsentiert werden können. Die Faktorenanalyse basiert auf diesem Sachverhalt; sie kann als Versuch aufgefaßt werden, die Konstanten a_i, b_j so zu schätzen, dass eine plausible inhaltliche Interpretation der Konstanten möglich ist.

3.5.2 Bedingte Erwartungswerte

In Abschnitt 3.3 sind die bedingten Verteilungen eingeführt worden. Dazu korrespondierend lassen sich die bedingten Erwartungswerte definieren.

Definition 3.5.4 *Es sei $f(x|A)$ die in (3.20) eingeführte bedingte Dichte von X, d.h. die Dichte von X unter der Bedingung, dass das zufällige Ereignis A eingetreten ist. Dann heißt*

$$E(X|A) = \int_{-\infty}^{\infty} x f(x|A)\, dx \tag{3.101}$$

die bedingte Erwartung *von X. Ist X diskret, so ist die bedingte Erwartung gemäß*

$$E(X|A) = \sum_k x_k P(X = x_k|A) \tag{3.102}$$

definiert.

Insbesondere kann das Ereignis A natürlich bezüglich eines Wertebereiches von X definiert werden:

Beispiel 3.5.3 Es sei X eine zufällige Veränderliche und $A = \{X \leq x_0\}$. Dann ist gemäß (3.19) und (3.20)

$$E(X|A) = \frac{\int_{-\infty}^{x_0} x f(x)\, dx}{\int_{-\infty}^{x_0} f(x)\, dx} \tag{3.103}$$

□

3.5.3 Der mehrdimensionale Fall

Es sei $X = (X_1, \cdots, X_n)'$ ein zufälliger Vektor (vergl. Definition 3.2.3). Die Begriffe des Erwartungswerts, der Varianz bzw. Kovarianz können auf zufällige Vektoren übertragen werden:

Definition 3.5.5 *Es sei X ein zufälliger Vektor und $\mu_i = E(X_i)$ sei der Erwartungswert der zufälligen Veränderlichen X_i, d.h. der i-ten Komponente von X. Dann heißt*

$$E(X) = \mu = (\mu_1, \cdots, \mu_n)' \tag{3.104}$$

der Vektor der Erwartungswerte *oder auch* Erwartungswertvektor. *Es sei weiter $\sigma_{ij} = Kov(X_i, X_j)$ die Kovarianz zwischen der i-ten und der j-ten Komponente von X. Dann heißt*

$$\Sigma = E((X-\mu)(X-\mu)') = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{n1} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{n2} \\ \vdots & \vdots & \cdots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{pmatrix} \tag{3.105}$$

die Matrix der Kovarianzen *von X, oder einfach die* Kovarianzmatrix *von X. Schließlich seien σ_i^2 und σ_j^2 die Varianzen von X_i und X_j, so dass $\rho_{ij} = \sigma_{ij}/(\sigma_i\sigma_j)$ die Korrelation zwischen X_i und X_j ist. Dann heißt*

$$R = \begin{pmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ \rho_{n1} & \rho_{n2} & \cdots & \rho_{nn} \end{pmatrix} \tag{3.106}$$

die Matrix der Korrelationen *oder einfach die* Korrelationsmatrix *von X.*

Sicherlich ist $\rho_{ii} = 1$ für alle i.

Satz 3.5.6 *Es seien X und Y zwei n-dimensionale zufällige Vektoren, α und β seien reelle Zahlen, a und b seien ebenfalls n dimensionale Vektoren und A und B seien geeignet dimensionierte Matrizen.*

$$E(X+Y) = E(X) + E(Y) \tag{3.107}$$

$$E(\alpha X + \beta) = \alpha E(X) + \beta \tag{3.108}$$

$$\Sigma(X) = E(XX') - \mu\mu' \tag{3.109}$$

$$Var(\vec{a}'X) = \vec{a}'\Sigma(X)\vec{a}a = \sum_{i,j=1}^{n} a_i a_j \sigma_{ij} \tag{3.110}$$

$$\Sigma(AX+b) = A\,\Sigma(X)\,A' \tag{3.111}$$

$$\Sigma(X,X) = \Sigma(X), \quad \Sigma(X,Y) = \Sigma(Y,X)' \tag{3.112}$$

$$\Sigma(AX+a, BY+b) = A\,\Sigma(X,Y)\,B' \tag{3.113}$$

Beweis: Die Aussagen folgen sofort aus den Rechenregeln für Vektoren und Matrizen.

□

3.6 Erzeugende Funktionen

3.6.1 Allgemeine Definition

Für den Spezialfall, dass die zufällige Veränderliche *nur ganzzahlige Werte* annehmen kann, erweist sich der Begriff der *erzeugenden Funktion* als sehr nützlich für die Bestimmung der Momente der Verteilung. Darüber hinaus können erzeugende Funktionen | wie die charakteristischen Funktionen, die für stetige Verteilungen bestimmt werden können | bei der Herleitung von Verteilungen, z.B. der Verteilung der Summe von zufälligen Veränderlichen, hilfreich sein. Die folgende Darstellung entspricht der von Feller (1968).

Definition 3.6.1 *Es sei $a_0, a_1, a_2, \cdots$ eine Folge reeller Zahlen. Wenn die Summe*

$$A(s) = a_0 + a_1 s + a_2 s^2 + \cdots = \sum_{k=0}^{\infty}$$

in einem bestimmten Intervall $-s_0 < s < s_0$ konvergiert, so heißt $A(s)$ die erzeugende Funktion *der Folge $a_0, a_1, a_2, \cdots$.*

Die Variable s hat gewissermaßen nur eine Hilfsfunktion, wie gleich deutlich werden wird. Es sei X eine zufällige Veränderliche, die nur die Werte 0, 1, 2, $\cdots$ annehmen möge. Weiter sei $p_j = P(X = j)$, $q_j = P(X > j)$, so dass $q_k = p_{k+1} + p_{k+2} + \cdots$. Dann sind $\{p_j\}$ und $\{q_j\}$ Zahlenfolgen im Sinne von Definition 3.6.1 und die zugehörigen erzeugenden Funktionen sind durch

$$A(s) = p_0 + p_1 s + p_2 s^2 + \cdots \tag{3.114}$$

$$Q(s) = q_0 + q_1 s + q_2 s^2 + \cdots \tag{3.115}$$

gegeben. $A(s)$ konvergiert sicherlich für $-1 \leq s \leq 1$, denn $P(1) = 1$ (die Summe der Wahrscheinlichkeiten ist stets 1). $Q(s)$ konvergiert sicherlich für $-1 < s < 1$, dass die $q_j < 1$. Es gilt dann der folgende

Satz 3.6.1 *Es sei $-1 < s < 1$; dann ist*

$$Q(s) = \frac{1 - A(s)}{1 - s}. \tag{3.116}$$

Beweis: Die Aussage ist richtig, wenn $(1-s)Q(s) = 1 - A(s)$. Multipliziert man $Q(s)$ mit $1 - s$, so erhält man nach (3.116)

$$\begin{aligned}
(1-s)Q(s) &= q_0(1-s) + q_1 s(1-s) + q_2 s^2(1-s) + \cdots \\
&= q_0(1-s) + q_1(s - s^2) + q_2(s^2 - s^3) + \cdots \\
&+ q_{n-1}(s^{n-1} - s^n) + q_n(s^n - s^{n+1}) + \cdots
\end{aligned}$$

Multipliziert man diesen Ausdruck weiter aus und faßt die Terme mit gleichem Exponenten von s zusammen, so sieht man, dass sich

$$(1-s)Q(s) = q_0 + s(q_1 - q_0) + s^2(q_2 - q_1) + \cdots + s^n(q_n - q_{n-1}) + \cdots$$

ergibt. Für $n = 1, 2, \cdots$ ist aber

$$q_n - q_{n-1} = p_{n+1} + p_{n+2} + \cdots - p_n - p_{n+1} - \cdots = -p_n, \quad n \geq 1.$$

Für $n = 0$ ist

$$q_0 = p_1 + p_2 + p_3 + \cdots = 1 - p_0$$

und somit

$$(1-s)Q(s) = 1 - p_0 - p_2 - p_3 - \cdots = 1 - A(s),$$

und das wurde behauptet. □

3.6.2 Erzeugende Funktionen und die Momente einer Verteilung

Der Zusammenhang zwischen den Momenten und der erzeugenden Funktion einer Wahrscheinlichkeitsverteilung ergibt sich, wenn man die Ableitungen der erzeugenden Funktion bezüglich s betrachtet. Mit $p_k = P(K = k)$ ist die erzeugende Funktion

$$A(s) = \sum_{k=1}^{\infty} p_k s^k$$

Dann ist

$$A'(s) = \frac{dA(s)}{ds} = \sum_{k=1}^{\infty} k\, p_k s^{k-1} \tag{3.117}$$

$$A''(s) = \frac{d^2A(s)}{ds^2} = \sum_{k=1}^{\infty} k(k-1)p_k s^{k-2} = \sum_{k=1}^{\infty} k^2 p_k s^{k-2} - \sum_{k=1}^{\infty} k p_k s^{k-2} \tag{3.118}$$

etc. Setzt man in (3.117) und (3.118) $s = 1$, so erhält man

$$A'(1) = \sum_{k=1}^{\infty} k\, p_k \tag{3.119}$$

$$A''(1) = \sum_{k=1}^{\infty} k^2 p_k - \sum_{k=1}^{\infty} k p_k \tag{3.120}$$

d.h. $A'(1)$ ist offenbar gerade das erste Moment der Verteilung von K, und das zweite Moment ist durch

$$\sum_{k=1}^{\infty} k^2 p_k = A''(1) + A'(1) \tag{3.121}$$

gegeben. Da die Varianz von K allgemein durch $Var(K) = E(K^2) - E^2(K)$ definiert ist, hat nan zusammenfassend

$$\begin{aligned} E(K) &= A'(1) && (3.122) \\ Var(K) &= A''(1) + A'(1) - (A'(1))^2 = A''(1) + A'(1)(1 - A'(1)) && (3.123) \end{aligned}$$

Erwartungswert und Varianz einer Verteilung von ganzzahligen zufälligen Veränderlichen lassen sich also über die erzeugende Funktion ausdrücken.

Die Momente bzw. der Erwartungswert und die Varianz lassen sich auch über die Funktion $Q(s)$ (vergl. (3.116)) ausdrücken. Aus (3.116) folgt

$$\begin{aligned} A(s) &= 1 - (1-s)Q(s) = 1 - Q(s) + sQ(s), \text{ und damit} \\ A'(s) &= Q(s) - (1-s)Q'(s) \\ A''(s) &= 2Q'(s) - (1-s)Q''(s). \end{aligned}$$

Man rechnet leicht nach, dass dann auch

$$Var(K) = 2Q'(1) + Q(1) - Q^2(1) \tag{3.124}$$

gilt.

Im folgenden Kapitel werden die erzeugenden Funktionen einer Reihe diskreter Verteilungen zur Bestimmung der jeweiligen Erwartungswerte angewandt.

Beispiel 3.6.1 X sei eine Bernoulli-Variable mit $P(X = 0) = p_0 = 1 - p$, $P(X = 1) = p_1 = p$. Die erzeugende Funktion ist

$$\begin{aligned} A(s) &= \sum_{k=0}^{\infty} p_k s^k \\ &= 1 - p + p\,s \\ &= 1 + p(s-1) \end{aligned}$$

Dann ist

$$\begin{aligned} A'(s) &= p \quad \text{und dementsprechend} \\ A''(s) &= 0 \quad \text{für alle } s \end{aligned}$$

Also folgt

$$\begin{aligned} E(X) = A'(1) &= p && (3.125) \\ Var(X) = A''(1) + A'(1)(1 - A'(1)) &= p(1-p) && (3.126) \end{aligned}$$

□

3.7 Charakteristische Funktionen

Für stetige zufällige Veränderliche ist das äquivalent für die erzeugende Funktion die charakteristische Funktion:

Definition 3.7.1 *Es sei X eine stetige zufällige Veränderliche mit der Dichtefunktion $f(x)$, die absolut integrierbar sei, d.h. es soll $\int |f(x)|dx < \infty$ gelten. Dann heißt*

$$F(\omega) = \int_{-\infty}^{\infty} f(x)e^{-i\omega x}dx \tag{3.127}$$

die charakteristische Funktion *(c.F.) von $f(x)$.*

Offenbar ist die charakteristische Funktion gerade die Fouriertransformierte (s. Anhang) der Dichtefunktion f. Aus der c.F. lassen sich die Momente von X wie folgt bestimmen: man bildet zunächst die Ableitung $dF(\omega)/d\omega$ und erhält

$$\frac{dF(\omega)}{d\omega} = F'(\omega) = -i\int_{-\infty}^{\infty} xf(x)e^{-i\omega x}dx$$

Für $\omega = 0$ ist aber $e^{-i\omega x} = 1$, so dass

$$F'(0) = -i\int_{-\infty}^{\infty} x\, f(x)\, dx$$

oder

$$-\frac{\mathbf{F}'(0)}{i} = E(X)$$

Um das k-te Moment einer Verteilung zu bestimmen, bildet man die k-te Ableitung von $\mathbf{F}(\omega)$:

$$\frac{d^k\mathbf{F}(\omega)}{d\omega^k} = (-i)^k\int_{-\infty}^{\infty} x^k f(x)e^{-i\omega}\, dx$$

und erhält das Resultat

$$E(x^k) = \frac{\mathbf{F}^k(0)}{(-i)^k}$$

Die Resultate können in dem folgenden Satz zusammengefaßt werden:

Satz 3.7.1 *Es sei X eine stetige zufällige Veränderliche mit der absolut integrierbaren Dichte $f(x)$ und der charakteristischen Funktion $\mathbf{F}(\omega)$. Dann ist das k-te Moment von X durch*

$$E(X^k) = \frac{1}{(-i)^k}\frac{d^k\mathbf{F}(\omega)}{d\omega^k}\bigg|_{\omega=0}, \quad k = 1, 2, 3, \cdots \tag{3.128}$$

gegeben.

Beispiel 3.7.1 X sei zwischen den Werten $a < b$ gleichverteilt, $a < X < b$. Die Verteilungsfunktion ist dann durch $F(x) = x/(b-a)$ gegeben und die Dichte durch $f(x) = 1/(b-a)$. Die c.F. ist dann

$$\mathbf{F}(\omega) = \frac{1}{(b-a)} \int_a^b e^{-i\omega x} dx = \frac{1}{(b-a)} \frac{1}{i\omega} (e^{-i\omega a} - e^{-i\omega b}).$$

Um den Erwartungswert zu bestimmen, muß $\mathbf{F}'(\omega)$ gefunden werden. Differenziert man den Ausdruck auf der rechten Seite, so ergeben sich gewisse Schwierigkeiten für den Fall $\omega = 0$, so dass man besser unter dem Integralzeichen differenziert:

$$\mathbf{F}'(\omega) = -\frac{i}{b-a} \int_a^b x e^{-i\omega x} dx$$

und man erhält $\mathbf{F}'(0) = -i/(2(b-a))$, so dass

$$E(X) = -\frac{1}{i} \mathbf{F}'(0) = \frac{1}{2} \frac{1}{b-a}.$$

Die höheren Momente erhält man auf analoge Weise. □

Anwendung: Faltungen

Erzeugende und charakteristische Funktionen lassen sich gelegentlich zur Bestimmung der Verteilung von Summen unabhängiger zufälliger Veränderlicher anwenden. Die Basis dafür sind die folgenden Sätze:

Satz 3.7.2 *Es seien X und Y zwei stochastisch unabhängige, diskrete zufällige Veränderliche mit den Verteilungen $p_j = P(X = j)$, $p_k = P(Y = k)$. Die zu den Verteilungen korrespondierenden erzeugenden Funktionen seien $P_j(s)$ und $P_k(s)$. Weiter sei $Z = X + Y$. Die Verteilung von Z sei $p_r = P(Z = r)$. Dann hat Z die erzeugende Funktion*

$$P_r(s) = P_j(s) P(_k(s). \tag{3.129}$$

Beweis: Es ist

$$\begin{aligned} P(Z = r) &= P(0 + r \cup 1 + r - 1 \cup 2 + r - 2 \cup \cdots \cup r + 0) \\ &= \sum_{r=0}^{\infty} P(X = m \cap Y) = \sum_{r=0}^{\infty} P_j(X = m) P_k(Y = r - m) = p_r. \end{aligned}$$

Die erzeugende Funktion für Z ist dann

$$P_r(s) = \sum_{r=0}^{\infty} p_r s^r \sum_{r=0}^{\infty} \sum_{m=0}^{r} P_j(X = m) P_k(Y = r - m) s^r.$$

Die erzeugenden Funktionen für X un Y sind

$$P_j(s) = \sum_{j=0}^{\infty} p_j s^j, \quad P_k(s) = \sum_{k=0}^{\infty} p_k s^k$$

Dann ist

$$P_j(s)P_k(s) = \sum_{j=0}^{\infty} p_j s^j \sum_{k=0}^{\infty} p_k s^k = \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} p_j p_k s^{j+k}$$

Dies ist aber gerade die erzeugende Funktion für Z, wie man sieht, wenn man in der Summe die Terme mit gleichem Wert $j + k = r$ zusammenfaßt. □

Beispiel 3.7.2 X_1, X_2 seien stochastisch unabhängige Bernoulli-Variable mit identischer Verteilung $P(X_i = 0) = 1 - p$, $P(X_i = 1) = p$. Gesucht ist die Verteilung von $X = X_1 + X_2$. Aus Beispiel 3.6.1 ist die erzeugende Funktion für X_i, $i = 1, 2$ durch $P_i(s) = 1 - p + ps$ gegeben. Nach (3.129) ist dann

$$P_x(s) = (1 - p + ps)^2 = (1 - p)^2 + 2(1 - p)ps + p^2 s^2.$$

Andererseits ist die Verteilung von X die Binomialverteilung

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

$n = 2$, $k = 0$, 1, 2. Die Terme $P(X = k)$ für $k = 0, 1, 2$ sind demnach $(1 - p)^2$, $2p(1 - p)$, p^2. Offenbar sind dies die Koeffizienten in der Reihe $P_x(s)$. Für die allgemeine Binomialverteilung $P(X|n, p)$ erhält man dementsprechend die erzeugende Funktion

$$P_x(s) = (1 - p + ps)^n. \tag{3.130}$$

Für den Erwartungswert von X erhält man

$$\frac{dP_x(s)}{ds} = n(1 - p + ps)^{n-1} p$$

und folglich

$$E(X) = \left.\frac{dP_x(s)}{ds}\right|_{s=1} = n(1 - p + p)^{n-1} p = np$$

Weiter ist $P_x''(s) = n(n - 1)(1 - p + ps)^{n-2} p$ und

$$Var(X) = P''(1) + P'(1) - [P'(1)]^2,$$

mithin

$$\begin{aligned} Var(X) &= n(n-1)(1-p+p)^{n-2}p + n(1-p+p)^{n-1}p \\ &- (n(1-p+p)^{n-1}p)^2 \\ &= n(n-1)p + np - n^2 p = n^2 p - np + np - n^2 p^2 \\ &= np(1-p), \end{aligned}$$

vergl. (3.85) und (3.86). □

Satz 3.7.3 *Es seien X, Y zwei stochastisch unabhängige, stetige zufällige Veränderliche mit den Dichten f_x und f_y, und f_z sei die Dichte von $Z = X + Y$. Die charakteristischen Funktionen für X, Y und Z seien $\mathbf{F}_x(\omega)$, $\mathbf{F}_y(\omega)$ und $\mathbf{F}_z(\omega)$. Dann gilt*

$$\mathbf{F}_z(\omega) = \mathbf{F}_x(\omega)\mathbf{F}_y(\omega). \tag{3.131}$$

Beweis: Es ist

$$f_z(z) = \int_{-\infty}^{\infty} f_x(x)f(z-x)\,dx = \mathbf{F}_z(\omega) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_x(x)f(z-x)e^{-i\omega z}\,dx\,dz.$$

Es sei $y = z - x$, $dy/dx = -1$. Dann ist

$$\begin{aligned}\mathbf{F}_z(\omega) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_x(x)f(y)e^{-i\omega(y+x)}\,dx\,dy \\ &= \int_{-\infty}^{\infty} f_x(x)e^{-i\omega x}\int_{-\infty}^{\infty} f(y)e^{-i\omega y}\,dx\,dy \\ &= \mathbf{F}_x(\omega)\mathbf{F}_y(\omega). \end{aligned} \tag{3.132}$$

□

Beispiel 3.7.3 Es werde angenommen, X und Y seien stochastisch unabhängig mit den Verteilungen $f_x(x) = \lambda e^{-\lambda x}$, $f_y(y) = \mu e^{-\mu y}$. Gesucht ist die charakteristische Funktion von $Z = X + Y$.

Für f_x ist die charakteristische Funktion

$$\begin{aligned}\mathbf{F}_x(\omega) &= \lambda\int_0^{\infty} e^{-i\lambda x - i\omega x}\,dx \\ &= \int_0^{\infty} e^{-i(\lambda+\omega)x}\,dx \\ &= \frac{1}{i(\lambda+\omega)}, \end{aligned} \tag{3.133}$$

und analog dazu findet man für f_y die charakteristische Funktion

$$\mathbf{F}_y(\omega) = \frac{\mu}{i(\mu-\omega)};$$

dementsprechend ist die c.F. für f_z

$$\mathbf{F}_z(\omega) = -\frac{\lambda\mu}{(\lambda-\omega)(\mu-\omega)} \tag{3.134}$$

Der Erwartungswert von Z ergibt sich sofort aus $F_z'(0)$. Es ist

$$\frac{d\mathbf{F}_z(\omega)}{d\omega} = \lambda\mu\left[\frac{1}{(\lambda-\mu)^2(\mu-\omega)} + \frac{1}{(\lambda-\mu)(\mu-\omega)^2}\right] = \frac{\mu+\lambda}{\lambda\mu} = \frac{1}{\lambda} + \frac{1}{\mu}. \tag{3.135}$$

Hieraus läßt sich die Dichte von Z durch inverse Transformation finden,

$$f_z(z) = \int_{-\infty}^{\infty} \frac{\lambda\mu}{(\lambda-\omega)(\mu-\omega)} e^{i\omega z} d\omega = \frac{\mu}{\lambda-\mu}(e^{-\mu z} - e^{-\lambda z}). \tag{3.136}$$

□

Kapitel 4

Diskrete Wahrscheinlichkeitsverteilungen

In diesem Kapitel werden einige der gebräuchlichsten Wahrscheinlichkeitsverteilungen für diskrete zufällige Veränderliche zusammengestellt.

4.1 Die Gleichverteilung

Mit dieser Verteilung wird oft, wenn auch nicht immer zu recht, die "reine Zufälligkeit" assoziiert:

Definition 4.1.1 *Es sei Ω eine endliche Menge von Elementarereignissen, und es gelte insbesondere $|\Omega| = n < \infty$. Weiter sei X eine zufällige Veränderliche, die die Werte $x_1, \cdots, x_n$ annehme, mit $x_i = x(\omega_i)$. Gilt*

$$P(X = x_i) = P(\omega_i) = \frac{1}{n} \quad \textit{für alle} \;\; \omega_i \in \Omega, i = 1, \cdots, n, \tag{4.1}$$

so heißt P Gleichverteilung.

Für den Erwartungswert und die Varianz der Gleichverteilung folgt dann sofort

$$E(X) = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{4.2}$$

$$Var(x) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 \tag{4.3}$$

Ohne weitere Aussagen über die Werte x_i lassen sich diese Ausdrücke nicht weiter vereinfachen. Im folgenden Beispiel wird der Spezialfall $x_i = i$, $1 \leq i \leq n$ betrachtet:

Beispiel 4.1.1 Es sei $x_i = i$, $i = 1, \cdots, n$; dieser Fall bildet z.B. die Situation beim normalen, sechsseitigen Würfel ab. Allgemein gilt $p_i = P(X = x_i)$, und für die Gleichverteilung | alle Seiten haben die gleiche Wahrscheinlichkeit, nach einem Wurf oben zu liegen | gilt insbesondere $p_i = p = 1/n$. Für den Erwartungswert hat man

$$E(X) = \sum_{i=1}^{n} p_i x_i = \frac{1}{n} \sum_{i=1}^{n} i = \frac{1}{n} \frac{n(n+1)}{2} = \frac{n+1}{2} \tag{4.4}$$

und für die Varianz

$$Var(X) = E(X^2) - E^2(X) = \sum_{i=1}^{n} x_i^2 p_i - \left(\sum_{i=1}^{n} p_i x_i \right)^2$$

ergibt sich

$$\begin{aligned} Var(X) &= \frac{1}{n} \sum_{i=1}^{n} i^2 - \frac{1}{n^2} \left(\sum_{i=1}^{n} i \right)^2 \\ &= \frac{1}{n} \frac{n(n+1)(2n+1)}{6} - \frac{1}{n^2} \frac{n^2(n+1)^2}{4} \\ &= \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} \\ &= \frac{(n+1)(n-1)}{12} \end{aligned} \tag{4.5}$$

Zur Illustration sollen die Ausdrücke für $E(X)$ und $Var(X)$ über die erzeugende Funktion hergeleitet werden. Für die Gleichverteilung ist die erzeugende Funktion durch

$$P(s) = \sum_{i=1}^{n} p_i s^i = p \sum_{i=1}^{n} s^i = \frac{1}{n} \sum_{i=1}^{n} s^i \tag{4.6}$$

gegeben. Es ist

$$\begin{aligned} P'(s) &= \frac{1}{n} \sum_{i=1}^{n} i\, s^{i-1} \\ P''(s) &= \frac{1}{n} \sum_{i=1}^{n} i(i-1) s^{i-2} = \frac{1}{n} \sum_{i=1}^{n} i^2 s^{i-2} - \frac{1}{n} \sum_{i=1}^{n} i s^{i-2} \end{aligned}$$

und mithin

$$E(X) = P'(1) = \frac{1}{n} \sum_{i=1}^{n} i = \frac{n+1}{2} \tag{4.7}$$

Da für die Varianz

$$Var(X) = P''(1) + P'(1) - (P'(1))^2$$

gilt, folgt (vergl. (3.122), p. 106)

$$\begin{aligned} Var(X) &= \frac{1}{n}\sum_{i=1}^{n} i^2 - \frac{1}{n}\sum_{i=1}^{n} i + \frac{1}{n}\sum_{i=1}^{n} i - \left(\frac{1}{n}\sum_{i=1}^{n} i\right)^2 \\ &= \frac{1}{n}\left(\frac{n(n+1)(2n+1)}{6}\right) - \frac{1}{n^2}\left(\frac{n(n+1)}{2}\right)^2 \\ &= \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} = \frac{(n+1)(n-1)}{12}. \end{aligned} \tag{4.8}$$

vergl. (4.5). Für $n = 6$ findet man $E(X) = 3.5$, und $\mathrm{Var}(X) = 15.16667 - 3.5^2 = 2.91667$.

□

4.2 Die Bernoulli-Verteilung

Definition 4.2.1 *Es sei Ω eine beliebige, nicht notwendig endliche oder abzählbare Menge von Elementarereignissen. X sei eine zufällige Veränderliche mit $X(\omega) = 1$ für $\omega \in A$, $X(\omega) = 0$ für $\omega \in \bar{A}$. Dann heißt X* Bernoulli-Variable, *und*

$$P(X = 1) = p, \quad P(X = 0) = 1 - p \tag{4.9}$$

heißt Bernoulli-Verteilung

Die erzeugende Funktion der Bernoulli-Verteilung ist durch

$$A(s) = (1-p)s^0 + p \cdot s^1 = q + ps, \quad q = 1 - p. \tag{4.10}$$

Man erhält sofort $A'(s) = p$, $A''(s) = 0$, und mithin

$$E(X) = p \tag{4.11}$$

$$\mathrm{Var}(X) = p(1-p) \tag{4.12}$$

wie man natürlich auch ohne Betrachtung der erzeugenden Funktion direkt sieht. $E(X) = p \cdot 1 + (1-p) \cdot 0 = p$, $\mathrm{Var}(X) = E(X^2) - E^2(X) = p \cdot 1^2 + (1-p) \cdot 0^2 - p^2 = p - p^2 = p(1-p)$.

4.3 Die Binomialverteilung

4.3.1 Definition

Definition 4.3.1 *Es seien $X_1, \cdots, X_n$ stochastisch unabhängige Bernoulli-Variable mit $P(X_i = 1) = p$ für alle i und es sei $X = X_1 + \cdots + X_n$. Dann heißt die Verteilung $P(X = k)$, $k = 0, 1, \cdots, n$* Binomialverteilung *mit den Parametern n und p; man schreibt auch $B(X; n, p)$ für diese Verteilung.*

Einen expliziten Ausdruck $P(X = k|n, p)$ erhält man, wenn man bedenkt, dass die Wahrscheinlichkeit einer *bestimmten* Folge von k "Erfolgen" und "Misserfolgen" bei n Versuchen durch

$$p^k(1-p)^{n-k}$$

gegeben ist; dies folgt sofort aus der Unabhängigkeit der einzelnen Versuche und der Tatsache, dass es auf die Reihenfolge der Multiplikation nicht ankommt. Man ist aber nicht an der Wahrscheinlichkeit einer bestimmten Folge interessiert, sondern an der Wahrscheinlichkeit, überhaupt k Erfolge, unabhängig von der Reihenfolge, zu erhalten. Nun gibt es gerade $\binom{n}{k}$ Möglichkeiten, k Erfolge in einer Folge von n Versuchen zu beobachten, und nach dem Summensatz für Wahrscheinlichkeiten (verschiedene Folgen sind einander ausschließende Ereignisse) folgt

$$P(X = k|n, p) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{4.13}$$

Eine andere Herleitung der Verteilung ergibt sich über die erzeugende Funktion der Verteilung. Offenbar ist die Binomialverteilung die n-fache Faltung der Bernoulli-Verteilung. Dementsprechend muß die erzeugende Funktion der Binomialverteilung durch die n-te Potenz (n-faches Produkt mit sich selbst) der erzeugenden Funktion der Bernoulli-Verteilung sein; man hat dementsprechend

$$P(s) = \sum_{k=1}^{n} p_k s^k = (q + p\,s)^n \tag{4.14}$$

wobei $p_k = P(X = k)$. Nun kann man das Binom $(q + ps)^n$ entwickeln, und der k-te Term muß dann gleich p_k sein. Man erhält dementsprechend

$$(q + ps)^n = \sum_{k=1}^{n} \binom{n}{k} (p\,s)^k q^{n-k} = \sum_{k=1}^{n} \binom{n}{k} p^k q^{n-k} s^k$$

so dass

$$P(X = k|n, p) = \binom{n}{k} p^k (1-p)^{n-k},$$

also (4.13) folgt. □

Für den Erwartungswert und die Varianz einer binomialverteilten Variablen ergibt sich sofort

$$E(X) = np \tag{4.15}$$

$$\mathrm{Var}(X) = n\,p(1-p)) \tag{4.16}$$

da ja X die Summe stochastisch unabhängiger Bernoulli-Variablen ist ($E(X_i) = p$, $E(X_1 + \cdots + X_n) = np$, $Var(X_i) = p(1-p)$, $Var(X_1 + \cdots + x_n) = np(1-p)$).

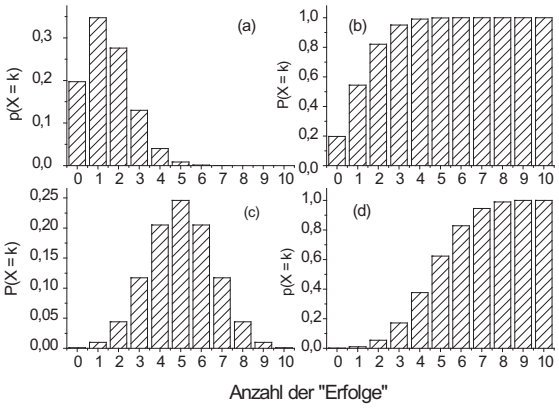
Erwartungswert und Varianz sind also bei einer binomialverteilten zufälligen Veränderlichen miteinander verkoppelt. Es ist aber nicht so, dass die Varianz

monoton mit dem Erwartungswert wächst. Vielmehr nimmt sie für $p = .5$ ein Maximum an. Denn man kann ja $Var(X)$ als Funktion von p auffassen; differenziert man also $Var(X) = Var(X; p)$ nach p und setzt die Ableitung gleich Null, so erhält man ein Extremum für $Var(X)$:

$$\frac{dVar(X;p)}{dp} = n(1-p) - np = n - 2np$$

Für $n-2np = 0$ folgt sofort $p = 1/2$, und da die zweite Ableitung $d^2Var(X;p)/dp^2 = -2n$ negativ ist, muß das Extremum ein Maximum sein. Die Abbildungen in Abb.

Abbildung 4.1: Binomialverteilungen für $p = .25$, $n = 10$ ((a) und (b)) und $p = .5$, $n = 10$ ((c) und (d))



4.1 zeigen die Wahrscheinlichkeitsverteilung und die zugehörige Verteilungsfunktion für $p = .25$ ((a) und (b)) und für $p = .5$ ((c) und (d)) für jeweils $n = 10$. Für $p = 1/2$ sind die Wahrscheinlichkeiten $p(X = k)$ symmetrisch zu $E(X) = np$; für $p \neq 1/2$ ist die Wahrscheinlichkeitsverteilung schief.

Die Binomialverteilung spielt in empirischen und in theoretischen Untersuchungen eine außerordentlich bedeutsame Rolle:

Beispiel 4.3.1 Bei einer Meinungsumfrage über die Wahlchancen der einzelnen Parteien werden genau n Personen befragt, ob sie eine bestimmte Partei wählen werden (A) oder nicht ($\bar{A}$). Die Stichprobe wird so angelegt, dass die Antworten der einzelnen Personen stochastisch unabhängig voneinander sind, d.h. die Befragten sollen über ihre Antworten vor der Befragung nicht über die Befragung kommunizieren. Weiter soll eine Person aus der Population wie eine Kugel aus einer Urne "gezogen" werden, wobei die Wahrscheinlichkeit, eine rote Kugel zu ziehen, bei jedem Zug gleich groß, nämlich gleich p ist, und p soll dem Anteil der Wähler der interessierenden Partei in der Population entsprechen. Dann ist die Anzahl X der Personen, die sich zur Wahl der Partei bekennen, binomialverteilt.

Ist insbesondere $X = k$, so ist die relative Häufigkeit "bekennender" Wähler gleich $\hat{p} = k/n$. Dann ist $E(\hat{p}) = E(k/n) = E(k)/n = np/n = p$, und analog findet man $Var(\hat{p}) = p(1-p)/n$. Je größer also der Wert von n ist, desto näher wird im allgemeinen $\hat{p}$ bei p liegen. Für eine spezielle Befragung kann sich $\hat{p}$ allerdings sehr von p unterscheiden, auch wenn die genannten Bedingungen der Stichprobenbildung alle erfüllt wurden. □

Beispiel 4.3.2 In einem psychophysischen Experiment soll die Intensität eines Reizes (Ton, Licht,...) bestimmt werden, mit der der Reiz an der Schwelle wahrgenommen, d.h. entdeckt wird. Dies bedeutet, dass er mit einer Intensität m_0 dargeboten wird, die gerade so groß ist, dass die Wahrscheinlichkeit des Entdeckens einen bestimmten Wert, etwa $p_o = 1/2$, hat. m_0 hängt von den experimentellen Bedingungen ab und muß experimentell bestimmt werden. Dazu wählt man einen Wert m, der in mutmaßlicher Nachbarschaft vom m_0 liegt, und bietet den Reiz mit dieser Intensität n-mal dar. Kann man annehmen, dass die Reaktionen der Versuchspersonen auf die verschiedenen Reizdarbietungen stochastisch unabhängig sind, so ist die Anzahl der Entdeckungen des Reizes $B(X; n, p)$-, d.h. binomialverteilt. Dabei ist der Parameter p, die Wahrscheinlichkeit des Entdeckens in einem Versuchsdurchgang, von der Intensität m abhängig, $p = p(m)$. Für $X = k$ findet man die Schätzung $\hat{p}(m) = k/n$. Aus verschiedenen Schätzungen $\hat{p}_i = \hat{p}(m_i)$ kann man dann eine Schätzung $\hat{p}_o = \hat{p}(\hat{m}_0)$ gewinnen, die eine Schätzung von $\hat{m}_0$ liefert. □

4.3.2 Verallgemeinerung

Eine Verallgemeinerung der Binomialverteilung ergibt sich, wenn man annimmt, dass die Wahrscheinlichkeit für das Ereignis A für jeden Versuch verschieden ist. Im i-ten Versuch sei also die Wahrscheinlichkeit für A durch p_i gegeben. Die einzelnen Versuche seien aber nach wie vor stochastisch unabhängig. Gesucht ist dann wieder $P(X = k)$ für $k = 0, \cdots, n$. Die Wahrscheinlichkeit für eine bestimmte Folge der Ereignisse A und $\bar{A}$ ist dann durch das zu der Folge korrespondierende Produkt der p_i bzw. $1 - p_i$ gegeben. So ist z.B. $p_1 p_2 (1 - p_3)$ die Wahrscheinlichkeit für die Folge $AA\bar{A}$, und es ist $X = 2$. Die Wahrscheinlichkeit für die Folge $A\bar{A}A$ ist $p_1(1-p_2)p_3$ mit ebenfalls $X = 2$. Für den Fall $p =$ konstant sind die Ausdrücke für $P(X = k)$ durch die Entwicklung von $(p + q)^n$ gegeben. Man verifiziert leicht, dass sich die Wahrscheinlichkeiten für $P(X = k)$ für den Fall $p \neq$ konstant durch die Terme der Entwicklung

$$\prod_{j=1}^{n} (p_j + q_j) = 1 \tag{4.17}$$

ergeben. Gilt (4.17) so ist der Erwartungswert für X durch

$$E(X) = n\bar{p}, \tag{4.18}$$

mit

$$\bar{p} = \frac{1}{n}\sum_{k=1}^{n} p_k$$

gegeben, und die Varianz ist

$$Var(X) = n\bar{p}\bar{q} - \sigma_p^2 \tag{4.19}$$

mit

$$\overline{q} = \frac{1}{n}\sum_{k=1}^{n} q_k, \quad \sigma_p^2 = \frac{1}{n}\sum_{k=1}^{n} p_k^2 - \overline{p}^2$$

(Kendall und Stuart (1968a), p. 127). Im Falle ungleicher p_i-Werte ist die Varianz von X kleiner als bei konstantem $p = \bar{p}$.

4.4 Die Multinomialverteilung

Eine zweite Verallgemeinerung der Binomialverteilung ergibt sich, wenn bei jedem der n stochastisch unabhängigen Versuche eines von k möglichen zufälligen Ereignissen A_i auftreten kann. Es sei $p_i = p(A_i) = konstant$ für alle Versuche. Diese Situation hat man z. B., wenn man in Beispiel 4.3.1, Seite 117, die Personen nicht nur befragt, ob sie eine bestimmte Partei wählen werden oder nicht, sondern welche von k möglichen Parteien sie wählen werden. Es gilt

Definition 4.4.1 *Es werden n stochastisch unabhängige Versuche durchgeführt. In jedem Versuch kann das zufällige Ereignis A_i, $i = 1, \cdots, k$ mit der Wahrscheinlichkeit $p_i = p(A_i)$ auftreten. Die zufällige Veränderliche X_i repräsentiere die Häufigkeit, mit der das Ereignis A_i insgesamt auftritt, und x_i sei der Wert, den X_i nach den n Versuchen angenommen hat; es gilt*

$$\sum_{i=1}^{k} p_i = 1, \quad \sum_{i=1}^{k} x_i = n$$

gilt. Dann heißen die zufälligen Veränderlichen $X_1, \cdots, X_k$ multinomialverteilt..

Satz 4.4.1 *Sind die $x_1, \cdots, x_k$ multinomialverteilt so gilt*

$$P(X_1 = x_1, \ldots, X_k = x_k) = \frac{n!}{x_1! x_2! \cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k} \tag{4.20}$$

mit $0 \leq x_i \leq n$.

Beweis: Die Anzahl $N(x_1, \ldots, x_k)$ der Möglichkeiten, n Elemente auf k Kategorien so aufzuteilen, dass gerade die Verteilung $X_1 = x_1, \ldots, X_k = x_k$ resultiert, ist

gerade $n!/(x_1! \cdots x_k!)$ (vergl. Kap.1, Multinomialkoeffizienten), und jede dieser Folgen hat die gleiche Wahrscheinlichkeit; wegen der stochastischen Unabhängigkeit folgt sofort, dass eine spezielle Folge mit $X_1 = x_1, \ldots, X_k = x_k$ die Wahrscheinlichkeit $p_1^{x_1} \cdots p_k^{x_k}$ hat. Dies führt unmittelbar zu (4.20). □

Für die Erwartungswerte und Varianzen der X_i gilt

$$E(X_i) = np_i, \quad Var(X_i) = np_i(1 - p_i). \tag{4.21}$$

Dies sind natürlich die Erwartungswerte und Varianzen der Binomialverteilung, denn für eine bestimmtes Ereignis A_i gilt ja, dass es in einem bestimmten Versuch entweder eintritt, und zwar mit der Wahrscheinlichkeit p_i, oder nicht mit der Wahrscheinlichkeit $1 - p_i$.

Es sei noch darauf hingewiesen, dass die Variablen X_i nicht stochastisch unabhängig sind; dies wird sofort klar, wenn man bedenkt, dass ja $\sum_i x_i = n$ sein muß. Also ist etwa $x_k = n - x_1 - \cdots - x_{k-1}$ und $p_k = 1 - p_1 - \cdots - p_{k-1}$, x_k kann also keinen Wert unabhängig von den übrigen x_j, $j = 1, \ldots, x_{k-1}$, annehmen.

Man findet durch eine leichte Rechnung, dass

$$Kov(X_i, X_j) = \begin{cases} np_i(1 - p_i), & i = j \\ -np_ip_j, & i \neq j \end{cases} \tag{4.22}$$

d.h. $\mathrm{Kov}(X_i, X_j) = 0$ nur dann, wenn $p_i = 0$ oder $p_j = 0$.

Beispiel 4.4.1 Die Bundesregierung beauftragt ein Meinungsforschungsinstitut, die Einstellung der Bevölkerung zur Arbeitsmarktpolitik der Regierung zu erfassen. Das Institut will dazu Personen aus m verschiedenen Berufsgruppen befragen; aus jeder Gruppe werden n zufällig ausgewählte Personen befragt. Die Personen sollen die Politik auf einer Skala von -3 (miserabel) über 0 (weder gut noch schlecht) bis +3 (ausgezeichnet) beurteilen. Es gibt also insgesamt 7 Kategorien. Es sei π_{ij} die Wahrscheinlichkeit, dass eine Person aus der i-ten Gruppe die j-te Kategorie ankreuzt, $i = 1, \cdots, m$, $j = 1, \cdots, 7$. Dann ist die Wahrscheinlichkeit, für die i-te Gruppe gerade r_1 Beurteilungen -3, r_2 Beurteilungen -2 und schließlich r_7 Beurteilungen +3 zu erhalten, durch die Verteilung (4.20) gegeben, wobei jetzt p_j durch π_{ij} zu ersetzen ist. □

4.5 Die geometrische Verteilung

Häufig befindet man sich in der Lage, einen Versuch so lange wiederholen zu müssen, bis sich ein "Erfolg" eingestellt hat. So kann es geschehen, dass man im Dunkeln vor der Tür steht und die Schlüssel eines Schlüsselbundes so lange ausprobieren muß, bis man den richtigen gefunden hat, wobei es vorkommen kann, dass man den gleichen, aber falschen Schlüssel mehrfach probiert. Manche Menschen versuchen, sich die Zukunft durch einen Münzwurf zu erschließen: liegt

z.B. "Kopf" oben, so tritt von zwei möglichen Ereignissen das angenehmere ein, liegt die "Zahl" oben, so ist einem das unangenehmere vorherbestimmt. Liegt beim ersten Wurf die Zahl oben, so hat man allerdings nicht "richtig" geworfen, man versucht es besser noch einmal. Liegt wiederum die "Zahl" oben, so hat man beim Werfen der Münze wieder etwas falsch gemacht, und man versucht es aufs Neue, - so lange, bis nach einem Wurf endlich der "Kopf" oben liegt und man getrost der Zukunft entgegensehen kann $\cdots$

Man kann nun fragen, wie groß die Wahrscheinlichkeit ist, dass erst nach dem k-ten Misserfolg der "Erfolg" beim $(k+1)$-ten Versuch eintritt. Die Frage führt auf die folgende Wahrscheinlichkeitsverteilung:

Definition 4.5.1 *Es werde eine Folge von stochastisch unabhängigen Versuchen durchgeführt, bei denen das Ereignis A jeweils mit der Wahrscheinlichkeit $p = P(A)$ und das Ereignis $\bar{A}$ mit der Wahrscheinlichkeit $1-p$ beobachtet wird. Es sei X die Anzahl der Versuche, in denen $\bar{A}$ eintritt, bevor zum ersten Male A beobachtet wird. Dann heißt X* geometrisch *oder* Laplace-verteilt.

Für die Wahrscheinlichkeit, dass X gerade den Wert k annimmt, dass also beim $(k+1)$-ten Versuch der erste "Erfolg" auftritt, findet man sofort

$$P(X=k) = P(\bar{A} \cap \bar{A} \cap \cdots \cap \bar{A} \cap A) = (1-p)^k p = q^k p \tag{4.23}$$

wegen der stochastischen Unabhängigkeit der einzelnen Versuche und wegen der Konstanz der Wahrscheinlichkeit p.

Es sei noch einmal angemerkt, dass man insgesamt $k+1$ Versuche betrachtet. Gelegentlich wird diese Gesamtzahl von $k+1$ Versuchen als geometrisch bezeichnet; die Form der erzeugenden Funktion und damit die Momente der Verteilung sind natürlich für die beiden Ansätze ein wenig verschieden. Ist z.B. $E(X)$ der Erwartungswert für den Ansatz, dass X gleich der Anzahl der Misserfolge vor dem ersten Erfolg ist, so ist $E(X)+1$ der Erwartungswert für den Fall, dass der $(k+1)$-te Versuch mitgezählt wird.

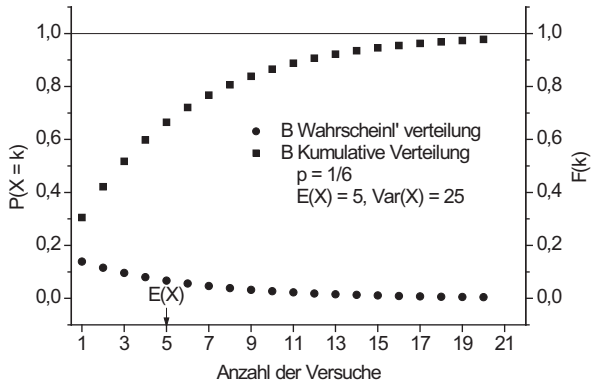
Die Verteilungsfunktion $P(X \leq k)$ ist durch

$$P(X \leq k) = p \sum_{j=0}^{k} (1-p)^j = 1 - (1-p)^{k+1}) \tag{4.24}$$

gegeben. Es soll zunächst werde die erzeugende Funktion für den Fall, dass X die Anzahl der Misserfolge vor dem ersten Erfolg repräsentiert, bestimmt werden. Dazu werde $p_k = q^k p$ gesetzt. Dann erhält man

$$\begin{aligned} A(s) &= \sum_{k=0}^{\infty} p_k s^k = p \sum_{k=0}^{\infty} q^k s^k \\ &= p \sum_{k=0}^{\infty} (qs)^k = \frac{p}{1-qs} \end{aligned} \tag{4.25}$$

Abbildung 4.2: Geometrische Verteilung, $p = 1/6$



Die Ausdrücke für den Erwartungswert und die Varianz können wieder über die erzeugende Funktion der Verteilung hergeleitet werden. Es gilt

$$E(X) = \frac{1-p}{p} \tag{4.26}$$

$$Var(X) = \frac{1-p}{p^2} = \frac{q}{p^2} \tag{4.27}$$

Zum Beweis rechnet man nach, dass der Ausdruck (4.25) für die erzeugende Funktion

$$\begin{aligned} A'(s) &= \frac{pq}{(1-qs)^2} \\ A''(s) &= \frac{2pq^2}{(1-qs)^3} \end{aligned}$$

liefert, woraus man

$$E(X) = A'(1) = \frac{pq}{(1-q)^2} = \frac{q}{1-q} = \frac{1-p}{p}$$

erhält, und

$$\begin{aligned} Var(X) &= A''(1) + A'(1) - (A'(1))^2 \\ &= \frac{2(1-p)^2}{p^2} - \frac{(1-p)^2}{p^2} + \frac{1-p}{p} \\ &= \frac{1-p}{p^2} = \frac{q}{p^2}. \end{aligned}$$

Zur Illustration werde noch der Fall betrachtet, dass X die Gesamtzahl der Versuche, also die Anzahl einschließlich des Versuches, der den ersten Erfolg ge-

bracht hat, repräsentiert. Es sei nun $X = k$, wobei die ersten $k-1$ Versuche Misserfolge waren. Die erzeugende Funktion ist dann, wegen $p_k = P(X = k) = q^{k-1}p$,

$$\begin{aligned} A(s) &= \sum_{k=0}^{\infty} p_k s^k = \frac{p}{q}\sum_{k=0}^{\infty}(qs)^k \\ &= \frac{p}{q(1-qs)}. \end{aligned}$$

Dann ist

$$\begin{aligned} A'(s) &= \frac{p}{(1-qs)^2} \\ A''(s) &= \frac{2pq}{(1-qs)^3} \end{aligned}$$

Für $s = 1$ folgt

$$E(X) = A'(1) = \frac{p}{(1-q)^2} = \frac{1}{p}.$$

Da X hier die Anzahl der Misserfolge *plus* 1 ist, muß die Anzahl der Misserfolge den Erwartungswert $E(X-1) = 1/p - 1 = (1-p)/p$ haben; dies ist der in (4.26) angegebene Wert.

Für die Varianz erhält man

$$\begin{aligned} Var(X) &= A''(1) + A'(1) - (A'(1))^2 = \frac{2q}{(1-q)^2} - \frac{1}{(1-q)^2} + \frac{1}{1-q} \\ &= \frac{q}{(1-q)^2} = \frac{1-p}{p^2} \end{aligned}$$

Dieser Ausdruck ist der gleiche wie der in (4.27) gegebene; dies muß auch so sein, denn man betrachtet hier die Varianz der zufälligen Veränderlichen $X + 1$, wenn X die Anzahl der Misserfolge ist, und $Var(X + 1) = Var(X)$; – die additive Konstante ändert nichts an der Varianz der Werte.

Beispiel 4.5.1 (Russisches Roulette) Ein 6-schüssiger Trommelrevolver wird mit nur einer Patrone bestückt. Dann wird die Trommel willkürlich (= zufällig) gedreht und man schießt auf sich selbst. Wie groß ist die Wahrscheinlichkeit, dass man genau nach dem fünften Schuß "Erfolg" hat, d.h. defunkt ist? Interessanter ist vermutlich aber die Frage, wie groß die Wahrscheinlichkeit ist, dass man eine Folge von fünf Schüssen überlebt.

Es sei A das Ereignis, dass die Trommelposition tödlich ist, und $\bar{A}$ ist das Ereignis, dass eine der übrigen Positionen eingestellt ist. Die Zufälligkeit der Trommelposition soll insbesondere bedeuten, dass alle Positionen gleichwahrscheinlich sind. Also ist $p = P(A) = 1/6, p(\bar{A}) = 5/6$. Um genau mit dem fünften Schuß "erfolgreich" zu sein, muß man vorher vier Misserfolge hingenommen haben. Dann ist $P(X = 4) = (1-p)^4 p = (5/6)^4(1/6) \approx .0804$.

Die Wahrscheinlichkeit, fünf Schüsse zu überleben, ist die Wahrscheinlichkeit von mindestens fünf Misserfolgen, also

$$P(X \geq 5) = P(X = 6 \cup X = 7 \cup \cdots).$$

Da nun

$$\begin{aligned} P(X \geq 5) &= 1 - P(X < 5) = 1 - (P(X = 1) + P(X = 2) + \cdots + P(X = 5)) \\ &= 1 - [(5/6 + (5/6)^2 + \cdots + (5/6)^5)/6] = .5685 > 1/2, \end{aligned} \tag{4.28}$$

d.h. man überlebt mit größerer Wahrscheinlichkeit, als würfe man die Münze.

Berechnet man noch den Erwartungswert und die Varianz von X, so findet man $E(X) = 5$, $Var(X) = 30$. Die Abbildung 4.2 zeigt die Wahrscheinlichkeitsverteilung und die Verteilungsfunktion für den Parameter $p = 1/6$. □

Beispiel 4.5.2 Jan Rapp aus Vegesack will durch einen Lottogewinn seine Rente aufbessern. Er meint, dass er keine Sorgen mehr haben wird, wenn er 7 "Richtige" aus 49 Zahlen tippt.

Nun ist bekannt, dass die Wahrscheinlichkeit, bei einmaligen Spielen 7 Richtige zu tippen, gerade $p = .0000000116414$ ist (vergl. Beispiel 4.8.1, Seite 136.Jan Rapp spielt jede Woche; wieviele Wochen muß er ansetzen, damit er innerhalb des durch diese Wochen definierten Zeitraums mit der Wahrscheinlichkeit $p_0 = .95$ mindestens einmal 7 Richtige tippt?

Die Wahrscheinlichkeit von k Mißerfolgen (keine 7 Richtige) ist durch (4.24) gegeben; es soll also $P(X \leq k) = .95$ gelten, also

$$1 - (1 - p)^{k+1} = .95.$$

Daraus findet man

$$k + 1 = \frac{\log(1 - .95)}{\log(1 - p)} = 2.57334 \times 10^8 \text{ Wochen},$$

was 4.5956×10^6 Jahren entspricht.

Der Erwartungswert liegt bei $E(X) = 8.59003 \times 10^7$ Wochen bzw. bei 1.53393×10^6 Jahren. Jan Rapp gibt nicht auf. Er weiß, dass auch Ereignisse mit der Wahrscheinlichkeit 0 nicht unmöglich sein müssen. □

Viele Fragestellungen, die auf die Anwendung der geometrischen Verteilung hinauslaufen, lassen sich paradigmatisch auf das Ziehen aus einer Urne mit Zurücklegen zurückführen:

Beispiel 4.5.3 McGill (1963) betrachtete die Suche nach Wörtern im Langzeitgedächtnis als Ziehen aus einer Urne mit Zurücklegen. So besteht im Experiment

von Bousfield und Sedgewick (1944) die Aufgabe der Vpn darin, alle zu einem vorgegebenen Oberbegriff gehörenden Wörter zu finden, etwa die Namen europäischer Städte. Die im Langzeitgedächtnis (LZG) gespeicherten Städtenamen werden nun als Kugeln in einer Urne — die das LGZ repräsentiert — aufgefaßt. Die Suche nach einem Städtenamen findet statt, indem zufällig irgendeine Kugel gewählt wird. Jede Kugel hat dabei die gleiche Wahrscheinlichkeit, gewählt zu werden. Repräsentiert die Kugel einen Städtenamen, der noch nicht genannt worden ist, so war der Zug ein "Erfolg". Wenn nicht, wird die Kugel zurückgelegt und die Suche geht weiter, wobei die eben gezogene Kugel wieder gezogen werden kann. Weiter sind die einzelnen Züge stochastisch unabhängig. Damit bleibt die Wahrscheinlichkeit p, dass eine bestimmte Kugel gezogen wird, konstant. Offenbar ist die Anzahl der Züge bis zum ersten "Erfolg" geometrisch verteilt. □

4.6 Die Poisson-Verteilung

Vorbemerkung zur e-Funktion: Die e-Funktion oder Exponentialfunktion ist durch $f(x) = \exp(x) = e^x$ defininiert. Die Ableitung (der Differentialquotient) ist durch $f/dx = f'(x) = e^x$ gegeben, d.h. sie reproduziert sich bei Differentiation. Sie gehört damit zu den Funktionen, die sich ein eine Taylor-Reihe

$$f(x) = 1 + xf'(0) + \frac{x^2}{2!}f''(0) + \frac{x^2}{3!}f^{(3)} + \cdots = \sum_{k=0}^{\infty} \frac{x^k}{k!} \tag{4.29}$$

entwickeln lassen:

$$f(x) = e^x = \frac{1}{0!} + \frac{x}{1!} + \frac{x^2}{2!} + \cdots = \sum_{k=0}^{\infty} \frac{x^k}{k!} \tag{4.30}$$

Für e^{-x} erhält man

$$f(x) = e^{-x} = 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \cdots = \sum_{k=0}^{\infty} \frac{(-x)^k}{k!}. \tag{4.31}$$

Für hinbreichend kleine x erhält man daraus die nützlichen Approximationen $e^x \approx 1 + x$, $e^{-x} \approx 1 - x$.

Zur Poisson-Verteilung: In vielen empirischen Situationen zählt man die Häufigkeit des Auftretens eines bestimmten zufälligen Ereignisses A, ohne dass die Gesamtzahl der Beobachtungen von vornherein festgelegt ist. Dann kann die folgende Verteilung von Interesse sein

Definition 4.6.1 *Es sei X eine zufällige Veränderliche, die die Häufigkeit des Ereignisses A repräsentiert. Es gelte*

1. *Die verschiedenen Realisierungen von A sind unabhängig voneinander,*

2. *$p(A) = p$ für alle Realisierungen des Ereignisses*

3. *Die Wahrscheinlichkeit, dass $X = k$-mal das Ereignis A beobachtet wurden, ist durch*

$$P(X = k|\lambda) = e^{-\lambda}\,\frac{\lambda^k}{k!}, \tag{4.32}$$

gegeben. Dann heißt X Poisson-verteilt.

Wegen (4.30) sieht man sofort, dass

$$\lim_{k\to\infty} P(X = k|\lambda) = e^{\lambda}e^{-\lambda} = 1. \tag{4.33}$$

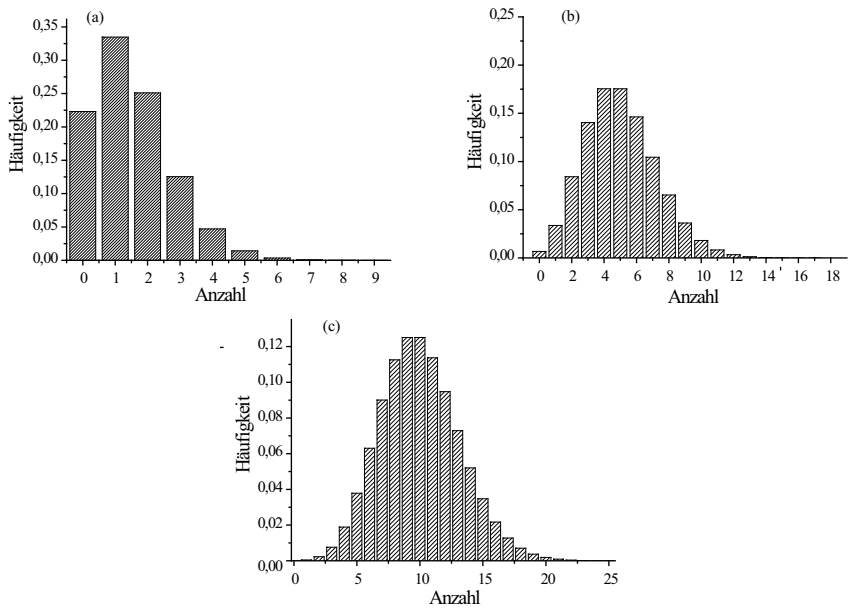
Der Erwartungswert und die Varianz der Verteilung sind durch

$$E(X) = \lambda \tag{4.34}$$

$$Var(X) = \lambda \tag{4.35}$$

gegeben (Beweis über die erzeugende Funktion s. unten).

Abbildung 4.3: Poisson-Verteilungen, (a) $\lambda = 1.5$, (b) $\lambda = 5.0$, (c) $\lambda = 10.0$



Die Verteilung (4.32) ist nicht aus den Annahmen 1. und 2. gefolgert worden, sondern wurde in 3. postuliert. In der Tat implizieren die ersten beiden Annahmen noch nicht den Ausdruck (4.32) für $P(X = k|\lambda)$, so dass dieser Ausdruck mit zur

Definition der Verteilung gehört. Er kann aber als Grenzwert der Binomialverteilung hergeleitet werden, wie Poisson[1] als erster zeigte, wenn man annimmt, dass (i) p klein ist, d.h. $p \ll 1$, und dass (ii) n groß ist ($n \to \infty$). Diese Annahmen impliziere, wie gleich gezeigt wird, den Ausdruck (4.32), aber dieser Ausdruck muß nicht bedeuten, dass in einer konkreten Anwendung die wahre Verteilung die Binomialverteilung ist, die durch (4.32) nur approximiert wird.

Zur Herleitung der Poisson-Verteilung aus der Binomialverteilung wird insbesondere angenommen, dass $p \to 0$ und $n \to \infty$ gelten derart, dass $pn \to \lambda$, wobei λ eine endliche reelle Zahl ist (vergl. z. B. Feller (1968), p. 153). Für die Binomialverteilung mit kleinem p hat man dann

$$P(K = 0|n, p) = (1 - p)^n = \left(1 - \frac{\lambda}{n}\right)^n,$$

und da $(1 - \lambda_n)^n \to e^{-\lambda}$ für[2] $n \to \infty$, folgt

$$P(K = 0|n, p) \approx e^{-\lambda}$$

für hinreichend großen Wert von n. Weiter rechnet man leicht nach, dass

$$\frac{P(K = k|n, p)}{P(K = k - 1|n, p)} = \frac{(n - k + 1)}{k(1 - p)} = 1 + \frac{(n + 1)p - k}{k(1 - p)} = \frac{\lambda - (k - 1)p}{k(1 - p)} \approx \frac{\lambda}{k}.$$

Hieraus folgt

$$\begin{aligned} P(X = 1|n, p) &= \lambda P(K = 0|n, p) \approx \lambda e^{-\lambda} \\ P(X = 2|n, p) &= \frac{1}{2} P(K = 1|n, p) \approx \frac{1}{2}\lambda^2 e^{-\lambda} \\ P(X = 3|n, p) &= \frac{1}{2 \cdot 3}\lambda^3 e^{-\lambda} \\ &\text{etc} \end{aligned}$$

so dass man induktiv

$$P(X = k|n, p) \approx e^{-\lambda}\frac{\lambda^k}{k!} \tag{4.36}$$

folgern kann.

Auf den ersten Blick mögen die Annahmen der Definition 4.6.1 einen seltenen Spezialfall darstellen. Man macht sich aber leicht klar, dass die Annahmen als gute Näherung für viele nahezu alltägliche Situationen aufgefaßt werden können. Man denke etwa an Situationen, in denen das Auftreten zufälliger, voneinander unabhängiger Ereignisse in der Zeit betrachtet wird. Ein Beispiel hierfür sind Unfälle pro Monat oder Jahr auf einem gegebenen Straßenabschnitt. Die Annahme der perfekten Unabhängigkeit ist sicherlich eine Idealisierung, denn Autofahrer,

[1] Siméon Denis Poisson, 1781-1840

[2] vergl. etwa Courant, R.: Vorlesungen über Differential- und Integralrechnung (1961), p.38

die gerade Zeuge eines Unfalls wurden, fahren anschließend vorsichtiger, allerdings verliert sich dieser Effekt bei den meisten relativ schnell, so dass die Annahme der Unabhängigkeit eine gute Approximation an die Wirklichkeit darstellen kann. Die Gesamtdauer des Beobachtungszeitraums sei T, die Wahrscheinlichkeit eines Unfalls im Intervall $(t, t + \Delta t)$ sei p, und $n\Delta t = T$. Für $\Delta t \to 0$ wird $p \to 0$ und gleichzeitig $n \to \infty$ gelten. Für die Häufigkeit von Unfällen im Zeitraum $(0, T)$ ist dann wichtig, *wie* der Wert von p mit Δt gegen Null strebt, d.h. gegen welchen Wert λ das Produkt np strebt: λ kann durchaus sehr groß sein. Ob es in solchen Situationen eine vernünftige Entscheidung ist, (4.32) zur "Erklärung" heranzuziehen, ist dann eine empirische Frage. Wenn die Wahrscheinlichkeiten der verschiedenen Autofahrer, einen Unfall zu verursachen, unterschiedlich sind, kann die Poisson-Verteilung keine gute Wahl sein, *obwohl* sie für jeden individuellen Fahrer die korrekte Verteilung sein kann. Die Fahrer unterscheiden sich dann durch ihre Parameter λ. Die Verteilung der Unfälle ist in diesem Fall durch eine *zusammengesetzte* Poisson-Verteilung gegeben (vergl. Abschnitt 4.9). Letzlich muß man empirisch entscheiden, ob (4.32) die adäquate Verteilung ist oder nicht.

Die erzeugende Funktion der Poisson-Verteilung: Für die erzeugende Funktion

$$A(s) = \sum_{k=0}^{\infty} p_k s^k$$

ergibt sich mit $p_k = e^{-\lambda}\lambda^k/k!$

$$A(s) = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda s)^k}{k!} = e^{-\lambda} e^{\lambda s}. \tag{4.37}$$

Für den Erwartungswert und die Varianz erhält man

$$\begin{aligned} A'(s) &= \lambda e^{-\lambda+\lambda s} \\ A''(s) &= \lambda^2 e^{-\lambda+\lambda s} \end{aligned}$$

und damit die Ausdrücke $E(X) = \lambda$, d.h.(4.34) für den Erwartungswert und $Var(X) = \lambda$, d.h. (4.35) für die Varianz, denn $A'(1) = \lambda$ und $A''(1) + A'(1) - (A'(1))^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$. □

Die Poisson-Verteilung spielt, wie die Binomial- und auch die geometrische Verteilung, in theoretischen und praktischen Anwendungen der Statistik eine hervorragende Rolle. Zunächst eine eher formale Eigenschaft der Poisson-Verteilung: da X auf der Menge aller4 natürlichenb Zahlen definiert ist, würde man intuitiv vermuten, dass $X = k$ mit geradem und ungeradem k mt Wahrscheinlichkeit 1/2 auftreten, – aber das ist nicht der Fall. Aber dies ist nicht der Fall:

Beispiel 4.6.1 Die Wahrscheinlichkeit, dass $X = k$ gerade bzw ungerade ist, ist durch

$$P_g = P(X = k|k \text{ gerade}, \lambda) = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^{2k}}{(2k)!}, \tag{4.38}$$

$$P_{ug} = P(X = x|k \text{ ungerade}, \lambda) = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^{2k+1}}{(2k+1)!} \tag{4.39}$$

gegeben. Dann ist

$$P_g - P_{ug} = e^{-\lambda} \sum_{k=0}^{\infty} \left(\frac{\lambda^{2k}}{(2k)!} - \frac{\lambda^{2k+1}}{(2k+1)!} \right).$$

Schreibt man den Ausdruck aus, so erhält man

$$\frac{\lambda^{2k}}{(2k)!} - \frac{\lambda^{2k+1}}{(2k+1)!} = 1 - \lambda + \frac{\lambda^2}{2!} - \frac{\lambda^3}{3!} + \frac{\lambda^4}{4!} - \frac{\lambda^5}{5!} = \sum_{k=0}^{\infty} \frac{(-\lambda^k}{k!} = e^{-\lambda},$$

nach (4.31), so dass

$$P_g - P_{ug} = e^{-\lambda} > 0, \quad \lambda < \infty. \tag{4.40}$$

Gerade Anzahlen sind also wahrscheinlicher als ungerade Anzahlen; die Differenz ist um so größer, je kleiner der Wert von λ ist. Für hinreichend kleine Werte von λ gilt insbesondere $e^{-\lambda} \approx 1 - \lambda$, d.h. die Wahrscheinlichkeit, dass eine gerade Anzahl beobachtet wird, geht gegen 1 mit $\lambda \to 0$.

□

Das folgende Beispiel soll nur eine einfache Anwendung illustrieren.

Beispiel 4.6.2 Ein erfahrener Psychotherapeut entnimmt seinen Unterlagen, dass er pro Jahr durchschnittlich drei Personen behandelt, die über Panikanfälle klagen. Er stellt fest, dass sich die Personen untereinander nicht kannten und vermutet daher, dass sie stochastisch unabhängig voneinander eine Neigung zu Panikanfällen entwickeln. Im Laufe eines bestimmten Jahres melden sich nun 6 miteinander nicht bekannte Personen bei ihm, die unter Panikanfällen leiden. Ist es vernünftig, von einer Panikepidemie zu sprechen?

Insgesamt sind, den Beobachtungen des Psychotherapeuten entsprechend, Personen mit Panikanfällen "seltene Ereignisse", denn man muß ihre Anzahl im Vergleich zur Gesamtzahl der Personen in der Population sehen. Zusammen mit der Annahme der stochastischen Unabhängigkeit führt dies zu der Hypothese, dass die Anzahl der unter Panikanfällen leidenden Personen pro Jahr Poisson-verteilt ist. Der Therapeut kann $\lambda = 3$ annehmen. Die Wahrscheinlichkeit, dass 6 oder mehr Personen mit Panikanfällen zum Therapeuten gehen, ist

$$P(X > 5) = 1 - P(X \leq 5)$$

Es ist

$$\begin{aligned}
P(X=0) &= e^{-3}\cdot 3^0/0! = e^{-3} = .04979, \quad P(X=1) = e^{-3}3^1/1! = .14936\\
P(X=2) &= e^{-3}3^2/2! = .22404, \quad P(X=3) = e^{-3}3^3/3! = .22404\\
P(X=4) &= e^{-3}3^4/4! = .16803\\
P(X=5) &= e^{-3}3^5/5! = .10082
\end{aligned}$$

Also ist $P(X \leq 5) = .91608$, so dass $P(X > 5) = 1 - .91608 = .0839$; — diese Wahrscheinlichkeit ist in der Tat gering. Vergleicht man sie mit dem Wert der Wahrscheinlichkeit für exakt $X = 6$, nämlich $P(X = 6) = e^{-3}\, 3^6/6! = .05041$, so findet man dass $P(X > 5)$ nur unwesentlich größer als $P(X = 6)$ ist, d.h. noch größere Anzahlen von Panikpatienten pro Jahr wird man nur sehr selten finden, wenn die Anzahl X zufällig, d.h. hier gemäß der Poisson-Verteilung, auftritt. Man kann argumentieren, dass die Anzahl von mehr als 5 Panikpatienten in einem Jahr so wenig wahrscheinlich ist - nämlich eben nur $P(X > 5) = .0839$, dass es nicht unvernünftig ist, die Hypothese einer Epidemie aufzustellen. Einen Beweis für eine Epidemie stellt der Wert von $X = 6$ aber noch nicht dar.

Es ist $E(X) = \lambda = 3$ und $Var(X) = \sigma^2 = \lambda = 3$, d.h. die Standardabweichung beträgt $\sigma = 1.732$; 5 Panikpatienten pro Jahr repräsentieren also ungefähr eine Abweichung von einer Standardabweichung vom Erwartungswert, 6 Patienten pro Jahr bedeuten eine Abweichung von etwas weniger als 2 Standardabweichungen vom Erwartungswert. □

4.7 Die Negative Binomialverteilung

Die Zufällige Veränderliche X ist geometrisch verteilt, wenn sie die Anzahl der Ereignisse $\bar{A}$ bis zum ersten Eintreffen von A repräsentiert, wobei $X = K$, wenn K die Anzahl der "Misserfolge" ($\bar{A}$) ist, oder $X = K + 1$, wenn der Versuch, bei dem zum ersten Mal A eintritt, mitgezählt wird. Dabei gilt $p = P(A)$ für alle Versuche, die überdies als stochastisch unabhängig vorausgesetzt werden. Die geometrische Verteilung kann verallgemeinert werden.

Definition 4.7.1 *Es sei $X = K + r$, wobei $r > 1$ eine festgelegte Anzahl von "Erfolgen" ist. Die zufällige Veränderliche X heißt dann* negativ binomialverteilt *mit den Parametern p und r. r heißt auch* stopping parameter.

Anmerkungen:

1. In bezug auf die negative Binomialverteilung wird auch von der *inversen Stichprobenentnahme* (*inverse sampling*) geredet. Während man ja üblicherweise eine Stichprobe vom Umfang n erhebt und dann nachschaut, wie groß die Anzahl r der "Erfolge" ist, wird ja bei der in Definition der negativen

Binomialverteilung beschriebenen Situation der Wert von n offengelassen und der von r festgelegt.

2. Die negative Binomialverteilung ist eine diskrete Verteilung für Wartezeiten: man beobachtet so lange, bis man r "Erfolge" beobachtet hat, und dieser Prozess endet, wenn bei einem Versuch die gewünschte Anzahl r von Erfolgen erreicht wurde; dieser Sachverhalt wird durch den Ausdruck 'stopping parameter' wiedergegeben.

□

Gesucht ist ein Ausdruck für die Wahrscheinlichkeit $P(X = n|p, r)$. Sicherlich ist $n \geq r$. Es werde Fall betrachtet, dass bereits $n - 1$ Versuche durchgeführt worden sind und dabei $(r - 1)$-mal das Ereignis A beobachtet wurde. Die Wahrscheinlichkeit dieses Falls ist durch die Binomial-Verteilung gegeben: Es sei Y die Häufigkeit, mit der das Ereignis A bei $n - 1$ Versuchen beobachtet wurde. Dann ist insbesondere

$$P(Y = r - 1|p, n - 1) = \binom{n-1}{r-1} p^{r-1}(1-p)^{n-1-(r-1)}.$$

Die Wahrscheinlichkeit, dass bei $n-1$ Versuchen gerade $(r-1)$-mal A beobachtet wurde *und* beim n-ten Versuch wieder das Ereignis A beobachtet wird, ist – wegen der Unabhängigkeit der Versuche – dann $P(Y = r - 1|p, n - 1)p$, so dass

$$P(X = n|p, r) = \binom{n-1}{r-1} p^r(1-p)^{n-r} \tag{4.41}$$

Für $r = 1$ geht dieser Ausdruck in den für die geometrische Verteilung über, da dann $n = k + 1$ und $\binom{n-1}{r-1} = (k + 1)!/(0!(k + 1)!) = 1$. Da $X = n = k + r$, kann man ebensogut nach der Wahrscheinlichkeit, dass $K = k$ ist fragen, denn K ist die eigentliche zufällige Veränderliche in diesem Experiment. K ist die Anzahl der Misserfolge, also der Ereignisse $\bar{A}$. Sicherlich gilt

$$P(K = k|p, r) = P(X = n|p, r),$$

und wegen der allgemeinen Beziehung

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

hat man

$$\binom{n-1}{r-1} = \binom{k+r-1}{r-1} = \frac{(k+r-1)!}{(r-1)!(k+r-1-(r-1))!} = \frac{(k+r-1)!}{(r-1)!k!} = \binom{k+r-1}{k}.$$

Substituiert man diesen Ausdruck in (4.41) und berücksichtigt noch, dass $n-r = k$, so erhält man

$$P(K = k|p, r) = \binom{k+r-1}{k} p^r(1-p)^k \tag{4.42}$$

Die Ausdrücke (4.41) und (4.42) werden gleichermaßen als Definition der negativen Binomialverteilung betrachtet.

Anmerkung: Es ist

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!}.$$

Auf $\binom{k+r-1}{k}$ angewendet erhält man wegen $-(r+k-1) = -r-k+1$, $-(r+k-1-1) = -r-k+2$ etc

$$\binom{k+r-1}{k} = \binom{-r}{k} = \frac{(-r)(-r-1)\cdots(-r-k+1}{k!}(-1)^k$$

und damit die Beziehung

$$P(K=k|p,r) = (-1)^k\binom{-r}{k}p^r(1-p)^k = (-1)^k\binom{-r}{k}p^r q^k = \binom{-r}{k}p^r(-q)^k \tag{4.43}$$

die den Ausdruck "Negative Binomialverteilung" erklärt. □

Erwartungswert und Varianz: Die Gesamtzahl K der Misserfolge ist die Summe stochastisch unabhängiger, geometrischer verteilter Variablen $K_1, \cdots, K_r$: K_1 ist die Anzahl der Versuche bis zum ersten Erfolg, K_2 die Anzahl der Versuche bis zum zweiten Erfolg, etc. Ist r die gewünschte Zahl der Erfolge (das Eintreten von A), so hat man also

$$K = K_1 + K_2 + \cdots + K_r \tag{4.44}$$

und

$$X = K + r. \tag{4.45}$$

Damit ist der Erwartungswert von X durch $E(X) = E(K+r) = E(K)+r$, denn r ist ja eine vorgegebene Konstante. Die zufälligen Veränderlichen K_j, $j = 1, \ldots, r$ sind jeweils geometrisch verteilt, und nach (4.26) und (4.27), Seite p. 122, sind der Erwartungswert von K_j durch $(1-p)/p$ und die Varianz von K_j durch $(1-p)/p^2$ gegeben. Demnach hat man

$$E(K) = \frac{r(1-p)}{p} \tag{4.46}$$

$$E(X) = E(K) + r = \frac{r - rp + rp}{p} = \frac{r}{p} \tag{4.47}$$

$$Var(X) = \frac{r(1-p)}{p^2} \tag{4.48}$$

Erzeugende Funktion: Wegen (4.44) und $X = K + r$, r eine Konstante ist die "wirkliche" zufällige Veränderliche K, wobei die K_j jeweils geometrisch verteilt

sind. Die erzeugende Funktion für die geometrische Verteilung ist nach (4.25)

$$A(s) = \frac{p}{1-qs}$$

und damit hat man sofort die erzeugende Funktion der negativen Binomialverteilung: da die Verteilung von K die r-fache Faltung der geometrischen Verteilung mit sich selbst ist, muß (vergl. Abschnitt 4.5, Gleichung (4.25), Seite 121)

$$A(s) = \left(\frac{p}{1-qs}\right)^r \tag{4.49}$$

die erzeugende Funktion für K sein. Hieraus folgen natürlich wieder der Erwartungswert und die Varianz von K: Es ist

$$A'(s) = k\left(\frac{p}{1-qs}\right)^{r-1}\frac{pq}{(1-qs)^2} \tag{4.50}$$

$$A''(s) = r(r-1)\left(\frac{p}{1-qs}\right)^{r-2}\left(\frac{pq}{(1-qs)^2}\right)^2 + r\left(\frac{p}{1-qs}\right)^{r-1}\frac{2pq^2}{1-qs)^2} \tag{4.51}$$

Aus (4.50) folgt dann

$$E(k) = A'(1) = r\left(\frac{p}{1-q}\right)^{k-1}\frac{pq}{(1-q)^2} = \frac{r(1-p)}{p} \tag{4.52}$$

Für $X = K + r$ erhält man den Erwartungswert $E(X) = E(k) + r$, d.h.

$$E(X) = \frac{r(1-p)}{p} + r = \frac{r - rp + rp}{p} = \frac{k}{p} \tag{4.53}$$

Die Varianz ergibt sich dann ebenfalls als die Varianz einer Summe geometrisch verteilter Variablen, d.h. nach (4.27), p. 122,

$$Var(X) = Var(K) = \frac{r(1-p)}{p^2}. \tag{4.54}$$

Poisson-Verteilung als Limit der negativen Binomialverteilung Für $r \to \infty$ und $p \to 1$ derart, dass der Erwartungswert von K (vergl. (4.46)) durch

$$\lambda = r\frac{1-p}{p} = \text{konstant} \tag{4.55}$$

gegeben ist, strebt die negative Binomialverteilung gegen eine Poisson-Verteilung. Der eleganteste Weg zu dieser Behauptung führt über die erzeugende Funktion der negativen Binomialverteilung (Feller (1968, p. 281)[3]. Die Gleichung (4.55)

[3]Man kann auch von (4.42) ausgehen, muß dann aber zeigen, dass $\binom{r+k-1}{k}/(r+\lambda)^k \to 1/k!$, was rechnerisch aufwändig ist.

impliziert dann

$$p = \frac{r}{r+\lambda} = \frac{1}{1+\lambda/r} \tag{4.56}$$

$$q = 1-p = 1 - \frac{1}{1+\lambda/r} = \frac{1}{1+r/\lambda} \tag{4.57}$$

Offenbar folgt dann

$$\lim_{r\to\infty} rq = \lim_{r\to\infty} \frac{r}{1+r/\lambda} = \lim_{r\to\infty} \frac{1}{1/r+1/\lambda} = \lambda, \tag{4.58}$$

d.h. für hinreichend großen Wert von r erhält man $q \approx \lambda/r$. Für die erzeugende Funktion (4.49) erhält man für hinreichend großen Wert von r

$$\left(\frac{1-q}{1-qs}\right)^r \approx \left(\frac{1-\lambda/r}{1-\lambda s/r}\right)^r,$$

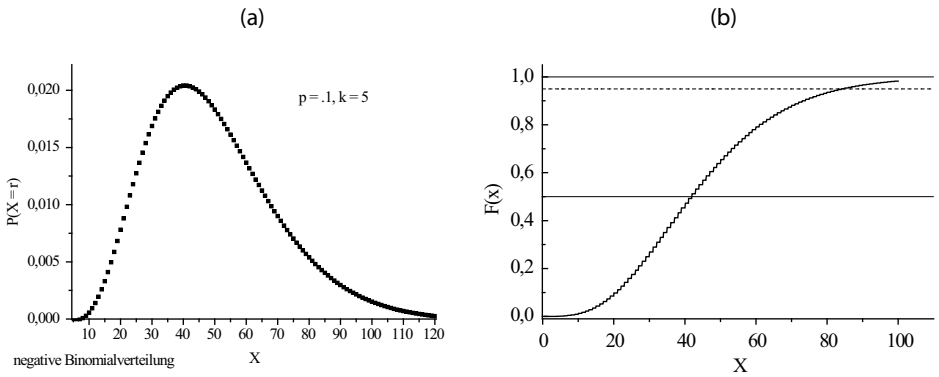
und mithin

$$\lim_{r\to\infty} \left(\frac{1-q}{1-qs}\right)^r = \lim_{r\to\infty} \frac{(1-\lambda/r)^r}{(1-\lambda s/r)^r} = \frac{e^{-\lambda}}{e^{-s\lambda}} = e^{-\lambda+s\lambda}. \tag{4.59}$$

Dies ist aber die erzeugende Funktion der Poisson-Verteilung (vergl. (4.37), Seite 128), womit die Behauptung, dass unter den angegebenen Bedingungen die negative Binomialverteilung gegen eine Poisson-Verteilung konvergiert, bewiesen ist. □

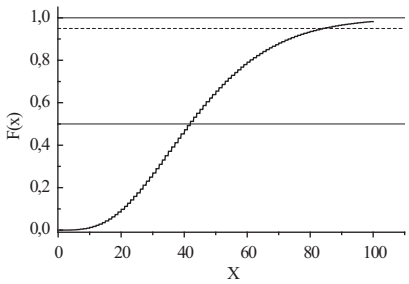
Beispiel 4.7.1 Ein Student benötigt für seine Diplomarbeit 5 Versuchsperso-

Abbildung 4.4: Negative Binomialverteilung, $p = .1$, $r = 5$, Wahrscheinlichkeitsverteilung



nen. Um diese zu finden, befragt er Kommilitonen – nicht notwendig aus seinem

Abbildung 4.5: Negative Binomialverteilung, $p = .1$, $r = 5$, Verteilungsfunktion



Fach, sondern irgendwelche zufällig ausgewählte Kommilitonen – aus der Universität, ob sie bereit sind, an seiner Untersuchung teilzunehmen. Da die Untersuchungen lange dauern und ermüdend sind, ist nur jeder zehnte Kommilitone dazu bereit. Es gibt 45 000 Studierende an der Universität, so dass die Wahrscheinlichkeit, bei zufälliger Auswahl einen Kommilitonen zu finden, der bereit ist, der Wissenschaft zu dienen, mit $p = 1/10 = .1$ angesetzt werden kann. Wegen der großen Zahl von Studierenden kann angenommen werden, dass die negative Binomialverteilung eine gute Näherung für die Verteilung der Anzahl von Studierenden, die insgesamt befragt werden müssen, bis man 5 Hilfswillige gefunden hat, ist. Die erwartete Anzahl von Befragten ist dann nach (4.52) $r(1-p)/p = 5 \times .9/.1 = 45$, die Varianz ist nach (4.54) $r(1-p)/p^2 = 45/.1 = 450$. Die Abbildungen 4.4 und 4.5 zeigen die Wahrscheinlichkeitsverteilung und die Verteilungsfunktion der negativen Binomialverteilung für $p = .1$, $k = 5$. Für den Erwartungswert $r = 45$ gilt $F(r) = .569$, d.h. die Wahrscheinlichkeit, dass die Anzahl der zu befragenden Kommilitonen kleiner oder gleich 45 ist, bis die 5 Vpn gefunden worden sind, beträgt .569. Mit der Wahrscheinlichkeit $p = .95$ muß man höchstens 84 Studierende befragen, bis man die 5 Vpn beieinander hat. □

4.8 Die hypergeometrische Verteilung

Definition 4.8.1 *Eine Menge Ω mit N Elementen enthalte M Elemente mit dem Merkmal $\mathcal{M}$ und $N - M$ Elemente die das Merkmal nicht aufweisen. Es wird eine Stichprobe mit $n \leq N$ Elementen aus Ω ohne Zurücklegen gebildet und es sei X die Anzahl der Elemente in der Stichprobe, die das Merkmal $\mathcal{M}$ zeigen. Dann heißt $P(X = k)$, $k = 0, 1, \ldots, n$ die* hypergeometrische Verteilung.

Die Verteilung ist durch

$$P(X = k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}, \tag{4.60}$$

und der Erwartungswert und die Varianz sind durch

$$E(X) = \frac{M \cdot n}{N} \tag{4.61}$$

$$Var(X) = \frac{Mn(N-M)(N-n)}{N^2(N-1)} \tag{4.62}$$

gegeben.

Beweis: Die Anzahl der möglichen Stichproben vom Umfang n ist gleich der Anzahl der möglichen Teilmengen von Ω. also ist ihre Anzahl gemäß (1.9) $\binom{N}{n}$. Die Anzahl der Mengen von genau k Elementen mit dem Merkmal $\mathcal{M}$ ist durch $\binom{M}{k}$ gegeben, und die Anzahl der Mengen mit genau $n-k$ Elementen, die das Merkmal nicht haben, ist $\binom{N-M}{n-k}$. Jede Menge mit k Elementen mit dem Merkmal $\mathcal{M}$ kann nun mit jeder Menge von $n-k$ Elementen mit dem Merkmal $\neg\mathcal{M}$ kombiniert werden, also gibt es genau $\binom{M}{k}\binom{N-M}{n-k}$ Stichproben vom Umfang n mit gerade k Elementen, die das Merkmal $\mathcal{M}$ haben. Alle diese Stichproben sind gleichwahrscheinlich, mithin folgt die Behauptung.

Der Beweis von (4.61) und (4.62) ist ein wenig länglich; Details findet man z.B. in Plachky et al. (1983), p. 150. □

Beispiel 4.8.1 Jan Rapp aus Vegesack fragt nach der Wahrscheinlichkeit, mit der er im Lotto k "Richtige" aus insgesamt 49 Zahlen gewählt hat, wobei $1 \leq k \leq 7$ ist.

Es ist $n = 7$, und insgesamt gibt es $M = 7$ aus $\Omega = \{1, 2, \cdots, 49\}$ "richtige" Zahlen (die am Wochenende öffentlich bestimmt werden); es ist hier also $n = M$. Es ist $|\Omega| = N = 49$. Damit gibt es dann $\binom{M}{k}$ Möglichkeiten, aus den M gerade k auszuwählen, und $\binom{N-M}{M-k}$ Möglichkeiten, $M-k$ "falsche" Zahlen auszuwählen. Das Ereignis, gerade k "Richtige" zu wählen, kann also auf $\binom{M}{k}\binom{N-M}{M-k}$ Weisen zustande kommen. Insgesamt gibt es $\binom{N}{M}$ Möglichkeiten, $M = 7$ Zahlen aus $N = 49$ auszuwählen. Also ist die Wahrscheinlichkeit, gerade k "Richtige" zu haben, durch

$$P(X = k) = \frac{\binom{7}{k}\binom{49}{7}}{\binom{49}{7}}$$

gegeben. Für $k = 3$ findet Jan Rapp, dass die Wahrscheinlichkeit gerade

$$\frac{35 \times 111930}{85900584} = .0456056$$

beträgt. Die Wahrscheinlichkeit, dass Jan Rapp auch einmal 7 "Richtige" hat, beträgt nur

$$P(7 \text{ Richtige}) = \frac{1}{1/\binom{49}{7}} = .0000000116414. \tag{4.63}$$

Wird er durch das Lottospiel je seine Rente aufbessern können? Die erwartete Anzahl "richtiger" Zahlen ist nach (4.61)

$$E(X) = \frac{Mn}{N} = \frac{7 \times 7}{49} = 1,$$

Jan Rapp wird also im Durchschnitt 1 Zahl korrekt wählen, bei einer Varianz von, nach (4.62),

$$\frac{Mn(N-M)(N-n)}{N^2(N-1)} = \frac{7 \times 7(49-7)(49-7)}{49^2(49-1)} = .75.$$

□

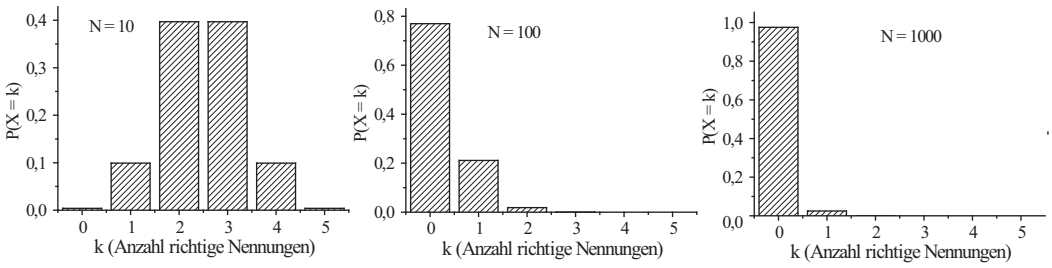
Beispiel 4.8.2 In Beispiel 1.3.4 wurde ein Experiment beschrieben, bei der eine Vp 5 Blumennamen erinnern sollte, die in einer vorgelesenen Geschichte genannt worden waren. Wie groß ist die Wahrscheinlichkeit, dass sie gerade $0 \leq k \leq 5$ tatsächlich vorgekommene Blumennamen nennt, wenn sie (i) die Namen zufällig aus den insgesamt N Blumennamen, die sie in ihrem Langzeitgedächtnis hat, auswählt?

Es ist also $M = 5$, und die Wahrscheinlichkeit, $0 \leq k \leq M$ Namen korrekt, wenn auch rein zufällig zu nennen, beträgt

$$P(X = k) = \frac{\binom{5}{k}\binom{N-5}{5-k}}{\binom{N}{5}}$$

Abb. 4.6 zeigt die Verteilungen für die Anhzahl k richtig genannter Namen für

Abbildung 4.6: Hypergeometrische Verteilungen für richtige Nennungen



(i) $N = 10$,(ii) $N = 100$ und (iii) $N = 1000$ gespeicherte Blumennamen. □

Beispiel 4.8.3 Ein Dozent für statistische Verfahren soll aus einer Gruppe von zwölf Bewerbern vier aussuchen, die die Tutorien für seine Statistik-Vorlesung

übernehmen sollen. Ein Bewerber ist entweder geeignet (A) oder nicht ($\bar{A}$). Erfahrungsgemäß sei nur ein Drittel der Bewerber für die Aufgabe geeignet. Wie groß ist die Wahrscheinlichkeit, dass er mindestens 3 geeignete Tutoren bekommt, wenn er vier Tutoren nach dem Zufall (jede Bewerber hat die gleiche Chance, gewählt zu werden) aus den 12 Bewerbern auswählt?

Es ist $P(\text{mindestens 3 geeignete Bewerber}) = P(X = 3 \cup X = 4) = P(X = 3) + P(X = 4)$, und $P(X = 3)$, $P(X = 4)$ sind durch die hypergeometrische Verteilung gegeben. Dann ist $N = 12$, $n = 4$, $M = 4$

$$\begin{aligned} P(X = 3) &= \frac{\binom{4}{3}\binom{12-4}{4-3}}{\binom{12}{4}} = \frac{4 \cdot 8}{495} = .06465 \\ P(X = 4) &= \frac{\binom{4}{4}\binom{12-4}{4-4}}{\binom{12}{4}} = \frac{1 \cdot 8}{495} = .01616 \end{aligned}$$

und

$$P(X = 3) + P(X = 4) = .08081.$$

□

Beispiel 4.8.4 (Blut, Aids, und Stichprobenumfänge.) Gegeben sei eine Menge von N Blutkonserven. Sie enthalte $M \geq 0$ HIV-positive Elemente. M ist unbekannt und kann als zufällige Veränderliche aufgefaßt werden. Es wird eine Stichprobe vom Umfang n gebildet und man findet, dass sie *kein* HIV-positives Element enthält.

1. Wie sicher kann man sein, dass die Population kein positives Element enthält, also $M = 0$ ist?

2. Wie groß muß n sein, damit man mit vorgegebener Wahrscheinlichkeit sagen kann, dass die Population kein positives Element enthält?

Natürlich gilt generell, dass man überhaupt nicht sicher sein kann, dass die Menge nicht mindestens ein HIV-positives Element enthält, so lange man nicht alle Elemente untersucht hat. Andererseits zeigt sich, dass die Art und Weise, in der die Wahrscheinlichkeit, dass kein solches Element in der Menge enthalten ist, als Funktion von n gegen 1 geht, von der Annahme über die Verteilung von M abhängt; dies soll im Folgenden illustriert werden.

Es sei X die Anzahl der positiven Elemente in der Stichprobe. Die Wahrscheinlichkeit, dass X den Wert k, $0 \leq k \leq n$ annimmt, ist durch die hypergeometrische Verteilung

$$P(X = k) = \frac{\binom{M}{k}\binom{N-M}{n}}{\binom{N}{n}} \tag{4.64}$$

gegeben (sampling without replacement). Man kann $P(X = k)$ als bedingte Wahrscheinlichkeit auffassen:

$$P(X = k) = P(X = k|M, n) \tag{4.65}$$

Die Gleichung (4.64) gilt also nur für den zunächst zufälligen, aber dann festgehaltenen Wert von M. n ist nicht zufällig, stellt aber gleichwohl eine Bedingung dar. Der Spezialfall $k = 0$, also $P(X = 0|M, n) \neq 0$ ist also nur möglich, wenn $n < N - M$ gilt. Man kann nun

$$P(M|X = 0, n) = \frac{P(M \cap X = 0|n)}{P(X = 0|n)} \tag{4.66}$$

betrachten. Für $M = 0$ ist $P(M = 0|X = 0, n)$ eine Funktion von n und liefert eine Antwort auf die Fragen 1 und 2.

Es ist

$$P(M \cap X = 0|n) = \frac{P(X = 0|M, n)P(M, n)}{P(X = 0|n)}$$

$P(M, n)$ kann als *a-priori-Wahrscheinlichkeit* für den Wert von M aufgefaßt werden. Da notwendig $n < N - M$ vorausgesetzt werden muß (der Fall $X = 0$ kann sonst nicht auftreten), folgt für gewählten Wert von n die Bedingung $M < N - n$, d.h. M kann maximal den Wert $N - n - 1$ annehmen. Die a-priori-Verteilung für die möglichen Werte von M ist also auf $[0, N - n - 1]$ definiert. Jedenfalls gilt dann für (4.66)

$$P(M|X = 0, n) = P(X = 0|M, n)\frac{P(M, n)}{P(X = 0|n)} \tag{4.67}$$

Es muß $P(X = 0|n)$ bestimmt werden. Nach dem Satz von der totalen Wahrscheinlichkeit gilt aber

$$P(X = 0|n) = \sum_{i=0}^{N-n-1} P(X = 0|M = i, n)P(M = i) \tag{4.68}$$

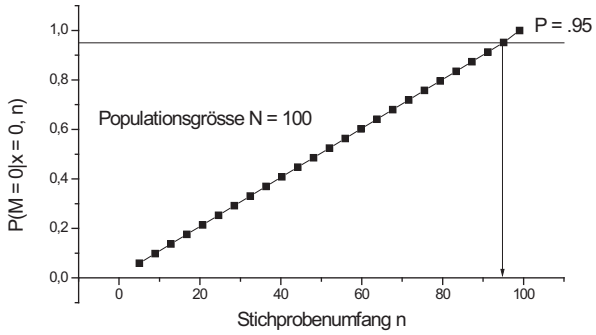
Man muß also eine Annahme über die a-priori-Wahrscheinlichkeiten $P(M = i)$ machen. Es gibt (mindestens) zwei Möglichkeiten:

Die Gleichverteilungsannahme

Nimmt man an, dass gar nichts über diese Wahrscheinlichkeiten bekannt ist, so kann man annehmen, dass die Werte von M im Bereich von 0 bis $M - n - 1$ gleichverteilt sind, so dass

$$P(M = i) = \frac{1}{M - n}, \quad 0 \leq i \leq M - n - 1 \tag{4.69}$$

Abbildung 4.7: Wahrscheinlichkeiten für $M = 0$ für verschiedene Stichprobenumfänge n, M binomialverteilt



Dann folgt

$$P(X = 0|n) = \frac{1}{M-n} \sum_{i=0}^{N-n-1} P(X = 0|M = i, n) \tag{4.70}$$

Jetzt kann man (4.66) anschreiben:

$$\begin{aligned} P(M|X = 0, n) &= P(X = 0|M, n) \frac{P(M, n)}{P(X = 0|n)} \\ &= \frac{\binom{N-M}{n}}{\binom{N}{n} \sum_{i=0}^{N-n} P(X = 0|M = i, n)}, \end{aligned} \tag{4.71}$$

denn

$$\binom{M}{0} = 1$$

und für $M = 0$ erhält man

$$P(M = 0|X = 0, n) = \frac{\binom{N}{n}}{\binom{N}{n} \sum_{i=0}^{N-n} \binom{N-i}{n} / \binom{N}{n}} = \frac{\binom{N}{n}}{\sum_{i=0}^{N-n} \binom{N-i}{n}} \tag{4.72}$$

$P(M = 0|X = 0, n)$ ist eine Funktion von n; man berechnet sie für alle Werte von n, bis die gewünschte Wahrscheinlichkeit erreicht wird. Das Bemerkenswerte an diesem Ansatz ist, dass $P(M = 0|X = 0, n)$ linear in n ist.

Die Binomialverteilungsannahme

Eine andere Annahme über $P(M, n)$ ergibt sich, wenn man davon ausgeht, dass ein bestimmter Anteil p der blutspendenden Bevölkerung HIV-positiv ist. Die Anzahl der HIV-positiven Blutspenden ist dann binomialverteilt, und zwar, wegen der Voraussetzung $X = 0$, in der Menge der nicht in der Stichprobe enthaltenen Spenden. Also gilt für die a-priori-Wahrscheinlichkeit $P(M, n)$

$$P(M, n) = \binom{N}{M} p^M (1-p)^{N-M}, \quad M < N - n \tag{4.73}$$

Dementsprechend gilt dann für $P(M|X=0,n)$

$$P(M|X=0,n) = \frac{\binom{N-M}{n}\binom{N}{M}p^M(1-p)^{N-M}}{\binom{N}{n}\sum_{i=0}^{N-M-1} P(X=0|i,n)\binom{N}{i}p^i(1-p)^{N-i}} \tag{4.74}$$

Für $M=0$ vereinfacht sich dieser Ausdruck zu

$$P(M=0|X=0,n) = \frac{(1-p)^N}{\sum_{i=0}^{N-M-1} P(X=0|i,n)\binom{N}{i}p^i(1-p)^{N-i}} \tag{4.75}$$

Dieser Ausdruck hängt nicht mehr nur von n, sondern vom Parameter p ab. Man braucht also eine gute Schätzung für p.

In Abbildung 4.7 werden die Wahrscheinlichkeiten für die Parameter $p = .05$ und $p = .01$ dargestellt, d.h. für einen Anteil von 5% bzw. für einen Anteil von 1% von HIV-Positiven in der Population. Sicher kann man – trivialerweise – stets nur sein, wenn man die gesamte Population testet. Für $p = .05$ ist die Unsicherheit allerdings wesentlich größer als für $p = .01$. □

4.9 Zusammengesetzte Verteilungen

In vielen Untersuchungen wird die Anzahl n der Versuche von vornherein festgelegt. Handelt es sich um Bernoulli-Versuche, d.h. ist $X_i = 1$ mit Wahrscheinlichkeit p oder $X_i = 0$ mit der Wahrscheinlichkeit $1-p$, so ist die Anzahl $k = X_1 + \cdots + X_n$ der "Erfolge" binomialverteilt, und wir wissen, dass $\hat{p} = k/n$ im Durchschnitt gleich dem oft unbekannten Parameter p der Binomialverteilung ist, d.h. $E(\hat{p}) = p$ (man erinnere sich: es gilt ja allgemein $E(\overline{x}) = E(X)$).

In anderen Situationen ist aber die Anzahl n nicht von vornherein bestimmt. So liegt gelegentlich der *Zeitraum* einer Untersuchung, nicht aber die *Anzahl* der Beobachtungen fest. Die Anzahl n der Beobachtungen ist zufällig. Gesucht ist die Verteilung der Anzahl k von Beobachtungen, bei denen ein bestimmtes Ereignis registriert wird. So kann man etwa einen Monat lang die Einweisungen in eine Klinik registrieren, deren Anzahl sei n. n ist sicherlich eine zufällige Veränderliche, deren Wert von Monat zu Monat schwankt. k sei die Anzahl der Personen, die wegen einer bestimmten Krankheit eingewiesen werden. Die Verteilung des Quotienten k/n hängt jetzt nicht nur von der zufälligen Veränderlichen k, sondern auch von der der zufälligen Veränderlichen n ab.

Man fragt damit nach der Verteilung der Summe $S_n = X_1 + \cdots + X_n$, wenn n selbst eine zufällige Variable ist. Es wird zunächst ein allgemeiner Ausdruck für die Verteilung von S_n hergeleitet. N sei die zufällige Veränderliche, deren spezielle Werte durch n repräsentiert werden. Dann ist die Wahrscheinlichkeit, dass $X = X_1 + \cdots + X_n$ den Wert k annimmt, durch

$$P(\cup_n\{X = k \cap N = n\}) \quad = \quad \sum_{n=1}^{\infty} P(X = k \cap N = n)$$

$$= \sum_{n=1}^{\infty} P(X = k|N = n)P(N = n) \qquad (4.76)$$

gegeben. Die Verteilung hängt jetzt offenbar von der Verteilung $P(X = k|N = n)$ einerseits und von der Verteilung $P(N = n)$ andererseits ab.

Definition 4.9.1 *Die durch (4.76) gegebene Verteilung heißt* zusammengesetzte Verteilung *(engl.: compound distribution).*

Haben die X_i die Wahrscheinlichkeitsverteilung f, so ist für festes n die Verteilung von $X = X_1 + \cdots + X_n$ durch die n-fache Faltung f^n der f gegeben. Gesucht ist nun die erzeugende Funktion für eine zusammengesetzte Verteilung. Nach (4.76) ist der allgemeine Ausdruck für die erzeugende Funktion durch

$$A(s) = \sum_{k=0}^{\infty} P(X = k)s^k = \sum_{k=0}^{\infty}\sum_{n=1}^{\infty} P(X = k|N = n)P(N = n)s^k \qquad (4.77)$$

gegeben. Setzt man

$$A_n(s) = \sum_{k=0}^{\infty} P(X = k)$$

für die erzeugende Funktion für X bei festem n, so erhält man durch Umordnen der Terme

$$\begin{aligned} A(s) &= \sum_{n=1}^{\infty} P(N = n) \sum_{k=0}^{\infty} P(X = k|N = n)s^k \\ &= \sum_{n=1}^{\infty} P(N = n|A_n(s)) \qquad (4.78) \end{aligned}$$

Da $P(N = n)$ eine Wahrscheinlichkeitsverteilung ist, existiert dafür wiederum eine erzeugende Funktion. Sie sei durch

$$B(s) = \sum_{n=0}^{\infty} P(N = n)s^n \qquad (4.79)$$

gegeben. Nun kann $B(s)$ in eine Taylor-Reihe entwickelt werden:

$$B(s) = B(0) + sB'(0) + \frac{s^2}{2!}B''(0) + \cdots$$

Es sei weiter $f(s)$ die erzeugende Funktion der X_i. Ersetzt man in $B(s)$ die Variable s durch $f(s)$, so erhalten man

$$B(f(s)) = B(0) + f(s)B'(s) + \frac{f^2(s)}{2!}B''(0) + \frac{f^3(s)}{3!}B'''(0) + \cdots$$

Nun ist f^2 die erzeugende Funktion der 2-fachen Faltung der Verteilung von X_i, f^3 die erzeugende Funktion der 3-fache Faltung, etc., d.h. aber $f^2(s) = A_2(s)$, $f^3(s) = A_3(s)$, etc. Damit folgt aber

$$\sum_{n=1}^{\infty} P(N = n)A_n(s) = B(f(s)) = A(s) \tag{4.80}$$

Aufgrund dieser Beziehung kann die erzeugende Funktion einer zusammengesetzten Verteilung u. U. leicht gefunden werden.

Beispiel 4.9.1 (Fortsetzung von Beispiel 4.6.2) Um die erzeugende Funktion für die in (4.76) gegebene Verteilung zu finden, wir die erzeugende Funktion (i) der geometrischen Verteilung und (ii) die der Binomialverteilung benötigt. für die Bernoulli-Variablen ist aber $A_n(s) = (q + ps)$, $q = 1 - p$. Für die geometrische Verteilung ist nach (4.25)

$$B(s) = \frac{p_0}{q_0(1 - q_0 s)}$$

Um $A(s)$ zu finden, muß hierin s durch $A_n(s)$ ersetzt werden:

$$A(s) = \frac{p_0}{q_0(1 - q_0(q + ps))},$$

woraus sich dann

$$E(X) = \frac{p}{p_0}$$

ergibt. □

Einen Spezialfall erhält man, wenn $P(N = n)$ durch die Poisson-Verteilung gegeben ist:

Definition 4.9.2 *Es sei $P(N = n)$ in (4.76) durch die Poisson-Verteilung gegeben, dh es sei $P(N = n) = e^{-\lambda}\lambda^n/n!$. Die erzeugende Funktion der Poisson-Verteilung ist nach (4.37) $A(s) = \exp(-\lambda + \lambda s)$. Ersetzt man nun s durch die erzeugende Funktion $f(s)$ der X_i, so erhält man*

$$A(s) = e^{-\lambda + \lambda f(s)}. \tag{4.81}$$

Die zu dieser erzeugenden Funktion korrespondierende Wahrscheinlichkeitsverteilung heißt zusammengesetzte Poisson-Verteilung.

Beispiel 4.9.2 Die X_i seien Bernoulli-verteilt mit dem Parameter p, d.h. $P(X_i = 1) = p$, $P(X_i = 0) = 1 - p$. Die erzeugende Funktion für diese Verteilung ist $f(s) = ps$. Es sei N Poisson-verteilt; dann ist die erzeugende Funktion Für X durch

$$A(s) = e^{-\lambda + \lambda ps} \tag{4.82}$$

gegeben. Dies ist die erzeugende Funktion der Poisson-Verteilung mit dem Erwartungswert $E(X) = \lambda p$ und der Varianz $Var(X) = \lambda p$; die Summe (d.h. die Anzahl der "Erfolge") $X = X_1 + \cdots + X_n$ mit $X_i = 0$ oder $X_i = 1$ ist also Poisson-verteilt! □

Die Restriktion $X_i = 0$ oder 1 ist nicht notwendig, damit die Summe X zusammengesetzt Poisson-verteilt ist. Haben die X_i die erzeugende Funktion $f(s)$, so ist der Erwartungswert der Summe X durch $E(X) = A'(1)$ gegeben. Mit $A(s) = \exp(-\lambda + \lambda f(s))$ erhält man $A'(s) = \lambda f'(s)A(s)$ und damit allgemein

$$E(X) = \lambda f'(1)A(1). \tag{4.83}$$

die erzeugende Funktion $f(s)$ wird den Parameter λ im allgemeinen nicht enthalten, er charakterisiert ja die Verteilung für die Anzahl N der Varialen X_i. Der Erwartungswert für dieSumme X hängt dann im allgemeinen von λ ab. Dies bedeutet, dass bei einer Stichprobenbildung, bei der das N zufällig bestimmt wird, der Erwartungswert von $X = X_1 + \cdots + X_n$, $N = n$ durch den Parameter des Sampling-Schemas mitbestimmt wird.

4.10 Produkte und Quotienten

In vielen Anwendungen ist die Verteilung von Produkten und Quotienten von zufälligen Veränderlichen von Interesse. Ein spezielles Problem ist die Schätzung des Parameters p einer Binomialverteilung; ist die Gesamtzahl N der Versuche bzw. der Beobachtungen festgelegt, so ist k/N eine Schätzung. Die Anzahl N der Versuche kann aber ebenfalls zufällig sein. Man muß sich dann fragen, ob k/N immer noch eine gute Schätzung für p ist.

4.10.1 Die Verteilung eines Quotienten von natürlichen Zahlen

Es seien U und V zwei diskrete zufällige Veränderliche[4], und es sei M_u die Menge der möglichen Werte von U, und M_v sei die Menge der möglichen Werte von V. Gesucht sei die Verteilung der zufälligen Veränderlichen $R = U/V$. Es ist dann $U = RV$; damit hat man

$$P(R = r) = \sum_{v \in M_v} P(U = rv \cap V = v) \tag{4.84}$$

und wegen $P(U = rv \cap V = v) = P(U = rv|V = v)P(V = v)$ erhält man

$$P(R = r) = \sum_{v \in M_v} P(U = rv|V = v)P(V = v) \tag{4.85}$$

wobei natürlich $P(U = rv|V = v) = 0$ für $rv \notin M_u$. Sind U und V stochastisch unabhängig, so ist $P(U = rv|V = v) = P(U = rv)$, $rv \in M_u$.

[4] $\mathbb{N}$ bezeichnet die Menge der natürlichen Zahlen.

Beispiel 4.10.1 Ein Verkehrsforscher ist an der Anzahl alkoholbedingter Unfälle interessiert. Dazu betrachtet er die Anzahl der Unfälle in einer Stadt an einem bestimmten Tag und zählt aus, bei wievielen der Unfälle Alkohol eine Rolle gespielt hat. Er nimmt an, dass für gegebene Zahl V von Unfällen die Anzahl alkoholbedingter Unfälle binomialverteilt ist, und dass die Anzahl V wiederum Poisson-verteilt ist.

U und V sind hier nicht stochastisch unabhängig. Gesucht ist die Verteilung von $R = U/V$. Es ist $M_u = \{1, \cdots, V\}$, $M_v = \{1, 2, \cdots\}$. Nach (4.9.96) ist mit $r = U/v$ und $U = rv$

$$P(U = rv|V = v)P(V = v) = \begin{cases} e^{-\lambda}\binom{n}{rv}p^{rv}(1-p)^{v-rv}\frac{\lambda^v}{v!}, & rv \in \mathbb{N} \\ 0, & rv \notin \mathbb{N} \end{cases} \tag{4.86}$$

Für den Erwartungswert erhält man

$$E(R) = e^{-\lambda}\int_0^1 r\binom{n}{rv}p^{rv}(1-p)^{v-rv}\frac{\lambda^v}{v!}dr, \tag{4.87}$$

wobei der Integrand natürlich nur für $0 \neq rv \in \mathbb{N}$ von Null verschieden ist. Offenbar ist im allgemeinen $E(R) \neq p$ und von λ abhängig. □

4.10.2 Die Verteilung eines Produktes von natürlichen Zahlen

Die Verteilung des Produktes zweier diskreter zufälliger Veränderlicher ergibt sich leicht wie folgt: $\{R = r\}$ genau dann, $\{U = u \cap V = r/u\}$; mit ist

$$\begin{aligned} P(R = r) &= \sum_{u=1}^{\infty} P(U = u \cap V = r/u) \\ &= \sum_{u=1}^{\infty} P(U = u|V = r/v)P(V = r/v) \end{aligned} \tag{4.88}$$

für $r \geq 1$ und $P(R = 0) = P(U = 0 \cup V = 0 \cup (U = 0 \cap V = 0))$.

Beispiel 4.10.2 U und V seien stochastisch unabhängig und Poisson-verteilt. Dann hat das Produkt $r = UV$ die Verteilung

$$P(R = r) = e^{-\lambda}e^{-\mu}\sum_{u=0}^{\infty}\frac{\lambda^u\mu^{r/u}}{u!(r/u)!} \tag{4.89}$$

□

Kapitel 5

Beispiele für stetige Verteilungen

5.1 Die Gleichverteilung

Definition 5.1.1 *Es sei $[a,b] \subset \mathbb{R}$ ein abgeschlossenes Intervall, d.h. a und b gehören zum Intervall. X sei eine zufällige Veränderliche derart, dass für jedes $\omega \in \Omega$, $X(\omega) \in [a,b]$. Es werde angenommen, dass die ω alle mit gleicher Wahrscheinlichkeit auftreten; dementsprechend soll für die Dichte $f(x)$*

$$f(x) = \frac{1}{b-a} \tag{5.1}$$

für $x \in [a,b]$, $f(x) = 0$ für $x \notin [a,b]$ gelten. Dann heißt X auf $[a,b]$ gleichverteilt.

Satz 5.1.1 *Es gelten die folgenden Beziehungen:*

$$F(x) = P(X \leq x) = \begin{cases} 0, & x \leq a \\ (x-a)/(b-a), & a < x \leq b \\ 1, & x > b \end{cases} \tag{5.2}$$

$$E(X) = \frac{a+b}{2} \tag{5.3}$$

$$Var(X) = \frac{(b-a)^2}{12} \tag{5.4}$$

$$Median = E(X) \tag{5.5}$$

Beweis: Es ist

$$F(x) = \int_{-\infty}^{x} f(\xi)d\xi = \frac{1}{b-a}\int_{a}^{x} d\xi = \frac{x-a}{b-a}$$

also (5.2). Für den Erwartungswert erhält man

$$\begin{aligned} E(X) &= \frac{1}{b-a}\int_{-\infty}^{\infty}\xi d\xi = \frac{1}{b-a}\int_a^b \xi d\xi \\ &= \frac{1}{2}\frac{b^2-a^2}{b-a} = \frac{1}{2}\frac{(a+b)(a-b)}{b-a} = \frac{1}{2}(a+b). \end{aligned}$$

Für die Varianz muß $E(X^2) = \int_a^b \xi^2 f(\xi)\xi$ bestimmt werden. Es ist

$$E(X^2) = \frac{1}{b-a}\int_a^b \xi^2 d\xi = \frac{1}{3}\frac{b^3-a^3}{b-a}$$

und somit ist

$$Var(X) = \frac{1}{3}\frac{b^3-a^3}{b-a} - \frac{(a+b)^2}{4} = \frac{4(b^3-a^3)-(a+b)^2(b-a)}{12(b-a)}.$$

Nun ist $(b-a)^3 = b^3 - a^3 - 3(a^2b - ab^2)$, wie man durch Ausrechnen bestätigt, und wenn man noch $(a+b)^2(b-a)$ ausmultipliziert, so erkennt man, dass der Zähler gerade gleich $(b-a)^3$ wird; die Behauptung folgt dann sofort. Die Aussage (5.5) folgt wegen der Definition des Medians sofort aus $E(X) = (a+b)/2$. □

Von einem intuitiven Standpunkt aus gesehen drückt die Annahme der Gleichverteilung der X-Werte einen vollständigen Mangel an Wissen über das Eintreffen bestimmter Werte von X aus. Aber es kann auch eine Eigenschaft des betrachteten Mechanismus' sein, dass keine Region des Intervalles [a, b] "bevorzugt" wird.

Beispiel 5.1.1 Bei der Berechnung des Medians bei gruppierten Daten wird angenommen, dass die Meßwerte innerhalb eines Intervalles gleichverteilt sind. □

Der Ansatz, "reine Zufälligkeit" über die Gleichverteilung auszudrücken, setzt aber die Kenntnis der Intervallgrenzen a und b voraus; a und b sollten überdies *endlich* sein, da sonst die Dichte identisch Null, der Erwartungswert aber - ebenso wie die Varianz - unendlich wird. Der Begriff der "reinen Zufälligkeit" ist überdies nicht so klar, wie es auf den ersten Blick scheinen mag. Dies wird besonders deutlich, wenn man z.B. Wartezeiten betrachtet:

Beispiel 5.1.2 Bekanntlich hat Kurt seiner Freundin Lisa versprochen, sie zwischen 14^{00} und 15^{00} anzurufen (vergl. Beispiel **??**), p. **??**. Lisa nimmt an, dass Kurt jeden Zeitpunkt dieser Stunde mit gleicher Wahrscheinlichkeit für einen Anruf wählt. Dementsprechend ist etwa die Wahrscheinlichkeit, dass Kurt zwischen 14^{30} und 14^{40} anruft, durch

$$P(\tau \in (14^{30}, 14^{40}]) = 10/60 \approx .167$$

gegeben; τ ist die Wartezeit. Lisa hat nun bis 14^{30} gewartet und fragt sich *jetzt*, sie groß die Wahrscheinlichkeit ist, dass Kurt im Intervall $(14^{30}, 14^{40}]$ anruft. Sie

fragt damit nach der bedingten Wahrscheinlichkeit $P(\tau \in (14^{30}, 14^{40}] | \tau > 14^{30})$. Es ist

$$P(\tau \in (14^{30}, 14^{40}] | \tau > 14^{30}) = \frac{\tau \in (14^{30}, 14^{40}]}{P(\tau > 14^{30})} = \frac{10}{60 \times 30/60} \approx .333$$

Nehmen wir an, Kurt habe bis 14^{50} nicht angerufen. Die *a priori* Wahrscheinlichkeit, dass er zwischen 14^{50} und 15^{00} anruft, ist wieder $\approx .167$. Andererseits ist

$$P(\tau \in (14^{50}, 15^{00}] | \tau > 14^{50}) = \frac{10}{60 \times 10/60} = 1$$

was natürlich auch ohne Rechnung klar ist, denn wenn Kurt his 14^{50} nicht angerufen hat, dann *muß* er in den restlichen 10 Minuten anrufen, sofern er sich an seine Ankündigung hält. Die zufälligen Mechanismen, die den Zeitpunkt seines Anrufs bestimmen, müssen den Zeitpunkt also entweder vor Beginn der Stunde $[14^{00}, 15^{00}]$ festlegen, oder aber ganz bestimmten Gesetzmäßigkeiten gehorchen, wenn der Zeitpunkt erst während der Stunde gewählt wird. Diese Gesetzmäßigkeiten müssen von der Art sein, dass die Wahrscheinlichkeit des Anrufs im Intervall $(t, 15^{00}]$ in einer genau definierten, vom Wert von t abhängenden Weise größer wird, wenn der Anruf bis zum Zeitpunkt t *noch nicht* eingetreten ist. □

Es liegt also nahe, zu vermuten, dass Wartezeiten kaum gleichverteilt sein werden; betrachtet man etwa wie in Beispiel 4.6.7 die Zeit, die ein Autofahrer bis zur Reaktion benötigt, wenn plötzlich, d.h. zur Zeit a, ein Hindernis auftaucht, so kann man kaum eine sinnvolle obere Grenze b angeben derart, dass der Fahrer innerhalb des Intervalles $(a, b]$ mit Sicherheit reagiert. Die Eigenheiten der Gleichverteilung legen nahe, noch andere Modelle für "rein" zufällige Ereignisse zu betrachten; ein wichtiges Modell wird durch die Exponentialverteilung beschrieben, die jetzt besprochen werden soll.

5.2 Die Exponentialverteilung

5.2.1 Definition

Definition 5.2.1 *Die zufällige Veränderliche X sei auf $[0, \infty)$ definiert mit der Dichte*

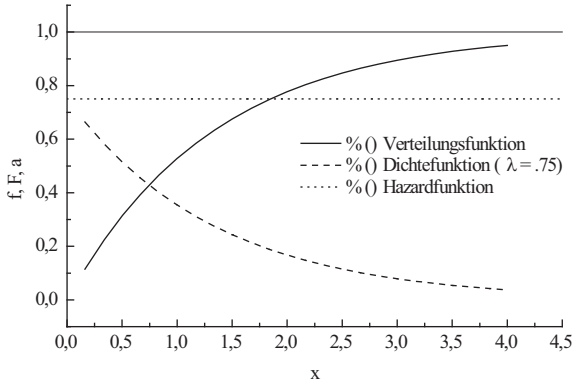
$$f(x) = \lambda e^{-\lambda x}, \quad 0 \leq x < \infty. \tag{5.6}$$

Dann heißt X (negativ) exponentialverteilt.

Es gilt

$$\mathbf{F}(\omega) = \frac{\lambda}{\lambda + i\omega} \tag{5.7}$$

Abbildung 5.1: Exponentialverteilungen



$$F(x) = 1 - e^{-\lambda x} \tag{5.8}$$

$$E(X) = \frac{1}{\lambda} \tag{5.9}$$

$$Var(X) = \frac{1}{\lambda^2} \tag{5.10}$$

Beweis: Es ist

$$\begin{aligned} \mathbf{F}(\omega) &= \int_0^\infty e^{-\lambda x - i\omega x} dx = \lambda \int_0^\infty e^{-(\lambda + i\omega)x}\, dx \\ &= \frac{\lambda}{\lambda + i\omega} \end{aligned}$$

und

$$F(x) = \int_{-\infty}^x f(\xi)\, d\xi = \lambda \int_0^x e^{-\lambda\xi}\, d\xi = \lambda \left(\frac{1}{\lambda} e^{-\lambda\xi} \right) \Big|_0^x = 1 - e^{-\lambda x}$$

Das k-te Moment der Exponentialverteilung ist durch $i^{-k}\mathbf{F}(0)$ gegeben, wobei $\mathbf{F}^k(0)$ die k-te Ableitung der charakteristischen Funktion an der Stelle $\omega = 0$ ist (vergl. (4.8.38)). Wir benötigen insbesondere die erste und die zweite Ableitung von $\mathbf{F}(\omega)$: es ist $d\mathbf{F}(\omega)/d\omega = -i\lambda/(\lambda + i\omega)^2$, und $E(X) = d\mathbf{F}(\omega)/d\omega|_{\omega=0} = (-i\lambda/\lambda^2)/(-i) = 1/\lambda$. Für das zweite Moment erhält man $F''(\omega) = 2\lambda i^2/(\lambda + i\omega)^3$, so dass $E(X^2) = 2\lambda/\lambda^3 = 2/\lambda^2$. Dementsprechend ist $Var(X) = 2/\lambda^2 - 1/\lambda^2 = 1/\lambda^2$. □

Die Abbildung 5.1 zeigt Exponentialverteilungen für die Parameter $\lambda = .5$, $\lambda = 1.0$ und $\lambda = 1.5$.

In der Anmerkung 2.2.2 auf Seite 39 wurde darauf hingewiesen, dass bei einer Messung von Werten aus einem Kontinuum die Wahrscheinlichkeit, dass ein Wert aus $\mathbb{Q}$ auftritt, gleich Null ist. Dies gilt für alle Verteilungen von Wartezeiten, die Exponentialverteilung ist nur ein Beispiel von Vielen.

5.2.2 Die Exponentialverteilung als Grenzverteilung

Es werde angenommen, dass die Zeitachse $[0, \infty)$ in Abschnitte gleicher Länge Δt aufgeteilt werden, und dass das interessierende Ereignis A in einem Abschnitt mit der Wahrscheinlichkeit p auftritt. A kann in verschiedenen, auch benachbarten, Abschnitten auftreten, und die Ereignisse "A tritt in Abschnitt j auf", für $j = 1, 2, \cdots$ seien stochastisch unabhängig. Die Wahrscheinlichkeit p ist sicherlich von der Länge Δt der Intervalle abhängig, je größer der Wert von Δt, desto größer wird p sein. Umgekehrt wird p um so kleiner sein, je kleiner Δt ist. $p(\Delta t)$ ist eine zunächst nicht explizit gegebene Funktion von Δt, um also fortfahren zu können, muß eine Annahme über $p(\Delta t)$ gemacht werden, die allerdings so allgemein, d.h. so wenig einschränkend wie möglich sein soll. Eine sehr allgemeine Annahme ist, dass $p(\Delta t)$ in eine Taylor-Reihe (s. Gl. (4.29), Seite 125) entwickelt werden kann, und das heißt hier, dass p mindestens einmal in Bezug auf Δt differenzierbar ist, d.h. es soll für hinreichend kleinen Wert von Δt

$$p(\Delta t) = p(0) + \Delta t \left.\frac{dp(\Delta t)}{d\Delta t)}\right|_{\Delta t=0} + o(\Delta t) \tag{5.11}$$

gelten, wobei $o(\Delta t)$ eine Funktion von Δt ist, die schneller gegen Null geht als Δt). Man kann weiter annehmen, dass $p(0) = 0$ ist, denn $\Delta t = 0$ bedeutet, dass das Ereignis in einem Zeitintervall der Länge Null eintritt. Ist $F(t)$ die Verteilungsfunktion von t, so gilt ja in der Tat

$$p(0) = \lim_{\Delta t \to 0} (F(t + \Delta t) - F(t)) = 0 \text{ für alle } t$$

Mithin kann man für hinreichend kleinen Wert von Δt den Ausdruck (5.11) in der Form $p(\Delta t) = \lambda \Delta t + o(\Delta)$ mit $\lambda = dp(\Delta t)/dt$, $\Delta t = 0$ schreiben. Man betrachtet nun die Anzahl K der Versuche, die notwendig sind, bis zum ersten Male A auftritt. K ist offenbar geometrisch verteilt, und es gilt $P(K \leq k) = 1 - (1-p)^k$. Die bis zum Ende des k-ten Intervalles verstrichene Zeit ist $t = k\Delta t$, so dass $\Delta t = t/k$. Für festes t folgt aus $\Delta t \to 0$ nun $k \to \infty$. Für $0 < \Delta t \to 0$ kann man aber $p = p(\Delta t) \approx \lambda \Delta t$ setzen und erhält somit

$$P(K \leq k) = 1 - (1-p)^k = 1 - (1 - \lambda \Delta t)^k = 1 - \left(1 - \lambda \frac{t}{k}\right)^k$$

Aber aus der Analysis ist bekannt, dass

$$\lim_{k \to \infty} \left(1 - \frac{\lambda t}{k}\right)^k = e^{-\lambda t}$$

so dass man für $k \to 0$, $\Delta t \to 0$ das Resultat $P(K \leq k) = 1 - e^{-\lambda t}$ erhält. $\square$

5.2.3 Zum Begriff der reinen Zufälligkeit

Nun ist $f(x) = F'(x) = dF(x)/dx = \lambda e^{-\lambda x}$, und der Vergleich mit (5.7) liefert dann

$$F'(x) = \lambda(1 - F(x)).$$

Dies ist eine Differentialgleichung für die Verteilungsfunktion der zufälligen Veränderlichen X. Nun gilt allgemein $F(x) \to 1$ für $x \to \infty$, und (5.8) impliziert dann für diesen Fall $F'(x) \to 0$. Andererseits gilt allgemein $F(x) \to 0$ für $x \to -\infty$; da $F(x) = 0$ für $x \leq 0$ (dies folgt aus der Definition der Dichte f, denn f ist nur $x \geq 0$ erklärt), und (5.8) impliziert dann $f(0) = F'(0) = \lambda$. An dieser Stelle muß f auch maximal sein, denn aus (5.6) folgt, dass f eine fallende Funktion von x ist. Dieser Sachverhalt impliziert, dass das Ereignis A mit größerer Wahrscheinlichkeit zu Beginn der Wartezeit eintritt als in späteren Abschnitten. Dies kann man sich direkt anhand der Abbildungen in Fig. 4.10.1 veranschaulichen: der Integrand $\lambda e^{-\lambda x}$ ist ja um so größer, je kleiner x. Man mag diesen Sachverhalt als Einwand gegen die Repräsentation "reiner Zufälligkeit" durch die Exponentialverteilung vorbringen. Man muß sich dabei aber überlegen, was man denn unter "reiner Zufälligkeit" verstehen will.

Beispiel 5.2.1 Im Beispiel 4.5.3 wurde ein Urnenmodell für die Suche im Langzeitgedächtnis vorgestellt. Es werde nun angenommen, dass die Zeit t, die für die überprüfung eines zufällig gewählten Wortes auf Zugehörigkeit zu einer bestimmten Klasse benötigt, zufällig, insbesondere exponentialverteilt mit dem Parameter λ ist. Die durchschnittlich benötigte Zeit betrage .1 Sekunde; also ist $1/\lambda = .1$ und damit $\lambda = 1/.1 = 10$. Die Varianz der Zeiten beträgt dann $\sigma^2 = 1/\lambda^2 = 1/100 = .01$, und dementsprechend ist die Standardabweichung $\sigma = .1$ Sekunden.

Die Wahrscheinlichkeit, dass die Suche nach einer Zeit $\tau \in (t, t + \Delta t]$ zum Erfolg führt, ist durch

$$\begin{aligned} P(\tau \in (t, t + \Delta]) &= F(t + \Delta t) - F(t) \\ &= 1 - e^{-\lambda(t+\Delta t)} - (1 - e^{-\lambda t}) \\ &= e^{-\lambda t} - e^{-\lambda(t+\Delta t)} \\ &= e^{-\lambda t}(1 - e^{-\lambda \Delta t}) \end{aligned}$$

gegeben. Für größer werdenden Wert von t geht diese Wahrscheinlichkeit also gegen Null. Betrachtet man nun

$$\begin{aligned} P(\tau \in (t, t + \Delta t] | \tau > t) &= \frac{P(\tau \in (t, t + \Delta t])}{P(\tau > t)} \\ &= \frac{e^{-\lambda t}(1 - e^{-\lambda \Delta t})}{e^{-\lambda t}} \\ &= 1 - e^{-\lambda \Delta t} \end{aligned}$$

so sieht man, dass dieser Ausdruck unabhängig von t ist! Anders als bei der Gleichverteilung verändert sich die Wahrscheinlichkeit, dass das zufällige Ereignis A in einem Zeitintervall $(t, t + \Delta t]$ eintritt, wenn es bis zum Zeitpunkt t noch nicht eingetreten ist, nicht mit dem Wert von t. Man kann dies so interpretieren, dass die zufälligen Mechanismen unabhängig von der Vergangenheit arbeiten; ihre Zufälligkeit ist in jedem Augenblick die gleiche. Will man also diese Art von reiner Zufälligkeit ausdrücken, so darf man nicht die Gleichverteilung, sondern muß die Exponentialverteilung zur Beschreibung der Prozesse wählen. □

Die Dichte der Exponentialverteilung ist maximal für $X = 0$. Empirisch gewonnene Häufigkeitsverteilungen zeigen dieses Merkmal häufig nicht. Gleichwohl können exponentialverteilte Variable konstituierende Komponenten der betrachteten Variablen sein, wie im folgenden ausgeführt wird.

5.3 Die Gamma-Verteilung

Definition 5.3.1 *Es sei $X \geq 0$ und es gelte*

$$f(x) = \frac{\lambda^p}{\Gamma(p)} x^{p-1} e^{-\lambda x} \tag{5.12}$$

wobei

$$\Gamma(p) = \int_0^\infty \xi^{p-1} e^{-\xi}\, d\xi \tag{5.13}$$

die Gamma-Funktion mit den Eigenschaften $\Gamma(p) = (p-1)\Gamma(p-1)$ und $\Gamma(p+1) = p!$ für $p \in \mathbb{N}$, $\Gamma(1) = 1$, $\Gamma(1/2) = \sqrt{\pi}$ ist. Dann heißt X gamma-verteilt *mit den Parametern p und λ.*

Anmerkung: Für $p = 1$ erhält man offenbar die Dichte $g(x) = \lambda e^{-\lambda x}$, d.h. die Exponentialverteilung ist ein Spezialfall der Gamma-Verteilung.

Satz 5.3.1 *Es sei X gamma-verteilt mit den Parametern $p \in \mathbb{N}$ und λ. Dann ist X als Summe $X = X_1 + \cdots + X_p$ von p stochastisch unabhängigen, mit dem Parameter λ exponentialverteilten zufälligen Veränderlichen $X_1, \cdots, X_p$ darstellbar, und dementsprechend gilt*

$$\mathbf{F}(\omega) = \left(\frac{\lambda}{\lambda + i\omega}\right)^p \tag{5.14}$$

$$E(X) = \frac{p}{\lambda} \tag{5.15}$$

$$Var(X) = \frac{p}{\lambda^2} \tag{5.16}$$

Beweis: Die charakteristische Funktion der Gamma-Verteilung ist

$$\mathbf{F}(\omega) = \frac{\lambda^p}{\Gamma(p)} \int_0^\infty x^{p-1} e^{-\lambda x - i\omega x} dx = \frac{\lambda^p}{\Gamma(p)} \int_0^\infty x^{p-1} e^{-(\lambda + i\omega x)} dx$$

Die Substitution $u = (\lambda + i\omega)x$ liefert wegen $dx/du = 1/(\lambda + i\omega)$

$$\mathbf{F} = \left(\frac{\lambda}{\lambda + i\omega}\right)^p \frac{1}{\Gamma(p)} \int_0^\infty u^p e^{-u}\, du = \left(\frac{\lambda}{\lambda + i\omega}\right)^p$$

denn das Integral, dividiert durch $\Gamma(p)$, muß gleich 1 sein, denn es ist nichts weiter als das Integral über die Gamma-Verteilung mit den Parametern 1 und p. $\lambda/(\lambda + i\omega)$ ist aber die charakteristische Funktion der Exponentialfunktion, und nach Satz 3.7.3 ist die p-te Potenz einer c.F. gerade die c.F. der p-fachen Faltung.

Der Erwartungswert von X ist dann gemäß Satz 3.5.6, p. 103, durch

$$E(X) = E(X_1) + \cdots + E(X_p)$$

gegeben. Aber nach 5.9, p. 150, ist $E(X_i) = 1/\lambda$, so dass $E(X) = p/\lambda$. Für die Varianz $Var(X)$ gilt wegen der Unabhängigkeit der X_i nach Satz 3.5.6

$$Var(X) = Var(X_1) + \cdots + Var(X_p)$$

und da nach 5.10 $Var(X_i) = 1/\lambda^2$ folgt $Var(X) = p/\lambda^2$. (Der Weg über die charakteristische Funktion wäre hier umständlicher.)

Die Gamma-Verteilung ist eine schiefe Verteilung, nimmt aber für größer werdendes p eine in guter Näherung symmetrische Gestalt an: kleinere Werte von X sind wahrscheinlicher als größere. Mit größer werdendem Wert von p nimmt aber auch die Varianz von X zu. Dies ist plausibel, wenn man bedenkt, dass X als Summe stochastisch unabhängiger Variabler aufgefaßt werden kann; jede Variable X_i trägt ihre eigene Variation zur Gesamtvarianz von X bei. Weiter gilt der folgende

Satz 5.3.2 *Die Variablen $X_1, \cdots, X_n$ seien stochastisch unabhängig, und X_i sei gamma-verteilt mit dem Parametern λ und p, $i = 1, \cdots, n$. Dann ist die Summe $X = X_1 + \cdots + X_n$ wieder gamma-verteilt mit dem Parametern np und λ und hat demnach den Erwartungswert $E(X) = np/\lambda$ sowie die Varianz $Var(X) = np/\lambda^2$.*

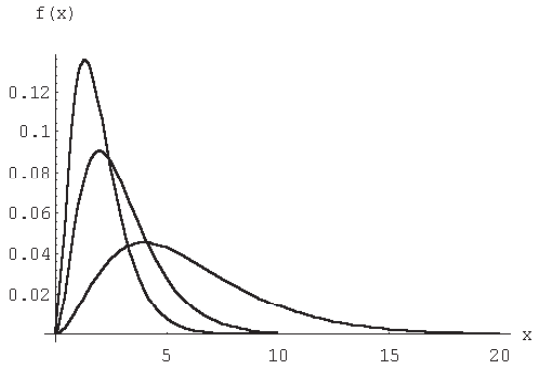
Beweis: Die Aussage folgt sofort aus der Tatsache, dass die charakteristische Funktion der Verteilung einer Summe identisch verteilter, stochastisch unabhängiger Variablen durch $\mathbf{F}^n$ gegeben ist, wobei $\mathbf{F}$ die charakteristische Funktion der Verteilung der X_i ist. $\mathbf{F}$ ist aber durch (5.14) gegeben, so dass die c.F. von X durch

$$\mathbf{F}^n = \left(\frac{\lambda}{\lambda + i\omega}\right)^{pn}$$

gegeben ist. Dies ist aber die c.F. einer gamma-verteilten Variablen mit dem Parametern np und λ. □

Der Satz ist natürlich schon aufgrund von Satz 5.3.1 plausibel, denn wenn eine gamma-verteilte als Summe von exponentialverteilten Variablen dargestellt werden kann, so muß eine Summe von gamma-verteilten Variablen wieder eine Summe von exponentialverteilten Variablen sein, - und diese Summe ist gamma-verteilt. Die Gamma-Verteilung gehört damit zur Klasse der *stabilen Verteilungenfunktionen.*[1] Abbildung 5.2 zeigt drei Gammaverteilungen für $p = 3$, $\lambda = .5$,

Abbildung 5.2: Gammaverteilungen



$\lambda = 1.0$ und $\lambda = 1.5$.

Beispiel 5.3.1 Es werde angenommen, dass die Dauer t, die Kunden am Postschalter bedient werden, exponentialverteilt ist mit dem Parameter λ. Die durchschnittliche Bedienungszeit ist dann $1/\lambda$, und die Varianz ist $1/\lambda^2$. Muß man sich also im Postamt in eine Warteschlange einordnen, so wächst die durchschnittliche Wartezeit ebenso wie die Varianz proportional zur Anzahl der Personen, die bereits in der Schlange warten. □

Beispiel 5.3.2 Wird die Reaktionszeit bis auf eine motorische Komponente, die als unabhängig von den Entscheidungsprozessen angenommen werden soll, durch k seriell auszuführende Entscheidungsschritte bestimmt, für die die Zeit jeweils exponentiell ist, so wächst die durchschnittliche Reaktionszeit ebenfalls proportional zu k, und die Varianz der Reaktionszeit muß ebenfalls proportional zu k wachsen. □

[1] Allgemein heißt eine Verteilung F *stabil*, wenn die Verteilung der Summe zweier zufälliger Veränderlicher, die jede den Verteilungstyp F haben, wieder vom Verteilungstyp F ist. Die Parameter der Verteilung der Summe werden sich aber i.a. von denen der Verteilung der Summanden unterscheiden.

Die Annahme, dass alle X_i mit dem gleichen Parameter λ exponentialverteilt sind, ist unter Umständen eine zu weitgehende Simplifikation. Im vorangegangenen Beispiel ist es z.B. denkbar, dass die einzelnen Identifikationszeiten zwar exponentialverteilt sind, aber mit unterschiedlichen Parametern λ_i. Es sei etwa $X = X_1 + X_2$, und X_1 habe die Dichte $f_1(x_1) = \mu \exp(-\mu x_1)$, und x_2 habe die Dichte $f_2(x_2) = \lambda \exp(-\lambda x_2)$; X_1 und X_2 seien stochastisch unabhängig. Die Dichte g von X ergibt sich dann gemäß

$$g(x) = \int_{-\infty}^{\infty} f_1(x - x_2) f_2(x - 2)\, dx_2 = \mu\lambda e^{-\lambda\mu} \int_0^{\infty} e^{(\mu-\lambda)x_2}\, dx_2$$

d.h.

$$g(x) = \frac{\mu\lambda}{\mu - \lambda}(e^{-\mu x} - e^{-\lambda x}), \quad \mu > \lambda \tag{5.17}$$

man spricht von der *verallgemeinerten Gamma-Verteilung.*

Beispiel 5.3.3 *xyxy* McGill (1963) schlug zur Interpretation der Daten von Bousfield und Sedgewick (1944) vor, dass die Suche nach einem Wort im Langzeitgedächtnis (LZG) dem Ziehen von Kugeln aus einer Urne vergleichbar sei (vergl . Beispiele 4.5.3, 5.2.1). Es werde nun angenommen, dass für die überprüfung eines Wortes daraufhin, ob es einem bestimmten Kriterium entspricht, eine exponentialverteilte Zeit benötigt wird, wobei der Parameter λ der Verteilung für alle Wörter gleich sei. Unter der Bedingung, dass n Wörter überprüft werden müssen, bis ein dem Kriterium genügendes Wort gefunden wird, ist die Wartezeit bis zum Auffinden eines solchen Wortes Gamma-verteilt mit dem Parametern n und λ. □

Die Summe stochastisch unabhängiger, exponentialverteilter Variablen ist nicht notwendig Gamma-verteilt, wenn die Anzahl der Summanden selbst eine zufällige Veränderliche ist:

Satz 5.3.3 *Die Variablen X_i, $i = 1, \cdots, n$ seien stochastisch unabhängig und exponentialverteilt mit dem Parameter λ. Weiter sei die Anzahl n geometrisch verteilt mit dem Parameter p_o. Dann ist $X = X_1 + \cdots + X_n$ wieder exponentialverteilt mit dem Parameter $p_0\lambda$, d.h.*

$$P(X \leq x) = 1 - \exp(-p_0\lambda x).$$

Beweis: X kann nach Definition 4.9.1, (Abschnitt 4.9, p. 142) als eine zusammengesetzte Verteilung betrachtet werden. Es gilt jedenfalls

$$P(X \leq x) = \sum_{i=1}^{\infty} P(X \leq x|n)P(n)$$

wobei $P(n)$ die Wahrscheinlichkeit bezeichnet, dass n die Anzahl der Summanden ist.

Es ist

$$P(X \leq x|n) = \int_0^x \frac{\lambda^n}{\Gamma(n)} \xi^{n-1} e^{-\lambda\xi}\, d\xi$$

und demnach ist

$$\begin{aligned} P(X \leq x) &= p_0 \sum_{n=1}^{\infty} (1-p_0)^{n-1} \int_0^x \frac{\lambda^n}{\Gamma(n)} \xi^{n-1} e^{-\lambda\xi}\, d\xi \\ &= p_0 \int_0^x e^{-\lambda\xi} \sum_{n=1}^{\infty} q_0^{n-1} \frac{\lambda^n}{\Gamma(n)} \xi^{n-1} d\xi \\ &= p_0 \lambda \int_0^x e^{-\lambda\xi} \sum_{n=1}^{\infty} \frac{(q_o\lambda\xi)^{n-1}}{(n-1)!}\, d\xi \end{aligned}$$

Aber die Summe ist gleich

$$e^{q_0\lambda\xi} = e^{(1-p_0)\lambda\xi} = e^{\lambda\xi - p_0\xi}$$

so dass

$$P(X \leq x) = p_0\lambda \int_0^x e^{-\lambda\xi+\lambda\xi-p_0\lambda\xi}\, d\xi = p_0\lambda \int_0^\infty e^{-p_o\lambda\xi} d\xi$$

Dies ist aber die Exponentialverteilung mit dem Parameter $p_o\lambda$. □

Das Beispiel zeigt, dass die Verteilungsfunktion für sich genommen noch keinen Aufschluß über die Art der Zusammensetzung der betrachteten Variablen zu geben braucht. Findet man, dass die Verteilung einer Variablen der gamma-Verteilung entspricht, so macht es Sinn, zu vermuten, dass sich die Variable als Summe exponentialverteilter Variablen auffassen läßt. Wird die Verteilung aber eher durch die Exponentialverteilung repräsentiert, so kann man noch nicht ausschließen, dass die Variable nicht doch als Summe von exponentialverteilten Variablen angeschrieben werden kann, deren Anzahl aber geometrisch verteilt ist.

Weitere Beispiele für die Gamma-Verteilung werden bei der Behandlung von Stichprobenverteilungen gegeben.

5.4 Die Normalverteilung

Definition 5.4.1 *Es sei X eine zufällige Veränderliche mit der Dichte*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad -\infty < x < \infty \tag{5.18}$$

Dann heißt X normalverteilt *(auch:* Gauß-verteilt*) mit den Parametern μ und σ. Gilt insbesondere $\mu = 0$ und $\sigma = 1$, so heißt X* standardnormalverteilt*.*

Statt "normalverteilt mit den Parametern μ und σ" wird auch einfach "$N(\mu, \sigma)$-verteilt" geschrieben.

Satz 5.4.1 *Es sei X normalverteilt mit den Parametern μ und σ. Dann gilt*

$$\begin{aligned} \mathbf{F}(\omega) &= e^{i\mu\omega-\sigma^2\omega^2/2} && (5.19)\\ E(X) &= \mu && (5.20)\\ Var(X) &= \sigma^2 && (5.21)\\ Median &= Modalwert = E(X) && (5.22) \end{aligned}$$

Beweis: $\mathbf{F}(\omega) = \int_{-\infty}^{\infty} f(x)\exp(-i\omega x)dx$; die Auswertung des Integrals kann hier übergangen werden (Integraltafeln!). Der Rest folgt durch Bestimmung von $d\mathbf{F}(\omega)/d\omega$ und $d^2\mathbf{F}(\omega)/d\omega^2$ an der Stelle $\omega = 0$; (5.22) folgt dann aus der Symmetrie der Dichte. □

Für die Verteilungsfunktion $P(X \leq x)$ ist auch die Schreibweise $\Phi(x)$ gebräuchlich, also

$$\Phi(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)\,dx \tag{5.23}$$

Satz 5.4.2 *Die zufällige Veränderliche X sei normalverteilt mit dem Erwartungswert μ und der Varianz σ^2. Dann ist $Y = aX + b$, $a \neq 0$ normalverteilt mit dem Erwartungswert $\mu_y = a\mu + b$.*

Beweis: Es ist $P(Y \leq y) = P(aX + b \leq y) = P(X \leq (y-b)/a)$. Dann ist die Dichtefunktion für Y durch

$$\begin{aligned} f_y(y) &= \frac{d\Phi_x((y-b)/a)}{dy}\\ &= \frac{1}{a^2} f_x\left(\frac{y-b}{a}\right)\\ &= \frac{1}{(a\sigma)^2\sqrt{2\pi}}\exp\left(-\frac{(y-b-a\mu)^2}{2a^2\sigma^2}\right) \end{aligned}$$

Die rechte Seite ist aber die Dichte für eine normalverteilte Variable mit dem Erwartungswert $a\mu + b$ und der Varianz $a^2\sigma^2$. □

Es sei $Z = (X-\mu)/\sigma$. Dann ist Z standardnormalverteilt. Denn es ist ja $E(Z) = 0$ und $Var(Z) = 1$, und dass Z normalverteilt ist folgt aus Satz 5.4.2 mit $a = 1/\sigma$, $b = -\mu/\sigma$.

Das Integral für $\Phi(x)$ ist nicht in geschlossener Form darstellbar, Näherungsformeln findet man etwa in Abramowitz und Stegun (1965). Wegen $\Phi(x) = \Phi(z)$ mit $z = (x-\mu)/\sigma$ genügt es aber, Φ nur für die Standardvariable Z zu tabellieren.

Allgemein gilt

$$P(a \leq X \leq b) = \Phi(z_b) - \Phi(z_a) \tag{5.24}$$

Es sei insbesondere $a = \mu - z\sigma$, $b = \mu + z\sigma$; a sei also ein Wert, der z σ-Einheiten oberhalb von μ liegt, während b gerade z σ-Einheiten unterhalb von μ liege. Dann ist

$$z_b = (b - \mu)/\sigma = (z\sigma + \mu - \mu)/\sigma = z$$

und

$$z_a = (a - \mu)/\sigma = (\mu - z\sigma - \mu)/\sigma = -z$$

Aus (5.24) folgt dann

$$P(\mu - z\sigma \leq X \leq \mu + z\sigma) = \Phi(z) - \Phi(-z) \tag{5.25}$$

Die Differenz $\Phi(z) - \Phi(-z)$ ist aber unabhängig von μ und σ, so dass man Aussagen über die Wahrscheinlichkeit machen kann, mit der eine beliebige, normalverteilte Variable mit dem Erwartungswert μ und der Varianz σ^2 innerhalb des Intervalles $(\mu - z\sigma, \mu + z\sigma)$, also innerhalb eines Intervalles mit der Länge von $2z$ σ-Einheiten liegt, ohne dass explizit auf die Werte von μ und σ bezug genommen werden müßte. So erhält man für $z = 1$ (vergl. die Tabelle im Anhang)

$$P(\mu - \sigma \leq X \leq \mu + \sigma) = \Phi_z(1) - \Phi_z(-1) = .841 - .159 = .683 \tag{5.26}$$

Für $z = 2$ findet man

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = \Phi_z(2) - \Phi_z(-2) = .977 - .023 = .954 \tag{5.27}$$

Für $z = 3$ - also für 3 σ-Einheiten - erhält man schließlich

$$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = .999 - .0006 = .998 \tag{5.28}$$

d.h. die Wahrscheinlichkeit, dass X um mehr als 3 σ-Einheiten nach oben oder unten von μ abweicht, ist

$$P(X \notin (\mu - 3\sigma, \mu + 3\sigma)) = 1 - .998 \approx .002$$

Die Normalverteilung gehört zur Klasse der *stabilen Verteilungen*, denn es gilt der

Satz 5.4.3 *Die Variablen X_i, $i = 1, \cdots, n$ seien stochastisch unabhängig und normalverteilt mit den Erwartungswerten μ_i und den Varianzen σ_i^2. Dann ist die Summe $X = a_1X_1 + \cdots + a_nX_n$,wobei die a_k, $k = 1, \cdots, a_n$ reelle Zahlen sind, ebenfalls normalverteilt mit dem Erwartungswert $\mu = a_1\mu_1 + \cdots + a_n\mu_n$ und der Varianz $\sigma^2 = a_1^2\sigma_1 + \cdots + a_n^2\sigma_n^2$*

Beweis: Es werde zunächst die charakteristische Funktion für $Y = aX$, wobei $X \sim N(\mu, \sigma^2)$ ist, bestimmt. Die Dichte von Y ergibt sich aus $P(Y \leq y) =$

$P(aX \leq y) = P(X \leq y/a)$, so dass $f_y(y) = dP(Y \leq y)/dy = dP(X \leq y/a)/dy = f_x(y/a)/a$. Also ist

$$\begin{aligned}
\mathbf{F}_y(\omega) &= \int_{-\infty}^{\infty} f_y(y) e^{-i\omega y}\, dy \\
&= \frac{1}{a} \int_{-\infty}^{\infty} f_x\left(\frac{y}{a}\right) e^{-i\omega y}\, dy \\
&= \frac{1}{a\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{(y/a-\mu)^2}{2\sigma^2}\right) e^{-i\omega y}\, dy
\end{aligned}$$

Substituiert man $\xi = y/a$, so ergibt sich

$$\mathbf{F}_y(\omega) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{(\xi-\mu)^2}{2\sigma^2}\right) e^{-i\omega a\xi}\, d\xi$$

und dies ist die c.F. der zufälligen Variablen X für $a\omega$; aus (5.19) erhält man

$$\mathbf{F}_y(\omega) = e^{i\mu a\omega - \sigma^2 (a\omega)^2/2} \tag{5.29}$$

Die charakteristische Funktion einer Summe unabhängiger Variabler ist durch das Produkt der charakteristischen Funktionen der Variablen, die als Summanden auftreten, gegeben. Mit (5.29) ergibt sich dann

$$\mathbf{F}_y(\omega) = \prod_{k=1}^{n} e^{ia_k\mu_k\omega - \sigma_k^2 (a_k\omega)^2/2} = \exp\left(i\omega \sum_{k=1}^{n} a_k\mu_k - \frac{\omega^2}{2} \sum_{k=1}^{n} (a_k\sigma_k)^2\right)$$

gegeben sein, mit $\mu = \sum_k a_k\mu_k$ und $\sigma^2 = \sum_k (a_k\sigma_k)^2$ ist dies aber gerade wieder die c.F. einer Gauß-Verteilung mit dem Erwartungswert μ und der Varianz σ^2. □

Beispiel 5.4.1 Es sei

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

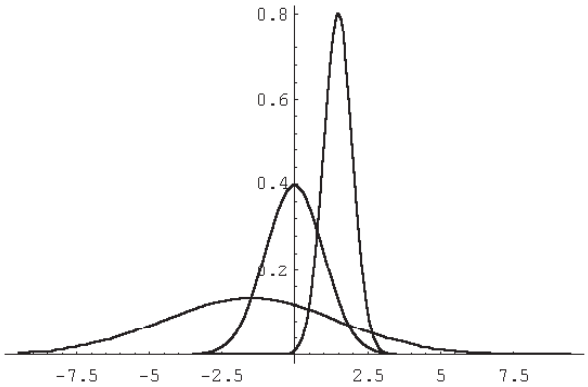
wobei die X_i Realisierungen normalverteilter Variablen seien. Nach Satz 5.4.3 ist dann jedenfalls $X = \sum_i X_i$ normalverteilt mit dem Erwartungswert $\mu = \sum_i \mu_i$ und der Varianz $\sigma^2 = \sum_i \sigma_i^2$. Weiter gilt nach Satz 5.4.2, dass die Variable aX normalverteilt ist mit dem Erwartungsert $aE(X)$ und der Varianz $a^2\sigma^2$. Mit $a = 1/n$ folgt also, dass $\overline{x} = X/n$ normalverteilt ist mit den Parametern

$$E(\overline{x}) = \frac{1}{n} \sum_{i=1}^{n} \mu_i \tag{5.30}$$

$$Var(\overline{x}) = \frac{1}{n^2} \sum_{i=1}^{n} \sigma_i^2 \tag{5.31}$$

□

Abbildung 5.3: Gaußsche Dichten



Der Stichprobenmittelwert normalverteilter Variablen ist also wieder normalverteilt. Die *Varianz* einer Stichprobe normalverteilter Variabler ist dagegen *nicht* normalverteilt: da die Varianz über die Quadrate der Abweichungen vom arithmetischen Mittel definiert ist, kann die Varianz nur Werte größer/gleich Null annehmen, was bereits eine Normalverteilung ausschließt. Die Verteilung der Stichprobenvarianz wird später hergeleitet.

Die Fehlerfunktion Zum Abschluß soll noch die *Fehlerfunktion* angeführt werden:

Definition 5.4.2 *Die Funktion*

$$erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{t^2} dt \tag{5.32}$$

heißt Fehlerfunktion.

Die Bezeichnung "erf(x)" hat sich wegen der englischen Bezeichnung "error function" eingebürgert.

Wegen der Definition von erf(x) als Integral von $\exp(-t^2)$ und der häufig gemachten Annahme, dass Meßfehler normalverteilt seien, liegt nahe, dass eine Beziehung zwischen der Normalverteilung und erf(x) existieren muß. Tatsächlich läßt sich die Fehlerfunktion aus der Normalverteilung herleiten und umgekehrt. Es sei wieder

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp(-\frac{1}{2}z^2)dz$$

Weiter werde $t = z/\sqrt{2}$ gesetzt, bzw. $z = t\sqrt{2}$. Dann ist $dz/dt = \sqrt{2}$ oder $dz = \sqrt{2}dt$. Für $z = x$ ist dann $t = x/\sqrt{2}$, und man erhält

$$\Phi(x) \quad = \quad \frac{\sqrt{2}}{2\pi} \int_{-\infty}^{x/\sqrt{2}} \exp(-t^2)dt$$

$$= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{x/\sqrt{x}} \exp(-t^2)dt \tag{5.33}$$

Für $x = 0$ gilt $\Phi(x) = 1/2$. Dann folgt aus (5.33)

$$\Phi(0) = \frac{1}{2} = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{0} \exp(-t^2)dt,$$

so dass für $x > 0$ aus (5.33)

$$\Phi(x) = \frac{1}{\sqrt{\pi}} \int_{0}^{x/\sqrt{x}} \exp(-t^2)dt + \frac{1}{2} = \frac{1}{2}\mathrm{erf}(x/\sqrt{2}) + \frac{1}{2}$$

folgt, d.h. aber

$$\mathrm{erf}(x/\sqrt{2}) = 2\Phi(x) - 1. \tag{5.34}$$

5.5 Die logistische Verteilung

Definition 5.5.1 *Es sei X eine zufällige Veränderliche mit der Dichte*

$$f(x) = \frac{\exp(-(x-\mu)/\beta)}{\beta(1+\exp(-(x-\mu)/\beta))^2}, \quad -\infty < x < \infty \tag{5.35}$$

Dann heißt X logistisch verteilt. *(5.5.1) heißt insbesondere die* parametrische Form *der Dichte.*

Satz 5.5.1 *Die zufällige Veränderliche X sei logistisch verteilt. Weiter sei*

$$g = \pi/\sqrt{3} \tag{5.36}$$

Dann gelten die Beziehungen

$$F(x) = \frac{1}{1+\exp(-(x-\mu)/\beta)} \tag{5.37}$$

$$E(X) = \mu \tag{5.38}$$

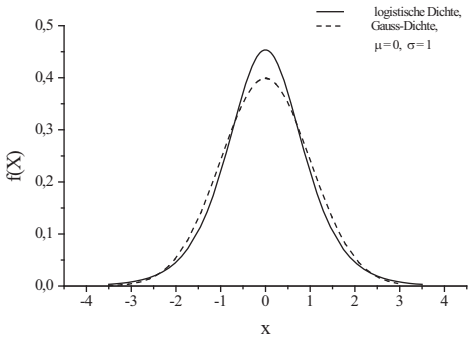
$$Var(X) = \sigma^2 = \beta^2 g^2 \tag{5.39}$$

Beweis: s. Johnson und Kotz (2) (1970), p.6 □

Aus (5.39) folgt $\beta = \sigma/g$. Unter Berücksichtigung der Ausdrücke für $E(X)$ und $Var(X)$ läßt sich (5.5.1) auch in der Form

$$f(x) = \frac{g\ \exp(-g(x-\mu)/\sigma)}{\sigma(1+\exp(-g(x-\mu)/(\sigma))^2} \tag{5.40}$$

Abbildung 5.4: Logistische Verteilungen



schreiben; für die Verteilungsfunktion ergibt sich

$$F(x) = \frac{1}{1 + \exp(-g(x-\mu)/\sigma)} \tag{5.41}$$

Die Beziehung zwischen der Dichte und der Verteilungsfunktion ist bei der logistischen Verteilung durch die Beziehung

$$f(x) = \frac{1}{\beta} F(x)[1 - F(x)], \quad \beta > 0 \tag{5.42}$$

wie man durch Einsetzen der Ausdrücke für f und F bestätigt. Hieraus wird deutlich, dass $f(x) \to 0$ für $F(x) \to 0$ oder $F(x) \to 1$; die Verteilung ist symmetrisch.

Die Abbildung 5.4 zeigt logistische Verteilungen für die Parameter $\beta = 1$, $\mu = 0$ und $\mu = 5$ (erste Reihe), und $\beta = 3$, $\mu = 0$ und $\mu = 5$ (zweite Reihe). Die ähnlichkeit mit der Normalverteilung ist deutlich.

Beispiel 5.5.1 Es soll die Wirkung eines schmerzlindernden Präparates untersucht werden. Für eine gegebene Dosis μ des Präparates sei die Wwirkung durch eine zuffällige Veränderliche X gegeben. Ein Patient bezeichnet sich als schmerzfrei, wenn $X > S$, wobei S ein Schwellenwert ist. X sei durch die logistische Funktion gegeben; dann gilt

$$P(X > S|\mu) = 1 - F(S|\mu) = 1 - \frac{1}{1 + \exp(-g(S-\mu)/\sigma)} \tag{5.43}$$

Aus (5.43) folgt sofort

$$P(X > S|\mu) = \frac{e^{-g(S-\mu)/\sigma)}}{1 + e^{-g(S-\mu)/\sigma}} \tag{5.44}$$

Man kann nun $P(X > S|\mu)$ als Funktion von μ betrachten, d.h. man betrachtet

$$G(\mu) := \frac{e^{-g(S-\mu)/\sigma)}}{1 + e^{-g(S-\mu)/\sigma}} \tag{5.45}$$

Dann ist

$$\frac{dG(\mu)}{d\mu} = \frac{g}{\sigma} \frac{e^{g(S-\mu)/\sigma}}{(1 + e^{-g(S-\mu)/\sigma})^2}$$

woraus

$$\frac{dG(\mu)}{d\mu} = \frac{g}{\sigma} G(\mu)(1 - G(\mu)) \tag{5.46}$$

folgt. Diese Gleichung entspricht in formaler Hinsicht der Gleichung (5.42). Sie wird in Zusammenhang mit Dosiswirkungen diskutiert und beschreibt, als deterministische Gleichung, den Anteil von Reaktionen in Abhängigkeit von der Dosis μ. Eine Gleichung dieses Typs wurde zuerst von Verhulst (1838) zur Beschreibung von Wachstumsprozessen mit Sättigung bzw. bei begrenzten Rescourcen diskutiert. Gleichung (5.46) stellt auch eines der einfachsten epidemiologischen Modelle dar: sie beschreibt die Ausbreitung einer Infektion bei einer gut gemischten Population, in der infizierte und nicht infizierte oder bereits immune Individuen zufällig aufeinander treffen. Je mehr bereits infiziert sind, desto geringer wird die Wahrscheinlichkeit, ein noch nicht infiziertes oder nicht immunes Individuum zu treffen; dies erklärt die "Sättigung". □

Luce und Galanter (1963) diskutieren Modelle des Entdeckens von Unterschieden zwischen Reizen, bei denen die Wahrscheinlichkeit des Entdeckens durch die logistische Funktion (5.46) gegeben ist. Dabei wird kein bezug auf zufällige Veränderliche genommen, die interne Größen repräsentieren. Die psychometrische Funktion wird durch (5.46) repräsentiert, wobei μ jetzt ein Maß für die Differenz der Stimuli ist; man scheibt dann natürlich $P(\mu)$ statt $G(\mu)$.

Setzt man in (5.35) $\mu = 0$, $\beta = 1/g$, so erhält man für die Dichte

$$f(x) = \frac{g\, e^{-gx}}{(1 + e^{-gx})^2} \tag{5.47}$$

X hat hier den Erwartungswert $\mu = 0$ und die Varianz $\sigma^2 = 1$. Dementsprechend wird X als $L(0, 1)$-verteilt bezeichnet. Die logistische Verteilung ähnelt sehr der Normalverteilung. Bezeichnet man mit

$$G_1(x) = \frac{1}{2\pi} \int_{-\infty}^{x} e^{-u^2(2}\, du$$

die Verteilungsfunktion einer $N(0, 1)$-verteilten Variablen an der Stelle x und mit

$$G_2(x) = \int_{-\infty}^{x} f(u)\, du$$

die Verteilungsfunktion für die in (5.47) gegebene Dichte einer $L(0,1)$-Variablen, so läßt sich zeigen, dass

$$\max_x (G_1(x) - G_2(x)) \approx .0228 \tag{5.48}$$

d.h. die maximale Differenz der Verteilungsfunktion einer $N(0,1)$- und einer $L(0,1)$-verteilten Variablen beträgt .0228; diese maximale Differenz ergibt sich an der Stelle $x = .7$. Da die logistische Funktion numerisch sehr viel leichter zu handhaben ist als die Gauß-Funktion kann man sich diesen Sachverhalt zunutze machen und die logistische Funktion als Approximation für die Gauß-Verteilung benützen. Unterschiede zwischen der logistischen und der Normalverteilung ergeben sich aber bei den höheren Momenten: die logistische Verteilung geht langsamer gegen Null als die Normalverteilung. Die Unterschiede zwischen den beiden Verteilungen werden also bei Extremwerten relevant.

5.6 Die Weibull-Verteilung

5.6.1 Allgemeine Definition

Definition 5.6.1 *Gegeben sei die zufälligen Veränderlichen X mit $-\infty < X \leq \eta_0$ und Y mit $\eta_0 \leq Y < \infty$ Die Verteilungsfunktionen von X bzw. Y seien durch*

$$F_X(x) = P(X \leq x) = \begin{cases} \exp\left[-\left(\frac{\eta_0 - x}{a}\right)^\beta\right], & x \leq \eta_0. \\ 1, & x < \eta_0 \end{cases} \tag{5.49}$$

$$F_Y(y) = P(Y \leq y) = \begin{cases} 1 - \exp\left[-\left(\frac{y - \eta_0}{a}\right)^\beta\right], & y \geq \eta_0 \\ 0, & y < \eta_0 \end{cases} \tag{5.50}$$

gegeben. Dann heißen X bzw. Y Weibull-verteilt *mit den Parametern η_0, $a > 0$, und $\beta > 0$.*

Anmerkungen:

1. Der Name Weibull-Verteilung geht auf den schwedischen Ingenieur Ernst Hjalmar Waloddi Weibull (1887 – 1979) zurück, der diese Verteilung im Zusammenhang mit Fragen der Bruchfestigkeit spröden Materials diskutierte. Fellers (1966), p. 54 Anmerkung, dass insbesondere die Verteilung (5.50) in Reliabilitätsuntersuchungen "rather mysteriously" unter dem Namen Weibull-Verteilung auftaucht, ist vermutlich damit zu erklären, dass die Verteilung schon vor Weibulls Arbeiten bekannt war.

2. Die zufälligen Veränderlichen X und Y in Definition 5.6.1 sind durch die Beziehung $Y = 2\eta_0 - X$ miteinander verbunden. Ist die Verteilungsfunktion von X durch (5.49) gegeben, so folgt die Verteilungsfunktion (5.50) für Y, denn

$$P(Y \leq y) = 1 - P(X \leq 2\eta_0 - y) = 1 - \exp\left[-\left(\frac{y-\eta_0}{a}\right)^{\beta}\right].$$

3. Der Hintergrund für die Definition der Weibull-Verteilung durch (5.49) und (5.50) ist, dass beide Formen als Definition Weibull-Verteilung in der Literatur vorkommen. Die Form (5.49) wird in Johnson, Kotz und Balakrishna (1994) und Kotz und Nadarajah (2000) als Verteilung vom Weibull-Typ bezeichnet. Die Beziehung zwischen einer Dosis, etwa eines Medikaments, oder einer Stimulusstärke und der Wahrscheinlichkeit einer Reaktion wird gelegentlich durch eine *Weibull-Funktion* beschrieben. Sowohl (5.49) als auch (5.50) können auf die gleiche Weibull-Funktion führen (s. unten).

4. Die Form (5.50) geht für $\beta = 1$ in die Exponentialverteilung $P(Y \leq y) = 1 - \exp(-\lambda y)$ über, mit $\lambda = a^{-1/\beta}$.

5. Die Entscheidung zwischen den Formen (5.49) und (5.50) wird davon abhängen, wie man die zufällige Veränderliche definiert, von deren Werten das jeweils interessierende zufällige Ereignis E abhängt. Will man ausdrücken, dass E eintritt, wenn die zufällige Veränderliche einen Wert *größer* als S, S ein kritischer Wert, ist, wird man auf (5.49) geführt, sofern für die zufällige Veränderliche eine Verteilung vom Weibull-Typ indiziert ist. Macht es Sinn, anzunehmen, dass E eintritt, wenn die zufällige Veränerliche einen Wert kleiner als ein kritischer Wert S annimmt, wird man auf (5.50) geführt, – vorausgesetzt, die zufällige Veränderliche ist vom Weibull-Typ. □

Für die Dichtefunktionen folgen die Ausdrücke

$$f_X(x) = \begin{cases} \frac{\beta}{a}\left(\frac{\eta_0 - x}{a}\right)^{\beta-1} \exp\left[-\left(\frac{\eta_0-x}{a}\right)^{\beta}\right], & x \leq \eta_0 \\ 0, & x > \eta_0 \end{cases} \tag{5.51}$$

$$f_Y(y) = \begin{cases} \frac{\beta}{a}\left(\frac{y - \eta_0}{a}\right)^{\beta-1} \exp\left[-\left(\frac{y-\eta_0}{a}\right)^{\beta}\right], & y \geq \eta_0 \\ 0, & y < \eta_0 \end{cases} \tag{5.52}$$

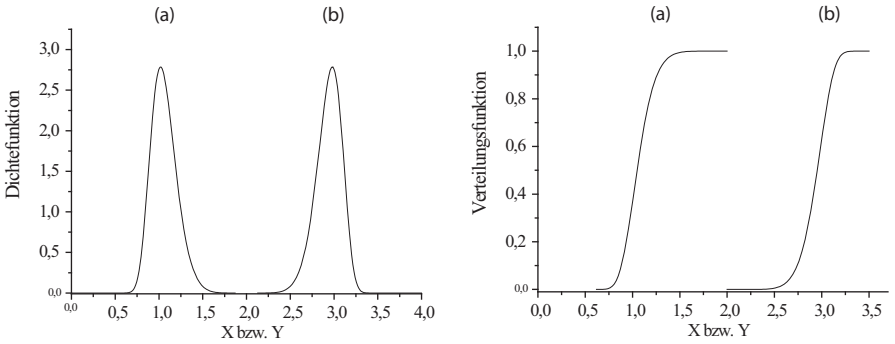
und für die Erwartungswerte und Varianzen findet man

$$\mathbb{E}(X) = \eta_0 - \frac{1}{a}\Gamma(1 - 1/\beta), \quad \mathbb{E}(Y) = \eta_0 + \frac{1}{a}\Gamma(1 - 1/\beta) \tag{5.53}$$

sowie

$$\mathbb{V}(X) = \mathbb{V}(Y) = \frac{1}{a^2}\left[\Gamma(1 + 2/\beta) - \Gamma^2(1 + 1/\beta)\right]. \tag{5.54}$$

Abbildung 5.5: Weibulldichten und -verteilungen – (a) nach (5.50), (b) nach (5.49); $\eta_0 = 2$, $a = 1$, $\beta = 7.5$.



Verteilungen vom Weibull-Typ ergeben sich auch dann, wenn die Verteilungen der Maxima und Minima unabhängiger zufälliger Größen $X_1, \ldots, X_n$ bzw. $Y_1, \ldots, Y_n$ gesucht werden; Verteilungen vom Weibull-Typ sind damit auch Extremwertverteilungen, vergl. Abschnitt 5.6.2.

Verteilungen vom Weibull-Typ treten nicht nur bei der Verteilung der Bruchfestigkeitein spröden Materials auf, sondern auch, wenn etwa Lebensdauern bzw. Wartezeiten oder Wirkungen der Dosis von Giften bzw. Medikamenten bzw. der Intensität von externen Stimuli auf Sinnesorgane (pychometrische Funktionen) diskutiert werden. Dabei kann es interessant sein, die Verteilungen von Dosiswirkungen und Wartezeiten, oder Stimuluswirkung und Reaktionszeit, zueinander in Beziehung zu setzen, vergl. Abschnitt 5.6.3.

5.6.2 Die Verteilung extremer Werte

Es soll die Verteilung des Maximums bzw. Minimums stochastisch unabhängiger, Weibull-verteilter Variablen betrachtet werden. Obwohl die allgemeine Theorie der Verteilung extremer Werte in Kapitel 7 behandelt wird, ist es nützlich, einige allgemeine Betrachtungen an den Anfang zu stellen.

Es seien $U_1, \ldots, U_n$ zufällige Veränderliche, und es sei

$$U_+ = \max(U_1, \ldots, U_n), \quad U_- = \min(U_1, \ldots, U_n).$$

Maximum: Es gilt $U_+ \leq u_+$, wenn alle $U_i \leq u_+$ sind, so dass

$$P(U_+ \leq u_+) = P(U_1 \leq u_+ \cap \cdots \cap U_n \leq u_+). \tag{5.55}$$

Mimimum: Es gilt $U_- > u_-$, wenn $U_i > u_-$ für alle U_i gilt. Dementsprechend folgt

$$P(U_- \leq u_-) = 1 - P(U_1 > u_- \cap \cdots \cap U_n > u_-). \tag{5.56}$$

Maxima unabhängiger zufälliger Variablen Die zufälligen Veränderlichen $X_1, \ldots, X_n$ seien unabhängig voneinander und Weibull-verteilt gemäß (5.49). Gesucht ist die Verteilung des Maximums der X_i, $i = 1, \ldots, n$, d.h. die Verteilung von $X = \max(X_1, \ldots, X_n)$. Mit $X_i = U_i$ folgt aus (5.55) und der Unabhängigkeit der X_i

$$P(X \leq x) = \prod_{i=1}^{n} \exp\left[-\left(\frac{\eta_{0i} - x}{a_i}\right)^{\beta_i}\right] = \exp\left[-\sum_{i=1}^{n}\left(\frac{\eta_{0i} - x}{a_i}\right)^{\beta_i}\right] \tag{5.57}$$

Sind die X_i außerdem identisch verteilt mit $a_i = a$ und $\beta_i = \beta$ und $\eta_{0i} = \eta_0$, erhält man

$$P(X \leq x) = \exp\left[-n\left(\frac{\eta_0 - x}{a}\right)^{\beta}\right] = \exp\left[-\left(\frac{\eta_0 - x}{b}\right)^{\beta}\right], \quad b = an^{-1/\beta}, \tag{5.58}$$

d.h. das Maximum unabhängig und identisch (5.49)-verteilter zufälliger Veränderlicher ist wieder (5.49)-verteilt, allerdings mit einem anderen Skalenfaktor.

Minima unabhängiger zufälliger Variablen Die zufälligen Veränderlichen $Y_1, \ldots, Y_n$ seien unabhängig voneinander und Weibull-verteilt nach (5.50). Gesucht ist die Verteilung von $Y = \min(Y_1, \ldots, Y_n)$. Mit $U_i = Y_i$ folgt aus (5.56) und der Unabhängigkeit der Y_i

$$P(Y \leq y) = 1 - \prod_{i=1}^{n} \exp\left[-\left(\frac{\eta_{0i} - y}{a_i}\right)^{\beta_i}\right] = 1 - \exp\left[-\sum_{i=1}^{n}\left(\frac{\eta_{0i} - y}{a_i}\right)^{\beta_i}\right] \tag{5.59}$$

Sind die Y_i überdies identisch verteilt mit $a_i = a$, $\beta_i = \beta$ und $\eta_{0i} = \eta_0$, erhält man

$$P(Y \leq y) = 1 - \exp\left[-n\left(\frac{\eta_0 - y}{a}\right)^{\beta}\right] = 1 - \exp\left[-\left(\frac{\eta_0 - y}{b}\right)^{\beta}\right], \quad b = an^{-1/\beta}. \tag{5.60}$$

Das Mimimum (5.50)-verteilter Variablen ist wiederum (5.50)-verteilt, allerdings mit anderem Skalenfaktor.

Das Mimimum (5.49)- und das Maximum (5.50)-verteilter Variablen Für das Minimum (5.49)-verteilter Variablen folgt aus (5.56) und (5.49) für den Fall unabhängiger und identisch verteilter Variablen

$$P(X_- \leq x) = 1 - \left(1 - \exp\left[-\left(\frac{\eta_0 - x}{a}\right)^{\beta}\right]\right)^n. \tag{5.61}$$

Für das Maximum (5.50)-verteilter Variablen folgt nach (5.55)

$$P(Y_+ \leq y) = \left(1 - \exp\left[-\left(\frac{y - \eta_0}{a}\right)^{\beta}\right]\right)^n. \tag{5.62}$$

5.6.3 Verweildauern und Warte- bzw. Reaktionszeiten

Lebensdauern, Verweildauern und Wartezeiten sind u. U. Weibull-verteilt. Es gibt zwei Möglichkeiten: (i) da Zeiten i. A. nur positive Werte annehmen können, kann man für ihre Verteilung gleich die Verteilungsfunktion (5.50) annehmen, oder (ii) man kann versuchen, von der Verteilung (5.49) zur Verteilung der Lebensdauer bzw. Wartezeit zu gelangen. Es wird hier von (ii) ausgegangen, weil dadurch noch einmal die Beziehung zwischen den beiden Verteilungsfunktionen illustriert wird.

Die zufällige Veränderliche X repräsentiere die Belastung eines Materials, die Belastung durch eine toxische Substanz oder durch ein Medikament, oder die Aktivierung eines neuronalen Systems. Für $X > S$ wird ein bestimmtes Ereignis E registriert: E kann ein Materialbruch sein, das Ableben durch Einwirkung eins Gifts, das Aufhören eines Kopfschmerzes nach Einnahme geeigneter Pillen, oder das Entdecken eines Stimulus. X sei Weibull-verteilt gemäß (5.49).

τ sei die Wartezeit bis zum Eintreten von E. τ sei eine Funktion von X. Im einfachsten Fall sind X und τ linear aufeinander bezogen; intuitiv plausibel ist aber auch eine nichtlineare Beziehung:

$$\tau = \begin{cases} b(2\eta_0 - X) + k, \quad b > 0,\ \eta_0 \geq X, \ \text{oder} \\ \left(\frac{2\eta_0 - X}{b}\right)^p, \quad X \leq \eta_0, \quad p > 0, \quad b > 0. \end{cases} \tag{5.63}$$

Diese Beziehungen zwischen τ und X sind mögliche Modelle für die Kopplung von Dosis oder Stimulusstärke und Reaktionszeit.

Fall 1: $\tau = b(2\eta_0 - X) + k$

Es ist

$$P(\tau \leq t) = P(b(2\eta_0 - X) + k \leq t) = 1 - P\left(X \leq 2\eta_0 - \frac{t-k}{b}\right),$$

wobei X (5.49)-verteilt sei. Dann ist

$$P(\tau \leq t) = 1 - \exp\left[-\left(\frac{t - (k + b\eta_0)}{ab}\right)^\beta\right] = 1 - \exp\left[-\left(\frac{t - \eta_1}{c}\right)^\beta\right], \tag{5.64}$$

mit $c = ab$, $t \geq \eta_1 = k + b\eta_0$. Die in der Literatur zur Verteilung von Verweildauern oft vorgefundene Verteilung

$$P(\tau \leq t) = 1 - \exp\left[-\left(\frac{t}{c}\right)^\beta\right], \quad t \geq 0$$

entspricht dem Spezialfall $\eta_1 = 0$, d.h. $k = -b\eta_0$. Für $\eta_1 > 0$ gibt es eine "Totzeit", während der keine Reaktion eintritt.

Fall 2: $\tau = ((2\eta_0 - X)/b)^p$

Es ist

$$\frac{d\tau}{dX} = -\frac{p}{b}\left(\frac{2\eta_0 - X}{b}\right)^{p-1} \leq 0,$$

d.h. je größer der Wert von X, desto kleiner ist der Wert der Wartezeit τ. Dann ist

$$\begin{aligned} P(\tau \leq t) &= P\left(\left(\frac{2\eta_0 - X}{b}\right)^p \leq y\right) \\ &= P\left(2\eta_0 - X \leq bt^{1/p}\right) \\ &= P(2\eta_0 - bt^{1/p} \leq X) = 1 - P\left(X \leq 2\eta_0 - bt^{1/p}\right) \\ &= 1 - \exp\left[-\left(\frac{bt^{1/p} - \eta_0}{a}\right)^\beta\right] \end{aligned} \tag{5.65}$$

Es sei $0 < p \neq 1$. Dann geht (5.65) in eine Weibull-Verteilung über genau dann, wenn $\eta_0 = 0$, denn dann folgt

$$P(\tau \leq t) = 1 - \exp\left[-\alpha t^\gamma\right], \quad \alpha = (b/a)^{1/\beta}, \quad \gamma = \beta/p. \tag{5.66}$$

Dann ist aber $X \leq 0$, d.h. Belastbarkeit, Aktivierung etc. wird durch eine negative Größe repräsentiert. Will man diese Situation vermeiden, muß $p = 1$ gesetzt werden, so dass die Beziehung zwischen X und der Wartezeit τ linear ist. Für diesen Fall erhält man

$$P(\tau \leq t) = 1 - \exp\left[-\left(\frac{t - \eta_1}{c}\right)^\beta\right], \quad t \geq \eta_1, \quad \eta_1 = \eta_0/b, \quad c = a/b, \tag{5.67}$$

also eine Weibull-Verteilung vom Typ (5.50). η_1 repräsentiert eine Minimalzeit; für Werte $t \leq \eta_1$ kann kein Bruch und keine Reaktion einsetzen. Der in der Literatur oft angenommene Fall $\eta_1 = 0$ ist dann mit dem Ansatz, die Verteilung von τ aus der Verteilung von X zu gewinnen, nur vereinbar, wenn ausschließlich negative Werte von X sinnvoll zu interpretieren sind. In dem Fall kann dann aber der allgemeinere Fall einer nichtlinearen Beziehung zwischen X und τ betrachtet werden.

5.6.4 Dosiswirkungen und Psychometrische Funktionen

Es sei m die Dosis eines Medikaments (oder, was das gleiche sein kann, eines Gifts), oder die Intensität eines Stimulus, der auf ein Sinnesorgan einwirkt. Gesucht ist eine Funktion, die die Beziehung zwischen m und der Wahrscheinlichkeit einer Reaktion abbildet, wobei die Reaktion im Ableben (im Falle einer Gabe von Gift), im Aufhören des Kopfschmerzes oder im Entdecken des Stimulus besteht.

Es sei g eine deterministische Funktion, die die Särke der Reaktion auf die Dosis bzw. Intensität m abbildet. Im Allgemeinen wird g eine nichtlineare Funktion sein, wobei aber eine Linearisierung der Form

$$g(m) \approx g(0) + mg'(0), \text{ mit } g'(0) = \left.\frac{dg(m)}{dm}\right|_{m=0} \tag{5.68}$$

möglich sei. Es sei $U = X + g(m)$ die Reaktion auf die Dosis bzw. den Stimulus, und das in Frage stehende Ereignis E trete ein, wenn $U > S$; X sei Weibull-verteilt vom Typ (5.49). Dann gilt

$$P(X + g(m) > S) = 1 - P(X \leq S - g(m)),$$

d.h.

$$\psi(m) = P(X > S - g(m)) = 1 - \exp\left[-\left(\frac{\eta_0 - (S - g(m))}{a}\right)^\beta\right], \quad \eta_0 - (S - g(m)) \geq 0. \tag{5.69}$$

Insbesondere sei g sei linearisierbar (d.h. es gelte (5.68), so dass $g(m)/a = k\,m$ mit $g(0) = 0$ und $k = g'(0)$. Man erhält

$$\psi(m) = 1 - \exp\left[-(\eta_1 + \alpha\, m)^\beta\right], \quad \eta_1 = (\eta_0 - S)/a, \quad \alpha = k/a. \tag{5.70}$$

Repräsentiert ψ eine psychometrische Funktion, so ist $\psi(0) > 0$, sofern $\eta_1 > 0$, d.h. es existieren dann genuine falsche Alarme (die Versuchsperson nimmt etwas wahr, obwohl kein Stimulus gezeigt wurde); das oft in Zusammenhang mit der Weibullfunktion postulierte Hochschwellenmodell – $\psi(0) = 0$ – beruht auf der stillschweigenden Annahme $\eta_0 = S$, denn dann ist $\eta_1 = 0$.

Wollte man zur Definition der psychometrischen Funktion von der Verteilung (5.50) ausgehen, muß man die Entscheidungsvariable so definieren, dass sie kleiner oder gleich S ist, damit der Stimulus entdeckt wird. Dazu sei Y das Rauschen im entdeckenden System, und die Reaktion $g(m)$ auf den Stimulus wirke rauschunterdrückend. Der Stimulus werde entdeckt, wenn $Y - g(m) \leq S$. Dann ist

$$\psi(m) = P(Y \leq S + g(m)) = 1 - \exp\left[-\left(\frac{S + g(m) - \eta_0}{a}\right)^\beta\right], \tag{5.71}$$

und für den Spezialfall $S = \eta_0$, $g(m)/a = km$ erhält man wieder

$$\psi(m) = 1 - \exp\left[-\alpha m^\beta\right], \quad \alpha = k^{1/\beta}. \tag{5.72}$$

Für den Fall $S - \eta_0 > 0$ sind echte falsche Alarme möglich. Die Funktionen (5.70) und (5.72) sind identisch, d.h. $\psi(m)$ ist mit der Interpretation des Entdeckens durch eine überschwellige Aktivierung oder durch die Reduktions des Rauschens unter eine Schwelle kompatibel.

5.7 Die log-Normalverteilung

5.7.1 Definition der Verteilung

Es sei Y eine normalverteilt zufällige Veränderliche, und es sei $X = \log Y$, womit der natürliche Logarithmus (also der zur Basis e) gemeint ist; dann heißt Y *log-normalverteilt.* Die log-Normalverteilung ergibt sich in vielen natur- und sozialwissenschaftlichen Untersuchungen; einige Mechanismen, die auf diese Verteilung führen, werden im Anschluß an die Definition gegeben.

Definition 5.7.1 *Es sei Y eine normalverteilte zufällige Veränderliche mit dem Erwartungswert $E(Y) = \mu$ und der Varianz $Var(Y) = \sigma^2$. Weiter sei die zufällige Veränderliche $X > 0$ durch $Y = \log X$ erklärt, wobei* log *der natürliche Logarithmus sei. Dann heißt X* log-normalverteilt.

Satz 5.7.1 *Die Dichte f_x ist durch*

$$f_x(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\log x - \mu}{\sigma}\right)^2\right), \quad 0 < \infty \tag{5.73}$$

gegeben. Weiter ist

$$E(X) = e^{\mu+\sigma^2/2} \tag{5.74}$$

$$Var(X) = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1) \tag{5.75}$$

$$Median = Modus = e^{\mu} \tag{5.76}$$

Beweis: Die Dichte ergibt sich leicht wie folgt: die Verteilungsfunktion von X sei $F(x) = P(X \leq x)$. Aber $X = \exp(Y)$, mithin ist $P(X \leq x) = P(e^Y \leq x) = P(Y \leq \log x)$, denn der Logarithmus ist eine monotone Funktion. Damit ist die Verteilungsfunktion von X auf die von Y zurückgeführt, und die ist bekannt. Die Dichte ergibt sich durch Differentiantion nach x: es ist

$$\frac{dP(Y \leq \log x)}{dx} = f_Y(\log x)\frac{1}{x}$$

und dies ist gleichbedeutend mit (5.73). Die Herleitung des Erwartungswertes und der Varianz findet man z.B. in Kendall und Stuart (1969), p.169. □

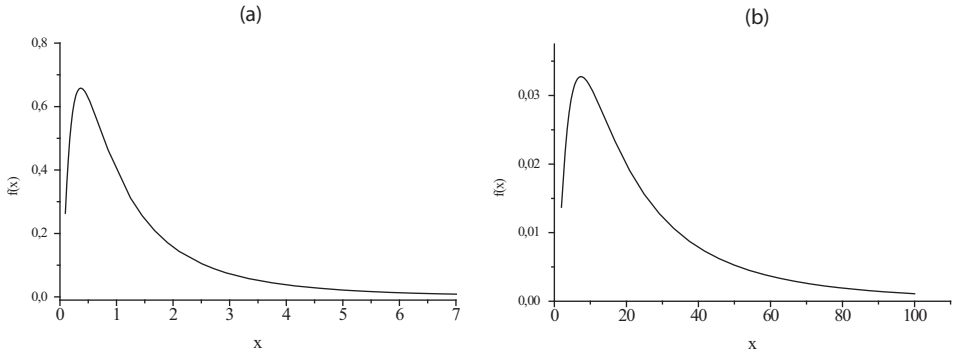
Die Dichte wird gelegentlich auch in etwas anderer Form geschrieben. Es sei $\mu = \log b$, und x werde durch $x - a$, $a \in \mathbb{R}$, ersetzt. Dann ist

$$\log(x - a) - \mu = \log(x - a) - \log b = \log(x - a)/b$$

und (5.73) kann in der Form

$$f(x) = \frac{1}{\sigma(x - a)} \exp\left(-\frac{1}{2\sigma^2}\left(\log(\frac{x - a}{b})\right)^2\right) \tag{5.77}$$

Abbildung 5.6: log-Normalverteilungen, (a) $\mu = 0, \sigma = 1$, (b) $\mu = 3, \sigma = 1$



geschrieben werden. Man schreibt auch abkürzend hierfür, X sei $LN(a, b, \sigma)$-verteilt. Mit bezug auf (5.77) ergibt sich für den Erwartungswert und die Varianz

$$E(X) = a + b \exp\left(\frac{\sigma^2}{2}\right) \tag{5.78}$$

$$Var(X) = b^2 \exp(\sigma^2)\left(\exp(\sigma^2) - 1\right) \tag{5.79}$$

$$Median = Modus = a + b \tag{5.80}$$

Offenbar bewirkt a - abhängig vom Vorzeichen - eine Verschiebung der Verteilung nach rechts oder links und kann deshalb als ein *Lokationsparameter* betrachtet werden. b bzw. $1/b$ wirkt als Skalenfaktor und bestimmt damit die Form der Verteilung. Für $a = 0$ geht die Dichte (5.77) in die Dichte (5.73) über. (5.77) ist die allgemeinere Form der Dichte einer log-normalverteilten Variablen. Die Abbildung 5.6 zeigt eine log-Normalverteilung für die Parameter $\mu = 3$ und $\sigma = .5$.

Beispiel 5.7.1 Gemäß Fechners Gesetz gilt für die Beziehung zwischen Reizintensität x und Empfindungsstärke y die Beziehung

$$y = \alpha \log \frac{x}{x_0} \tag{5.81}$$

wobei α eine Konstante und x_0 die Absolutschwelle ist. Nun hat aber für fixen Wert $X = x$ der Intensität die Wahrnehmungsintensität Y *nicht* stets den gleichen Wert $y(x)$, wie es die deterministische Beziehung (5.81) suggeriert. Es sei x_i die Stimulusintensität im i-ten Versuchsdurchgang. Die wahrgenommene Intensität sei - dem Fechnerschen Ansatz entsprechend - durch $\log x_i + \xi_i$ gegeben, wobei nun ξ_i der Wert einer normalverteilten zufälligen Größe ist mit dem Erwartungswert $E(\xi) = 0$ und der Varianz $Var(\xi) = \sigma^2$. ξ repräsentiert den Wert des sensorischen Rauschens im i-ten Versuchsdurchgang. Für festes x_i ist dann die Wahrnehmungsintensität über die Versuchsdurchgänge normalverteilt. Man kann nun die Werte X_i bestimmen derart, dass jeweils gerade $Y > S$ gilt, S

ein Schwellenwert. Da Y zufällig variiert, wird nun auch X_i variieren, und da $Y = \log X_i + \xi_i$, folgt $\log X_i = Y - X_i$, d.h. $\log X_i$ ist normalverteilt. Damit ist aber X_i dann log-normalverteilt.

Es sei angemerkt, dass hier angenommen wurde, dass das Rauschen *additiv* auf den Ausgang eines sensorischen Systems wirkt, das die Stimulusintensität logarithmisch transformiert. Diese Transformation ist selbst rauschfrei. Diese Annahme ist sicherlich nur eine erste Näherung. Weiter existiert eine Zahl η_i derart, dass $\log \eta_i = \xi_i$ ist. Demnach ist aber $\log X_i + \xi_i = \log(X_i \eta_i)$, so dass die Annahme des additiven Rauschen *nach* der Transformation äquivalent ist der Annahme eines *multiplikativen* Rauschens *vor* der Transformation. □

Sind die Variablen X_1 und X_2 $LN(a_1, b_1, \sigma_1)$- bzw. $LN(a_2, b_2, \sigma_2)$-verteilt, so ist die Variable $Y = (X_1 - a_1)(Y_2 - a_2)$ $LN(0, b_1 b_2, \sqrt{\sigma_1^2 + \sigma_2^2})$-verteilt. Dieses Resultat ergibt sich aus der Tatsache, dass die *Summe* normalverteilter Variablen wieder normalverteilt ist, und die Summe der Logarithmen zweier Zahlen gleich dem Logarithmus ihres Produktes ist.

Die log-Normalverteilung ist offenbar eng mit der Normalverteilung verwandt. Es gibt Autoren (z.B. Koch(1975)), die argumentieren, dass die Log-Normalveteilung für Anwendungen der Statistik in der Biologie und verwandten Gebieten - eben auch der Psychologie - fundamental sei. Koch asrghumentiert, dass die meisten Größen in der Natur normalverteilt seien und Organismen häufig eine logarithmische Transformation der auf sie einwirkenden Größen vornehmen. Diese Argumentation ist in ihrer Allgemeinheit sicherlich nicht besonders zwingend, aber gelegentlich naheliegend. Eine sehr allgemeine Diskussion der log-Normalverteilung und ihrer Anwendungen findet man in Aitchinson und Brown (1963). Weitere Modelle aus der Biologie, etwa von Wachstumsprozessen oder Dosiswirkungen, die auf die log-Normalverteilung beobachtbarer Daten führen, wird in Koch (1966) diskutiert. Hofstätter (1964) betrachtet die log-Normalverteilung im Kontext gruppendynamischer Prozesse, er spricht dann von der *Kapteynschen Verteilung*. So sei etwa die Häufigkeit von Diskussionsbeiträgen in sozialen Gruppen log-normalverteilt. Seine Betrachtungen über die Mechanismen, die gerade zu dieser Verteilung führen, bleiben allerdings diffus. Eine Reihe von Anwendungen der log-Normalverteilung wird in Sun (2004) gegeben[2], s. a. Limpert, Stahel & Abbt (2001).

Im Folgenden werden einfache Modelle für die log-Normalverteilung vorgestellt:

5.7.2 Mechanismen, die auf Log-Normalverteilungen führen

Herleitungen der log-Normalverteilung machen oft vom *Zentralen Grenzwertsatz* Gebrauch. Dazu seien $X_1, \ldots, X_n$ unabhängig verteilte Variable; die X_j müssen

[2]Dieser Artikel wird auf meiner Webseite – Statistik – verfügbar gemacht.

nicht normalverteilt sein. Dann ist die Summe $X = X_1 + \cdots + X_n$ approximativ normalverteilt mit dem Erwartungswert

$$\mathbb{E}(X) = \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n) \tag{5.82}$$

und der Varianz

$$\mathbb{V}(X) = \sigma_1^2 + \cdots + \sigma_j^2, \tag{5.83}$$

$\sigma_j = \mathbb{V}(X_j)$. Dies gilt streng für $n \to \infty$, es zeigt sich aber, dass $n > 5$ bereits eine sehr gute Approximation liefert.

Gibrats Gesetz

Gibrat (1931) untersuchte das Wachstum von Firmen und stellte die Hypothese auf, dass sie proportional zu ihrer Größe wachsen, wobei die Wachstumsrate zufällig verteilt ist. Demnach soll gelten

$$x_t - x_{t-1} = \varepsilon_t x_{t-1}, \tag{5.84}$$

wobei x_{t-1} und x_t die Firmengrößen zu den Zeiten $t-1$ und t sind und ε_t die zufällige Wachstumsrate zur Zeit t ist. Es folgt

$$x_t = x_{t-1} + \varepsilon_t x_{t-1} = (1 + \varepsilon_t) x_{t-1}$$

und daraus durch sukzessives Einsetzen schließlich

$$x_t = x_0(1 + \varepsilon_1)(1 + \varepsilon_2) \cdots (1 + \varepsilon_t). \tag{5.85}$$

Ist das Zeitintervall $[t-1, t)$ hinreichend klein, so dass $\log(1 + \varepsilon_t) \approx \varepsilon_t$, so folgt

$$\log x_t = \log x_0 + \varepsilon_1 + \varepsilon_2 + \cdots + \varepsilon_t \tag{5.86}$$

und unter den Bedingungen des Zentralen Grenzwertsatzes ist für hinreichend großen Wert für t und $\mu = \mathbb{E}(\varepsilon_t)$, $\sigma^2 = Var(\varepsilon_t)$

$$\log x_t - \log x_0 = \log\left(\frac{x_t}{x_0}\right) \sim N(t\mu, t\sigma^2). \tag{5.87}$$

Dies bedeutet, dass die Firmengrößen für hinreichend großen t-Wert log-normalverteilt sind.

Natürlich gilt dieses Resultat nicht nur für Firmengrößen, sondern für jede Größe, für die der Ansatz (5.84) gemacht werden kann. Unter anderem sind die Vermögen in einer Population (in einem Land) in guter Näherung log-normalverteilt, – bis auf große Vermögen, für die die Pareto-Verteilung gilt. Der Grund für die Log-Normalverteilung ist, dass "normale" Einkommen bzw Vermögen in erster Näherung proportional zum vorhandenen Vermögen wachsen, was zunächst eine Exponentialverteilung der Vermögen impliziert. Da aber die Zinssätze zufällig und in guter Näherung unabhängig vom Vermögen variieren, ergibt sich eine Log-Normalverteilung der Einkommen.

Erfolge von Maßnahmen: Shockleys Modell

Vielfach hat eine Maßnahme nur dann Erfolg, wenn eine Reihe von Teilmaßnahmen erfolgreich abgeschlossen wird. Der Erfolg bei jeder Teilmaßnahme sei *probabilistisch*, dh er trete nur mit einr bestimmten Wahrscheinlichkeit $0 \leq p \leq 1$ ein (der Fall $p = 1$ wird also nicht ausgeschlossen, aber der Fall $p = 0$ auch nicht). Weiter seien die Erfolge bei den einzelnen Teilmaßnahmen statistisch unabhängig voneinander. Gibt es n Teilmaßnahmen, so ist die Wahrscheinlichkeit eines Gesamterfolges durch

$$p = p_1 p_2 \cdots p_n = \prod_{i=1}^{n} p_i \tag{5.88}$$

gegeben. Die p_i-Werte können ihrerseits zufällig variieren: Ein Kunde ruft an und hat ein Problem, der Kundenberater (ein Techniker etc) muß nun eine Reihe von Teilproblemen lösen – vom Kunden eine Spezifikation des Problems bekommen, die es ermöglicht, überhaupt einen Lösungsansatz zu finden, etc. Für den Kundenberater gilt jetzt (5.88), ob er also Erfolg hat, hängt von seinen p_i-Werten ab. Für den j-ten Kundenberater hat man damit

$$p(E_j) = \prod_{i=1}^{n} p_{ij}, \tag{5.89}$$

wobei $p(E_j)$ die Wahrscheinlichkeit (p) des Erfolges (E_j) eben des Beraters B_j ist. Die p_{ij} können für jeden Berater m ir der Zeit variieren (Montags sind einige nicht so gut drauf, andere sind Freitags nicht so gut drauf, oder B_j kriegt an einem Mittwoch Zahnschmerzen und ist speziell an *diesem* Mittwoh nicht gut drauf, etc etc Man könnte also auch noch einen Zeitindex einführen: $p_{ij} = p_{ij}(t)$. Das Wesentliche ist aber etwas anderes: es sei $y = \log x$, $x > 0$; dann ist $e^y = \exp(\log x) = x$. Dann hat man von (5.88)

$$\log p = \sum_{i=1}^{n} \log p_i,$$

und nach der vorangegangenen Überlegung hat man

$$p = \exp(\log p) = \exp\left(\sum_{i=1}^{n} \log p_i\right). \tag{5.90}$$

Nun seien die p_i aber zufällige Veränderliche. Dann kann man den Zentralen Grenzwertsatz anwenden und findet

$$y = \sum_{i=1}^{n} \log p_i \sim N(\mu_y, \sigma_y^2), \tag{5.91}$$

d.h. die Summe der Logarithmen ist approximativ normalverteilt mit Erwartungswert (= Mittelwert über *alle* Wert von y) μ_y und Varianz σ_y^2. p, die Wahrscheinlichkeit des Gersamterfolgs, ist dann natürlich ebenfalls eine zufällige Variable, und da $p = e^y$ ist folgt, dass p log-normalverteilt ist.

Zur Erläuterung: p ist – im Sinne von (5.89) – eine zufällige Veränderliche. Die Anzahl der Erfolge bei einer Stichprobe ist dann $Np(E_j)$, sie wird also von Berater zu Berater und möglicherweise auch von Tag zu Tag, Woche zu Woche etc zufällig variieren. (N ist die Größe der Stichprobe von Erfolgen, die ihr betrachtet).

> Das hier vorgeschlagene Modell ist ursprünglich von Shokley[3] (1957) entwickelt worden. Shockley betrachtete die Anzahl der Publikationen zuerst seiner Mitarbeiter, dann allgemein, und fand, dass diese Anzahlen in sehr guter Näherung log-normalverteilt sind. Shockley ging davon aus, dass es mindestens acht Fähigkeiten oder Eigenschaften bedarf, um einen wissenschaftlichen Artikel zu veröffentlichen: (1) die Fähigkeit, ein geeignetes Problem zur Bearbeitung auszusuchen, (2) die Kompetenz, das Problem zu bearbeiten, (3) die Fähigkeit, ein relevantes Resultat als solches zu erkennen, (4) die Fähigkeit, einen geeigneten Schlußpunkt zu setzen und mit dem Verfassen des Manuskripts zu beginnen, (5) die Fähigkeit, die Resultate und Schlußfolgerungen adäquat zu präsentieren, (6) die Fähigkeit, die Kritiken anderer zu integrieren, die von anderen Wissenschaftlern gemacht werden, (7) den festen Willen zu haben, das Manuskript fertigzustellen und einzureichen, (8) die Fähigkeit haben, positiv auf die Kritiken der Gutachter zu reagieren. Es sei p_i die Wahrscheinlichkeit, über die i-te dieser Komponenten zu verfügen, und weiter seien die Komponenten stochastisch unabhängig voneinander. Die Wahrscheinlichkeit, dass ein Wissenschaftler einen Artikel publiziert, ist dann durch
>
> $$P = p_1 p_2 \cdots p_8 \tag{5.92}$$
>
> gegeben, – der Rest folgt wie oben.

Autokatalytische Prozesse

Allgemein kann man einen autokatalytischen Prozess der Form

$$\frac{dX}{dt} = \pm kX, \tag{5.93}$$

betrachten. Es folgt

$$\frac{dX}{X} = kdt.$$

[3] William Bradford Shockley Jr. (1910 – 1989), Physiker; erfand zusammen mit John Bardeen und Walter Houser Brattain den Transistor. Die drei bekamen dafür 1956 den Nobel-Preis für Physik. Shockley kommerzialisierte seine Erfindung, aber seine Mitarbeiter rebellierten und machten sich ihrerseits selbstständig. Dies war der Startpunkt für das Silicon Valley. Er war auch Anhänger eugenischer Ideen und schlug vor, das genetisch Benachteiligten eine finanzielle Belohnung bekommen sollten, wenn sie sich freiwillig sterilisieren ließen, – generell hielt er Schwarze für genetisch bedingt weniger intelligent als Weiße, und da sie sich stärker vermehrten als Weiße, befürchtete Shockley eine "retrogression in human evolution", vergl. den Nachruf in der New York Times http://www.nytimes.com/learning/general/onthisday/bday/0213.html.

Nun ist $\int dX/X = \log X + c$, dh

$$\frac{1}{X} = \frac{d\log X}{dX},$$

so dass unter Berücksichtigung von (5.93)

$$\frac{d\log X}{dt} = \frac{d\log X}{dX}\frac{dX}{dt} = \pm\frac{1}{X}kX = \pm k \Rightarrow \log X(t) = \log X_0 \pm kt \tag{5.94}$$

Die Konstante k sei nicht wirklich konstant, sondern variiere zufällig, so dass zur Zeit t

$$(\log X)_{t+1} = (\log X)_t \pm k_t, \tag{5.95}$$

wobei k_t der zufällige Wert von k zur Zeit t sei. Man hat wieder

$$(\log X)_{t+1} = (\log X)_0 \pm \sum_{s=0}^{t} k_s, \tag{5.96}$$

und der Zentrale Grenzwertsatz liefert eine Log-Normalverteilung, auch wenn die k_t nicht Gauß-verteilt sind.

5.8 Die Cauchy-Verteilung

Definition 5.8.1 *Es sei X eine zufällige Veränderliche mit der Dichte*

$$f(x) = \frac{1}{\pi}\frac{\lambda}{\lambda^2 + (x-\mu)^2}, \quad -\infty < x < \infty \tag{5.97}$$

Dann heißt X Cauchy-verteilt *mit den Parametern λ und μ.*

Satz 5.8.1 *Die charakteristische Funktion der Cauchy-Verteilung ist durch*

$$\mathbf{F}(\omega) = e^{i\mu\omega - \lambda|\omega|} \tag{5.98}$$

gegeben.

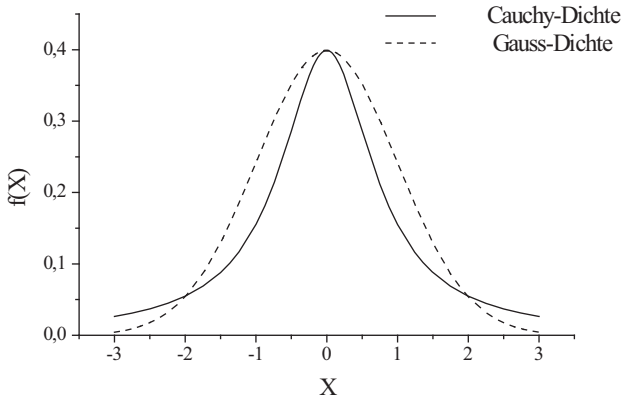
Beweis: Wie üblich durch Fourier-Transformation der Dichte $f(x)$. □

Man kann zunächst aus (5.8.1) folgern, dass $f(x)$ symmetrisch ist; dies folgt aus der Tatsache, dass x über den Ausdruck $(x-\mu)^2$ in die Definition von f eingeht. Da f für $|x| \to \mu$ maximal wird, folgt, dass der *Modalwert* von f durch μ gegeben ist.

μ ist aber nicht gleichzeitig auch der Erwartungswert von X. Um den zu bestimmen, muß man $d\mathbf{F}(\omega)/d\omega|_{\omega=0}$ finden. Nach (5.98) findet man

$$\frac{d\mathbf{F}(\omega)}{d\omega} = \begin{cases} (i\omega - \lambda\omega)e^{i\mu\omega - \lambda\omega}, & \omega > 0, \\ (i\mu + \lambda\omega)e^{i\mu\omega + \lambda\omega}, & \omega < 0. \end{cases} \tag{5.99}$$

Abbildung 5.7: Gaußsche und Cauchysche Dichten



Für $\omega = 0$ ist aber $d\mathbf{F}(\omega)/d\omega$ *gar nicht definiert.* Dies bedeutet, dass der Erwartungswert einer Cauchy-verteilten Variablen X nicht erklärt ist! Der Erwartungswert ist ja auch gemäß

$$E(X) = \lim_{a\to\infty} = \int_{-a}^{a} x f(x)\, dx$$

definiert. Wenn $E(X)$ nicht existiert, so heißt dies, dass das Integral auf der rechten Seite nicht existiert, d.h. für $a \to \infty$ keinem endlichen Wert zustrebt. Zwar folgt aus (5.8.1), dass zwar $f(x) \to 0$ für $|x| \to \infty$, aber diese Eigenschaft von f ist offenbar nicht hinreichend dafür, dass das Integral einen endlichen Wert annimmt: für $|x| \to \infty$ strebt $xf(x)$ nicht gegen Null. Man kann sagen, dass im Gegensatz zur Normal- und zur logistischen Verteilung die *Enden* der Cauchy-Verteilung zu langsam gegen Null gehen. Aus dieser Argumentation folgt sofort, dass auch die Varianz einer Wie aus der Abb. 5.7 ersichtlich ist, hat die Cauchysche Dichte eine der Gaußschen Dichte ähnliche Form: die beiden Dichten sind glockenförmig. Sie unterscheiden sich aber hinsichtlich der Art, mit der sie gegen Null gehen. Die Cauchysche Dichte geht wesentlich langsamer gegen Null als die Normalverteilung. Diese Aussagen gelten für alle Werte des Parameters λ der Cauchy-Verteilung; hier wure ein Wert von λ gewählt, für der maximale Wert von $f(x)$ für beide Verteilungen gleich groß ist.

Die Summe stochastisch unabhängiger Cauchy-verteilter Variablen ist ebenfalls Cauchy-verteilt. Dies läßt sich sofort aus (5.98) ablesen. Denn die charakteristische Funktion der Summe ist durch das Produkt der charakteristischen Funktionen gegeben, d.h.

$$\mathbf{F}(\omega) = \prod_{i=1}^{n} e^{i\mu_i\omega - \lambda_i|\omega|} = \exp\left(i\omega \sum_{i=1}^{n} \mu_i - |\omega| \sum_{i=1}^{n} \lambda_i\right) \tag{5.100}$$

und dies ist die charakteristische Funktion einer Cauchy-verteilten Variablen mit den Parametern $\sum_i \mu_i$ und $\sum_i \lambda_i$. Damit gehört die Cauchy-Verteilung zur Klasse der *stabilen Verteilungen.*

Die Cauchy-Verteilung ergibt sich gelegentlich aus dem Umgang mit Normalverteilungen, vergl. Abschnitt 5.11.2.

5.9 Die Beta-Verteilungen

Definition 5.9.1 *Die zufällige Veränderliche habe die Dichte*

$$f(x) = \frac{1}{B(p,q)} x^{p-1}(1-x)^{q-1}, \quad 0 < x < 1 \tag{5.101}$$

mit

$$B(p,q) = \int_0^1 \xi^{p-1}(1-\xi)^{q-1}\, d\xi = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}, \tag{5.102}$$

wobei $\Gamma(p) = \int_0^\infty y^{n-1}e^{-y}dy, n > 0$ *die Gamma-Funktion ist. Dann hat* X *die* Beta-Verteilung 1. Art.

Satz 5.9.1 *X habe die Beta-Verteilung 1.Art. Dann hat X den Erwartungswert*

$$E(X) = \frac{p}{p+q} \tag{5.103}$$

und die Varianz

$$Var(X) = \frac{pq}{(p+q)^2(p+q+1)} \tag{5.104}$$

Beweis: Der Beweis besteht wieder in der Anwendung des üblichen Verfahrens.

Definition 5.9.2 *die zufällige Veränderliche X habe die Dichte*

$$f(x) = \begin{cases} \frac{1}{B(p,q)} \frac{x^{p-1}}{(1+x)^{p+2q}}, & x > 0 \\ 0, \quad x \leq 0 \end{cases} \tag{5.105}$$

Dann heißt X Beta-verteilt 2. Art

Satz 5.9.2 *X habe die Beta-Verteilung 2. Art. Dann hat X den Erwartungswert und die Varianz*

$$E(X) = \frac{p}{q-1} \tag{5.106}$$

$$Var(X) = \frac{p(p+q-1)}{(q-1)^2(q-2)} \tag{5.107}$$

Beweis: Anwendung der üblichen Techniken. □

Der Erwartungswert ist also nur für den Fall $q > 1$ definiert, und die Existenz der Varianz setzt $q > 2$ voraus.

Die Beta-Verteilungen ergeben sich als Funktionen von Gamma-verteilten Variablen. So seien die Variablen X und Y stochastisch unabhängig und Gamma-verteilt mit dem Parameter $\lambda = 1$ und p_1, p_2. Es wird nun die Verteilung des Quotienten $Z = X/(X + Y)$ gesucht. Da $X + Y$ wieder Gamma-verteilt ist, muß der Quotient zweier unabhängiger, Gamma-verteilter Variablen hergeleitet werden. Man findet, dass Z die Beta-Verteilung 2. Art hat mit den Parametern $p = p_1, q = p_2$. Weiter läßt sich zeigen, dass, wenn Z Beta-verteilt 2. Art ist, $1/Z$ ebenfalls die Beta-Verteilung 2. Art hat.

5.10 Stichprobenverteilungen

Anhand von Stichproben geschätzte Parameter wie Mittelwerte, die Varianzen etc sind selbst zufällige Veränderliche, da sie von den Realisierungen der untersuchten zufälligen Variablen abhängen. Für inferenzstatistische Fragestellungen ist es wichtig, die Wahrscheinlichkeitsverteilungen der Statistiken zu kennen.

5.10.1 Die χ^2-Verteilung

Es seien $X_1, \ldots, X_n$ unabhängig und identisch verteilte Variablen; die Dichtefunktion sei

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{x^2}{\sigma^2}\right), \tag{5.108}$$

d.h. die X_j seien $N(0, \sigma^2)$-verteilt. Die Annahme $\mathbb{E}(X) = \mu = 0$ ist für die folgenden Betrachtungen keine Einschränkung, da man dann statt X die Variable $Y = X - \mu$ betrachten könnte. Gesucht ist die Verteilung der Summe

$$\chi_n^2 = X_1^2 + X_2^2 + \cdots + X_n^2. \tag{5.109}$$

Gesucht ist die Verteilungs- bzw. Dichtefunktion von χ_n^2.

Offenbar ergibt sich die Verteilung von χ_n^2 als Faltung der Verteilungen der X_j^2. Die Verteilung von X^2 ist in Beispiel 3.4.4 (Seite 89) bereits hergeleitet worden: für $y = x^2$ hat man

$$g(y) = \begin{cases} \frac{1}{2\sqrt{y}}[f(\sqrt{y}) + f(-\sqrt{y}], & y > 0 \\ 0, & y \leq 0 \end{cases}$$

und angewandt auf (5.108) erhält man

$$g(y) = \frac{1}{\sigma\sqrt{2\pi}} y^{-1/2} \exp\left(-\frac{1}{2}\frac{y}{\sigma^2}\right). \tag{5.110}$$

$g(y)$ ist eine Gamma-Verteilung (vergl. Abschnitt 5.3):

$$f(y) = \frac{\lambda^p}{\Gamma(p)} y^{p-1} e^{\lambda y}$$

wird mit $p = 1/2$, $\lambda = 1/\sigma^2$ und wegen $\Gamma(1/2) = \sqrt{\pi}$ gerade $g(y)$ in (5.110). Y hat den Erwartungswert und die Varianz

$$\mathbb{E}(Y) = \frac{p}{\lambda}, \quad Var(Y) = \frac{p}{\lambda^2}. \tag{5.111}$$

Wie in Abschnitt 5.3 gezeigt wurde ist die Faltung von n unabhängigen Gamma-Verteilungen wieder eine Gamma-Verteilung; man erhält die

χ^2-Verteilung mit n Freiheitsgraden:

$$f(\chi^2) = \begin{cases} \frac{1}{2^{n/2}\sigma^n \Gamma(n/2)} \exp\left(-\frac{\chi^2}{2\sigma^2}(\chi^2)^{\frac{n}{2}-1}\right), & \chi^2 > 0 \\ 0, & \chi^2 < 0. \end{cases} \tag{5.112}$$

Dies ist eine Gamma-Verteilung mit $p = n$, $\lambda = 1/\sigma^2$.

Die Gamma-Verteilung hat den Erwartungswert $\mathbb{E}(X) = p/\lambda$ und die Varianz $Var(X) = p/\lambda^2$. Für die χ^2-Verteilung (die ja eine Gamma-Verteilung ist) ergibt sich aus (5.111)

$$\mathbb{E}(\chi^2) = n\sigma^2, \quad var(\chi^2) = 2n\sigma^4. \tag{5.113}$$

Der Fall $\sigma^2 = 1$: standardnormalverteilte Variable: Die χ^2-Verteilung wird auch als Verteilung von

$$\chi^2 = Z_1^2 + \cdots + Z_n^2, \tag{5.114}$$

wobei die Z_j unabhängig und $N(0,1)$-verteilt sind. Dann gilt

$$\mathbb{E}(\chi^2) = n, \quad Var(\chi^2) = 2n. \tag{5.115}$$

Dieser Fall ist in (5.112) für $\sigma^2 = 1$ enthalten.

Beispiel 5.10.1 Es seien $X_1, \ldots, X_n$ stochastisch unabhängige Realisierungen einer $N(\mu, \sigma^2)$-verteilten zufälligen Veränderlichen. Dann ist

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{5.116}$$

der Stichprobenmittelwert, und

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{x})^2 \tag{5.117}$$

ist die Stichprobenvarianz. Diese Definition der Stichprobenvarianz entspricht der der Populationsvarianz $\sigma^2 = \mathbb{E}(X - \mu)^2$.

$\bar{x}$ ist eine zufällige Veränderliche mit dem Erwartungswert

$$\mathbb{E}(\bar{x}) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i)\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \frac{n}{n}\mu = \mu, \tag{5.118}$$

d.h. die Schätzung $\bar{x}$ für μ ist verzerrungsfrei (bias-frei). Für die Varianz von $\bar{x}$ erhält man

$$Var(\bar{x}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}. \tag{5.119}$$

Um eine mögliche Verzerrung der Schätrzung s^2 zu sehen muß, der Definition von σ^2 entsprechend, noch die Variation der X_i um μ in die Formel für s^2 eingebracht werden. Es ist

$$\begin{aligned}
X_i - \bar{x} &= X_i - \mu + \mu - \bar{x} = (X_i - \mu) - (\bar{x} - \mu) \\
(X_i - \bar{x})^2 &= (X_i - \mu)^2 + (\bar{x} - \mu)^2 - 2(\bar{x} - \mu)(X_i - \mu) && (5.120)\\
\sum_{i=1}(X_i - \bar{x})^2 &= \sum_{i=1}^{n}(X_i - \mu)^2 + n(\bar{x} - \mu)^2 - 2n(\bar{x} - \mu)^2 && (5.121)\\
&= \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{x} - \mu)^2 && (5.122)
\end{aligned}$$

Es folgt

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - (\bar{x} - \mu)^2,$$

und wenn man die Erwartung $\mathbb{E}(s^2)$ nimmt, ergibt sich

$$\mathbb{E}(s^2) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i - \mu)^2 - \mathbb{E}(\bar{x} - \mu)^2 = \sigma^2 - \frac{\sigma^2}{n} = \sigma^2\left(1 - \frac{1}{n}\right)$$

Der Erwartungswert von s^2 ist also nicht gleich σ^2, sondern um den Faktor $(n-1)/n$ verzerrt. Multiplikation mit $n/(n-1)$ liefert dann

$$\frac{n}{n-1}\mathbb{E}(s^2) = \sigma^2,$$

d.h.

$$\hat{s}^2 = \left(\frac{n}{n-1}\right)\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{x})^2 \tag{5.123}$$

ist eine unverzerrte Schätzung von σ^2.

Um die Stichprobenverteilung von $\hat{s}^2$ zu finden, betrachtet man noch einmal (5.122), d.h.

$$\sum_{i=1}(X_i - \bar{x})^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{x} - \mu)^2.$$

Offenbar ist

$$\begin{aligned}\sum_{i=1}(X_i-\bar{x})^2 &= \sigma^2\sum_{i=1}^{n}\frac{(X_i-\mu)^2}{\sigma^2} - n\sigma_{\bar{x}}^2\frac{(\bar{x}-\mu)^2}{\sigma_{\bar{x}}} = \sigma^2\sum_{i=1}^{n}z_i^2 - \sigma^2 z^2 \\ &= \sigma^2\sum_{i=1}^{n-1}z_i^2 = \sigma^2\chi_{n-1}^2 \qquad (5.124)\end{aligned}$$

Daraus folgt, dass

$$\frac{ns^2}{\sigma^2} = \frac{(n-1)\hat{s}^2}{\sigma^2} = \chi_{n-1}^2, \qquad (5.125)$$

d.h. die Quotienten auf der linken Seite sind χ^2-verteilt mit $n-1$ Freiheitsgraden.

□

5.11 Die mehrdimensionale Normalverteilung

5.11.1 Der allgemeine Fall

Definition 5.11.1 *Es sei $X = (X_1, \cdots, X_n)'$ ein n-dimensionaler Vektor, dessen Komponenten X_i, $i = 1, \cdots, n$ zufällige Variable sind. Es sei $\mu = (\mu_1, \cdots, \mu_n)'$ der Vektor der Erwartungswerte, d.h. $\mu_i = E(X_i)$, und es sei $\Sigma = (\sigma_{ij})$ die Matrix der Kovarianzen σ_{ij} zwischen den Variablen X_i, X_j bzw. der Varianzen $\sigma_{ii} = \sigma_i^2$ der X_i. ist die Dichte von X durch*

$$f(x_1, \cdots, x_n) = k^{-1}\exp\left(-\frac{1}{2}(X-\mu)'\Sigma^{-1}(X-\mu)\right) \qquad (5.126)$$

gegeben, so heißt X n-dimensional normalverteilt. *Dabei ist*

$$k = (2\pi)^{n/2}\left(\prod_{i=1}^{n}\lambda_i\right)^{1/2}$$

eine Normierungskonstante; λ_i, $i = 1, \cdots, n$ sind die Eigenwerte von Σ. Man schreibt dann auch abkürzend $X \sim N_n(\mu, \Sigma)$.

Anmerkung: Man kann die Dichte von X durch den Ausdruck (5.126) definieren und dabei lediglich fordern, dass Σ eine positiv-definite Matrix ist. Dann kann man *folgern*, dass Σ die Varianz-Kovarianz-Matrix und μ der Vektor der Erwartungswerte ist, vergl. etwa Seber (1977), p. 22.

Die charakteristische Funktion der n-dimensionalen Normalverteilung ist durch

$$\mathbf{F}(\omega) = \exp\left(i\omega'\mu - \frac{1}{2}\omega'\Sigma\omega\right) \qquad (5.127)$$

gegeben, wobei ω ein n-dimensionaler Vektor ist (vergl. Pruscha (1989), p. 20)

Es werde noch der Spezialfall, dass die Komponenten von X, d.h. die x_i, $i = 1, \cdots, n$, stochastisch unabhängig sind. Dann verschwinden die Kovarianzen zwischen den x_i, so dass

$$\Sigma = \mathrm{diag}(\sigma_1^2, \cdots, \sigma_n^2)$$

. Dann ist

$$\Sigma^{-1} = \mathrm{diag}(\sigma_1^{-1}, \cdots, \sigma_n^{-1})$$

und (5.126) kann in der Form

$$f(x_1, \cdots, x_n) = k^{-1} \exp\left(-\sum_{i=1}^{n} \frac{x_i^2}{2\sigma_i^2}\right) \tag{5.128}$$

geschrieben werden.

Jetzt sei Σ wieder beliebig. Da Σ positiv-definit ist, existiert eine orthogonale Matrix T derart, dass

$$T'\Sigma T = \mathrm{diag}(\lambda_1, \cdots, \lambda_n) \tag{5.129}$$

Die Spalten von T sind die Eigenvektoren von Σ; sie sind orthogonal, und ohne Einschränkung der Allgemeinheit können sie als normiert vorausgesetzt werden, so dass $T'T = I$, I die Einheitsmatrix. Es sei nun Y ein zufälliger Vektor derart, dass $X - \mu = TY$. Dann folgt

$$YY' = T'(X-\mu)(X-\mu)'T = T'\Sigma\, T = \Lambda$$

Aber Λ ist eine Diagonalmatrix, also müssen die Komponenten von Y unkorreliert sein. In der Tat gilt

Satz 5.11.1 *Es sei X ein zufälliger Vektor mit der Dichte (5.126). Ist T die Matrix der normierten Eigenvektoren von Σ und ist $Y = T'(X-\mu)$ so ist die Dichte von Y durch*

$$f(y_1, \cdots, y_n) = \prod_{i=1}^{n} (2\pi\lambda_i)^{-1/2} \exp\left(-\frac{y_i^2}{2\lambda_i}\right) \tag{5.130}$$

gegeben.

Beweis: Es sei Σ_y die Varianz-Kovarianz-Matrix der Komponenten von Y, und Σ_x sei die entsprechende Matrix für X, d.h. es sei $\Sigma_y = YY'$, $\Sigma_x = (X-\mu)(X-\mu)'$. Dann ist aber $YY' = T'(X-\mu)(X-\mu)'Y = T'\Sigma_x T = \Lambda = \mathrm{diag}\lambda_1, \cdots, \lambda_n)$. Dann sind aber die Komponenten von Y stochastisch unabhängig mit Varianzen $Var(y_i) = \lambda_i$ und die Dichte von Y muß die Form (5.128) haben; dies wird aber in (5.130) behauptet. $\square$

$f(y_1, \cdots, y_n)$ ist das Produkt von Dichten für normalverteilte Variable y_i. Ein Vergleich mit (5.18) zeigt, dass die Eigenwerte λ_i offenbar die Varianzen der y_i sind. In Kapitel 10 (Band I) wurde die Hauptachsentransformation einer Datenmatrix Z eingeführt; sie liefert eine Darstellung der (Spalten-)Vektoren der Matrix Z als Linearkombination orthogonaler Vektoren, die sich als Eigenvektoren von $Z'Z$ ergeben. Es war dort gezeigt worden, dass die Projektionen der Punkte, die etwa die "Tests" (Spalten von Z) repräsentieren, auf die einzelnen Achsen Varianzen haben, die durch die entsprechenden Eigenwerte der Eigenvektoren gegeben sind. Faßt man die Elemente (z_{ij}) der Matrix Z als Realisierungen von zufälligen Veränderlichen auf, die als Komponenten eines zufälligen Vektors definiert sind, und nimmt man weiter an, dass die Verteilung dieser Komponenten durch eine n-dimensionale Normalverteilung gegeben ist, so zeigt (5.130), dass die Hauptachsentransformation von der Varianz-Kovarianzmatrix Σ der Transformation von X auf einen zufälligen Vektor Y mit unabhängigen Komponenten mit den Varianzen λ_i führt. Die latenten Variablen in Kapitel 10, Band I, können also auch als unabhängige, normalverteilte Variable angesehen werden, aus denen sich die beobachteten Variablen durch Linearkombination ergeben.

Im Kapitel 10, Band I wurde die Hauptachsentransformation einer Matrix X von Meßwerten diskutiert. Variablen, insbesondere latente Variable, deren Korrelation den Betrag Null hat, wurden dabei als unabhängig interpretiert. "Unabhängig" hieß dabei zunächst einmal "lineare Unabhängigkeit" der Vektoren, die die latenten Variablen repräsentieren. Wird darüber hinaus noch Orthogonalität dieser Vektoren gefordert, so können die Vektoren unabhängig voneinander interpretiert werden. Wahrscheinlichkeitstheoretische Betrachtungen gingen dabei nicht in die Diskussion ein. Nimmt man nun an, dass die Messungen durch zufällig verteilte Fehler verrauscht sind und damit als zufällige Veränderliche betrachtet werden müssen, so führt die Forderung, dass $\rho = 0$ die Unabhängigkeit der entsprechenden Variablen bedeuten soll, auf die Annahme bzw. Forderung, dass die Messungen mehrdimensional (n-dimensional) normalverteilt sein sollen.

5.11.2 Der Fall n = 2

Es sei insbesndere $n = 2$, d.h. $X = (x_1, x_2)'$. Es sei weiter $\mu = (\mu_1, \mu_2)'$ und

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2, & \sigma_{12} \\ \sigma_{21}, & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2, & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2, & \sigma_{22} \end{pmatrix}$$

und ρ ist die (Produkt-Moment-) Korrelation zwischen den Komponenten x_1 und x_2. Dann ist

$$\mathbf{\Sigma}^{-1} = \begin{pmatrix} 1/\sigma_1^2(1-\rho^2), & -\rho/\sigma_1\sigma_2(1-\rho)^2 \\ \\ -\rho/\sigma_1\sigma_2(1-\rho)^2, & 1/\sigma_2^2(1-\rho^2) \end{pmatrix}$$

Dann ist

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \right.\right.$$
$$\left.\left. + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}\right)\right] \tag{5.131}$$

Die Randverteilungen $f_1(x_1)$ und $f_2(x_2)$ ergeben sich durch Integration von f

Abbildung 5.8: 2-d-Normalverteilung

über x_1 bzw. über x_2:

$$f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)\, dx_2 = \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2\right) \tag{5.132}$$

$$f_2(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2)\, dx_1 = \frac{1}{\sigma_2\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right) \tag{5.133}$$

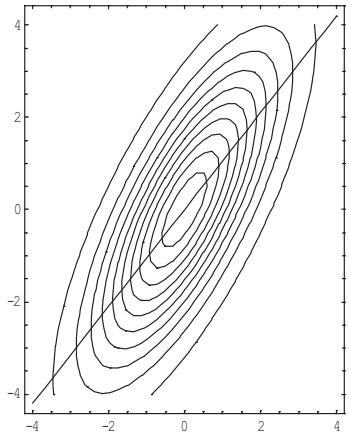
Es sei $z_x = (x_1 - \mu_1)/\sigma_1$, $z_2 = (x_2 - \mu_2)/\sigma_2$. Aus (5.131) folgt, dass

$$f(x_1, x_2) = c = \text{konstant} \;\Leftrightarrow \tilde{c} = \frac{z_1^2 + z_2^2 - 2\rho z_1 z_2}{2(1-\rho^2)} = \text{konstant} \tag{5.134}$$

Der Ausdruck (5.134) beschreibt eine Ellipse, deren Mittelpunkt die Koordinaten (μ_1, μ_2) hat. Die Orientierung der Hauptachse wird durch den Wert von ρ bestimmt. Die rechte Seite dieser Gleichung ist gerade die quadratische Form $Q = (X - \mu)'\mathbf{\Sigma}^{-1}(X - \mu)$. Für $Q = \tilde{c}$ beschreibt die quadratische Form also in der Tat ein Ellipsoid, und wegen $n = 2$ eben eine Ellipse. Abb. 5.8 zeigt die 2-dimensionale Normalverteilung. Es läßt sich zeigen, dass die bedingte Erwartung $E(y|x)$ durch

$$E(y|x) = \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x) + \mu_y \tag{5.135}$$

Abbildung 5.9: Normalverteilung: $f(x, y) = \text{const}$ und $E(y|x)$; $\sigma_x = 1.5$, $\sigma_y = 2.1$, $\rho = .75$.



gegeben ist; diese Gleichung entspricht der Regressionsgleichung

$$\hat{y} = r\frac{s_y}{s_x}(x - \overline{x}) + \overline{y},$$

die man bei Anwendung der Methode der Kleinsten Quadrate erhält. Die Abbildung 5.9 zeigt für die Dichte $f(x, y)$ mit den Parametern $\mu_x = 3, \mu_y = 5, \sigma_x = 2, \sigma_y = 3$ die Konturen gleicher Dichte, also die bereits erwähnten Ellipsen, sowie die Gerade (5.135). Speziell gilt

$$f(x_1, x_2) = f_1(x_1)f_2(x_2) \quad \text{genau dann, wenn} \quad \rho = 0 \tag{5.136}$$

Die Unabhängigkeit der Komponenten x_1 und x_2 impliziert nicht nur $\rho = 0$, sondern $\rho = 0$ impliziert auch die Unabhängigkeit der Variablen x_1 und x_2. Es läßt sich zeigen, dass die Normalverteilung die einzige Verteilung ist, für die die Beziehung (5.136) gilt.

Es sei noch auf eine Beziehung zwischen der Cauchy-Verteilung und der 2-dimensionalen Normalverteilung hingewiesen:

Beispiel 5.11.1 Es seien X und Y zwei normalverteilte zufällige Veränderliche mit der in (5.131) gegebenen gemeinsamen Dichte. Gesucht ist die Verteilung des Quotienten $Z = X/Y$. Die allgemeine Formel für die Dichte eines Quotienten wurde in (3.59) gegeben. Die Dichte für $f(x, y)$ ist symmetrisch bezüglich der Erwartungswerte μ_1 und μ_2, d.h. es gilt $f(x, y) = f(-x, -y)$. Dadurch vereinfacht sich (3.59) zu

$$g(z) = 2\int_0^\infty y f(zy, y)\, dy$$

Substituiert man hier (5.131) und führt die Integration durch, so erhält man

$$g(z) = \frac{1}{\pi} \frac{(\sigma_1/\sigma_2)\sqrt{1-r^2}}{(\sigma_1/\sigma_2)^2(1-r^2) + (z - r\sigma_1\sigma_2)}, \quad -\infty < z < \infty \tag{5.137}$$

(vergl. Papoulis (1968), p. 198). Setzt man hier

$$\lambda = \frac{\sigma_1}{\sigma_2}\sqrt{1-r^2}, \quad \mu = r\sigma_1\sigma_2$$

so zeigt der Vergleich mit (5.97), dass z Cauchy-verteilt ist. Dies bedeutet, dass die zufällige Veränderliche $Z = X/Y$, wenn X und Y normalverteilt sind, *keinen* Erwartungswert hat und *keine* Varianz besitzt. □

5.11.3 Lineare Transformationen und Randverteilungen

Es sollen einige Aussagen über die Verteilung von linearen Transformationen eines zufälligen Vektors X, $X \sim N(\mu, \Sigma)$, angegeben werden, sowie über die Randverteilungen. Die Beweise können in der angegeben Literatur gefunden werden.

Satz 5.11.2 *Es sei X ein n-dimensionaler Vektor, $X \sim N(\mu, \Sigma)$. Weiter sei $Y = CX + b$, wobei C eine $(m \times n)$-Matrix und b ein m-dimensionaler Vektor ist. Dann gilt $Y \sim N(C\mu + b, C\Sigma C')$.*

Beweis: Tong (1990), p. 32. □

Dieser Satz ist offenbar eine Verallgemeinerung des Resultats, dass $aX + b$ normalverteilt ist mit dem Erwartungswert $a\mu + b$ und der Varianz $a^2\sigma^2$, wenn X $N(\mu, \sigma^2)$-verteilt ist.

Satz 5.11.3 *Es sei X ein n-dimensionaler Vektor. X ist n-dimensional normalverteilt genau dann, wenn $\vec{a}\,'X$ für einen beliebigen rellen, n-dimensionalen Vektor $\vec{a}$ normalverteilt ist.*

Beweis: Pruscha (1989), p. 20. □

$\vec{a}\,'X$ ist ein Skalar, d.h. $Y = \vec{a}\,'X = a_1x_1 + a_2x_2 + \cdots + a_nx_n$ ist eine 1-dimensionale zufällige Veränderliche. Ist Y also normalverteilt, so kann man schließen, dass $x_1, \cdots, x_n$ gemeinsam normalverteilt sind. Sind umgekehrt die $x_1, \cdots, x_n$ gemeinsam normalverteilt, so ist auch die "gewichtete" Summe Y normalverteilt. Man beachte, dass nicht gefordert wird, dass die Komponenten x_i voneinander unabhängig sind.

Ist $\vec{x}$ ein $N(\mu, \Sigma)$-verteilter Vektor, so ergibt sich die Randverteilung für die x_i-te Komponente gemäß (3.15) durch Integration über alle $x_j, j \neq i$. Dieser Begriff

der Randverteilung läßt sich allgemeiner auffassen: man integriert über *einige* Komponenten, über einige nicht. Hat man über $n-p$ Komponenten integriert, so erhält man einen zufälligen Vektor $\vec{x}_p$. Für die Verteilung von $\vec{x}_p$ gilt

Satz 5.11.4 *Es sei $\vec{x}$ ein $N(\mu, \Sigma)$-verteilter n-dimensionaler Vektor und es werde daraus der Vektor $\vec{x}_p$ gebildet, indem über $n-p$ Komponenten integriert wird. Dann gilt $\vec{x}_p \sim N(\mu_p, \Sigma_p)$, wobei die Komponenten des Vektors μ_p die zu $\vec{x}_p$ korrespondierenden Komponenten von μ sind, und Σ_p ist die entsprechende Teilmatrix von Σ.*

Beweis: Anderson (1958), p. 24. □

Es sei $\vec{x}$ ein n-dimensionaler $N(\mu, \Sigma)$-Vektor. Es werde nun eine Stichprobe von m identisch normalverteilten verteilten Variablen Y erhoben; die Y_k seien Realisierungen der Komponenten x_k von $\vec{x}$. Der Vektor $\vec{\hat{y}} = (\hat{y}_1, \cdots, \hat{y}_n)'$ mit

$$\overline{y}_i = \frac{1}{m} \sum_{k=1}^{m} y_{ik}$$

ist eine Schätzung für μ, und die Matrix $S = S_{ij}$ der Stichprobenvarianzen und -kovarianzen ist eine Schätzung für Σ, wobei

$$S_{ij} = \frac{1}{m-1} \sum_{k=1}^{m} (Y_{ik} - \overline{Y}_i)(Y_{jk} - \overline{Y}_j)$$

Es gilt der

Satz 5.11.5 *Die Y_k seien identisch und unabhängig $N(\mu, \Sigma)$-verteilt. Dann sind $\overline{Y}$ und S stochastisch unabhängig voneinander.*

Beweis: Tong (1990), p. 48

Kapitel 6

Ungleichungen und Grenzwertsätze

6.1 Einige Ungleichungen

Der Begriff des Erwartungswertes gestattet es, einige Ungleichungen herzuleiten, mit denen Werte von Verteilungsfunktionen nach oben oder unten abgeschätzt werden köznnen.oder mit denen die Wahrscheinlichkeit von Abweichungen einer zufälligen Veränderlichen von ihrem Erwartungswert abgeschätzt werden kann. Diese Ungleichungen erweisen sich sowohl für praktische wie auch für theoretische Zwecke als nützlich und sollen deshalb kurz dargestellt werden.

Satz 6.1.1 *Es sei X eine zufällige Veränderliche mit der Dichtefunktion f, die die Bedingung $f(x) = 0$ für $x < 0$ erfüllt, und mit dem Erwartungswert $E(X)$. Weiter sei $\alpha \in \mathbb{R}$ eine Konstante. Dann gilt*

$$P(X \geq \alpha) \leq \frac{1}{\alpha} E(X). \tag{6.1}$$

Beweis: Sicherlich ist $E(X) = \int_0^\infty xf(x)dx > \int_\alpha^\infty xf(x)dx$, da ja beim rechten Integral nur über einen Teil des Intervalles, auf dem X definiert ist, integriert wird. Weiter ist $\int_\alpha^\infty xf(x)dx \geq \int_\alpha^\infty \alpha f(x)dx$, da ja $x \geq \alpha$. Also folgt $E(X) \geq \alpha \int_\alpha^\infty f(x)dx = \alpha P(X \geq \alpha)$. □

Die Gleichung (6.1) liefert eine Abschätzung von $P(X \geq \alpha)$ anhand des Erwartungswertes. Es ist aber $P(X \geq \alpha) = 1 - P(X < \alpha) \leq E(X)/\alpha$, mithin ist $1 - E(X)/\alpha \leq P(X < \alpha)$, so dass man auch eine Abschätzung für $P(X < \alpha)$ hat.

Satz 6.1.2 *(Ungleichung von Bienaymé)Es sei X eine nicht notwendig auf [0, ∞) beschränkte zufällige Veränderliche, und es sei $0 < \alpha \in \mathbb{R}$, $0 < \varepsilon \in \mathbb{R}$. Dann*

gilt

$$P(|X-\alpha| \geq \varepsilon^n) \leq \frac{1}{\varepsilon^n} E(|X-\alpha|^n). \tag{6.2}$$

Beweis: Es sei $Y := |X-\alpha|^n$ mit $0 < n \in \mathbb{N}$. Y ist eine zufällige Veränderliche, die die Bedingung des Satz 6.1.1 erfüllt, da ja Y nur Werte auf $[0, \infty)$ annehmen kann. Es existiert nun stets ein $\varepsilon > 0$ mit $\alpha = \varepsilon^n$, und (6.1) führt unmittelbar zu (6.2). □

Da $|X-\alpha|^n \geq \varepsilon^n$ genau dann, wenn $|X-\alpha| \geq \varepsilon$, so erhält man die zu (6.2) äquivalente Ungleichung

$$P(|X-\alpha| \geq \varepsilon) \leq \frac{1}{\varepsilon^n} E(|X-\alpha|^n) \tag{6.3}$$

für alle $n \leq 1$.

Satz 6.1.3 *(Ungleichung von Tchebyscheff) Es sei X eine zufällige Veränderliche mit dem Erwartungswert $E(X) = \mu < \infty$ und der Varianz $Var(X) = \sigma^2 < \infty$. Dann gilt*

$$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}. \tag{6.4}$$

Anmerkung: Die Tchebyscheffsche Ungleichung wird gelegentlich in etwas anderer Form geschrieben: es sei $c = k\sigma$. Dann ist $k^2 = c^2/\sigma^2$ und man hat

$$P(|X-\mu| \geq c) \leq \frac{\sigma^2}{c^2}. \tag{6.5}$$

Beweis: (6.4) folgt sofort aus (6.3), wenn man $\alpha = \mu$, $\varepsilon = k\sigma$ mit $0 < k \in \mathbb{R}$ und $n = 2$ setzt.

□

Die Ungleichung (6.4) liefert eine obere Abschätzung der Wahrscheinlichkeit, dass eine zufällige Veränderliche um mehr als k σ-Einheiten von μ abweicht. Das Interessante an den Ungleichungen ist, dass die genaue Form der Verteilungsfunktion der zufälligen Veränderlichen nicht eingeht, d.h. man braucht die explizite Form der Verteilungsfunktion gar nicht zu kennen! So gilt für alle Verteilungsfunktionen mit $\mu < \infty$, $\sigma < \infty$, dass $P(|X-\mu| \geq 3\sigma) \leq .111$, d.h. Abweichungen von μ die größer als 3 σ-Einheiten sind, treten nur in ungefähr 11 % der Fälle auf. Für spezielle Verteilungen, wie etwa die Normalverteilung, ist diese Wahrscheinlichkeit noch geringer.

Beobachtet man nun Werte der zufälligen Veränderlichen X, so tritt, für $0 < k \in \mathbb{R}$, entweder das Ereignis $\{|X-\mu \geq k\sigma\}$ oder $\{|X-\mu| < k\sigma\}$ ein, also ist

$1 - P(|X - \mu| < k\sigma) = P(|X - \mu| \geq k\sigma)$. Also gilt

$$P(|X - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}. \tag{6.6}$$

Es sei nun $\epsilon = k\sigma$. Das Ereignis $\{|X-\mu| < \epsilon\}$ tritt ein genau dann, wenn entweder $X - \mu < \epsilon$, d.h. $X < \epsilon + \mu$, oder $-\epsilon < X - \mu$, d.h. $\mu - \epsilon < X$ eingetreten ist. Daann ist $\{|X - \mu| < \epsilon\} = \{-\epsilon - \alpha < X < \alpha + \alpha\}$. Aus (6.6) folgt dann

$$P(\mu - \epsilon < X < \mu + \epsilon) \geq 1 - \frac{\sigma^2}{\epsilon^2} \tag{6.7}$$

Es sei insbesondere σ *klein* im Vergleich zu ϵ, d.h. $\sigma^2/\epsilon^2 \approx 0$. Dann ist die Wahrscheinlichkeit, dass X einen Wert im Intervall $(\mu - \epsilon, \mu + \epsilon)$ annimmt, ungefähr gleich 1. Die Bedeutung dieser Aussage liegt darin, dass sie für alle zufälligen Veränderlichen mit existierendem Erwartungswert und existierender Varianz gilt, und damit auch für für den Stichprobenmittelwert $\overline{x}$, der ja bekanntlich selbst eine zufällige Veränderliche ist. Kann man also annehmen, dass $E(\bar{x})$ und $Var(\bar{x})$ existieren, so kann man in (6.7) $\bar{x}$ für X und $Var(\bar{x})$ für σ substituieren. Nun ist aber $Var(\bar{A}) = \sigma^2/n$, $\sigma^2 = Var(X)$. Für vorgegebenes ϵ kann aber $Var(\bar{x})$ beliebig klein werden, wenn nur der Wert für n hinreichend groß gewählt wird. Dann bedeutet (6.7), dass für jedes $\epsilon > 0$

$$P(\mu - \epsilon < X < \mu + \epsilon) \approx 1$$

ist, wenn nur n hinreichend groß ist. Dieser Sachverhalt wird in den folgenden Sätzen präzisiert.

6.2 Gesetze der großen Zahlen

Die Gesetze der großen Zahlen sind für die Praxis von zentraler Bedeutung. So ist der Anteil p von Personen mit einem bestimmten Merkmal $\mathcal{M}$ (Krankheit, Meinung, etc.) im allgemeinen unbekannt. Aufgrund der Gesetze der großen Zahlen kann man aber sicher sein, dass für eine Stichprobe der Quotient k/n, k die Anzahl der Personen mit $\mathcal{M}$ und n der Stichprobenumfang, nahe an p liegt, wenn nur n hinreichend groß ist.

Satz 6.2.1 *(***Schwaches Gesetz der Großen Zahlen** *Es sei X_1, X_2, $X_3, \cdots$ eine Folge von zufälligen Veränderlichen, zwischen denen beliebige Abhängigkeiten bestehen können, mit den Erwartungswerten μ_1, μ_2, $\mu_3, \cdots$ und den Varianzen $\sigma_1^2, \sigma_2^2, \sigma_3^2, \cdots$, die die Bedingung*

$$\lim_{n\to\infty} Var\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = 0 \tag{6.8}$$

erfüllen. Weiter sei

$$\overline{x} = \frac{1}{n_i}\sum_{i=1}^{n} x_i, \quad \overline{\mu} = \frac{1}{n}\sum_{i=1}^{n} \mu_i$$

Dann gilt

$$\lim_{n\to\infty} P(|\overline{x} - \overline{\mu}| < \epsilon) = 1. \tag{6.9}$$

Beweis: Der Beweis ergibt sich aus der Tschebyscheffschen Ungleichung. Es werde $\epsilon = k\sigma$ gesetzt, wobei

$$\sigma = \sqrt{Var\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)}$$

sei. Dann ist $1/k^2 = \sigma^2/\epsilon^2$, und da $\sigma^2 \to 0$ mit wachsendem n folgt die Behauptung. □

Anmerkungen:

(i) Es ist $Var(\sum_i x_i/n) = Var(\sum_i x_i)/n^2$; die Bedingung (6.8) besagt, dass $Var(\sum_i x_i)$ langsamer wachsen soll als $1/n^2$ kleiner wird. Eine hinreichende, wenn auch nicht notwendige Bedingung hierfür ist, dass die Varianzen σ_i^2 kleiner oder gleich einer gewisse Konstanten sind.

(ii) Die praktische Bedeutung von (6.9) ist, dass dann für hinreichend großes n $P(|\bar{x} - \bar{\mu}| < \epsilon) \approx 1$ ist, bzw. dass ein Intervall $(\mu - \epsilon, \mu + \epsilon)$ angegeben werden kann, in dem X mit einer vorgegebenen Wahrscheinlichkeit, etwa .95, liegt. Dieser Sachverhalt wird später noch näher ausgeführt.

Mit wachsendem n wird also die Wahrscheinlichkeit, dass $\bar{x}$ von μ um mehr als ein vorgegebenes $\epsilon > 0$ abweicht, immer kleiner, und dabei wird nicht einmal vorausgesetzt, dass die X_i die gleiche Verteilung haben oder stochastisch unabhängig sind. Für die folgende Aussage ist aber die paarweise stochastische Unabhängigkeit Voraussetzung:

Satz 6.2.2 *(Tschebyscheff) Es sei $X_1, X_2, X_3, \cdots$ eine Folge paarweise stochastisch unabhängiger zufälliger Veränderlicher, deren Varianzen σ_i^2 der Bedingung*

$$\sigma_i \leq C, \quad \textit{für alle} \;\; i = 1, 2, 3, \cdots \tag{6.10}$$

genügen, wobei C eine Konstante ist. Weiter sei wieder

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \overline{\mu} = \frac{1}{n}\sum_{i=1}^{n} \mu_i \tag{6.11}$$

Dann gilt für alle $\epsilon > 0$

$$\lim_{n\to\infty} P(|\overline{x} - \overline{\mu}| < \epsilon) = 1 \tag{6.12}$$

Beweis: Es genügt wieder, (6.6) zu betrachten. Wegen der paarweisen stochastischen Unabhängigkeit ist für jedes n

$$\sigma^2 := Var(\bar{x}) = \frac{1}{n^2} Var\left(\sum_{i=1}^{n} x_i\right) \leq \frac{nC}{n^2} = \frac{C}{n} \tag{6.13}$$

Mit $\epsilon = k\sigma$ folgt wieder $1/k^2 = \sigma^2/\epsilon^2$, und da $C/n \to 0$ für $n \to \infty$ folgt aus (6.13) $\sigma^2 \to 0$ für $n \to 0$. □

Der Satz von Tschebyscheff kann als Spezialisierung des vorangegangenen Satzes von Markov aufgefaßt werden. Der folgende Satz ist eine weitere Spezialisierung des Satzes von Markov auf Bernoulli-Variable:

Satz 6.2.3 *(Poisson) Gegeben sei eine Folge stochastisch unabhängiger Versuche, in denen jeweils entweder das zufällige Ereignis A oder $\bar{A}$ eintritt; dabei sei p_k die Wahrscheinlichkeit von A im k-ten Versuch. Weiter sei X_n die Anzahl der "Erfolge" (Auftreten von A) in den ersten n Versuchen. Dann gilt*

$$P\left(\left|\frac{X_n}{n} - \frac{1}{n}\sum_{i=1}^{n} p_i\right| < \epsilon\right) = 1 \tag{6.14}$$

Beweis: Der Beweis erfolgt ebenfalls durch Rekurs auf (6.6). Man setzt wieder $\epsilon = k\sigma$, und es ist $\sigma^2 = (\sum_k \sigma_k^2)/n^2 = \sum_k p_k(1-p_k)/n$. Nun ist $\sigma_k = p_k(1-p_k)$ maximal für $p_k = 1/2$, d.h. $\sigma_k^2 \leq 1/2$ für alle k, so dass $1/k^2 = \sum_k p_k(1-p_k)/(n\epsilon^2) \leq 1/(2n\epsilon^2) \to 0$ für $n \to \infty$. □

Das erste Gesetz der großen Zahlen ist das sogenannte *schwache Gesetz der großen Zahlen.* Die bekannteste Form dieses Gesetzes Zahlen geht auf Bernoulli zurück:

Satz 6.2.4 *(Bernoulli) Gegeben sei eine Folge von Bernoulli-Versuchen (d.h. die Ergebnisse der Versuche sind stochastisch unabhängig voneinander, in jedem Versuch wird entweder A oder $\bar{A}$ auf und $p(A) = p = konstant$). Es sei X_n die Anzahl der "Erfolge", d.h. die Häufigkeit, mit der A beobachtet wurde. Dann gilt für jedes $\epsilon > 0$*

$$\lim_{n\to\infty} P\left(\left|\frac{X_n}{n} - p\right| < \epsilon\right) = 1 \tag{6.15}$$

Beweis: Der Beweis wird wieder durch Anwendung von (6.6) geführt. Es ist $\sigma^2 = Var(X_n/n) = Var(X_n)/n^2 = np(1-p)n^2 = p(1-p)/n$, und mit $\epsilon = k\sigma$ folgt wieder $1/k^2 = p(1-p)/(n\epsilon^2) \to 0$ für $n \to \infty$. □

Der Tschebyscheffsche, der Poissonsche und der Bernoullische Satz folgen durch Spezialisierung des Satzes von Markov. Nach Satz 6.2.3 müssen die $p(A)$ nicht

gleich groß sein, nach Satz 6.2.2 könne die p_n durch beliebige arithmetische Mittel ersetzt werden, und nach Satz 6.2.1 müssen die zufälligen Veränderlichen noch nicht einmal stochastisch unabhängig sein. Im Falle stochastischer Abhängigkeit muß aber sichergestellt sein, dass (6.8) gilt. Satz 6.2.4 liefert eine Rechtfertigung für den Schluß von den relativen Häufigkeiten eines Ereignisses auf dessen Wahrscheinlichkeit, denn (6.15) heißt ja, dass für hinreichend großes n $X_n/n \approx p$. Diese Aussage muß aber genauer betrachtet werden. Tatsächlich bedeutet (6.15) nur, dass für eine hinreichend große Anzahl n von Versuchen unter identischen Bedingungen, wobei n fest gewählt ist, die Wahrscheinlichkeit der Gültigkeit der einzelnen Ungleichung $|X_n/n - p| < \epsilon$ größer als $1 - \eta$ wird für jedes $\epsilon > 0$ (Gnedenko (1978) , p. 209); eine analoge Aussage gilt für die vorangegangenen allgemeineren Sätze. Wie Feller (1968), p. 203 ausführt kann diese Ungleichung aber nicht für jede Folge von n Versuchen gelten, da ja der Stichprobenraum auch eine Folge mit genau n "Erfolgen" enthält. Man hat noch zu zeigen, dass die Wahrscheinlichkeit für alle diejenigen Folgen, für die die Ungleichung *nicht* gilt, vernachlässigbar ist. Daß dies nun tatsächlich der Fall ist, ist die Aussage des *Starken Gesetzes der großen Zahlen*:

Satz 6.2.5 *Gegeben sei eine Folge $X_1, X_2, \ldots$ von zufälligen Veränderlichen mit identischem Erwartungswert μ. Dann gilt*

$$\lim_{n\to\infty} P\left(\bar{X}_n \to \mu\right) = 1. \tag{6.16}$$

Beweis: Der Beweis setzt den Begriff der Konvergenz einer Folge von zufälligen Veränderlichen ξ_n gegen eine zufällige Veränderliche ξ voraus, auf den hier nicht eingegangen werden kann, vergl. etwa Gnedenko (1978), wo auch der Beweis des Satzes geliefert wird. □

Anmerkung: Die Aussage (6.16) ist, dass die Konvergenz von $\bar{X}_n$ gegen μ *fast sicher* gilt; man erinnere sich, dass das sichere Ereignis die Wahrscheinlichkeit 1 hat, aber ein Ereignis mit Wahrscheinlichkeit 1 nicht notwendig sicher ist, – sondern eben i. A. nur fast sicher ist. □

Die Gleichungen (6.15) und (6.16) liefern, wie die übrigen, keine sehr scharfen Abschätzungen, sie sind vielmehr als eine grundsätzliche Aussage zu verstehen. Für numerische Abschätzungen ist der im folgenden Abschnitt behandelte Satz von DeMoivre-Laplace besser geeignet.

6.3 Grenzwertsätze

6.3.1 Der Poissonsche Grenzwertsatz

Viele statistische Fragestellungen führen auf die Binomialverteilung

$$P(X = k|n, p) = \binom{n}{k} p^k (1-p)^{n-k}, \tag{6.17}$$

die die Wahrscheinlichkeit der Anzahl von k "Erfolgen" bei n Bernoulli-Versuchen angibt. Dabei ist $p = p(A)$ die Wahrscheinlichkeit eines Erfolges, d.h. des Ereignisses A, und p ist für alle Versuche konstant ist; weiter sind die Versuche stochastisch unabhängig. Nun erweist sich die Berechnung der Binomialkoeffizienten $\binom{n}{k}$ für größere Werte von n, etwa für $n \geq 30$, als sehr lästig, so dass man nach Approximationen für $P(X = k|n, p)$ sucht. Die Poisson- und die Normalverteilung kommen dann als Näherungsverteilungen in Frage.

Satz 6.3.1 *(Poisson) Die zufällige Veränderliche X sei binomialverteilt mit den Parametern n und p. Für $n \to \infty$, $p \to 0$, $np \to \lambda$, $0 < \lambda \in \mathbb{R}$ folgt dann*

$$P(X = k|n, p) \to e^{-\lambda} \frac{\lambda^k}{k!}. \tag{6.18}$$

Beweis: vergl. Feller (1968), p. 153 □

Gelten die in Satz 6.3.1 angegebenen Bedingungen, so bedeutet (6.18), dass für hinreichend großes n und hinreichend kleines p der Wert von $P(X = k|n, p)$ in guter Näherung durch $e^{-\lambda}\lambda^k/k!$ gegeben ist.

Beispiel 6.3.1 Für eine Untersuchung im Rahmen der klinischen Psychologie werden psychiatrische Patienten mit einem bestimmten, aber seltenen Merkmal M gesucht; aus langjährig erhobenen Statistiken weiß man, dass der Anteil dieser Patienten an der Gesamtpopulation psychiatrischer Patienten $p = .05$ beträgt. Das Merkmal tritt unabhängig von anderen Eigenschaften auf. Im örtlichen Landeskrankenhaus befinden sich gerade 50 Patienten. Wie groß ist die Wahrscheinlichkeit, $X = 2$, $X = 5$, $X = 10$ Patienten mit dem Merkmal M zu finden?

Man kann davon ausgehen, dass X binomialverteilt ist mit den Parametern $p = .05$, $n = 50$. Folglich ist

$$\begin{aligned}
P(X = 2) &= \binom{n}{2} .05^2 .95^{48} = .261 \\
P(X = 5) &= \binom{n}{5} .05^5 .95^{45} = .066 \\
P(X = 10) &= \binom{n}{10} .05^{10} .95^{40} = .000129.
\end{aligned}$$

Die entsprechenden Poisson-Approximationen sind, mit $\lambda = np = 2.5$,

$$\begin{aligned} P(X=2) &\approx e^{-2.5}\frac{2.5^2}{2!} = .2565 \approx .26 \\ P(X=5) &\approx e^{-2.5}\frac{2.5^5}{5!} = .0668 \approx .067 \\ P(X=10) &\approx e^{-2.5}\frac{2.5^{10}}{10!} = .0002157 \approx .0002. \end{aligned}$$

Für $X = 2$ und $X = 5$ ist die Approximation recht gut, für $X = 10$ nicht so sehr; das Verhältnis zwischen n und p ist noch nicht optimal. □

Als Bedingung dafür, dass $P(k|\lambda) = e^{-\lambda}\lambda^k/k!$ eine approximierende Verteilung für die Binomialverteilung ist, wurde vorausgesetzt, dass p klein ist. Doe Poisson-Verteilung ist deshalb auch als Verteilung für seltene Ereignisse oder als Gesetz für kleine Zahlen (v. Bortkiewicz, 1898) bezeichnet worden. Diese Bezeichnung ist etwas irreführend. Denn es muß nicht nur p klein, sondern auch n groß sein, und der Wert von $\lambda = pn = E(X) = Var(X)$ (vergl. (4.9.27) und (4.9.28)) muß nicht klein sein. Ob man die Anzahl der Verkehrsunfälle in einer Woche eine kleine Zahl nennt ist vielleicht Geschmacksache, aber sie ist in guter Näherung Poisson-verteilt (Abweichungen von der Poisson-Verteilung kommen zum Teil durch die verschiedenen Parameter der einzelnen, am Verkehrsgeschehen beteiligten Personen zustande). Da bei Poisson-verteiltem X die Varianz $Var(X)$ mit $E(X) = \lambda$ wächst kann in speziellen Fällen der Wert von X durchaus von dem, was man "klein" nennt, abweichen.

6.3.2 Der DeMoivre-Laplacesche Grenzwertsatz

Die im Folgenden behandelte Approximation für (6.17) gilt zunächst für andere Randbedingungen als die, die für die Poissonsche angenommen wurden. Es gilt der

Satz 6.3.2 *(DeMoivre-Laplace) Es sei X binomialverteilt mit den Parametern n und p und es gelte*

$$z_1 = \frac{X_1 - np}{\sqrt{npq}}, \quad z_2 = \frac{X_2 - np}{\sqrt{npq}},$$

mit $q = 1 - p$. Dann gilt für $n \to \infty$

$$P(np + z_1\sqrt{npq} \le X \le np + z_2\sqrt{npq}) \to F(z_2) - F(z_1) \tag{6.19}$$

mit

$$F(z_i) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z_i} \exp\left(-\frac{z^2}{2}\right) dz, \quad i = 1, 2. \tag{6.20}$$

Beweis: Feller (1968), p. 186. □

Die Bedeutung dieses Satzes liegt darin, dass für hinreichend großes n die Differenz $F(z_2) - F(z_1)$ eine gute Näherung für die Wahrscheinlichkeit auf der linken Seite von (6.19) darstellt. Um zu sehen, unter welchen Bedingungen diese Näherung gut ist, sollen zunächst diejenigen Werte von X bestimmt werden, für die $P(X = k)$ in (6.17) einen maximalen Wert annimmt.

Beispiel 6.3.2 Es soll noch einmal die Situation in Beispiel 6.3.1 betrachtet werden: das Merkmal M komme bei 40 % aller Patienten vor. Es seien nun 30 Patienten vorhanden. Wir können davon ausgehen, dass die Anzahl der Patienten in einer Stichprobe binomialverteilt ist, d.h.

$$P(X = k) = \binom{n}{k}.4^k.6^{n-k}, \quad n = 30.$$

Wie groß ist nun die Wahrscheinlichkeit, dass mindestens 5 der 30 Patienten das Merkmal M haben?

Gesucht ist $P(X \geq 5) = 1 - P(X \leq 4)$. Exakt ist $P(X \leq 4)$ durch

$$P(X \leq 4) = \sum_{k=0}^{4} \binom{30}{k}.4^k.6^{30-k}$$

gegeben. Der Wert kann auch über die Normalverteilung abgeschätzt werden, d.h. durch $P(Z \leq z(4)) \approx P(X \leq 4)$. Es ist $z(4) = (4 - 30 \times .4)/\sqrt{30 \times .4 \times .6} = -2.981$. In der Tabelle findet man $P(Z \leq -2.981) = .0014$.

Die Approximation kann natürlich auch für die Berechnung der Wahrscheinlichkeit einzelner X-Werte verwendet werden. Wünschen man $P(X = k)$ zu berechnen, so muß man berücksichtigen, dass $P(X = k) = P(X \leq k) - P(X \leq k - 1)$, und damit hat $P(X = k) \approx F(z_k) - F(z_{k-1})$. Es sei z.B. wieder $n = 30$, und gewünscht sei $P(X = 10)$. Der exakte Wert gemäß der Binomialverteilung ist $= .2304$. Zur Approximation durch die Gauß-Verteilung müssen nun die z-Werte $z_{10} = (10 - 12)/2.683 = -.745$, $z_9 = (9 - 12)/2.683 = -1.12$; aus der Tabelle findet man $F(-.745) - F(-1.12) = .228 - .131 = .097$. In diesem Fall ist der Wert für n noch zu klein, um eine gute Approximation zu liefern. □

Satz 6.3.3 *Es bezeichne* $[\alpha]$, $\alpha \in \mathbb{R}$, *diejenige größte natürliche Zahl* $j \in \mathbb{N}$, *die kleiner als* α *ist. Dann gilt*

$$k_{max} = [(n + 1)p]. \tag{6.21}$$

Beweis: Zum Beweis werde der Quotient $Q = P(X = k-1)/P(k = k)$ betrachtet. Es ist

$$Q = \frac{\binom{n}{n-k} p^{k-1}(1-p)^{n-k+1}}{\binom{n}{k} p^k (1-p)^{n-k}} = \frac{k(1-p)}{(n-k+1)p}.$$

nun ist $P(X = k-1)/P(X = k) \leq 1$, d.h. aber $P(X = k-1) \leq P(X = k)$, genau dann, wenn $k(1-p) \leq (n-k+1)p$, und letztere Ungleichung gilt genau dann, wenn $k - kp \leq np - kp + p$, d.h. wenn $k \leq (n+1)p$. Für $k > (n+1)p$ wird also $P(X = k)$ wieder kleiner als $P(X = k-1)$, und damit ist die Behauptung bewiesen. □

Aus Fellers Beweis von Satz 6.3.3 folgt, dass die Annäherung besonders gut ist für Werte von k in der Nachbarschaft von $(n+1)p$ bzw. von np, dem Erwartungswert von X. Weiter ist die Annäherung der Wahrscheinlichkeit in der Nachbarschaft von $1/2$ schon für kleinere n gut im Vergleich zu Annäherungen für die Wahrscheinlichkeiten, wenn p näher bei 1 oder bei 0 liegt.

Aus (6.21) folgt

$$P\left(z_1 \leq \frac{X - np}{\sqrt{np(1-p)}} \leq z_2\right) = F(z_2) - F(z_1). \tag{6.22}$$

Es sei nun $z = (X - np)/\sqrt{np(1-p)}$; dann ist $X = \sqrt{np(1-p)} + np$, d.h. X ist eine lineare Funktion von z. Gemäß (6.21) hat z die Gaußsche Dichte $f(z) = \exp(-z^2/2)/\sqrt{2\pi}$, und da nach (6.20) die Dichte von X durch die von z angenähert werden kann, hat man nach (6.21) dann

$$f_x(X) = \frac{1}{\sqrt{2\pi np(1-p)}} \exp\left(-\frac{(X-np)^2}{np(1-p)}\right). \tag{6.23}$$

Nimmt nun X einen Wert in der Nachbarschaft von np oder $(n+1)p$ an, so ist $x - np \approx 0$ und es ist, für hinreichend großes n,

$$f_x(X) \approx \frac{1}{\sqrt{2\pi np(1-p)}}. \tag{6.24}$$

Für binomialverteiltes X ist bekanntlich $E(X) = np$, $Var(X) = np(1-p)$. Für gegebenes p wachsen also $E(X)$ und $Var(X)$ linear mit n; für $n \to \infty$ werden diese beiden Parameter selbst unendlich. (6.22) zeigt, dass X für jedes n linear skaliert wird mit dem Skalenfaktor $a_n = 1/\sqrt{\sigma_n}$ und der additiven Konstanten $b_n = -np/\sqrt{\sigma_n}$. Für jedes n geht man also von X über zum standardisierten Wert $z_n = a_n X + b_n$, für den jeweils $E(z_n) = 0$ und $Var(z_n) = 1$ gelten. Wir können nun die Differenz der Werte von z_n für $X = k$ und $X = k-1$ in Abhängigkeit von n betrachten:

$$\frac{k-np}{\sqrt{npq}} - \frac{k-1-np}{\sqrt{npq}} = \frac{k-np}{\sqrt{npq}} - \frac{k-np}{\sqrt{npq}} - \frac{1}{\sqrt{npq}} \to 0, \qquad \text{für } n \to \infty$$

d.h. die zu benachbarten Werten gehörenden standardisierten Werte rücken mit wachsendem n immer dichter aneinander; auf diese Weise kann die diskrete Variable X durch eine stetige approximiert werden.

Die Wahrscheinlichkeit, dass X innerhalb der Grenzen $np \pm 2\sqrt{np(1-p)}$ liegt, ist angenähert $F(2) - F(-2) = .9545$; für die Grenzen $np \pm 3\sqrt{np(1-p)}$ ergibt sich näherungsweise eine Wahrscheinlichkeit von $F(3) - F(-3) = .9973$, d.h. die Wahrscheinlichkeit, dass X außerhalb dieses Intervalles liegt, ist $\approx .0027$. Diese Ergebnisse mag man mit den Abschätzungen z.B. aufgrund der Ungleichung von Tschebyscheff (vergl. die Anmerkungen zum Satz 4.9.3) vergleichen; es war ja

$$P\left(|X - np| \geq 3\sqrt{np(1-p)}\right) \leq .11,$$

d.h. die Tschebyscheffsche Ungleichung liefert nur die obere Grenze .111 für diese Wahrscheinlichkeit. Die Approximation durch die Gauß-Verteilung ist genauer.

Die Annäherung einer Binomial- durch eine Gaußverteilung ist am besten für Werte von p nahe bei $1/2$; für Werte nahe bei 0 oder 1 muß der Wert von n erhöht werden, um eine gleich gute Annäherung zu erzielen. Allgemein kann man sagen, dass die Annäherung gut ist, wenn $np(1-p)$ groß ist. Ist nun p klein, so liegt es nahe zu vermuten, dass eine Poisson-Approximation besser als die Gauß-Approximation ist. Für die Poisson-Approximation ist aber der Wert von λ charakteristisch, und man kann umgekehrt vermuten, dass für großes λ die Gauß-Verteilung ebensogut zur Annäherung benutzt werden kann, d.h. dass die Poisson-Verteilung wiederum durch die Gauß-Verteilung angenähert werden kann. Diese Vermutung wird durch den folgenden Satz bestätigt.

6.3.3 Der Zentrale Grenzwertsatz

Satz 6.3.4 *(Zentraler Grenzwertsatz) Es seien $X_1, X_2, \cdots, X_n$ stochastisch unabhängige zufällige Veränderliche mit Dichten $f_i(X_i)$, Erwartungswerten $E(X_i) < \infty$ und Varianzen $\sigma_i^2 < \infty$ für alle i, und es sei $X = X_1 + \cdots + X_n$, $\mu = E(X)$, $\sigma^2 = Var(X)$. Für $n \to \infty$ gilt dann für die Dichte der Summe X*

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(X-\mu)^2}{2\sigma^2}\right) \tag{6.25}$$

falls die Bedingungen

(i) $\sigma_1^2 + \cdots + \sigma_n^2 \to \infty$, *und*

(ii) $\int_{-\infty}^{\infty} x^\alpha f_i(x)dx < C$,

C eine Konstante, für alle i erfüllt sind.

Beweis: Der Beweis ist nicht kurz und kann in Feller (1968) nachgelesen werden. □

Die hier angegebenen Bedingungen für eine Approximation durch die Normalverteilung sind nicht die allgemeinsten, vergl. Billingsley (1979). Die Voraussetzung der stochastischen Unabhängigkeit der X_i muß ebenfalls nicht in aller Strenge gefordert werden (Hoeffding, 19..)

Der Satz von DeMoivre-Laplace kann als Spezialfall des zentralen Grenzwertsatzes aufgefaßt werden. Sind X_1 und X_2 stochastisch unabhängig und Poisson-verteilt mit den Parametern λ_1 und λ_2, so kann gezeigt werden, dass die Summe $X_1 + X_2$ ebenfalls Poisson-verteilt ist mit dem Parameter $\lambda_1 + \lambda_2$. Daraus folgt, dass die Summe von n stochastisch unabhängigen Poisson-verteilten Variablen wiederum Poisson-verteilt ist mit dem Parameter $\lambda_1 + \cdots + \lambda_n$. Gleichzeitig erfüllen aber stochastisch unabhängige Poisson-verteilte Variablen die Vorausetzungen des zentralen Grenzwertsatzes, d.h. für $n \to \infty$ wird die Summe stochastisch unabhängiger Poisson-verteilter Variablen nicht nur Poisson-verteilt sein, sondern gleichzeitig approximativ normalverteilt, und zwar wird die Annäherung durch die Normalverteilung um so besser sein, je größer n ist.

Gemäß Beispiel 3.4.5 ist die Summe zweier exponentialverteilter Variablen gamma-verteilt. Allgemein ist die Summe X von n exponentialverteilten Variablen mit gleichem Parameter λ eine gamma-Verteilung mit der Dichte $f(x) = \lambda^n \exp(-\lambda x)/n!$. Da für exponentialverteilte Variable die Bedingungen des Zentralen Grenzwertsatzes gelten, ist X mit steigendem n aproximativ normalverteilt. Die gamma-Verteilung strebt also für $n \to \infty$ gegen eine Normalverteilung.

Es seien nun X_1 und X_2 zwei normalverteilte zufällige Veränderliche. Der Quotient $Y = X_1/X_2$ ist aber nicht normalverteilt, sondern Cauchy-verteilt. Die Cauchy-Verteilung hat weder einen Erwartungwert noch eine Varianz, und damit erfüllen Cauchy-verteilte Variablen nicht die Bedingungen des Zentralen Grenzwertsatzes. In der Tat kann gezeigt werden (vergl. Kendall und Stuart (1969)), dass die Summe Cauchy-verteilter Variablen ebenfalls Cauchy-verteilt und eben nicht normalverteilt ist. Dieses Beispiel macht deutlich, dass man nicht ohne weiteres annehmen kann, dass die Summe zufälliger Veränderlicher angenähert normalverteilt ist, wenn nur die Anzahl der Summanden hinreichend groß ist.

Die Binomial-, die Poisson- und die Normalverteilung können in vielerlei Hinsicht als die Basisverteilungen der Wahrscheinlichkeitstheorie betrachtet werden (Feller, 1968).

Kapitel 7

Die Verteilung extremer Werte

7.1 Einführung

Die Frage nach der Verteilung extremer Werte ergibt sich in verschiedenen Fragestellungen:

1. In einer Untersuchung wird eine Stichprobe von n Meßwerten $x_1, \cdots, x_n$ erhoben. Dabei können - relativ zur Größe der meisten der gemessenen x-Werte, sehr große oder sehr kleine Werte auftreten. Die Frage ist dann, ob diese Werte "Ausreißer" sind, also möglicherweise aufgrund fehlerhafter Messung gewonnen wurden, oder ob sie bei dem gegebenen Stichprobenumfang mit eine gewissen Wahrscheinlichkeit erwartet werden können.

2. Ein System besteht aus n Komponenten, und es arbeitet fehlerhaft oder gar nicht, wenn mindestens eine dieser Komponenten fehlerhaft arbeitet. Es gibt verschiedene Möglichkeiten, die fehlerhafte Funktionsweise der Komponenten durch zufällige Veränderliche und deren Extrema auszudrücken, so dass die hier beschriebene Situation der Struktur nach mit der unter 1. genannten identisch ist; so kann man etwa die Zeitdauern t_i bis zum Versagen der i-ten Komponente betrachten. Die Dauer bis zum Versagen des Systems ist dann durch die kleinste der Zeiten t_i, d.h. durch deren Minimum, gegeben. Beispiele mit psychologischen Entsprechungen dieses Paradigmas werden weiter unten gegeben. Allgemein kann man Fragestellungen, die sich auf Situationen dieser Art beziehen, mit dem Begriff *Zuverlässigkeit* oder *Reliabilität* verbinden.

Bevor erläuternde Beispiele gegeben werden, sollen einige Grundbegriffe eingeführt werden, die die Formulierung der Beispiele vereinfachen.

Mit X_+ werde das Maximum, mit X_- werde das Minimum der $x_1, \cdots, x_n$ bezeichnet, d.h. es sei

$$X_{n+} = \max(x_1, \cdots, x_n), \quad X_{n-} = \min(x_1, \cdots, x_n) \tag{7.1}$$

Es werde angenommen, dass die x_i, $1 \leq i \leq n$ stochastisch unabhängig voneinander sind. Dann können allgemeine Ausdrücke für die Verteilung von X_{n+} und X_{n-} angegeben werden.

Es sei

$$G_n(x) = P(X_{n+} \leq x) \tag{7.2}$$

und

$$H_n(x) = P(X_{n-} \leq x) \tag{7.3}$$

Das Maximum der x_i ist sicher kleiner als x, wenn *alle* x_i kleiner als x sind. Es sei F_i die Verteilung von x_i. Dann ist also wegen der angenommenen stochastischen Unabhängigkeit der x_i

$$G_n(x) = P\left(\bigcap_{i=1}^{n}\{x_i \leq x\}\right) = \prod_{i=1}^{n} F_i(x). \tag{7.4}$$

Sind die F_i überdies identisch, d.h. sind die x_i *i.i.d.*-verteilt (independent and identically distributed), so ergibt sich daraus sofort

$$G_n(x) = F^n(x). \tag{7.5}$$

Um die Verteilung des Minimums X_{n-} herzuleiten, kann man zunächst die Wahrscheinlichkeit betrachten, dass X_{n-} *größer* als x ist. Dazu müssen *alle* x_i größer als x sein, genau dann ist ja auch das Mimimum der x_i größer als x. Es ist aber

$$P(X_{n-} > x) = P\left(\bigcap_{i=1}^{n}\{x_i > x\}\right) = \prod_{i=1}^{n}(1 - F_i(x))$$

Daraus folgt sofort für die Verteilung von X_{n-}, d.h. für $H_n(x)$

$$H_n(x) = 1 - \prod_{i=1}^{n}(1 - F_i(x)) \tag{7.6}$$

und im i.i.d.-Fall

$$H_n(x) = 1 - (1 - F(x))^n \tag{7.7}$$

Zur Illustration sollen einige Beispiele betrachtet werden.

Beispiel 7.1.1 Ein gegebener Intelligenztest sei so geeicht, dass für eine betrachtete Population der durchschnittliche IQ $E(X) = \mu$ den Wert $\mu = 100$ und die Varianz der IQen den Wert $\sigma^2 = 100$ hat. Weiter seien die IQ-Werte normalverteilt. Man kann nun den Bereich von IQ-Werten charakterisieren, innerhalb

dessen 95 % der IQ-Werte liegen. Es ist $1.96\sigma + \mu = 1.96 \times 10 + 100 = 119.6$, und $-1.96\sigma + \mu = 80.4$, so dass

$$P(80.4 \leq X \leq 119.6) = .95$$

Wählt man nun *eine* Person zufällig aus, so wird ihr IQ mit der Wahrscheinlichkeit .95 innerhalb der Grenzen (80.4, 119.6) liegen, d.h. die Wahrscheinlichkeit, eine Person mit einem IQ größer als 119.6 zu wählen, beträgt nur .025; ebenso ist die Wahrscheinlichkeit, dass sie einen IQ kleiner als 80.4 ist, nur .025.

Es werde nun eine zufällige Stichprobe von 50 Personen gebildet. Ist x_i der IQ der i-ten Person, so hat man eine Stichprobe

$$x_1, x_2, \cdots, x_n, \quad n = 50$$

von Meßwerten. Man kann jetzt fragen, wie groß die Wahrscheinlichkeit ist, dass das Maximum dieser Werte größer als 119.6 ist, bzw. wie groß die Wahrscheinlichkeit ist, dass das Minimum kleiner als 80.4 ist. Nach (7.5) findet man sofort

$$F^{50}(119.4) = .975^{50} = .282$$

d.h. mit der Wahrscheinlichkeit $1 - F^n(x) = 1 - .282 = .718$ ist das Maximum X_{n+} der x_i *größer* als 119.6; diese Wahrscheinlichkeit ist deutlich größer als .025, der Wahrscheinlichkeit, bei einmaligem "Ziehen" aus der Population einen Wert größer als 119.4 zu erhalten.

Für die Wahrscheinlichkeit, dass das Minimum X_{n-} kleiner als 80.4 ist, ergibt sich nach (7.7)

$$1 - (1 - F(80.4))^{50} = 1 - (1 - .025)^{50} = .718$$

Die Wahrscheinlichkeit, dass das Minimum der x_i kleiner als 80.4 ist, ist also wiederum erheblich größer als .025; der Wert dieser Wahrscheinlichkeit entspricht dem der Wahrscheinlichkeit, dass das Maximum größer als 119.6 ist. Diese Entsprechung folgt aus der Symmetrie der Normalverteilung.

Es werde noch der Fall betrachtet, dass die Stichprobe nur den Umfang $n = 25$ habe. Dann ergibt sich

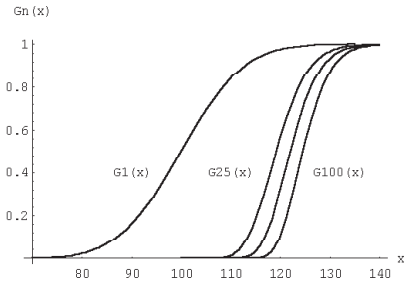
$$G_{25}(119.6) = .975^{25} = .531$$

und

$$H_{25}(80.4) = 1 - .975^{25} = .469$$

Bei dieser kleineren Stichprobe ist die Wahrscheinlichkeit, dass das Maximum und das Minimum außerhalb des Intervalls (80.4, 119.6) liegen, also kleiner als im Fall $n = 50$. Man kann deshalb vermuten, dass die Varianz einer Stichprobe im allgemeinen größer wird, je größer ihr Umfang ist; mit größer werdendem

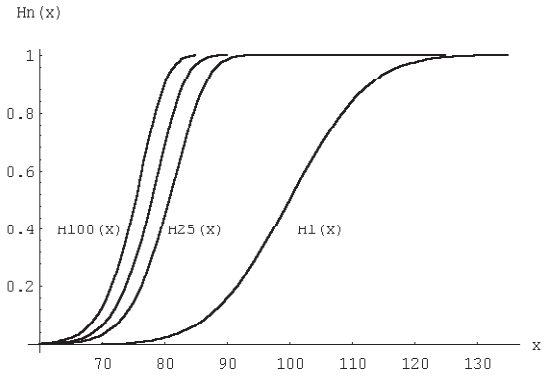
Abbildung 7.1: Maxima normalverteilter Variablen



Umfang wird ja die Wahrscheinlichkeit des Auftretens besonders großer und besonders kleiner Werte größer. Die Frage, um wieviel die Stichprobenvarianz mit steigendem n größer wird, ist damit aber noch nicht beantwortet, schließlich ist die Stichprobenvarianz s^2 ja eine Schätzung der Populationsvarianz σ^2, die mit größer werdendem n immer näher bei σ^2 liegen wird. Dieses scheinbare Paradox wird aufgelöst, wenn man die Verteilungen von X_{n+} und X_{n-} betrachtet, denn diese Verteilungen spezifieren, um *wieviel* größer das Maximum als 119.6 bzw. um *wieviel* kleiner das Minimum als 80.4 sein wird. Die Wahrscheinlichkeit, dass $X_{n+} > 119.6$ und $X_{n-} < 80.4$ ist, kann groß sein. Die Frage ist, ob sie mit zunehmendem Wert von n nicht nur wahrscheinlicher werden, sondern auch dem Betrag nach von diesen Werten immer mehr abweichen. □

Die Verteilungen der Extrema kann man sich anhand der Ausdrücke (7.2) und (7.3) verdeutlichen. Zur Illustration sollen die Verteilungen für die Maxima und Minima einer normalverteilten zufälligen Veränderlichen betrachtet werden, die den Erwartungswert $\mu = 100$ und die Standardabweichung $\sigma = 10$ haben möge (vergl. Beispiel 7.1.1). Die folgende Abbildung 7.1 zeigt die Verteilung der Maxima für (i) $n = 25$ ($G_{25}(x)$), (ii) $n = 50$ (dies ist die mittlere der Verteilungen auf der rechten Seite der Abbildung), und (iii) für $n = 100$ ($G100(x)$). Die Verteilung $G1(x)$ ist die Verteilung für X. Man sieht, dass die Verteilungen der Maxima rechts von der Verteilung $G1(x) = F(x)$ liegen. Je größer der Wert von n, desto mehr rückt die Verteilung der entsprechenden Maxima nach rechts; - allerdings entspricht der Betrag, um den sie nach rechts rutscht, nicht dem 2-dimen Zuwachs von n: der Betrag, um den die Verteilung für $n = 50$ von der für $n = 25$ entfernt ist, ist ungefähr genau so groß wie der, um den die Verteilung für $n = 100$ von der für $n = 50$ entfernt ist. Zudem werden die Verteilungen mit zunehmendem Wert von n steiler, d.h. die Varianz der Verteilungen der Maxima wird mit steigendem Wert von n kleiner. Die Maxima nehmen so gut wie nie Werte größer als 135 an. Mit steigendem Wert von n kann man also einigermaßen sicher sein, dass für die gegebenen Werte von μ und σ das Maximum zwischen 120 und 135 liegt, aber die Gesamtstichprobenvarianz wird deswegen nicht beliebig größer, denn die Masse der Werte der Stichprobe wird nach wie vor im Mittelfeld um μ liegen. Die Vertei-

Abbildung 7.2: Minima normalverteilter Variablen



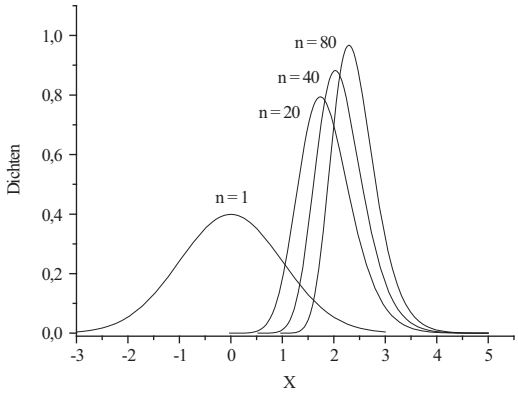
lung des Maximums gibt an, in welchem Bereich das Maximum einer Stichprobe liegen wird, aber nicht, wie häufig extreme Werte in der Stichprobe auftreten werden. Dies impliziert, dass s^2 tatsächlich gegen σ^2 strebt. Diese Betrachtung impliziert natürlich, dass analoge Betrachtungen für die Minima gelten. Zur Illustration werden in Abbildung 7.2 die entsprechenden Verteilungen für die Minima gezeigt. Diese Verteilungen entsprechen denen der Maxima; diese Eigenschaft der Verteilungen der Extrema ergibt sich aus der Symmetrie der Normalverteilung, für nichtsymmetrische Verteilungen ist die Entsprechung der Maxima- und Minimaverteilungen nicht gegeben.

Die Dichtefunktion für die Maximaverteilung ergibt sich sofort durch Differentiation von $G_n(x) = F^n(x)$: es ist

$$f_n(x) = \frac{dG_n(X)}{dx} = nF(x)^{n-1}f(x), \quad f(x) = \frac{dF(x)}{dx} \tag{7.8}$$

Zur weiteren Verdeutlichung werden in Abbildung 7.3 noch die Dichten der Maxima gezeigt: die linke Verteilung ist die Dichte der Variablen X, d.h. es ist $n = 1$, und die Dichten auf der rechten Seite entsprechen den Werten $n = 25$, $n = 50$ und $n = 100$. Generell gilt, dass die Dichten der Maxima nicht mehr symmetrisch sind; die Verteilungen sind linkssteil. Die Maxima der Dichten rücken zwar nach rechts (entsprechend der Tatsache, dass die Verteilungsfunktionen für steigenden Wert von n nach rechts rücken, gleichzeitig haben die Dichten mit steigendem Wert von n eine kleiner werdende Varianz; dies wird schon dadurch sichtbar, dass die Maxima der Dichten mit steigendem Wert von n größer werden. Da das Integral über die Dichte (d.h. die Fläche unter der Kurve) stets gleich 1 sein muß, bedeutet dieses Anwachsen des Maximalwertes von f_n, dass die Dichten schmaler werden, also eine kleinere Varianz reflektieren. Die Maxima werden also mit steigendem n in einem immer schmaler werdendem Bereich liegen. Man sieht weiterhin, dass der Modalwert dieser Dichten langsam gegen den Wert $x = 130$ wandert und dass größere Werte sehr schnell unwahrscheinlich werden. Man muß sich noch

Abbildung 7.3: Dichten der Maxima (standard-)normalverteilter Variablen



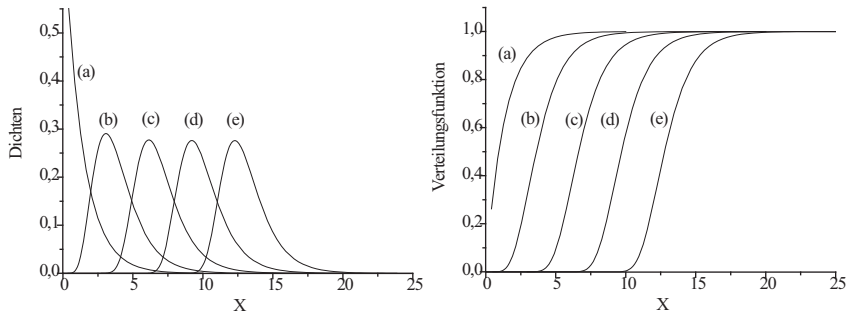
einmal verdeutlichen, dass es sich bei diesen Dichten um die Dichten der Maxima handelt, sie zeigen eben nur, in welchem Bereich der x-Skala das *Maximum* einer Stichprobe tatsächlich liegen wird. Dies ist tatsächlich der Extrembereich der ursprünglichen Dichte $f(x)$. Werte von X größer als 135 haben eine Wahrscheinlichkeit nahe bei Null, und dementsprechend treten sie eben auch nur als Maxima einer Stichprobe auf. Daß sie bei größer werdendem n tatsächlich in einer Stichprobe gefunden werden, drückt nur den Sachverhalt aus, dass auch Ereignisse mit kleiner Wahrscheinlichkeit beobachtet werden, wenn man ihnen nur hinreichend oft die Chance gibt, aufzutreten. Die Gesamtvarianz der Stichprobe wird dadurch nicht oder kaum erhöht.

Die Dichten für die Minima der Stichproben sind, wie die Verteilungsfunktionen für Minima, analog zu denen der Maxima und müssen deswegen nicht weiter diskutiert werden.

Zur weiteren Illustration werde noch kurz die Verteilung der Extrema von exponentialverteilten Variablen betrachtet. Dazu kann man eine Situation betrachten, die analog zu der unter Punkt 2. angesprochenen ist. Dort wurde das Minimum der Zeiten t_i als die Zeitdauer bis zum Ausfall des Gesamtsystems genannt. Man kann sich aber auch ein System aus unabhängig voneinander und parallel arbeitenden Komponenten vorstellen, das eine Aufgabe dann beendigt hat, wenn alle Komponenten ihre (Teil-)Aufgabe beendet haben.[1] Die Zeiten bis zur Beendigung der Teilaufgaben seien unabhängig und identisch verteilt. In Abb. 7.4 werden links die Dichten für $n = 1$, $n = 10$, $n = 10^2$, $n = 10^3$ und $n = 10^4$ gezeigt, rechts die zugehörigen Verteilungsfunktionen der Maxima. Für steigenden Wert von n rücken die Dichten und die entsprechenden Verteilungen weiter nach rechts; der ungefähr Gleichabstand sollte nicht täuschen, denn n wächst ja wie n^k, $k = 1, 2, \ldots$, – der Abstand zwischen den Verteilungen auf der x-Achse

[1] Ein Lehrer fährt erst dann mit dem Unterricht fort, wenn alle Kinder ihm zuhören.

Abbildung 7.4: Dichten und Verteilungen der Maxima exponentialverteilter Variablen, (a) $n = 1$, (b) $n = 10$, (c) $n = 100$, (d) $n = 1000$, (e) $n = 10000$.



für steigenden Wert von n wird tatsächlich kleiner.

Beispiel 7.1.2 Es sei x_i die Höhe der i-ten Flut an der ostfriesischen Küste. X ist sicherlich eine zufällige Veränderliche; bei Sturm bzw. Springfluten können die Bauern hinter den Deichen nur hoffen, dass x_i hinreichend klein wird, damit die Deiche nicht überspült werden. Umgekehrt muß sich der Deichgraf fragen, wie hoch die Deiche sein müssen, damit die Wahrscheinlichkeit, dass sie nicht überflutet werden, hinreichend groß ist. Es sei $X_+ = \max(x_1, \cdots, x_n)$, und n ist natürlich sehr groß (man will ja mindestens für die nächsten 20 oder 30 Jahre sicher sein). Es sei x_0 die Deichhöhe; dann soll also

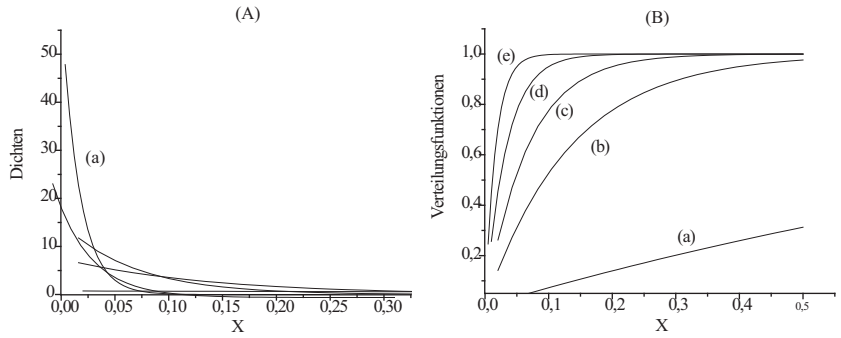
$$P(X_+ \leq x_0) \approx 1$$

gelten. Aus Kostengründen muß x_0 so niedrig wie möglich, aber natürlich so hoch wie nötig angesetzt werden. Man könnte vermuten, dass x_0 um so größer sein muß, je länger der Zeitraum ist, für den man Sicherheit verlangt. Es zeigt sich aber allgemein, dass x_0 nicht beliebig mit n wächst, sondern gegen einen Grenzwert strebt. Dieser läßt sich aus der Verteilung $\lim_{n\to\infty} F^n(x_{0n})$ bestimmen, d.h. $\lim_{n\to\infty} x_{0n} = x_0$. Die Fluthöhen sind sicher nicht exponentialverteilt; das in Abbildung 7.4 illustrierte Prinzip gilt auch für andere Verteilungen. □

Es ist noch von Interesse, sich die Verteilung der Minima exponentialverteilter Variablen anzusehen[2]; sie sind in Abbildung 7.5 für die Werte $n = 1, 3, 10, 25$ gegeben. Je größer der Wert von n, desto mehr wandert die Verteilung nach links, wobei aber der Betrag, um den sie nach links wandert, schnell sehr klein wird. Die Verteilung der Minima hat eine andere Form wie die der Maxima; dies ist eine Konsequenz der Tatsache, dass die Exponentialverteilung asymmetrisch ist.

[2]Der Lehrer beginnt mit dem Unterricht, sobald das erste Kind ihm zuhört.

Abbildung 7.5: Minima exponentialverteilter Variablen , (A) Dichten, (B) Verteilungsfunktionen; (a) $n = 1$, (b) $n = 10$, (c) $n = 20$, (d) $n = 40$, (e) $n = 80$.



Generell läßt sich also sagen, dass zwar die Verteilungen des Extremums mit steigendem n zwar immer weiter nach rechts bzw. nach links wandern, aber um stets kleiner werdende Beträge. Damit werden aber die Verteilungsfunktionen der jeweiligen Extrema für verschiedene Werte von n einander auch zunehmend ähnlicher. Man kann deswegen vermuten, dass sie gegen eine *Grenzverteilung* (*limiting distribution*) konvergieren. Die Grenzverteilung für die Maxima bzw. für die Minima wäre dann durch

$$G(x) = \lim_{n\to\infty} G_n(x), \quad H(x) = \lim_{n\to\infty} H_n(x) \tag{7.9}$$

definiert. Wie das Beispiel der Exponentialverteilung zeigt, sind die Grenzverteilungen für Maxima und Minima nicht notwendig von ähnlicher Form. Die Frage ist, ob für eine gegebene Verteilungsfunktion $F(x)$ tatsächlich solche Grenzverteilungen existieren und wie sich die Grenzverteilungen für verschiedene Ausgangsverteilungen F unterscheiden. Die überraschende, auf Fisher und Tippet (1928) zurückgehende Antwort ist, dass es zumindest für i.i.d. Variable nur drei solche Grenzverteilungen für Maxima, und dementsprechend auch nur drei Grenzverteilungen für Minima, gibt. Die Verteilung der Maxima (Minima) von Variablen mit verschiedenen Verteilungen F kann also durch die gleiche Grenzverteilung gegeben sein. Die Menge der Verteilungen mit der gleichen Grenzverteilung heißt der *Attraktionsbereich* dieser Grenzverteilung. Die meisten der gängigen Verteilungen gehören einem der drei Attraktionsbereiche an; eine Ausnahme ist z.B. die Poisson-Verteilung. Sie gehört keinem Attraktionsbereich an, wie Gnedenko (1943) gezeigt hat, d.h. die Verteilung der Maxima der Poisson-Verteilung konvergiert nicht mit steigendem Wert von n gegen eine der drei möglichen Grenzverteilungen.

Die Grenzverteilungen sind als *Extremwertverteilungen* bekannt. Sie sind für viele Betrachtungen von Interesse, nicht nur dann, wenn man etwa die Verteilung der extremen Werte einer Stichprobe diskutieren möchte, und auch nicht nur

dann, wenn man es mit i.i.d. Variablen zu tun hat. Die unter Punkt 2. angesprochenen Reliabilitätsfragen implizieren häufig, dass die zu betrachtende Verteilung der Meßwerte durch eine Extremwertverteilung gegeben ist und nicht etwa durch eine Normalverteilung. Im folgenden Abschnitt sollen die Extremwertverteilungen vorgestellt werden.

7.2 Extremwertverteilungen als Grenzverteilungen

7.2.1 Grenzverteilungen für Maxima

Wie im vorigen Abschnitt gezeigt wurde, rücken die Verteilungen F^n mit größer werdendem Wert von n nach rechts. Dies heißt, dass für einen gegebenen Wert $G_n = F^n(x) = p$, $0 < p < 1$ ein Wert x benötigt wird, der um so größer ist, je größer n ist. Hält man dagegen x fest und läßt n wachsen, so strebt $F^n(x)$ natürlich gegen Null. Damit F^n mit wachsendem n *nicht* gegen Null strebt, muß man eben für jeden Wert von n einen Wert x_n von x wählen, der entsprechend weiter rechts liegt als der Wert x, für den $F(x) = p > 0$. Da die Verteilungen F^n mit steigendem n zwar weiter nach rechts rücken, aber der Betrag, um den sie nach rechts rücken, mit steigendem n auch immer kleiner zu werden scheint und sich die Verteilungen F^n auch immer mehr ähneln - diese Beobachtung legt ja die Vermutung der Existenz einer Grenzverteilung nahe - scheinen sich auch die Bereiche von x-Werten, auf denen die F^n definiert sind, immer mehr anzunähern. Eine Möglichkeit, diese Bereiche zu definieren, besteht darin, x für jedes n geeignet zu *skalieren*. Dem entsprechend macht man die folgende

Annahme: *Es sei $F(x) = p$, $0 < p < 1$. Für jeden Wert von n existieren von p unabhängige Zahlen a_n und b_n derart, dass*

$$F^n(a_n x + b_n) = p. \tag{7.10}$$

Die Zahlen a_n und b_n heißen *Normierungskonstanten*; sie sind für eine gegebene Verteilung F charakteristisch. Sie müssen aber gewissen allgemeinen, d.h. von der Verteilung F unabhängigen Kriterien genügen. So ist klar, dass $F(x_n) = F(a_n + b_n)$ immer näher bei 1 liegen muß, damit $F^n(a_n x + b_n) = p$ ist. Die Folgen $\{a_n\}$ und $\{b_n\}$ müssen also so gewählt werden, dass

$$F(a_n x + b_n) \to 1 \text{ für } n \to \infty, \quad \lim_{n\to\infty} F^n(a_n x + b_n) = G(x), \tag{7.11}$$

wobei $G(x)$ die in (7.9) gegebene Grenzverteilung ist. Hieraus ergibt sich ein erster Ansatz, $G(x)$ zu charakterisieren. Für eine beliebige Zahl a gilt ja $a = e^{\log a}$, log der natürliche Logarithmus. Also kann man auch

$$F^n(a_n x + b_n) = e^{n \log F(a_n x + b_n)}$$

schreiben. Für wachsendes n strebt $F(a_n x+b_n)$ gegen 1, und damit $\log F(a_n x+b_n)$ gegen Null. Es sei z eine beliebige reelle Zahl; man kann die Funktion $\log z$ in eine Taylor-Reihe entwickeln, und für $z < 1$, aber hinreichend nahe bei 1, gilt dann

$$\log z \approx -(1-z) \tag{7.12}$$

Dann folgt

$$\lim_{n\to\infty} e^{-n(1-F(a_n x+b_n))} = G(x) \tag{7.13}$$

Da der Limes durch den Grenzwert des Exponenten $n(1-F(a_n x+b_n))$ bestimmt wird, kann man den in dieser Gleichung ausgedrückten Sachverhalt auch so ausdrücken, dass

$$\lim_{n\to\infty} n(1-F(a_n x+b_n)) = -u(x), \quad G(x) = e^{-u(x)} \tag{7.14}$$

gelten muß. In anderen Worten, die Grenzverteilung G ist, für alle x, durch den Grenzwert $u(x)$ definiert.

Die Frage ist nun, ob die $\{a_n\}$, $\{b_n\}$ für eine gegebene Verteilung F eindeutig festliegen. Hierzu gilt ein Satz von Khintchine:

Satz 7.2.1 *Es sei $\{F^n = F_n\}$ eine Folge von Verteilungsfunktionen und $a_n > 0$, b_n seien Normierungskonstanten, so dass*

$$F_n(a_n x + b_n) \to G(x).$$

Dann gilt für die Folgen $\alpha_n > 0$, β_n und die Verteilungsfunktion $G^*(x)$ die Aussage

$$F_n(\alpha_n x + \beta_n) \to G^*(x) \tag{7.15}$$

dann und nur dann, wenn

$$\frac{a_n}{\alpha_n} \to a > 0, \quad \frac{\beta_n - b_n}{a_n} \to b. \tag{7.16}$$

und

$$G^*(x) = G(ax+b) \tag{7.17}$$

Beweis: Leadbetter et al. (1983), p. 7.

Die Aussage dieses Satzes ist, dass die Normierungskonstanten eindeutig bis auf eine lineare Transformation sind, und dass sich die Grenzverteilungen, wenn sie für eine Ausgangsverteilung F existieren, eindeutig bis auf eine lineare Transformation von x sind. Die Beziehung (7.17) bedeutet, dass die Grenzverteilungen Elemente der in der folgenden Definition charakterisierten Verteilungen sind:

Definition 7.2.1 *Die Verteilung G heißt* max-stabil*, wenn für $n = 2, 3, \cdots$ Konstante $a_n > 0$, b_n existieren derart, dass*

$$G^n(a_n x + b_n) = G(x) \tag{7.18}$$

Gilt für die zwei Verteilungen G_1 und G_2 die Beziehung

$$G_2(x) = G_1(ax + b) \tag{7.19}$$

für reelle Zahlen $a > 0$ und b, so heißen die Verteilungen G_1 und G_2 vom gleichen Typ.

Offenbar ist eine Verteilung max-stabil, wenn G^n und G vom gleichen Typ sind. Der folgende Satz geht auf Fisher und Tippett (1928) zurück:

Satz 7.2.2 *Es sei G eine max-stabile Verteilung. Dann ist G durch eine der folgenden drei Verteilungstypen definiert:*

$$G(x) = \begin{cases} \exp\left(-e^{-x}\right), & -\infty < x < \infty & & \textit{Typ I: Extremwertverteilung} \\ \\ \begin{matrix} 0, \\ \exp\left(-x^{-\alpha}\right) \end{matrix}, & \begin{matrix} x > 0 \\ x \leq 0 \end{matrix} & \alpha > 0 & \textit{Typ II: Weibull} \\ \\ \begin{matrix} \exp\left(-(-x)^{\alpha}\right), \\ 1, \end{matrix} & \begin{matrix} x \leq 0 \\ x > 0 \end{matrix} & \alpha > 0 & \textit{Typ III: Pareto} \end{cases} \tag{7.20}$$

Ist ungekehrt G eine Extremwertverteilung, so folgt, dass sie auch max-stabil ist. Dann gilt

Beweis: Leadbetter et al. (1983), p. 10.

Die Aussage des Satzes ist, dass F^n für gegebene Verteilung F gegen eine der drei Verteilungstypen konvergiert, falls F^n überhaupt gegen eine Grenzverteilung konvergiert. Die Frage ist nun, wie man feststellt, dass F^n gegen eine der drei Verteilungen konvergiert. Dazu muß zunächst der Bereich gekennzeichnet werden, auf dem F definiert ist. Es sei

$$\alpha_F = \inf\{x|F(x) > 0\}, \quad \omega_F = \sup\{x|F(x) < 1\} \tag{7.21}$$

Dabei ist $\inf\{x|F(x) > 0\}$ das *Infimum* der Menge der x-Werte, auf denen $F(x) > 0$ ist, d.h. α_F ist der kleinste x-Wert, für den F noch größer als 0 ist. Mit $\sup\{x|F(x) < 1\}$ ist das *Supremum* der Menge der x-Werte, für die $F(x) < 1$ ist, d.h. ω_F ist der größte x-Wert, für den $F < 1$ ist. Die folgende Definition spezifiert eigentlich nur eine Redeweise, soll aber dennoch wegen ihren allgemeinen Gebrauchs explizit eingeführt werden:

Definition 7.2.2 *F^n konvergiere gegen eine Grenzverteilung. Dann gehört F zum* Attraktionsbereich *(*domain of attraction*) dieser Grenzverteilung.*

Der Attraktionsbereich einer Grenzverteilung ist also die Menge der Verteilungen F, für die F^n gegen eben diese Grenzverteilung strebt. Die Menge der überhaupt existierenden Grenzverteilungen kann also in vier disjunkte Teilmengen zerlegt werden: drei Teilmengen korrespondieren zu den drei Attraktionsbereichen der Grenzverteilungen, die vierte Teilmenge umfaßt diejenigen Verteilungen, die zu keinem der Attraktionsbereiche gehören. Daß die Attraktionsbereiche tatsächlich disjunkt sind, geht aus dem folgenden Satz 7.2.3 hervor. In diesem Satz werden Eigenschaften von Verteilungen F genannt, anhand derer entschieden werden kann, gegen welche der drei möglichen Grenzverteilungen F^n konvergiert, wenn letztere überhaupt gegen eine solche Verteilung konvergiert.

Satz 7.2.3 *Die Verteilung F gehöre zum Attraktionsbereich einer Grenzverteilung. Ein Attraktionsbereich ist durch bestimmte Eigenschaften definiert, die nur den Verteilungen, die zu diesem Bereich gehören, eigen sind, d.h. diese Eigenschaften sind notwendig und hinreichend für die Zugehörigkeit zu einem dieser Bereiche. Es sei*

$$\gamma_n = \inf\left\{x | F(x) \geq 1 - \frac{1}{n}\right\}.$$

Typ I: Es existiert eine streng positive Funktion $g(t)$ derart, dass

$$\lim_{t\to\omega_F} \frac{1 - F(t + xg(t))}{1 - F(t)} = e^{-x} \tag{7.22}$$

für alle reellen x.

Typ II: Es ist $\omega_F = \infty$ und es gilt

$$\lim_{t\to\infty} \frac{1 - F(tx)}{1 - F(t)} = x^{-\alpha}, \quad \alpha > 0 \tag{7.23}$$

Typ III: Es ist $\omega_F < \infty$ und es gilt

$$\lim_{h\to 0} \frac{1 - F(\omega_F - xh)}{1 - F(\omega_F - h)} = x^{\alpha}, \quad \alpha > 0, \text{ für alle} x > 0 \tag{7.24}$$

Liegt eine Verteilung F vor, deren Grenzverteilung vom Typ I ist, so gilt

$$\int_0^\infty (1 - F(u))\, du < \infty$$

und die Funktion g ist durch

$$g(t) = \int_t^{\omega_F} \frac{1 - F(u)}{1 - F(t)}\, du, \quad \text{für } t < \omega_F \tag{7.25}$$

gegeben. Die Normierungskonstanten für die einzelnen Typen sind durch die folgenden Ausdrücke gegeben:

$$\begin{array}{lll} \textit{Typ I:} & a_n = g(\gamma_n), & b_n = \gamma_n, \\ \textit{Typ II:} & a_n = \gamma_n, & b_n = 0, \\ \textit{Typ III:} & a_n = \omega_F - \gamma_n, & b_n = \omega_F; \end{array} \tag{7.26}$$

Dabei ist die Funktion g durch (7.25) gegeben.

Beweis: Leadbetter, Lindgren und Rootzen (1983), p. 17.

Will man also den Typ der Grenzverteilung für eine gegebene Verteilungsfunktion F finden, so muß man untersuchen, welche der in Satz 7.2.3 aufgeführten Eigenschaften für F gelten; das Vorgehen wird anhand der Beispiele in Abschnitt 7.2.2 illustriert.

Anmerkungen: Es sei kurz angemerkt, wie die Grenzverteilungen angewendet werden:

1. Allgemein zur Approximation der Wahrscheinlichkeit, dass bei n unabhängigen Versuchen der maximale beobachtete x-Wert kleiner als x_n ist:

$$P(X_+ \leq x_n) \approx G(x) = G\left(\frac{x_n - b_n}{a_n}\right), \tag{7.27}$$

   mit

$$x_n = a_n x + b_n, \quad \text{bzw. } x = (x_n - b_n)/a_n$$

2. Zur Bestimmung des Wertes x_n, der mit einer *vorgegebenen* Wahrscheinlichkeit $P(X_+ \leq x_n)$ bei n Versuchen nicht überschritten wird. Diese Art Fragestellung ergibt sich z.B. bei Reliabilitätsuntersuchungen: welche Deichhöhe x_n wird bei n Hochwassern mit einer Wahrscheinlichkeit von - etwa - .99 in den nächsten 10 Jahren (man rechne sich den entsprechenden n-Wert aus!) nicht überspült? Welche Reaktionszeit wird mit einer Wahrscheinlichkeit von .95 bei 1000 Versuchen nicht überschritten? Welcher IQ ist mit der Wahrscheinlichkeit .95 bei einer Stichprobe von 500 Personen maximal?

3. Die Grenzverteilungen können auch ohne expliziten Bezug auf Verteilungsfunktionen F als Verteilungen von zufälligen Veränderlichen betrachtet werden. Sie werden etwa in bestimmten Modellen von Beurteilungsprozessen angewandt: man nimmt an, dass sich Urteile u. U. an Extremwerten orientieren, kennt aber die Verteilung F nicht und nimmt deswegen gleich an, dass die Urteile etwa durch die Doppelte Exponentialverteilung (Typ I) gegeben sind.

7.2.2 Beispiele für die Verteilung von Maxima

Beispiel 7.2.1 F sei durch die Exponentialverteilung

$$F(t) = 1 - \exp(-\lambda t)$$

gegeben. Gesucht ist die Grenzverteilung für

$$X_+ = \max(x_1, \cdots, x_n)$$

Es wird zuerst der Typ der Grenzverteilung G bestimmt. Es ist $\omega_F = \infty$, folglich kann G nicht vom Typ III sein, da dieser Typ $\omega_F < \infty$ voraussetzt. Weiter ist

$$\frac{1 - F(tx)}{1 - F(t)} = \frac{\exp(-\lambda tx)}{\exp(-\lambda t)} \neq x^{-\alpha},$$

mithin kann G nicht vom Typ II sein. Also muß G vom Typ I sein. Um die Normierungskonstanten zu bestimmen, muß die Funktion g bestimmt werden. Nach (7.25) ergibt sich

$$g(t) = e^{\lambda t} \int_t^\infty e^{-\lambda u} du = \frac{1}{\lambda}, \text{ für alle } t$$

und

$$\gamma_n = \inf\{x | 1 - e^{-\lambda x} \geq 1 - 1/n\} = \inf\{x | 1/n \geq e^{-\lambda x}\}$$

so dass $-\log n = -\lambda x$, d.h.

$$\gamma_n = \frac{1}{\lambda} \log n$$

Dann folgt

$$a_n = \frac{1}{\lambda}, \quad b_n = \frac{1}{\lambda} \log n, \tag{7.28}$$

Also gilt

$$\begin{aligned} F^n(a_n x + b_n) &= \left(1 - \exp\left(-\lambda\left(\frac{1}{\lambda}x + \frac{1}{\lambda}\log n\right)\right)\right)^n \\ &= (1 - \exp(-x - \log n))^n = \left(1 - \frac{1}{n}e^{-x}\right)^n \\ &\approx G(x) = \exp\left(-e^{-x}\right) \end{aligned}$$

Man sieht, dass $F^n(a_n x + b_n)$ unabhängig vom Parameter λ wird; das muß so sein, da ja $G(x)$ unabhängig von λ ist. Gibt man allerdings den Wert von $x_n = a_n x + b_n$ vor (man möchte etwa wissen, wie groß die Wahrscheinlichkeit ist, dass bei 500 Versuchen die maximale Reaktionszeit nicht größer als 1 Sekunde ist), so muß man $G((x_n - b_n)/a_n)$ berechnen, und in diese Berechnung geht dann der Wert von λ ein, denn es ist

$$x = \frac{x_n - b_n}{a_n} = \frac{x_n - (1/\lambda)\log n}{1/\lambda} = \lambda x_n - \log n$$

Abbildung 7.6: Approximation der Verteilung der Maxima exponentialverteilter Variablen

In der Tat ist ja x ein Wert aus dem Wertebereich von F, der unter anderem durch den Wert des Parameters λ spezifiert wird, während x_n ein Wert aus dem Wertebereich von F^n ist, und der wird mit wachsendem n unabhängig von λ, wie aus der Konvergenz gegen $G(x)$ deutlich wird. In Abbildung 7.6 wird links die Verteilung der Maxima von 10 exponentialverteilter Variablen und rechts die von 30 ebenso verteilten Variablen (jeweils $\lambda = 1$) zusammen mit der Approximation durch die Doppelte Exponentialverteilung gezeigt. Man sieht, dass die Approximation bereits für $n = 10$ sehr gut ist und für $n = 30$ sind die "wahre" Verteilung F^n und die Approximation $G((X_+ - b_n)/a_n)$, praktisch identisch. □

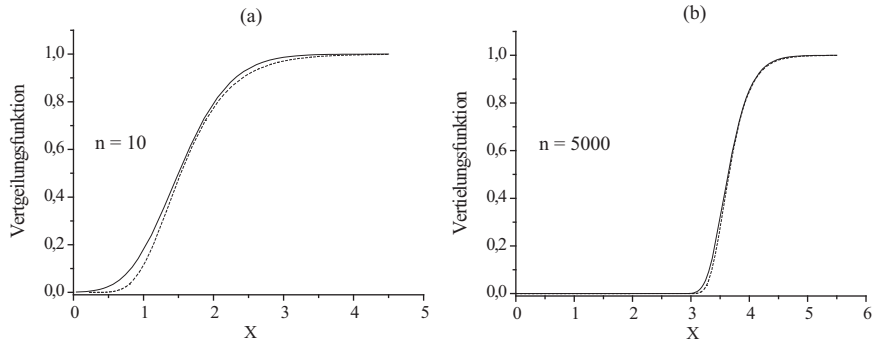
Beispiel 7.2.2 Es wird eine Stichprobe von n Personen gebildet. Jede Person P_i, $i = 1, \cdots, n$, besitzt das Merkmal M (etwa: Aggressivität) zu einem gewissen Ausmaß x_i. Man ist an der Wahrscheinlichkeit, dass der maximale Wert $X_+ = \max(x_1, \cdots, x_n)$ der Ausprägungen x_i nicht größer als ζ_n ist. Die x-Werte seien in der Population normalverteilt mit dem Erwartungswert $E(x) = \mu$ und der Varianz σ^2.

Es wird zunächst der Typ der Grenzverteilung bestimmt. Da $\omega_F = \infty$ kann G nicht vom Typ III sein. Es läßt sich zeigen, dass G vom Typ I ist (Leadbetter et al. (1983), p. 15), und dass für die *Standardnormalverteilung* die Normierungskonstanten durch

$$a_n = \frac{1}{\sqrt{2\log n}}, \quad b_n = \sqrt{2\log n} - \frac{\log\log n + \log 4\pi}{2\sqrt{2\log n}} \tag{7.29}$$

gegeben sind; die Details dieses Nachweises sind ein wenig länglich und sollen deswegen hier übergangen werden. Die Abb. 7.7 zeigt die Verteilung des Maximums einer $N(0,1)$-verteilten Variablen bei einem Stichprobenumfang von (a) $n = 10$ und (b) bei $n = 5000$. Eine Kurve ist dabei die "wahre" Verteilung, d.h. F^n, die

Abbildung 7.7: Maxima standardnormalverteilter Variablen II, (a) n = 10, (b) n = 5000. Gestrichelte Funktion: Approximation durch die doppelte Exponentialverteilung $\exp[-\exp(-(x-b_n)/a_n)]$.



zweite ist die Approximation durch die Grenzverteilung $G((x-b_n)/a_n)$ (gestrichelt). Offenbar rückt die Verteilung mit steigendem n nach rechts, entsprechend der Tatsache, dass mit größerem n auch die Wahrscheinlichkeit extremer Werte größer wird; die Verteilung wird aber auch steiler, d.h. die Varianz der Extreme wird kleiner. Die approximierende Verteilung G ist zwar in beiden Fällen nahe an F^n und zeigt vor allem einen ähnlichen Verlauf wie F^n, bei der Bestimmung der Wahrscheinlichkeiten ergeben sich aber gleichwohl Fehler. Dies entspricht der Tatsache, dass die Konvergenz von F^n gegen G bei der Normalverteilung ausgesprochen langsam verläuft.

Es sei nun $X \sim N(\mu, \sigma^2)$-verteilt. Dieser Fall wird wegen $z = (x-\mu)/\sigma$ leicht auf den eben betrachteten $N(0,1)$-Fall zurückgeführt. Es ist

$$F^n(a_n z + b_n) = F^n(\zeta_n) \approx G(z), \quad z = (x-\mu)/\sigma$$

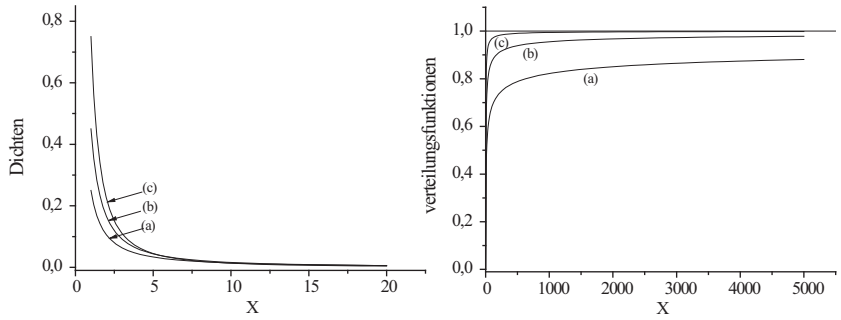
Da $z = (\zeta_n - b_n)/a_n$ wird G im *Bereich* der ζ_n-Werte dargestellt, allerdings für die entsprechenden z-Werte. Es sei angemerkt, dass für die Gauß-Verteilung die Approximation durch die Doppelte Exponentialverteilung für steigende n außerordentlich langsam ist.

Es soll noch angemerkt werden, dass die Lognormalverteilung ebenfalls die Doppelte Exponentialverteilung als Grenzverteilung hat. Die Normierungskonstanten ergeben sich aus denen der Normalverteilung:

$$a_n^{'} = a_n \exp(-b_n), \quad b_n^{'} = \exp(b_n) \tag{7.30}$$

wobei a_n und b_n die Normierungskonstanten der Normalverteilung sind. □

Abbildung 7.8: Pareto-Dichten und Verteilungen, $K = 1$, (a) $\alpha = .25$, (b) $\alpha = .45$, (c) $\alpha = .75$.



Beispiel 7.2.3 Pareto-Verteilung Es sei nun die Verteilung

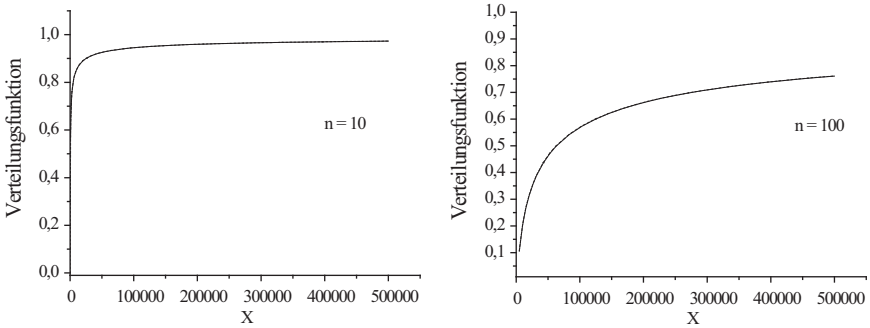
$$F(x) = 1 - K\,x^{-\alpha}, \quad \alpha > 0, \quad K > 0, \quad x \geq K^{1/\alpha} \tag{7.31}$$

gegeben; dies ist die Pareto-Verteilung. Die Dichtefunktion der Pareto-Verteilung ist

$$f(x) = \frac{dF(x)}{dx} = \frac{\alpha\,K}{x^{1+\alpha}}, \quad x \geq x_0. \tag{7.32}$$

Offenbar ist $f(x)$ maximal für $x = x_0$ und fällt für alle $x > x_0$ monoton gegen Null, und zwar um so schneller, je größer der Wert von α ist. für $K = .75$ und $\alpha = .25$ bzw. $\alpha = .75$ ist die Verteilung in Abbildung 7.8 gegeben. Diese Verteilung

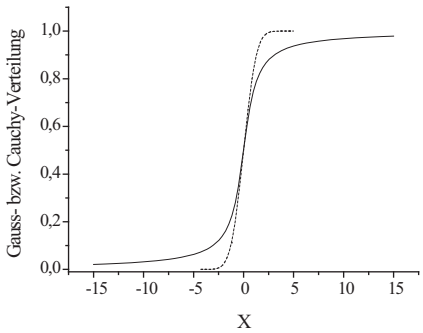
Abbildung 7.9: Pareto-Verteilungen, $K = 1$, $\alpha = .45$, für $n = 10$ und $n = 100$. Wahre und approximierende Verteilungen werden praktisch identisch.



zeichnet sich dadurch aus, dass sie *langsam* gegen 1 strebt; je kleiner der Wert von α ist, desto langsamer geht sie gegen 1.

Gegeben sei weiter eine Stichprobe von n Werten, die jeweils nach (7.31) verteilt sind, und es soll die Grenzverteilung für $P(X_+ \leq \zeta_n)$ bestimmt werden.

Abbildung 7.10: Die Cauchy-Verteilung im Vergleich zur Standardnormalverteilung (gestrichelt)



Es ist $\omega_F = \infty$, also kann die Verteilung nicht vom Typ III sein. Man findet

$$\frac{1 - F(tx)}{1 - F(t)} = x^{-\alpha}$$

woraus folgt, dass sie vom Typ II ist, d.h.

$$G(x) = 0 \text{ für } x \leq 0, \quad G(x) = \exp(-x^{-\alpha}) \text{ für } x > 0, \quad \alpha > 0$$

Es ist

$$\gamma_n = \inf\{x | F(x) \geq 1 - 1/n\}.$$

Für $1 - Kx^{-\alpha} = 1 - 1/n$ folgt $Kx^{-\alpha} = 1/n$, also $\gamma_n = (Kn)^{1/\alpha}$, so dass

$$a_n = \gamma_n = (Kn)^{1/\alpha}, \quad b_n = 0$$

nach (7.26). □

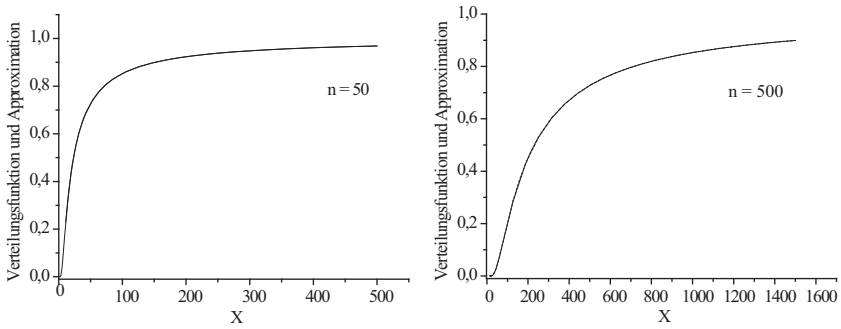
Ein Beispiel für die Typ III Grenzverteilung ist diese Verteilung selbst; man mag sich die Verhältnisse zur übung selbst klar machen.

Beispiel 7.2.4 Gesucht ist die Grenzverteilung von $X_+ = \max(x_1, \cdots, x_n)$, wenn F durch die Cauchy-Verteilung

$$F(x) = \frac{1}{2} + \frac{1}{\pi} \tan^{-1} x \tag{7.33}$$

gegeben ist; diese Verteilung ergibt sich u.a. als Verteilung des Quotienten zweier Normalverteilter Größen, vergl. Abschnitt 5.11.2 In Abbildung 7.10 ist die Cauchy-Verteilung zusammen mit der Gauss-Verteilung ($N(0, 1)$) dargestellt. Man sieht, dass die Cauchy-Verteilung wesentlich langsamer gegen 1 bzw. gegen 0 geht, als die Normalverteilung. Es ist dieses Verhalten an den "Enden", dass die Grenzverteilung für die Cauchy-Verteilung bestimmt. Bei dieser Verteilung ist $\omega_F = \infty$,

Abbildung 7.11: Die Verteilung der Maxima der Cauchy-Verteilung für $n = 50$ und $n = 500$



also kann die Grenzverteilung nicht vopm Typ III sein. Angenommen, sie ist vom Typ I; dann muß sie die Bedingung $\int_0^\infty (1 - F(x))dx < \infty$ erfüllen. Es ist aber

$$\int_0^\infty (1 - F(x))\, dx = (1 - F(x))|_0^\infty + \int_0^\infty \frac{x}{\pi(1 + x^2}) \, dx = \infty;$$

die Bedingung ist also nicht erfüllt, mithin kann die Grenzverteilung nicht vom Typ III sein. Damit bleibt nur noch die Grenzverteilung vpm Typ II übrig. In der Tat, es gilt

$$\lim_{h \to 0} \frac{1 - F(tx)}{1 - F(t)} = \lim_{h \to 0} \frac{(1 + t^2)x}{1 + (tx)^2} = x^{-1}$$

d.h. die Grenzverteilung ist vom Typ II mit $\alpha = 1$:

$$G(x) = \exp\left(-x^{-\alpha}\right), \quad \alpha > 0, \quad x > 0 \tag{7.34}$$

und es ist

$$a_n = \tan\left(\frac{\pi}{2} - \frac{\pi}{n}\right) \tag{7.35}$$

Die Verteilungen der Maxima zweier Stichproben von $n = 50$ und $n = 500$ i.i.d. Cauchy-verteilten Werten werden zusammen mit ihrer Approximation durch die Grenzverteilung in Abbildung 7.11 dargestellt; die Approximation ist exzellent, – die wahre Verteilung und ihre Approximation durch eine Grenzverteilung können nicht unterschieden werden. □

7.2.3 Grenzverteilungen der Minima

Ist $X_+ = \max(x_1, \cdots, x_n)$, so sieht man leicht, dass

$$X_- = \min(x_1, \cdots, x_n) = -\max(-x_1, \cdots, -x_n) \tag{7.36}$$

Diese Beziehung legt nahe, dass die Grenzverteilungen für die Minima sich aus denen für die Maxima herleiten lassen. Aus (7.7) ist bekannt, dass die Verteilung des Minimums durch

$$H_n(x) = 1 - (1 - F(x))^n$$

gegeben ist. Für gegebenes x, $0 < F(x) < 1$ folgt

$$\lim_{n\to\infty} (1 - F(x))^n = 0$$

d.h. dass die x wieder in geeigneter Weise normiert werden müssen, damit $H_n > 0$ sein kann. Offenbar muß $1-F(x)$ gegeben 1 streben, damit $(1-F(x))^n$ einen Wert ungleich Null annimmt, so dass $F(x)$ in geeigneter Weise gegen Null streben muß. x muß also durch $\theta_n = c_n + d_n$ ersetzt werden, wobei die Normierungskonstanten c_n und d_n derart sein müssen, dass

$$H(x) = \lim_{n\to\infty} 1 - (1 - F(c_n x + d_n))^n, \quad 0 < H(x) < 1 \tag{7.37}$$

erfüllt ist. Nun ist

$$H_n = 1 - (1 - F(c_n x + d_n))^n = 1 - \exp(n \log(1 - F(c_n x + d_n)))$$

und für $F(c_n x + d_n) \to 0$ folgt

$$\log(1 - F(c_n x + d_n)) \approx -F(c_n x + d_n),$$

so dass

$$H_n = 1 - \exp(-nF(c_n x + d_n)) \tag{7.38}$$

und

$$\lim_{n\to\infty} nF(c_n x + d_n) = v(x)$$

ist endlich; dies heißt, dass sich die Grenzverteilung in der Form

$$H(x) = \exp(-v(x)) \tag{7.39}$$

darstellen läßt. Es gilt der folgende

Satz 7.2.4 *Es sei* $X_- = \min(x_1, \cdots, x_n)$ *und die* x_i *seien i.i.d. verteilt. Die Grenzverteilungen sind dann durch*

$$H(x) = \begin{cases} 1 - \exp\left(-e^x\right), & -\infty < x < \infty & & \textit{Typ I: Extremwertverteilung} \\ 1 - \exp\left(-(-x)^{-\alpha}\right), & x < 0 & \alpha > 0 & \textit{Typ II: Pareto} \\ 1, & x \geq 0 & & \\ 1 - \exp\left(-x^{\alpha}\right), & x \geq 0 & \alpha > 0 & \textit{Typ III: Weibull} \\ 0, & x < 0 & & \end{cases} \tag{7.40}$$

Beweis: Es ist

$$-X_- = \max(-x_1, \cdots, -x_n)$$

und

$$\begin{aligned} P\left(\frac{-X_- - d_n}{c_n} < x\right) &= 1 - P\left(\frac{-X_- - d_n}{c_n} \geq x\right) \\ &= 1 - P\left(\frac{X_- + d_n}{c_n} \leq -x\right) \\ &= 1 - H_n(-x) \end{aligned}$$

Andererseits ist

$$H_n(x) = 1 - G_n(-x), \quad \lim_{n\to\infty} G_n = G$$

so dass sich die Grenzverteilungen

$$\lim_{n\to\infty} H_n = H$$

aus den Grenzverteilungen G ergeben. Einen detaillierteren Beweis findet man in Leadbetter et al. (1983), p. 29. □

Es bleibt noch festzustellen, wie die Grenzverteilung der Minima für eine gegebene Verteilung zu ermitteln ist:

Satz 7.2.5 *Es sei $x_1, \cdots, x_n$ eine Stichprobe von i.i.d. verteilten Größen mit der Verteilungsfunktion F. Die Grenzverteilung für X_- gehört zu*

Typ I, wenn für endliches a, $\int_{\alpha_F}^{a} F(x)dx < \infty$, für alle x,

$$\lim_{t\to\alpha_F} \frac{F(t + xr(t))}{F(t)} = e^x, \quad r(t) = \frac{1}{F(t)} \int_{\alpha_F}^{t} F(x)dx \tag{7.41}$$

erfüllt ist. Die Normierungskonstanten sind dann durch

$$d_n = \sup\{x | F(x) \leq \frac{1}{n}\}, \quad c_n = r(d_n) \tag{7.42}$$

Typ II, wenn $\alpha_F = -\infty$ und es existiert eine Konstante $\gamma > 0$, so dass für alle x

$$\lim_{t\to-\infty} \frac{F(tx)}{F(t)} = x^{-\gamma} \tag{7.43}$$

gilt; die Normierungskonstanten sind dann durch

$$c_n = \sup\left\{x | F(x) \leq \frac{1}{n}\right\}, \quad d_n = 0 \tag{7.44}$$

Typ III, wenn für $\alpha_F > -\infty$,

$$\lim_{t\to-\infty} \frac{F(\alpha_F - 1/tx)}{F(\alpha_F - 1/t)} = x^{-\gamma}, \quad x < 0 \tag{7.45}$$

gilt. Die Normierungskonstanten sind dann durch

$$c_n = \sup\left\{x | F(x) \leq \frac{1}{n}\right\} - \alpha_F, \quad d_n = \alpha_F \tag{7.46}$$

gegeben.

7.2.4 Beispiele für die Verteilung von Minima

Beispiel 7.2.5 Es sei wieder eine Stichprobe von Messungen x_i, $1 \leq i \leq n$ gegeben, und die x_i seien i.i.d. $N(\mu, \sigma^2)$-verteilt. Gesucht ist die Verteilung von $X_- = \min(x_1, \cdots, x_n)$.

□

Beispiel 7.2.6 Die $x_1, \cdots, x_n$ seien nun i.i.d. exponentialverteilt mit dem Parameter λ. Gesucht ist die Grenzverteilung $H(x)$ für das Minimum dieser Werte.

Es ist $\alpha_F = 0 > -\infty$, also kann die Grenzverteilung nicht vom Typ II sein. Die Bedingung für die Weibull-Verteilung (Typ III) ist leichter zu überprüfen als die für Typ I, also wird sie zuerst überprüft. Es ist, für $x < 0$,

$$\begin{aligned}
\lim_{t\to-\infty} \frac{F(\alpha_F - 1/tx)}{F(\alpha_F - 1/t)} &= \lim_{t\to-\infty} \frac{F(-1/tx)}{F(-1/t)} \\
&= \lim_{t\to-\infty} \frac{1 - \exp(-\lambda/tx)}{1 - \exp(-\lambda/t)} \\
&= \lim_{t\to-\infty} \frac{1 - 1 + \lambda/tx}{1 - 1 + \lambda/t} \\
&= \lim_{t\to-\infty} \frac{\lambda t}{\lambda t x} = x^{-1}
\end{aligned}$$

Dies ist aber die Charakterisierung für die Grenzverteilung Typ III mit $\gamma = 1$, d.h. die Grenzverteilung der Minima exponentialverteilter Variablen ist selbst eine Exponentialverteilung. Dieses Ergebnis kann man natürlich direkt aus Satz 7.2.5 ablesen: die Grenzverteilungen sind ja minimum-stabil, d.h. sie sind ihre eigenen Grenzverteilungen, und die Exponentialverteilung ist ein Spezialfall der Weibull-Verteilung mit $\gamma = 1$. Die Normierungskonstante c_n ergibt sich dann nach (7.46) gemäß

$$1 - \exp(-\lambda x) = \frac{1}{n}, \quad x = c_n$$

d.h.

$$c_n = \frac{1}{\lambda} \log\left(1 - \frac{1}{n}\right)$$

In Abbildung 7.5 wird die Verteilung der Minima exponentialverteilter Variablen illustriert; es ist deutlich, dass es sich dabei wieder um Exponentialverteilungen handelt. □

Beispiel 7.2.7 Die Weibull-Verteilung ist bei Anwendungen äußerst beliebt; als motivierendes Paradigma werden dabei oft Ankerketten genannt. Eine solche Kette besteht aus, sagen wir, n Gliedern. Die Stärke der Glieder ist aber, wegen der unvermeidlichen Zufälligkeiten bei der Fertigung, eine zufällige Größe, und man kann annehmen, dass die Stärken der Glieder identisch und unabhängig verteilt sind. Die Stärke muß größer als $-\infty$ sein, also kann die Verteilung der Stärke der Kette nicht vom Typ I sein. Die Stärke der Kette ist gleich der Stärke des schwächsten Gliedes. Dies heißt aber, dass die Verteilung der Stärke der Kette gleich der Verteilung von $X_- = \min(x_1, \cdots, x_n)$ ist, wobei x_i, $1 \leq i \leq n$, die Stärke des i-ten Gliedes ist. Aber die Verteilung von X_- ist in diesem Fall notwendig vom gleichen Typ wie die Verteilung der x_i, und da Stärke nicht negativ sein kann und $\alpha_F > -\infty$ und $X \geq 0$ nur von der Weibull-Verteilung erfüllt sind, muß die Verteilung der Stärke von Anker- (und anderer) Ketten die Weibull-Verteilung sein.

Der Verteilungstyp ist hier durch ein logisches Argument gefunden worden; spezielle mathematische Annahmen wurden nicht gemacht. Die Wahl spezieller Verteilungsfunktionen ist also keine mehr oder weniger plausible Beliebigkeit in einem mathematischen Modell. Es muß aber jedesmal sorgfältig überlegt werden, ob das Ankerkettenparadigma auch tatsächlich auf andere Situationen übertragen werden kann. So nimmt man z.B. in der Psychophysik oft an, dass Neurone (oder allgemeiner: "Kanäle") einen Stimulus unabhängig voneinander entdecken. Das System "bricht" (= entdeckt, d.h. bietet dem Stimulus keinen Widerstand mehr an), wenn mindestens ein Kanal den Stimulus entdeckt, wenn also der empfindlichste (der "schwächste") Kanal den Stimulus meldet. *Also* muß die Wahrscheinlichkeit des Entdeckens durch die Weibull-Verteilung gegeben sein, - oder vielleicht doch nicht? □

Beispiel 7.2.8 Es werde noch die Grenzverteilung des Minimums von unabhängig und identisch Cauchy-verteilten Variablen betrachtet. Sie ergibt sich wegen der Symmetrie der Cauchy-Verteilung sofort aus der für das Maximum, denn es ist ja

$$P(X_- < -x) = P(X_+ > x), \quad x > 0$$

so dass

$$H(x) = 1 - \exp\left(\frac{1}{x}\right) \tag{7.47}$$

mit

$$c_n = a_n = \tan^{-1}\left(\frac{\pi}{2} - \frac{\pi}{n}\right) \tag{7.48}$$

□

7.3 Unabhängige, nichtidentisch verteilte Variablen

Man kann noch den Fall betrachten, dass die x_i, $1 \leq i \leq n$ zwar alle den gleichen Verteilungstyp haben, die jeweiligen Verteilungen aber nicht die gleichen Parameter haben. So kann man den Fall haben, dass die x_i zwar alle normalverteilt sind, sich aber hinsichtlich ihres Erwartungswertes μ_i unterscheiden; es wird wieder die Unabhängigkeit der x_i angenommen. Dieser Fall läßt sich auf den i.i.d.-Fall zurückführen. So ist

$$P(X_+ \leq x) = \prod_{i=1}^{n} F_i(a_{ni}x + b_{ni}), \quad F_i(x) = P\left(Z \leq \frac{x - \mu_i}{\sigma}\right) \tag{7.49}$$

und es ist $Z \sim N(0,1)$, $X_i \sim N(\mu_i, \sigma)$. Da die Normierungskonstanten von den Parametern der Verteilung F_i abhängen können, müssen sie geeignet indiziert werden. Es ist

$$\prod_{i=1}^{n} F_i(a_{ni}x + b_{ni}) = \exp\left(\sum_{i=1}^{n} \log F_i\right)$$

und da $\log F_i \to 1$, damit $P(X_+ \leq x)$ nicht degeneriert, ergibt sich wegen $\log F_i \approx -(1 - F_i)$

$$\lim_{n\to\infty} P(X_+ \leq x) = \lim_{n\to\infty} \exp\left(-\sum_{i=1}^{n}(1 - F_i(a_{ni}x + b_{ni})\right). \tag{7.50}$$

Nun strebt, nach (7.14), $n(1 - F(a_n x + b_n))$ gegen einen Grenzwert $-u(x)$. Hier schreibt man dementsprechend

$$\lim_{n\to\infty} n(1 - F(a_{ni}x + b_{ni})) = -u_i(x), \quad G_i(x) = e^{-u_i(x)} \tag{7.51}$$

und wegen

$$1 - F_i(a_{ni}x + b_{ni}) = n(1 - F_i(a_{ni}x + b_{ni}))\frac{1}{n}$$

erhält man für die Summe in (7.50)

$$\sum_{i=1}^{n}(1 - F_i(a_{ni}x + b_{ni}) = \sum_{i=1}^{n} n(1 - F_i(a_{ni}x + b_{ni})\frac{1}{n}$$

Die Summe auf der rechten Seite kann dabei unter Umständen durch ein Integral ersetzt werden, dessen genaue Form vom Typ der Grenzverteilung einerseits, und von den speziellen Normierungskonstanten sowie weiteren Annahmen über die Größen, die die zufälligen Veränderlichen X_i repräsentieren, andererseits abhängt. Allgemein gilt jedenfalls für die Approximation von $\prod_i F_i(x)$:

$$\prod_{i=1}^{n} F_i(x) \approx \exp\left(-\sum_{i=1}^{n} u_i(a_{ni}x + b_{ni})\frac{1}{n}\right) \tag{7.52}$$

wobei $u_i(a_{ni}x + b_{ni})$ die Form der Grenzverteilung für F_i ist.

Kapitel 8

Elemente der Theorie stochastischer Prozesse

8.1 Allgemeine Begriffe

Die meisten der in der Psychologie betrachteten Merkmale von Personen sind nicht konstant in der Zeit. So gibt es systematische Schwankungen der Stimmung, aber auch von spezifischeren Merkmalen; die Periodizitäten der Schwankungen können von minütlich, stündlich, täglich circadiane Rythmen über monatliche bis zu Jahresrythmen gehen (Sinz, ...). Ihnen überlagert sind im allgemeinen noch zufällige Schwankungen. Misst man nun das Ausmaß etwa an ängstlichkeit eimal vor und einmal nach einer Therapie, so ist klar, dass die Variabilität dieser Messungen groß sein kann im Vergleich zu solchen, bei denen die systematische zeitliche Variation explizit erfaßt wird. Man wird also eine zufällige Veränderliche X als von der Zeit abhängig betrachten müssen. Dies führt zum Begriff des *stochastischen Prozesses*. Hierbei betrachtet man nicht nur eine zufällige Veränderliche X, sondern eine *Familie* X_t von zufälligen Variablen. Die zufälligen Veränderlichen dieser Familie können sich alle z.B. hinsichtlich ihres Erwartungswertes μ und ihrer Varianz σ^2 unterscheiden; μ und σ^2 können etwa Funktionen der Zeit sein, so dass man $\mu(t)$ und $\sigma^2(t)$ schreiben kann. Man ist offenbar gut beraten, z.B. die Wirkung irgendwelcher experimenteller Variablen oder "Behandlungen" nicht nur durch einzelne zufällige Veränderliche, sondern durch Familien X_t von zufälligen Veränderlichen zu repräsentieren; einige Details solcher Repräsentationen sollen im Folgenden und in späteren Kapiteln gegeben werden.

Es ist im Vorangegangenen implizit angenommen worden, dass die betrachteten Merkmale kontinuierlich in der Zeit variieren. Es gibt aber Situationen, in denen eine andere Beschreibung naheliegt. So wird man bei vielen Betrachtungen dazu geführt, das Auftreten bestimmter Ereignisse in der Zeit oder im Raum zu zählen. So kann man bei Untersuchungen zur Struktur des Gedächtnisses die As-

soziationen einer Versuchsperson (Vp) zu einem Oberbegriff notieren und dabei außählen, wie sich die Häufigkeiten von Assoziationen in aufeinanderfolgenden Zeitabschnitten verändern. Ebenso kann man die Zeiten zwischen den einzelnen Assoziationen bestimmen. Die äußerung einer Assoziation kann als "Ereignis" bezeichnet werden, und wir betrachten also das Auftreten von Ereignissen in der Zeit. Die Ereignisse treten nicht kontinuierlich, sondern zu diskreten Zeitpunkten auf, und die Folge der Ereignisse kann wieder als ein stochastischer Prozess (die äußerung einer Assoziation kann nicht exakt für einen bestimmten Zeitpunkt vorausgesagt werden, ebensowenig die Anzahl solcher Ereignisse in einem Zeitabschnitt) aufgefaßt werden, aber jetzt eben in diskreter Zeit. Weitere Beispiele für stochastische Prozesse in diskreter Zeit erhält man bei der Betrachtung von Unfällen innerhalb bestimmter Zeitabschnitte, bei der Betrachtung sozialer Akte psychiatrischer Patienten innerhalb einer Woche, etc.

Der Begriff des stochastischen Prozesses soll nun genauer definiert werden, damit die Anwendungen auf psychologische Fragestellungen weitgehender illustriert werden können.

Es werde zunächst an den Begriff der zufälligen Veränderlichen erinnert. Gegeben sei ein Wahrscheinlichkeitsraum (Ω, A, P), wobei Ω ein Stichprobenraum, d.h. eine Menge von Elementarereignissen ist, A ist eine σ-Algebra und P ist ein Wahrscheinlichkeitsmaß. Dann heißt bekanntlich die Abbildung $X : \Omega \to \mathbb{R}$ eine (reelle) zufällige Veränderliche, wenn $P(X(\omega) \in B) = P(\omega \in X^{-1}(B))$ mit $X^{-1}(B) \in A$, $B \subseteq \mathbb{R}$. Die Verteilungsfunktion von X ist durch

$$F(x) = P(X \leq x) = P(X \in (-\infty, x]) \tag{8.1}$$

gegeben. Die Dichte von X ist durch

$$f(x) = \frac{dF(x)}{dx} \tag{8.2}$$

gegeben.

Definition 8.1.1 *Mit $X_t = \{X(t),\ t \in I\}$ werde eine* Familie von zufälligen Veränderlichen *bezeichnet, wobei I ein Intervall aus $\mathbb{R}$ ist. Die Familie X_t heißt* stochastischer Prozeß*, wenn die $X(t) \in X_t$ einen gemeinsamen Wahrscheinlichkeitsraum (Ω, A, P) haben. Für jedes $\omega \in \Omega$ heißt die Abbildung $t \to X(t, \omega)$ eine* Realisierung *oder* Trajektorie *von X_t. Ist die Menge $I \subseteq \mathbb{R}$ abzählbar, so heißt X_t* stochastischer Prozeß mit diskreter Zeit*, ist $I \subseteq \mathbb{R}$ überabzählbar, so heißt X_t* stochastischer Prozeß mit stetiger Zeit*.*

Ist $X(t, \omega) \in \mathbb{R}^d$, $d \in \mathbb{N}$, so heißt X_t ein d-dimensionaler Prozeß. $X(t, \omega)$ repräsentiert dann einen Vektor zufälliger Funktionen.

Anmerkung: d-dimensionale Prozesse werden insbesondere dann betrachtet, wenn stochastische dynamische dynamische Systeme diskutiert werden. Für das

Folgende ist es hinreichend, $d = 1$ anzunehmen; alle Begriffsbildungen übertragen sich auf den Fall $d > 1$.

Für festes $\omega \in \Omega$ ist $X(t, \omega)$ eine Funktion der Zeit. Diese Funktion ist zufällig insofern, als sie ebenfalls von dem eben zufällig auftretenden ω abhängt. Diese Abhängigkeit von ω bedeutet noch nicht, dass $X(t, \omega)$ irregulär aussieht, sie kann auch eine Gerade über einem bestimmten Intervall sein. Die Abhängigkeit von ω kann sich *u.* U. nur auf die Werte der Parameter der Geraden $X = at + b$, $t_1 \leq t \leq t_2$, auswirken, d.h. etwa $a = a(\omega)$, $b = b(\omega)$. Im allgemeinen wird man aber nicht nur die Menge der Geraden zulassen, sondern *alle* Funktionen über einem Intervall $[t_1, t_2]$, und damit eben auch die "zerknitterten". Man kann auch den Wert von t festhalten; dann ist $X(t, \omega)$ eine von ω abhängige zufällige Veränderliche. Welche Interpretation man für $X(t, \omega)$ wählt - als Funktion der Zeit für zufällig gewähltes ω oder X als von ω anhängige zufällige Veränderliche für gegebenes t - hängt im Prinzip nur davon ab, welche der beiden Auffassungen die gerade vorliegende statistische Fragestellung leichter handhabbar macht.

Ist ein expliziter Bezug auf ω nicht notwendig, wird einfach $X(t)$ statt $X(t, \omega)$ geschrieben.

Beispiel 8.1.1 Im Rahmen hirnphysiologischer Untersuchungen werden unter bestimmten Bedingungen EEG-Verläufe erhoben. Misst man an einer bestimmten Schädelposition unter bestimmten Bedingungen, so werden die Potentialverläufe bei wiederholten Messungen unter identischen Bedingungen - kontinuierliche Messungen während eines Zeitraumes Δt - nicht identisch sein, sondern sich durch zufällige Fluktuationen unterscheiden. Der Verlauf des Potentials während des Intervalls Δ kann als Trajektorie eines stochastischen Prozesses betrachtet werden. Diese Trajektorie ist eine von vielen möglichen, sie ist eine Realisierung ω aus der Menge Ω der überhaupt möglichen Trajektorien.

Nehmen wir für den Moment an, dass das Potential X nur zu den Zeitpunkten $t_1 = t, t_2, t_3, \cdots, t_n = t + \Delta t$ gemessen werden kann. Die Messungen liefern dann Realisierungen der zufälligen Veränderlichen $X_1 = X(t_1)$, $X_2 = X(t_2)$, $\cdots$, $X_n = X(t_n)$; es sind dies die Werte der Trajektorie $X(t, \omega)$, die während eines Meßabschnittes der Länge Δt auftritt. □

Es sei $I_n = \{t_1, \cdots, t_n\}$ eine endliche Menge von Werten aus I und es sei $P(I_n)$ die gemeinsame Verteilung der zu den t_i, $i = 1, \cdots, n$ korrespondierenden zufälligen Veränderlichen $X_1, \cdots, X_n$. Weiter sei $\{I_n\}$ die Menge der möglichen Mengen von n (Zeit-)Punkten aus I. Dann heißt $\{P(I_n)\}$ die *Familie der endlich-dimensionalen Verteilungen* des stochastischen Prozesses X_t. Viele den stochastischen Prozess beschreibenden statistischen Größen lassen sich über diese Familie bestimmen. Die wichtigsten dieser statistischen Größen werden in der folgenden Definition eingeführt:

Definition 8.1.2 *Für festes $t \in I$ ist $F(x; t) = P(X(t) \leq x)$ die* Verteilungsfunktion von $X(t)$, *und $f(x; t) = dF(x; t)/dx$ die dazu korrespondierende Dichte-*

funktion. Dann ist

$$\mu(t) = E(X(t)) = \int_{-\infty}^{\infty} x f(x;t) dx \tag{8.3}$$

die Mittelwertsfunktion *von* X_t*, und*

$$R(t_1, t_2) = E[X(t_1), X(t_2)] = \int_{-\infty}^{\infty} x_1 x_2 f(x_1, x_2; t_1, t_2) dx_1 dx_2 \tag{8.4}$$

die Autokorrelationsfunktion *von* X_t*; dabei ist* $f(x_1, x_2; t_1, t_2)$ *die gemeinsame Dichte von* X_1*,* X_2*. Die Größe*

$$Kov(t_1,\, t_2) = E[(X(t_1) - \mu(t_1))(X(t_2) - \mu(t_2))] = R(t_1, t_2) - \mu(t_1)\mu(t_2) \tag{8.5}$$

heißt Kovarianzfunktion*. Dementsprechend ist die* Varianz *von* X_t *durch*

$$\sigma^2(t) = Kov(t, t) = R(t, t) - \mu^2(t) \tag{8.6}$$

erklärt.

Für $t_1 = t_2$ nimmt $R(t_1, t_2)$ ein Maximum an. Um dies zu sehen, betrachte man

$$E[(X(t_1) - X(t_2))^2] = E[X^2(t_1)] + E[X^2(t_2)] - 2E[X(t_1)X(t_2)] \geq 0$$

Nun ist aber $E[X^2(t_1)] = R(t_1, t_1)$, $E[X^2(t_2)] = R(t_2, t_2)$, $E[X(t_1), X(t_2)] = R(t_1 t_2)$, und es folgt

$$R(t_1, t_1) + R(t_2, t_2) \geq 2R(t_1, t_2) \tag{8.7}$$

für alle $t_1 \neq t_2$. Für $t_2 \to t_1$ bzw. $t_2 \to t_1$ folgt dann die Gleichheit der beiden Ausdrücke, und dies heißt ja, dass $R(t_1, t_2)$ nicht größer als $R(t_1, t_1)$ oder $R(t_2, t_2)$ werden kann.

Definition 8.1.3 *Es seien* X_t *und* Y_t *zwei stochastische Prozesse; dann heißt*

$$R_{xy}(t_1, t_2) = E[X(t_1)Y(t_2)] \tag{8.8}$$

Die Kreuzkorrelation *der Prozesse* X_t *und* Y_t*.*

Definition 8.1.4 *Es sei* $X_t = \{X(t),\, t \in (-\infty, \infty)\}$ *ein stochastischer Prozess.* X_t *heißt* stationär*, wenn für jede endliche Menge* $I_n = \{t_1, \cdots, t_n\}$ *die gemeinsame Verteilung von* $\{X_{t_1+t_0}, \cdots, X_{t_n+t_0}\}$ *von* t_0 *unabhängig ist.* X_t *heißt* stationär im weiten Sinne*, wenn gilt: (i)* $E(X^2(t)) > \infty$*, (ii)* $\mu = E(X(t)) =$ *konstant, (iii)* $Kov(s, t)$ *hängt nur von der Differenz* $t - s$ *ab.*

Für die Autokorrelation $R(t_1,\, t_2)$ bedeutet Stationärität und Stationärität im weiten Sinne, dass $R(t_1,\, t_2) = R(t_2 - t_1) = R(\tau)$ gilt, d.h. R hängt nur von der Differenz $\tau = t_2 - t_1$, nicht aber von den Werten t_1 und t_2 ab. Die Beziehung (8.7) läßt sich dann zu

$$R(0) = \max_{\tau} R(\tau) \tag{8.9}$$

spezifizieren.

Definition 8.1.5 *Der stochastische Prozess $X_t = \{X(t), t \in I\}$ sei stationär. Lassen sich alle Statistiken des Prozesses (Mittelwertsfunktion, Autokorrelationsfunktion, etc.) bereits aus einer Trajektorie $X(t, \omega)$ berechnen, so heißt der Prozeß* ergodisch.

Die Ergodizität eines Prozesses ist für empirische Untersuchungen von großer Bedeutung, da man eben im allgemeinen *nur eine* Trajektorie des Prozesses beobachten und deshalb nur sie zur Berechnung von Statistiken heranziehen kann. Dieser Punkt wird insbesondere bei der Betrachtung von Zeitreihen wieder aufgenommen werden.

Die Eigenschaften eines stochastischen Prozesses lassen sich weiter durch Bezug auf den Energiebegriff beschreiben. Um diesen Zusammenhang zu verdeutlichen, sei daran erinnert, dass viele Funktionen sich als überlagerung von Sinus-Funktionen dargestellt werden können; dazu muß nur die Amplitude $|F(\nu)|$ und die Phase $\varphi(\nu)$ für jede Frequenz ν geeignet gewählt werden; wie im Anhang hergeleitet wird, können diese beiden Größen durch Berechnung der Fourier-Transformierten $F(\nu)$ bestimmt werden. Wir nehmen an, dass die Fourier-Transformierte auch für jede Trajektorie $X(t)$ des stochastischen Prozesses X_t berechnet werden kann, dass also

$$F(\nu) = \int_{-\infty}^{\infty} X(t) e^{-i\nu t} dt, \tag{8.10}$$

$i = \sqrt{-1}$, existiert. $F(\nu)$ ist im allgemeinen eine komplexe Zahl, $F(\nu) = F_1(\nu) + iF_2(\nu)$, wobei F_1 der Real- und F_2 der Imaginärteil von F ist. F_1 und F_2 sind jeweils reelle Funktionen von ν. Die Amplitude für die Frequenz ν ergibt sich gemäß $|F(\nu)| = \sqrt{F_1^2(\nu) + F_2^2(\nu)}$, und $\varphi(\nu) = \tan^{-1}[F_2(\nu)/F_1(\nu)]$.

Weiter kann man jeder Trajektorie $X(t)$ entsprechend bestimmter Begriffsbildungen in der Physik eine "Energie" zuordnen, nämlich

$$E = \int_{-\infty}^{\infty} |X(t)|^2 dt, \tag{8.11}$$

Ist $F(\nu)$ die zu $X(t)$ gehörige Fourier-Transformierte, so gilt *Parsevals Theorem*, demzufolge

$$E = \int_{-\infty}^{\infty} |X(t)|^2 dt = \int_{-\infty}^{\infty} |F(\nu)|^2 d\nu. \tag{8.12}$$

Da $X(t) = X(t, \omega)$ eine zufällig gewählte Funktion ist, sind die Größen $F(\nu)$, $|F(\nu)|$, $\varphi(\nu)$ und E zufällige Veränderliche und haben deshalb möglicherweise einen Erwartungswert. Wir sind insbesondere an der durchschnittlichen Energie der Trajektorien und damit des Prozesses X_t interessiert. Es ergibt sich nun, dass sich die durchschnittliche Energie des Prozesses als Fourier-Transformierte der Autokorrelation $R(\tau)$ (bei einem stationären Prozeß) darstellen läßt. Dies sieht man wie folgt:

Es sei $\bar{F}$ die zu F konjugiert komplexe Zahl, d.h. es sei $\bar{F} = F_1(\nu) - iF_2(\nu)$; dann ist $|F(\nu)|^2 = F\bar{F}$. Ebenso kann man $X(t)$ vorübergehend als komplexe Funktion auffassen und $X\bar{X}$ für $|X(t)|^2$ schreiben. Nun ist $|F(\nu)|^2$ durch

$$|F(\nu)|^2 = F(\nu)\bar{F}(\nu) = \left|\int_{-\infty}^{\infty} X(t)e^{-i\nu t}dt\right|^2 = \int_{-\infty}^{\infty} X(u)e^{-i\nu u}du \int_{-\infty}^{\infty} \bar{X}(v)e^{i\nu v}\, dv$$

gegeben (dabei ist die Integrationsvariable t formal durch u und v ersetzt worden, um Verwechslungen zu vermeiden). Dann gilt aber

$$|F(\nu)|^2 = \int_{-\infty}^{\infty} X(u)e^{-i\nu u}du \int_{-\infty}^{\infty} X^*(v)e^{i\nu v}dv = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} X(u)X^*(v)e^{-i\nu(u-v)}\, dudv.$$

Da $F(\nu)$ eine zufällige Veränderliche ist, muß auch $|F(\nu)|^2$ eine zufällige Veränderliche sein, mithin kann man den Erwartungswert von $|F(\nu)|^2$ bilden, und da $e^{-i\nu(u-v)}$ keine zufällige Veränderliche ist, folgt

$$E(|F(\nu)|^2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} E(X(u)\bar{X}(v))e^{-i\nu(u-v)}\mathrm{dudv}. \tag{8.13}$$

Aber $E(X(u)\bar{X}(v)) = R(u-v)$, R die Autokorrelation eines stationären Prozesses (vergl. Definition 8.1.4). Diese Beziehung sollte gezeigt werden. Damit hat man

Definition 8.1.6 *Es sei X_t ein stationärer Prozeß und $F(\nu)$ seien die Fourier-Transformierten der Trajektorien $X(t)$ von X_t. Dann heißt*

$$S(\nu) = E(|F(\nu)|^2) = \int_{-\infty}^{\infty} R(t)e^{-i\nu t}dt \tag{8.14}$$

die Spektraldichtefunktion *des stochastischen Prozesses X_t.*

Die Spektraldichtefunktion ist also die Fourier-Transformierte der Autokorrelationsfunktion. Dann kann man $R(\tau)$ als inverse Transformation von $S(\nu)$ darstellen:

$$R(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} S(\nu)e^{i\nu t}\, d\nu. \tag{8.15}$$

Aus der Definition von R folgt aber, dass $R(0) = \sigma^2$ die Varianz des Prozesses X_t, und gemäß (8.15) ist dann

$$\sigma^2 = R(0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} S(\nu)d\nu; \tag{8.16}$$

σ^2 gleicht also der *durchschnittlichen Energie* des Prozesses (vergl (8.12) und (8.14).

Beispiel 8.1.2 Es sei X_t ein stationärer Prozeß mit der Autokorrelationsfunktion

$$R(\tau) = \begin{cases} \sigma^2(\tau) = \sigma^2, & \tau = 0 \\ 0, & \tau \neq 0 \end{cases} \tag{8.17}$$

Dies bedeutet, dass beliebig benachbarte Werte von $X(t)$ stochastisch unabhängig und damit unkorreliert sind. Man überlegt sich leicht, dass ein solcher Prozeß in der Natur kaum auftreten kann, denn jeder an die Gesetze der Physik gekoppelte Prozeß benötigt Zeit für eine Veränderung, und die Forderung, dass die Werte $X(t)$ und $X(t+\epsilon)$ für $\epsilon > 0$ *beliebig* klein unkorreliert sind für alle $t \in I$ bedeutet ja, dass die Trajektorien von X_t sich beliebig schnell verändern können. Soweit psychische Prozesse eine physikalische Basis haben gelten diese Bemerkungen auch für psychische Prozesse. Prozesse mit der Autokorrelation (8.17) heißen dann auch *generalisierte Prozesse.* Es zeigt sich aber, dass generalisierte Prozesse häufig gut als erste Annäherung an die tatsächlichen Prozesse gewählt werden können; ihre angenehmen mathematischen Eigenschaften bedeuten dabei eine wesentliche Vereinfachung des formalen Aufwandes bei der Diskussion der Daten. Dazu werde die zu (8.17) gehörende Spektraldichtefunktion betrachtet. Es ist

$$S(\nu) = \sigma^2 \int_{-\infty}^{\infty} R(t) e^{-i\nu t} dt = \sigma^2, \tag{8.18}$$

d.h. aber $S(\nu) = \sigma^2$ ist eine konstante für alle ν. Dies bedeutet, dass jede Frequenz ν gleich viel Energie zur durschnittlichen Energie beiträgt. Hat man eine Lichtquelle, in der jede Wellenlänge (Frequenz) mit gleicher Energie vorkommt, so erhält man bekanntlich weißes Licht. In Analogie hierzu spricht man dann auch von X_t als von einem *weißen Rauschen*; das Wort *Rauschen* erinnert dabei an das durch Zufallsprozesse im Radio oder Telefonhörer auftretende Rauschen, und es ist *weiß*, weil es wie das weiße Licht alle Frequenzen mit gleicher Energie enthält. Die Annahme weißen Rauschens ist überall dort angebracht, wo die Spektraldichte über einem hinreichend breiten Frequenzband angenähert konstant ist. □

Der in Beispiel 8.1.2 beschriebene Prozeß des Weißen Rauschens ist ein Spezialfall, der nicht für für alle Prozesse als angenäherte Beschreibung gewählt werden kann. Im allgemeinen sind benachbarte Werte $X(t)$ und $X(t+\Delta t)$ in einem von Δt abhängendem Ausmaß korreliert, d.h. die Werte $X(t)$ und $X(t+\Delta t)$ sind stochastisch abhängig. Man kann nun, zur weiteren Dikussion der Abhängigkeiten zwischen Werten von $X(t)$ zu verschiedenen Zeitpunkten $t_1, \cdots, t_n$, vom Begriff der bedingten Verteilung Gebrauch machen. Dazu faßt man $X_1 = X(t_1), X_2 = X(t_2), \cdots, X_n = X(t_n)$ als zufällige Veränderliche auf. Die Abhängigkeiten lassen sich dann über die bedingten Dichten

$$f(x_1, \cdots, x_k | x_{k+1}, \cdots, x_n) = \frac{f(x_1, \cdots, x_n)}{f(x_{n+1}, \cdots, x_n)} \tag{8.19}$$

behandeln.

Beim Weißen Rauschen sind die $x_n, \cdots, x_1$ *alle* stochastisch unabhängig voneinander, so lange nur die $t_1, \cdots, t_n$ verschieden sind. Das Weiße Rauschen fluktuiert also mit der höchstmöglichen Geschwindigkeit. Bei Prozessen mit Abhängigkeiten zwischen den x_i ist die Geschwindigkeit der Fluktuationen dann geringer. Es liegt nahe, ein Maß für diese Geschwindigkeit zu suchen; intuitiv ist klar, dass ein solches Maß mit der Autokorrelation zusammenhängen muß, da diese ja die Korrelationen der $X(t)$ für benachbarte Zeitpunkte angibt. Ein solches Maß kann wie folgt konstruiert werden.

Es seien $X(t)$ die Trajektorien des Prozesses X_t, und es werde angenommen, dass sie zumindest fast überall differenzierbar sind. Für eine spezielle Trajektorie sei dann

$$Y(t) = \frac{dX(t)}{dt} = X'(t); \tag{8.20}$$

$Y(t)$ ist wieder eine Funktion von t, und da $X(t)$ eine zufällige Funktion ist, muß auch $Y(t)$ eine zufällige Funktion sein. Ist also $X_t = \{X(t),\ t \in I\}$ ein stochastischer Prozeß mit differenzierbaren Trajektorien, so ist $Y_t = \{Y(t), t \in I\}$ ebenfalls ein stochastischer Prozeß, dessen Trajektorien eben die Differentiale der Trajektorien $X(t)$ sind.

Wir suchen nun die Autokorrelationsfunktion des Prozesses Y_t. die Autokorrelation von X_t ist $R(t_1, t_2) = E[X(t_1)X(t_2)]$. Die Kreuzkorrelation der Prozesse X_t und Y_t ist dann $R_{xy} = E[X(t_1)X'(t_2)]$. Nun ist

$$E\left[X(t_1)\frac{X(t_2+\epsilon) - X(t_2)}{\epsilon}\right] = \frac{R_{xx}(t_1, t_2+\epsilon) - R_{xx}(t_1, t_2)}{\epsilon},$$

und für $\epsilon \to 0$ wird hieraus

$$R_{xy}(t_1, t_2) = \frac{\partial R_{xx}(t_1, t_2)}{\partial t_2}. \tag{8.21}$$

Ebenso kann man

$$E\left[\frac{X(t_1+\epsilon) - X(t_1)}{\epsilon}X'(t)\right] = \frac{R_{xy}(t_1+\epsilon, t_2) - R_{xy}(t_1, t_2)}{\epsilon}$$

bilden, und für $\epsilon \to 0$ folgt hieraus

$$R_{yx}(t_1, t_2) = \frac{\partial R_{xy}(t_1, t_2)}{\partial t_1}. \tag{8.22}$$

Aus (8.21) und (8.22) folgt dann

$$R_{yy}(t_1, t_2) = \frac{\partial^2 R_{xx}(t_1, t_2)}{\partial t_1 \partial t_2}; \tag{8.23}$$

damit ist die Autokorrelation des Prozesses $Y_t = \{X'(t)|t \in I\}$ aus der Autokorrelation des Prozesses $X_t = \{X(t)|t \in I\}$ durch zweifache partielle Differentiation ableitbar. Insbesondere sei nun der Prozeß X_t stationär, dann gilt ja

$R(t_1, t_2) = R(\tau)$, $\tau = t_2 - t_1$. Dann folgt insbesondere

$$R_{yy}(\tau) = -\frac{d^2 R(\tau)}{d^2 \tau}. \tag{8.24}$$

Für $\tau = 0$ folgt dann

$$R_{yy}(0) = -\frac{d^2 R(0)}{d^2 \tau} = E[Y^2(t)] = E[(X'(t))^2], \tag{8.25}$$

d.h. $R_{yy}(0)$ ist die Varianz ${\sigma_y}^2$ des Prozesses $Y_t = \{X'(t)|t \in I\}$. Ist σ_y^2 groß, so heißt dies, dass die *Varianz* der Veränderungen $X'(t)$ der Trajektorien ds Prozesses X_t groß ist, dass die Trajektorien $X(t)$ des Prozesses X_t also stark fluktuieren bzw. dass die $X(t)$ keinen sehr regelmäßigen Verlauf haben. Die Spektraldichte von Y_t ist nun wieder durch die Fourier-Transformierte von R_{yy} gegeben, und die *durchschnittliche Energie des Prozesses* Y_t ist durch $R_{yy}(0)$ gegeben. Gemäßt (8.25) ist aber $R_{yy}(0)$ auch mit Eigenschaften des Prozesses X_t verbunden. Diese Verbindung soll noch etwas elaboriert werden.

Es sei X_t stationär; dann ist $R(t_1, t_2) = R(\tau)$ mit $\tau = t_2 - t_1$. R ist maximal für $\tau = 0$, und weiter gilt

$$R(-\tau) = R(\tau); \tag{8.26}$$

dies folgt leicht aus der Stationärität. Nun ist einerseits

$$S(\nu) = \int R(\tau) e^{-i\nu t}\, d\nu/2\pi,$$

so dass umgekehrt

$$R(\tau) = \int S(\nu) e^{i\nu\tau} d\nu/2\pi$$

geschrieben werden kann. Dann ist andererseits

$$R'(\tau) = \frac{dR(\tau)}{d\tau} = \int S(\nu) e^{i\nu\tau} d\nu/2\pi$$

und

$$R''(\tau) = \frac{d^2 R(\tau)}{d\tau^2} = \int (i\nu)^2 S(\nu) e^{i\nu\tau} d\tau/2\pi = -\int \nu^2 S(\nu) e^{i\nu\tau} d\tau/2\pi,$$

so dass

$$\left.\frac{d^2(R(\tau)}{d\tau^2}\right|_{\tau=0} = -\frac{1}{2\pi}\int_{-\infty}^{\infty} \nu^2 S(\nu) d\nu \tag{8.27}$$

folgt. Das Integral kann nun als zweites Moment der Spektraldichte $S(\nu)$ des Prozesses X_t aufgefaßt werden. Dies führt zu der

Definition 8.1.7 *Es sei X_t ein stochastischer Prozess mit differenzierbaren Trajektorien, und es sei $Y_t = \{X'(t)|X'(t) = dX(t)/dt,\ t \in I\}$. Dann heißt*

$$\lambda_2 := E[(X'(t))^2] = -\left.\frac{d^2 R_{xx}(\tau)}{d\tau^2}\right|_{\tau=0} = -R''(0) \tag{8.28}$$

das zweite Spektralmoment *des stochastischen Prozesses X_t.*

Natürlich ist $\lambda_2 = \sigma_y^2$. Notwendige Bedingung dafür, dass das zweite Spektralmoment λ_2 existiert, ist die Differenzierbarkeit von $R(\tau)$ an der Stelle $\tau = 0$. In Zusammenhang mit dynamischen Modellen werden allerdings oft spezielle Prozesse, die Diffusionsprozesse diskutiert, die *nirgends differenzierbare* Trajektorien haben. Man betrachtet hier einen anderen Begriff als das zweite Spektralmoment:

Definition 8.1.8 *Es sei X_t ein stationärer Prozeß mit nicht notwendig differenzierbaren Trajektorien und der Autokorrelationsfunktion $R(t)$. Dann heißt*

$$\tau_{cor} = \frac{1}{R(0)} \int_0^\infty R(\tau)\, d\tau \tag{8.29}$$

die Korrelationszeit *des Prozesses.*

Beispiel 8.1.3 Gegeben sei ein stationärer stochastischer Prozeß X_t mit der Autokorrelationsfunktion $R(\tau) = \exp(-\alpha\tau^2)$, $\alpha > 0$. Für große α geht $R(\tau)$ schnell gegen Null, d.h. die Trajektorien fluktuieren schnell; für kleines α fluktuieren sie langsam. Es ist $dR(\tau)/d\tau = -2\alpha\tau R(\tau)$, und $d^2R(\tau)/d\tau^2 = -2\alpha R(\tau) + 4\alpha^2\tau^2 R(\tau)$. Dann ist $d^2R(\tau)/d\tau^2|_{\tau=0} = -2\alpha R(0)$. Aber $R(0) = 1$, so dass das zweite Spektralmoment durch $\lambda = 2\alpha$ gegeben ist.

Nun sei X_t ein Prozeß mit nicht differenzierbaren Trajektorien und der Autokorrelationsfunktion $R(\tau) = \sigma^2 e^{-\alpha\tau}$, $\alpha > 0$. Die Korrelationszeit ist

$$\tau_{cor} = \frac{1}{R(0)} \int_0^\infty R(\tau) d\tau = \frac{1}{\sigma^2} \int_0^\infty e^{-\alpha\tau} d\tau = \frac{1}{\sigma^2}\frac{1}{\alpha}.$$

Wieder gilt, dass τ_{cor} um so kleiner ist, je größer α ist. Will man das zweite Spektralmoment λ berechnen, so muß die zweite Ableitung bestimmt werden. Es ist $dR(\tau)/d\tau = -\alpha R(\tau)$, $d^2R(\tau)/d\tau^2 = \alpha^2 R(\tau)$ und $d^2R(\tau)/d\tau^2|_{\tau=0} = \sigma^2\alpha^2 > 0$. Es ist aber

$$\lambda = -\left.\frac{d^2R(\tau)}{d\tau^2}\right|_{\tau=0},$$

so dass $\lambda < 0$ wäre. Offenbar macht die Berechnung des zweiten Spektralmoments hier keinen Sinn. □

In (8.8) war die bedingte Dichte

$$f(x_1, \cdots, x_k | x_{k+1}, \cdots, x_n) = \frac{f(x_1, \cdots, x_n)}{f(x_{k+1}, \cdots, x_n)}$$

eingeführt worden. Hieraus ergibt sich ein für die Anwendung der Theorie stochastischer Prozesse wichtiger Spezialfall, der durch eine eigene Definition charakterisiert werden soll:

Definition 8.1.9 *Es sei X_t ein stochastischer Prozeß und es gelte für jedes n und $t_1 < t_2 < \cdots < t_n$*

$$P(x(t_n) \leq x_n | x(t_{n-1}), \cdots, x(t_1)) = P(x(t_n) \leq x_n | x(t_{n-1})). \tag{8.30}$$

Dann heißt X_t Markov-Prozeß.

Insbesondere Markov-Prozesse lassen sich oft durch übergangswahrscheinlichkeiten charakterisieren. Es sei $p(x, t; x_0.t_0)$ die Wahrscheinlichkeit, dass der Prozeß zur Zeit t den Wert (Zustand) $x = x(t)$ annimmt, wenn er zur Zeit $t_0 \leq t$ den Wert (Zustand) $x_0 = x(t_0)$ angenommen hat. $p(x, t; x_0, t_0)$ heißt auch übergangswahrscheinlichkeit. $p(x, t; x_0, t_0)$ ist eine bedingte Dichtefunktion. Es gilt sicherlich

$$p(x, t; x_0, t_0) \to \delta(x - x_0) \text{ für } \ t \to t_0, \tag{8.31}$$

wobei δ die Dirac-Delta-Funktion ist, d.h. es ist $\delta(x - x_0) = 0$ für $x \neq x_0$ und $\delta(x - x_0) = \infty$ für $x = x_0$ unter der Nebenbedingung $\int \delta(x - x_0)dx = 1$. Da $p(x, t; x_0, t_0)$ eine Dichtefunktion ist, muß darüber hinaus

$$\int_{-\infty}^{\infty} p(x, t; x_0, t_0)\, dx = 1 \tag{8.32}$$

gelten.

Es sei $t_0 < t_1 < t$. Dann gilt die *Chapman-Kolmogorov-Gleichung*:

$$p(x, t; x_0, t_0) = \int_{-\infty}^{\infty} p(x, t; x_1, t_1) p(x_1, t_1; x_0, t_0)\, dx_1. \tag{8.33}$$

Die Gleichung ist eine direkte Konsequenz von (8.29). Denn sei $t_0 < t_1 < t_2$, und sei $p(x(t_2), x(t_1), x(t_0))$ die Wahrscheinlichkeit, dass der Prozeß zu den Zeiten t_0, t_1 und t_2 die Werte $x(t_0)$, $x(t_1)$ und $x(t_2)$ annimmt. Nach dem Satz über bedingte Wahrscheinlichkeiten gilt dann

$$p(x(t_2), x(t_1), x(t_0)) = p(x(t_2)|x(t_1), x(t_0)) p(x(t_1)|x(t_0)).$$

Aber die Markov-Eigenschaft bedeutet, dass $p(x(t_2)|x(t_1), x(t_0)) = p(x(t_2)|x(t_1))$ und mithin

$$p(x_2, t_2; x_1, t_1; x_0, t_0) = p(x_2, t_2; x_1, t_1) p(x_1, t_1; x_0, t_0).$$

Andererseits ist

$$p(x_2, t_2; x_0, t_0) = \int_{-\infty}^{\infty} p(x_2, t_2; x_1, t_1; x_0, t_0) dx_1;$$

setzt man den vorangehenden Ausdruck für den Integranden ein, so erhält man die Chapman-Kolmogorov-Gleichung.

Die Markov-Annahme unterliegt auch der Beschreibung dynamischer Systeme durch deterministische Differentialgleichungen. Ist nämlich $x(t) = (x_1(t), \cdots, x_n(t))^T$ der Zustandsvektor eines solchen Systems, so ist es im allgemeinen durch eine Differentialgleichung der Form $\dot{x} = f(x,t)$ beschreibbar, wobei f ein geeignet gewählter Vektor von Funktionen ist. Die Beschreibung durch eine solche Differentialgleichung besagt, dass die Veränderung $\dot{x}$ des Zustandsvektors durch den *momentanen* Wert $x(t)$ und den einer eventuell auf das System einwirkenden "Stör"funktion ist. Der Markov-Prozeß kann als Verallgemeinerung einer solchen Konzeption von Systemen für den Fall, dass zufällige Efekte berücksichtigt werden müssen, aufgefaßt werden. Die Annahme allerdings, dass Veränderungen momentan wirken und vergangene Zeitpunkte keine Rolle spielen sollen, kann genau genommen nur eine Approximation sein, da jede Veränderung Zeit benötigt. Die Markov-Annahme stellt deshalb eine Idealisierung dar, deren Vorteil eine erhebliche Erleichterung der mathematischen Analyse der jeweils interessierenden Prozesse ist. Der Nachteil, eben nur eine Approximation an die Wirklichkeit zu sein, spielt praktisch keine Rolle, wenn der Einfluß der Vergangenheit im Vergleich zu der interessierenden Zeitskala verschwindend klein ist.

8.2 Zählprozesse

In vielen empririschen Untersuchungen werden Folgen von Ereignissen betrachtet. So ist man in manchen neurophysiologischen Untersuchungen an der Folge der Aktionspotentiale (Spikes), die von einem Neuron erzeugt werden, interessiert. In Untersuchungen zur Gruppendynamik protokolliert man die Folge bestimmter sozialer Handlungen, die im Verlauf einer sozialen Interaktion vorkommen. In entwicklungspsychologischen Untersuchungen ist man an der Folge bestimmter Spielhandlungen unter gegebenen Bedingungen interessiert. Allgemein betrachtet man dann Folgen von Ereignissen, über deren Eintreten keine deterministischen Aussagen gemacht werden können und die in diesem Sinne zufällig sind. Eine bestimmte Ereignisfolge kann dann durch die Trajektorie $X(t)$ eines stochastischen Prozesses repräsentiert werden: $X(t)$ gibt die Anzahl der Ereignisse, die bis zu dem Zeitpunkt t eingetreten sind, an. Die $X(t)$ gehören dementsprechend zu einem *Zählprozeß*. Man kann die Ereignisse auch durch Punkte auf der Zeitachse repräsentieren; dementsprechend ist dann auch die Rede von einem *Punktprozeß*. Wir beginnen mit einer formalen Definition dieser Klasse von stochastischen Prozessen.

Definition 8.2.1 *Der stochastische Prozeß* $X_t = \{X(t),\ t \in I\}$ *heißt* Zählprozeß*, wenn für die Trajektorien die folgenden Aussagen gelten:*

(i) $X(t) \geq 0$

(ii) $X(t) \in \mathbb{N}$, *d.h.* $X(t)$ *hat stets den Wert einer natürlichen Zahl*

(iii) für $t_1 < t_2$ *ist* $X(t_1) \leq X(t_2)$

(iv) für $t_1 < t_2$ *ist* $X(t_2) - X(t_1)$ *die Anzahl der Ereignisse im Intervall* $(t_1, t_2]$.

Offenbar ist ein Zählprozeß ein stochastischer Prozeß *in stetiger Zeit*, denn die Ereignisse können zu einem beliebigen Zeitpunkt eintreten.

Die Definition des Zählpozesses enthält noch keine Aussagen über die Wahrscheinlichkeit, mit denen bestimmte Trajektorien auftreten können. Eine Einschränkung der Menge der möglichen Wahrscheinlichkeitsmaße ergibt sich, wenn bestimmte Forderungen bezüglich der Abhängigkeiten zwischen den Anzahlen der Ereignisse bzw. der Zeiten zwischen Ereignissen — dies sind die Zwischenzeiten (interarrival times) — gestellt werden. Solche Forderungen werden in der folgenden Definition formuliert:

Definition 8.2.2 *Es sei* X_t *ein Zählprozeß. Sind die Anzahlen von Ereignissen in beliebigen, nichtüberlappenden Zeitintervallen stochastisch unabhängig, so heißt* X_t *ein* Prozeß mit unabhängigen Zuwächsen. *Hängt die Verteilung der Anzahlen in einem Intervall* Δt *von der Länge des Intervalles ab, so heißt* X_t *ein* Prozeß mit stationären Zuwächsen.

Für einen Prozeß X_t mit stationären Zuwächsen ist also die Verteilung der Anzahl von Ereignissen in $(t_1, t_2]$ die gleiche wie in $(t_1 + t_0, t_2 + t_0], t_0 > 0$.

Zählprozesse lassen sich gelegentlich über die Verteilung der Zeiten zwischen den einzelnen Ereignissen, den sogenannten *Zwischenzeiten*, charakterisieren. Tritt ein Ereignis zur Zeit t_1, das nächste zur Zeit t_2 ein, so ist $T = t_2 - t_1$ die zugehörige Zwischenzeit. Im allgemeinen können die Zwischenzeiten T_1, $T_2, \cdots$ stochastisch abhängig oder stochastisch unabhängig sein, und sie können alle die gleiche Verteilung oder verschiedene Verteilungen haben. Die im folgenden definierte Klasse von Verteilungen ist für Anwendungen der Theorie der stochastischen Prozesse von großer Wichtigkeit.

Definition 8.2.3 *Es sei* X_t *ein Zählprozeß mit den Zwischenzeiten* $T_1, T_2, \cdots$ *Sind die* T_i *stochastisch unabhängig und identisch verteilt (i.i.d.), so heißt* X_t *auch* Erneuerungsprozeß *(*Renewal process*).*

Ereignisse werden deshalb auch *Erneuerungen* genannt. Diese Redeweise ergibt sich aus bestimmten Fragestellungen, aufgrund der die Theorie entwickelt wurde: man betrachtet Maschinen oder Organismen, bei denen nach zufällig verteilten Zeiten Teile defekt und erneuert werden.

Es seien $T_1, T_2, \cdots$ Zwischenzeiten bei einem Zählprozeß. Dann ist

$$S_n = \sum_{i=1}^{n} T_i, \quad n \geq 1, \quad S_0 = 0 \tag{8.34}$$

die Zeit bis zum ersten Eintreten des n-ten Ereignisses, und

$$X(t) = \max(n|S_n \leq t), \tag{8.35}$$

d.h. die Trajektorien von X_t sind für jedes t durch den maximalen Wert $n \in \mathbb{N}$ definiert, für den die Summe S_n der Zwischenzeiten kleiner oder gleich t ist; nach Verstreichen von T_i erhöht sich $X(t)$ um 1 an der Stelle $t = S_n$ und behält diesen Wert bei, bis T_{i+1} verstrichen ist.

8.2.1 Der homogene Poisson-Prozeß

Der Poisson-Prozeß ist ein stochastischer Prozeß, der als Ausgangspunkt für die Modellierungen vieler zufälliger Prozesse gewählt werden kann und der deshalb von zentraler Bedeutung für die Theorie der stochastischen Prozesse ist. Er reflektiert in einem bestimmten Sinn die "reine Zufälligkeit" des Auftretens zufälliger Ereignisse, wie bei der Diskussion der Eigenschaften der Zwischenzeiten deutlich werden wird.

Definition 8.2.4 *Es sei $X_t = \{X(t)|0 \leq t\}$ ein Zählprozeß. X_t heißt* homogener Poisson-Prozeß mit der Rate λ, $\lambda > \infty$, *wenn*

(i) $X(0) = 0$ *für alle* $X(t) \in X_t$,

(ii) X_t *hat unabhängige Zuwächse*

(iii) die Anzahl der Ereignisse in einem Intervall der Länge Δt ist Poisson-verteilt mit dem Parameter $\lambda\Delta t$, d.h.

$$P(X(t + \Delta t) - X(t) = k) = e^{-\lambda\Delta t}\frac{(\lambda\Delta t)^k}{k!}, \quad k = 0, 1, 2, \cdots.$$

Statt "homogener Poisson-Prozeß" wird häufig auch einfach "Poisson-Prozeß" gesagt; gemeint ist dann stets der homogene Prozeß, im Unterschied zum weiter unten definierten inhomogenen Poisson-Prozeß.

Aus (iii) folgt unmittelbar, dass $P[X(t + \Delta t) - X(t) = k]$ unabhängig von t ist, denn die rechte Seite, $\exp(-\lambda\Delta t)(\lambda\Delta t)^k/k!$ hängt ja nur von Δt ab. Also müssen die Zuwächse bei einem homogenen Poisson-Prozeß auch stationär sein. Durch eine analoge Betrachtung sieht man, dass der Poisson-Prozeß auch ein Markoff-Prozeß ist.

Verteilung der Zwischenzeiten: Es werde zunächst die Verteilung der Zwischenzeiten betrachtet. Die Zwischenzeiten bilden eine *Folge* $\{T_k,\ k \geq 1\}$ von zufälligen Veränderlichen. Um die Verteilung von T_k zu bestimmen, betrachtet man das zufällige Ereignis $\{T_1 > t\}$. Dieses tritt dann und nur dann ein, wenn kein Ereignis im Intervall $[0, t]$ eingetreten ist, denn T_1 ist ja die Zeit bis zum

ersten Auftreten eines Ereignisses. Also ist $P(T_1 > t) = P(X(t) = 0)$, und nach Definition 8.2.4, (iii) ist $P(X(t) = 0) = e^{-\lambda t}$. Damit gilt aber

$$P(T_1 \leq t) = 1 - e^{-\lambda t}, \tag{8.36}$$

d.h. T_1 *ist exponentialverteilt mit dem Erwartungswert* $E(T_1) = 1/\lambda$. Jetzt muß die Verteilung von T_2 bestimmt werden. Dazu betrachte man $P(T_2 > t|T_1 = s)$; es gilt

$$P(T_2 > t|T_1 = s) = P(\text{kein Ereignis in } (s, s+t]|T_1 = s)$$

$$P(\text{kein Ereignis in } (s, s+t]) = e^{-\lambda t}$$

denn nach Definition 8.2.4, (ii) sind die T_i stochastisch unabhängig, und nach (iii) muß $P(T_2 > t|T_1 = s) = e^{-\lambda t}$ gelten, da ja die Länge des Intervalles gleich t ist. Also ist T_2 ebenfalls exponentialverteilt mit dem Erwartungswert $1/\lambda$. Damit gilt

Satz 8.2.1 *Die Zwischenzeiten* $T_1, T_2, \cdots$ *eines Poisson-Prozesses sind unabhängig und identisch exponentialverteilt mit dem Erwartungswert* $E(T) = 1/\lambda$ *und der Varianz* $Var(T) = 1/\lambda^2$.

Der Poisson-Prozeß ist also sicherlich ein Erneuerungsprozeß. Die stochastische Unabhängigkeit bedeutet wie die Stationarität, dass der Prozeß nach jedem Ereignis wieder neu started. Die Eigenschaften der Exponentialverteilung sind bereits in Abschnitt 5.2 diskutiert worden; die mit dieser Verteilung assoziierte Hazardfunktion $\lambda = konstant$ bedeutet weiter, dass auch die Wartezeiten auf das jeweils nächste Ereignis bei jedem Zeitpunkt, bis zu dem das nächste Ereignis nocht nicht eingetreten ist, neu starten. Dieser Sachverhalt definiert die "reine Zufälligkeit", von der eingangs die Rede war (vergl. auch Abschn. 5.2.3, p. 152).

Es sei

$$S_n = \sum_{i=1}^{n} T_i$$

die Wartezeit bis zum n-ten Ereignis. S_n ist die Summe exponentialverteilter Variabler und ist mithin Gamma-verteilt (vergl. Abschnitt 5.3, Satz 5.3.1), p. 153; mit $t = S_n$ gilt

$$f(t) = \frac{\lambda^n}{\Gamma(n)} x^{n-1} \exp(-\lambda t), \qquad \text{mit } \Gamma(n) = \int_0^\infty \xi^{n-1} \exp(-\xi) d\xi.$$

Beispiel 8.2.1 (Spike-Folgen) Bei Untersuchungen zur Funktionsweise einzelner Neurone hat man die Zeiten zwischen den Spikes gemessen. Man stellt dann fest, dass sie in guter Näherung stochastisch unabhängig und exponentialverteilt sind. Dementsprechend kann die Folge von Spikes, die ein Neuron erzeugt, wiederum in erster Näherung, als Poisson-Prozeß betrachtet werden.

Exakt wird diese Art der Betrachtung aber nicht sein, denn dabei wird die Refraktärzeit des Neurons vernachlässigt. Außerdem sind die Zwischenzeiten nicht stets exponentialverteilt. Gelegentlich fand man, dass die Normal-(Gauß-)Verteilung eine bessere Beschreibung lieferte, oder die Gamma-Verteilung der Ordnung 2 (Summe zweier unabhängiger, exponentialverteilter Variablen). Auch multimodale Verteilungen sind beobachtet worden. Einen überblick über die Befunde findet man in Sampath und Srinivasan (1977) □

Beispiel 8.2.2 (Absolute Wahrnehmungsschwelle) In Untersuchungen zur Funktionsweise des menschlichen Auges kann man die Absolutschwelle bestimmen; dies ist die Lichtmenge, die man (in vollständig dunkel-adaptiertem Zustand) benötigt, um überhaupt eine Lichtwahrnehmung zu haben. Es scheint so zu sein, dass 2 bis 7 Lichtquanten in der Zeiteinheit (1 ms) auftreten müssen, damit es zur Lichtempfindung kommt. Die Abschätzung dieser Anzahl beruht auf der Tatsache, dass der Fluß von Lichtquanten als ein Poisson-Prozeß repräsentiert werden kann, d.h. dass die Anzahl der Lichtquanten in einem Zeitintervall Δt Poisson-verteilt ist. Die Lichtintensität wird dabei duch den Parameter λ (der hier nicht die Wellenlänge des Lichts darstellt!) repräsentiert. Weitere Details findet man in Baumgartner et al. (1978). □

Beispiel 8.2.3 In einer Fabrik arbeiten Personen an Maschinen, und gelegentlich kommt es zu Unfällen. Man kann annehmen, dass ein Unfall, den eine Person erleidet, die Wahrscheinlichkeit, dass eine andere Person einen Unfall hat, sich nicht ändert und dass die "Wartezeit" zwischen den Unfällen exponentialverteilt ist. Dies bedeutet z.B., dass die Wartezeit τ bis zu einem Unfall exponentialverteilt ist. Der Zeitpunkt eines Unfalls ist dann wieder der Anfang der exponentialverteilten Wartezeit bis zum nächsten Unfall. Dann ist die Anzahl $k = X(t)$, $t = \tau_1 + \cdots + \tau_k$, der Unfälle, die eine Person bis zum Zeitpunkt t erleidet, ein Poisson-Prozeß.

Die Annahmen wirken nicht sehr realistisch; *psychologisch* ist plausibel, dass das Erleiden eines Unfalls sich u.a. so auswirkt, dass die Person vorsichtiger, vielleicht auch unsicherer wird. Dieser Verhaltensänderung entspricht eine Veränderung der Parameter des Prozesses, und die Annahme der Unabhängigkeit aufeinanderfolgender Ereignisse (Unfälle) ist vermutlich verletzt. Trotzdem ist der Poisson-Prozeß häufig eine gute erste Näherung. □

Beim Poisson-Prozeß sind die Ereignisse *nicht* in der Zeit gleichverteilt. Es kann zu Häufungen von Ereignissen kommen; aus diesen Häufungen darf man nicht schließen, dass sich die Bedingungen, unter denen die Ereignisse beobachtet werden, geändert haben. Man verdeutlicht sich diesen Sachverhalt, indem man sich vorstellt, dass man Sandkörner auf eine Tischplatte rieseln läßt. Dabei wird es rein zufällig zu Häufungen von Sandkörnern an bestimmten, aber zufällig auftretenden Stellen kommen (man erinnere sich hier an das Beispiel **??**, p. **??**,

über Besetzungswahrscheinlichkeiten). Das gleiche gilt für die Rosinen in einem Kuchen: dort findet man ebenfalls gelegentlich Anhäufungen von Rosinen. Man spricht von *Poisson-Klumpungen* (Poisson-Clusterings).

Gelegentlich beobachtet man Prozesse, die sich als additive überlagerung von Prozessen auffassen lassen. Dies gilt z.B. für die Unfälle, die man insgesamt in eine Fabrik zählt. Ist die Anzahl der Unfälle, die ein einzelner Arbeiter erfährt, Poisson-verteilt, so ist die Gesamtzahl der Unfälle die Summe der einzelnen Unfallhäufigkeiten.

Es läßt sich nun zeigen, dass die Summe von Poisson-Prozessen wieder ein Poisson-Prozeß ist:

Satz 8.2.2 *Gegeben seien n stochastisch unabhängige Poisson-Prozesse mit den Parametern $\lambda_1, \lambda_2, \cdots, \lambda_n$, und es sei X_t die additive überlagerung der Prozesse. Dann ist X_t ein Poisson-Prozeß* mit dem Parameter $\lambda = \lambda_1 + \cdots + \lambda_n$.

Beweis: Es sei T_1 die Zeit des ersten Auftretens eines der beim Prozeß X_t betrachteten Ereignisse. Dann ist $T_1 = T_{11}$, d.h. das Ereignis ist aus dem ersten Prozeß, oder $T_1 = T_{12}$, d.h. es ist aus dem zweiten, etc. Dann ist $\{T_1 > t\}$ genau dann, wenn $\{T_{11} > t \cap T_{12} > t \cap \cdots \cap T_{1n} > t\}$. Wegen der stochastischen Unabhängigkeit der Prozesse sind die T_{1i} ebenfalls stochastisch unabhängig, mithin ist

$$P(\{T_{11} > t \cap T_{12} > t \cap \cdots \cap T_{1n} > t\}) = \prod_{i=1}^{n}(1 - P(T_{ti} \leq t)),$$

so dass

$$P(T_1 > t) = P(\{T_{11} > t \cap T_{12} > t \cap \cdots \cap T_{1n} > t\}) = \exp\left(-\sum_{i=1}^{n} \lambda_i t\right).$$

Dies heißt aber gerade, dass X_t ein Poisson-Prozeß mit dem Parameter

$$\lambda = \sum_{i=1}^{n} \lambda_i$$

□

Findet man also, dass ein Zählprozeß ein Poisson-Prozeß ist, so ist es vorstellbar, dass der Prozeß sich additiv aus stochastisch unabhängigen Poisson-Prozessen zusammensetzt; dieser Sachverhalt kann für inhaltliche Interpretationen interessant sein. So gibt es z.B. viele Ereignisse, die zu einem Unfall führen. Nehmen wir einmal an, E_i sei ein Ereignis der i-ten Klasse, die einen Unfall nach sich ziehen, und die E_i, $i = 1, \cdots, n$ folgen jeweils einem Poisson-Prozeß. Der stochastische Prozeß, der die Unfälle beschreibt, ist als überlagerung dieser Prozesse zu verstehen; die Unfälle können jeweils verschiedene Ursachen haben.

Dieses Argument läßt sich verallgemeinern. Die Folge der Ereignisse E_i muß nicht notwendig ein Poisson-Prozeß sein. Man kann auch annehmen, dass die E_i unabhängigen und identischen Erneuerungsprozessen allgemeinerer Art entsprechen: die Zwischenzeiten müssen also nicht exponentialverteilt sein. Man kann sogar die Forderung, dass die E_i durch Erneuerungsprozesse beschreibbar sind, fallenlassen, und statt dessen allgemeine *Punktprozesse* betrachten; dies sind Prozesse, bei denen Folgen von Ereignissen betrachtet werden, ohne dass die Zwischenzeiten identisch oder unabhängig verteilt sind. Basierend auf Arbeiten von Palm hat Khintchine (1960) gezeigt, dass die überlagerung einer großen Zahl stationärer Punktprozesse auf einen Poisson-Prozeß führt. Kann man also einen gegebenen stochastischen Prozeß durch einen Poisson-Prozeß beschreiben, so kann man aufgrund dieser Tatsache wegen der Vielzahl der Mechanismen, die auf einen Poisson-Prozeß führen, noch nicht auf einen speziellen Mechanismus, die Folge erzeugt, zurückschließen. Nimmt man, ohne weitere Spezifizierung der Mechanismen, die einem Prozeß zugrundeliegen, einen Poisson-Prozeß an, so kann diese Annahme durchaus von *deskriptivem* Wert sein, ohne gleichzeitig eine Charakterisierung des Mechanismus zu liefern.

Man findet aber ebenfalls Beispiele, wo ein homogener Poisson-Prozeß keine adäquate Beschreibung liefert. Die bereits erwähnten Spike-Folgen bei Neuronen, aber auch Folgen von Unfällen zeigen gelegentlich Abweichungen vom Poisson-Prozeß. Eine erste Verallgemeinerung ist der inhomogene Poisson-Prozeß.

8.2.2 Der inhomogene Poisson-Prozeß

Beim homogenen Poisson-Prozeß ist die Rate λ eine Konstante. Es liegt nahe, als erste Verallgemeinerung λ als Funktion der Zeit aufzufassen:

Definition 8.2.5 *Es sei X_t ein Zählprozeß. X_t heißt* nichtstationärer *oder* inhomogener *Poisson-Prozeß mit der* Intensitätsfunktion $\lambda(t)$, $t \geq 0$, *wenn die folgenden Bedingungen erfüllt sind:*

(i) $X(0) = 0$ *für alle* $X(t) \in X_t$

(ii) X_t *hat unabhängige Zuwächse*

(iii) $\varphi(h) := P(X(t+h) - X(t) \geq 2) = o(h)$

(iv) $P(X(t+h) - X(t) = 1) = \lambda(t)h + o(h)$.

Dabei bezeichnet $o(h)$ eine beliebige Funktion φ, die für $h \to 0$ schneller gegen 0 geht als h selbst, d.h. φ heißt $o(h)$, wenn $\lim_{h\to 0} \varphi(h)/h = 0$.

Es läßt sich nun zeigen, dass nun $P(X(t) \leq k)$ durch

$$P(X(t) \leq k) = \exp\left(-\int_0^t \lambda(\tau)d\tau\right) \frac{(\int_0^t \lambda(\tau)d\tau)^k}{k!} \tag{8.37}$$

gegeben ist. Die Mittelwertsfunktion ist dementsprechend

$$m(t) = \int_0^t \lambda(\tau) d\tau.; \tag{8.38}$$

Die Wahrscheinlichkeit (8.4) kann man sich leicht plausibel machen: dazu betrachtet man die Ereignisse als Zufallspunkte im Zeitintervall $(0, T)$. Die Wahrscheinlichkeit, dass ein Punkt von n in ein Intervall der Länge $\Delta t \in (0, T)$ fällt, sei $p = \Delta t/T$. Sind die Punkte alle stochastisch unabhängig, so ist die Wahrscheinlichkeit, dass k Punkte in das Intervall der Länge Δt fallen, durch die Binomialverteilung gegeben. Ist nun n groß und $\Delta t/T$ klein, so kann hierfür die Poisson-Verteilung angenommen werden; dies führt zum homogenen Poisson-Prozeß. Statt $p = (t_2 - t_1)/T$, $\Delta t = t_2 - t_1$, können wir nun fordern, dass eine "Dichte"funktion $\lambda(t)$ existiert derart, dass $p = \int_{t_1}^{t_2} \lambda(\tau) d\tau$ ist. Dann ergibt sich sofort (8.4).

Es soll noch die Autokorrelationsfunktion des Poisson-Prozesses angegeben werden. Es werde zunächst der homogene Prozeß betrachtet (vergl. Papoulis (1965), p. 285).

Es sei $t_2 > t_1$. Dann ist $X(t_2) - X(t_1)$ eine zufällige Veränderliche mit der Verteilung

$$P[X(t_2) - X(t_1) = k] = \exp(-\lambda(t_2 - t_1)) \frac{[\lambda(t_2 - t_1)]^k}{k!}$$

und dem Erwartungswert

$$E[X(t_2) - X(t_1)] = \lambda(t_2 - t_1) \tag{8.39}$$

und dem zweiten Moment

$$E[X(t_2) - X(t_1)]^2 = \lambda^2 (t_2 - t_1)^2 + \lambda(t_2 - t_1). \tag{8.40}$$

Es sei nun $t_1 > t_2 > t_3 > t_4$. Die zufälligen Veränderlichen $X(t_2) - X(t_1)$ und $X(t_4) - X(t_3)$ sind dann stochastisch unabhängig, da sie die Anzahl von Ereignissen in nichtüberlappenden Zeitintervallen repräsentieren. Deshalb ist der Erwartungswert ihres Produktes gleich dem Produkt ihrer Erwartungswerte, so dass

$$E[(X(t_2) - X(t_1))(X(t_4) - X(t_3)] = \lambda^2 (t_2 - t_1)(t_4 - t_3). \tag{8.41}$$

Nun gelte $t_2 > t_4 > t_1 > t_3$; dann überlappen sich die Intervalle und (8.8) gilt nicht. Es ist

$$\begin{aligned} X(t_2) - X(t_1) &= [X(t_2) - X(t_4)] + [X(t_4) - X(t_1)] \\ X(t_4) - X(t_3) &= [X(t_4) - X(t_1)] + [X(t_1) - X(t_3)] \end{aligned}$$

Setzt man dies in (8.6) und (8.7) ein, so erhält man

$$E[(X(t_2) - X(t_1))(X(t_4) - X(t_3)] = \lambda^2 (t_2 - t_1)(t_4 - t_3) + \lambda(t_4 - t_1). \tag{8.42}$$

Es folgt sofort

$$E(X(t)) = \lambda t \tag{8.43}$$

und

$$R(t_p, t_q) = \begin{array}{ll} \lambda t_q + \lambda^2 t_p t_q, & t_p > t_q \\ \lambda t_p + \lambda^2 t_p t_q, & t_p \leq t_q \end{array} \tag{8.44}$$

Dabei folgt (8.10) aus (8.6), und (8.11) folgt aus (8.9), wenn man $t_p = t_2$, $t_q = t_4$ sowie $t_1 = t_3 = 0$ setzt.

Das Resultat überträgt sich auf den inhomogenen Fall, wenn nur $\lambda(t_2 - t_1)$ durch $\int_{t_1}^{t_2} \lambda(t)dt$ ersetzt wird. Dann ist

$$E(X(t)) = \int_0^t \lambda(\tau)d\tau \tag{8.45}$$

und für $t_2 > t_1$

$$R(t_2, t_1) = \int_0^{t_1} \lambda(\tau)d\tau \left(1 + \int_0^{t_2} \lambda(\tau)d\tau\right). \tag{8.46}$$

Ist ein Poisson-Prozeß inhomogen, so sind die Zuwächse zwar unabhängig, aber nicht mehr stationär. Dies bedeutet, dass die Wahrscheinlichkeit von Ereignissen zu bestimmten Zeiten größer sein kann als zu anderen. Betrachtet man z.B. die Folge von Unfällen während einer Woche, so mag diese Folge durch einen Poisson-Prozeß beschreibbar sein. Montags und Freitags können aber u.U. Häufungen auftreten, die die Annahme eines homogenen Poisson-Prozeß ungerechtfertigt erscheinen lassen. Ein inhomogener Prozess ist möglicherweise die bessere Beschreibung. Während der Ski-Saison beobachtet man Häufungen von Unfällen insbesondere nachmittags *nach* von den Skifahrern eingelegten Pausen. Dies geht auf eine erhöhte Müdigkeit einerseits, andererseits auf die im Laufe des Tages zunehmende Neigung mancher Skifahrer, geistige Getränke zu sich zu nehmen, zurück. Ein inhomogener Poisson-Prozeß ist vermutlich eine bessere Beschreibung als ein homogener.

Die Definition 8.2.5 enthält explizit keine Annahmen über das Wahrscheinlichkeitsmaß des Prozesses; die Gleichung (8.37), p. 244, *folgt* aus den Annahmen. Die Annahmen haben den Vorteil, dass man gelegentlich aufgrund inhaltlicher Betrachtungen zu Aussagen der Form (i) bis (iv) kommen kann und dann *schließen* kann, dass der betrachtete Prozeß ein Poisson-Prozeß sein muß.

Beispiel 8.2.4 (Rekorde) Beobachtet werde eine Folge von zufälligen Veränderlichen $Y_1, Y_2, \cdots$. die identisch und unabhängig verteilt seien. Dabei sei $Y_i = Y(t_i)$, t_i Zeitpunkte. Die Y_i haben die Dichte f und die Verteilungsfunktion F, und damit die Hazardfunktion $\lambda(t) = f(t)/(1 - F(t))$.

Man kann nun den Begriff des *Rekords* einführen: Ein Rekord kommt zur Zeit $t = t_n$ vor genau dann, wenn $Y_n > \max(Y_1, \cdots, Y_{n-1})$ ist; $Y_0 = -\infty$. Y_n heißt dann

Rekordwert. Es sei $X(t)$ die Anzahl der Rekordwerte, die die bis zur Zeit t erreicht wurden. Dann ist $X_t = \{X(t),\ t \geq 0\}$ ein Zählprozeß. Das Ereignis, das hier gezählt wird, ist das Auftreten eines Wertes, der größer als alle vorangegangenen ist. Gesucht ist nun das Wahrscheinlichkeitsmaß für den Prozeß X_t.

Dazu betrachtet man das Intervall $(t,\ t+h)$. Ein Rekord tritt auf zwischen t und $t+h$ nur dann, wenn $X(t+h) > X(t)$ ist. Für die Hazardfunktion gilt dann also

$$P(t_n \in (t, t+h) | t_n > t) = \lambda(t)h + o(h).$$

Nach Definition 8.2.5, p. 244, (iv) charakterisiert dies aber den inhomogenen Poisson-Prozeß (und für $\lambda(t) = \lambda = \textit{konstant}$ den homogenen). Das Auftreten von Rekorden kann also allgemein durch einen inhomogenen Poisson-Prozeß beschrieben werden, wenn nur die stochastische Unabhängigkeit der Zwischenzeiten der ursprünglichen Ereignisse angenommen werden kann. Damit ist auch die Verteilung der Zeiten zwischen den Rekorden gegeben. □

Beispiel 8.2.5 Ehe- und andere Partner streiten gelegentlich miteinander. Es seien $Y_1, Y_2, \cdots$ die Zeiten zwischen den einzelnen Streits. Ein Rekord friedlichen Zusammenlebens tritt dann auf, wenn die letzte Zwischenzeit Y_n länger als alle anderen ist; $T_i = Y_n$, $n > i$ ist dann der i-te Rekord. Macht man die – vereinfachende – Annahme, dass die Y_k stochastisch unabhängig und identisch verteilt sind, so wird über die T_i ein vermutlich inhomogener Poisson-Prozeß erzeugt. Gelegentliche Häufungen der Rekorde, d.h. das gelegentliche Auftreten immer länger werdender Zwischenzeiten – muß also noch nicht notwendig als Verbesserung des ehelichen Klimas gewertet werden, denn der Zufall zieht bereits solche Häufungen nach sich. Genauere Aussagen lassen sich machen, wenn eine bestimmte Hazardfunktion für die Zwischenzeiten Y_j, d.h. die Zeiten zwischen den ehelichen Auseinandersetzungen, angenommen werden kann. Für $Y_j = y$ gelte insbesondere

$$\lambda(y) = \alpha\beta y^{\beta-1}, \quad \alpha > 0, \quad \beta > 0. \tag{8.47}$$

Dies ist die Hazardfunktion der Weibull-Verteilung (vergl. Definition 3.3.2, p. 84, Gleichung (3.30), p. 84), und Beispiel 7.2.7, p. 225, d.h. es soll gelten

$$P(Y_j \leq y) = 1 - \exp\left(-\alpha y^{\beta}\right). \tag{8.48}$$

Für $\beta > 1$ steigt $\lambda(y)$ mit wachsendem t. Die Zeit bis zum nächsten Rekord ist aber durch

$$P(X(t) \leq y) = \exp\left(-\int_0^y \lambda(\tau)d\tau\right)$$

gegeben; wird nun $\lambda(\tau)$ größer für alle τ, so wird $P(X(t) \leq y)$ kleiner, d.h. aber die Abstände zwischen den Rekorden werden kleiner. Dem entspricht, dass nach (8.16) $P(Y \leq y)$ für festes y eher bei 1 liegt, wenn β größer ist, als bei kleinerem Wert von β. Dieses wiederum bedeutet, dass die Ehestreits schneller aufeinander

folgen; kleinere Abweichungen der Zwischenzeiten nach oben bedeutet also auch schneller aufeinanderfolgende Rekorde. Für $\beta < 1$ kehrt sich die Betrachtung um, im Vergleich zum Fall $\beta > 1$ lassen Rekorde länger auf sich warten, und dafür sind die Zeiten zwischen den Krächen von vornherein größer.

Man muß hier bedenken, dass die Weibull-Verteilung *nicht* aus einen spezifischen Modell der Interaktion zwischen (Ehe-)Partnern gefolgert wurde, sondern dass sie einfach aufgrund ihres häufig guten "Fits" an entsprechende Daten gewählt worden ist. Damit ist gemeint, dass empirische Daten dieser Verteilung häufig gut entsprechen. In einem theoretisch anspruchsvolleren Modell wird man aber versuchen, die Hazardfunktion und damit die Verteilung der Zwischenzeiten aus den psychologischen Annahmen zu deduzieren. □

In bezug auf Unfälle lassen sich Betrachtungen anstellen, die zu denen in Beispiel 4.10.1, p. 145, analog sind. In der Tat findet man nun häufig, dass die Unfallzahlen selbst sich nicht gut durch einen homogenen oder allgemeiner inhomogenen Poisson-Prozeß beschreiben lassen. Betrachtet man z.B. Unfälle in einer Fabrik in einem Zeitabschnitt bestimmter Länge, so erhält man oft eine bessere Charakterisierung der Häufigkeitsverteilung, wenn man letztere als eine *Mischung* von Verteilungen, insbesondere von Poisson-Verteilungen, auffasst: für jede Person ist die Anzahl der Unfälle zwar Poisson-verteilt, aber die Personen haben verschiedene Parameter. Eine Diskussion solcher Verteilungen findet sich in Kendall und Stuart (1969), p. 129.

Zum Abschluß sollen einige aus Poisson-Prozessen abgeleitete Prozesse betrachtet werden.

8.2.3 Der zusammengesetzte Poisson-Prozeß; Shot-Noise

Vielfach kann der interessierende stochastische Prozeß als aus überlagerungen von Effekten resultierend betrachtet werden, die durch zufällig auftretende Impulse ausgelöst werden; man denke etwa an die Schocks, die während einer Autofahrt durch unvorhergesehene Verkehrssituationen bewirkt werden und die durch sie erzeugten Adrenalinausschüttungen. Die Folge der Schocks kann häufig als Poisson-Prozeß beschrieben werden. Wir benötigen zunächst die

Definition 8.2.6 *Es sei X_t ein Poisson-Prozeß und $\epsilon > 0$ sei eine reelle Zahl. Dann heißt der Prozeß $Y_t = \{Y(t), t \geq 0\}$ mit*

$$Y(t) = \frac{X(t+\epsilon) - X(t)}{\epsilon} \tag{8.49}$$

wobei $X(t)$ eine Trajektorie des Poisson-Prozesses ist, ein Poisson-Inkrement-Prozeß. *Weiter seien t_i die zufälligen Zeitpunkte, zu denen sich die Trajektorien $X(t)$ jeweils um 1 erhöhen; die t_i heißen* Zufallspunkte *(Papoulis, (1965)). Weiter*

werde der Prozeß $Z_t = \{Z(t),\, t \geq 0\}$ *dessen Trajektorien durch*

$$Z(t) = \frac{dX(t)}{dt} = \lim_{\epsilon \to 0} Y(t) \tag{8.50}$$

definiert werden, gebildet. Z_t *heißt dann* Poisson-Impuls-Prozeß.

Es sei $X(t+\epsilon) - X(t) = k$ die Anzahl der Ereignisse in $(t, t+\epsilon)$. Dann ist $Y(t) = k/\epsilon$ und es folgt

$$P(Y(t) = k/\epsilon) = e^{-\lambda\epsilon}\,\frac{(\lambda\epsilon)^k}{k!} \tag{8.51}$$

und für die Mittelwertsfunktion $E(Y(t))$ erhält man daraus sofort

$$E(Y(t)) = \frac{1}{\epsilon}E(X(t+\epsilon)) - \frac{1}{\epsilon}E(X(t)) = \lambda. \tag{8.52}$$

Die Autokorrelation von Y_t ergibt sich durch leichte Rechnung (vergl. Papoulis (1965), p. 287)

$$R(t_1, t_2) = \begin{cases} \lambda^2, & |t_1 - t_2| > \epsilon \\ \lambda^2 + \frac{\lambda}{\epsilon} - \frac{\lambda|t_1 - t_2|}{\epsilon^2}, & |t_1 - t_2| < \epsilon \end{cases} \tag{8.53}$$

Der Impulsprozeß kann auf die folgende Weise dargestellt werden. Es sei $\delta(t)$ die Dirac-Delta-Funktion, d.h. $\delta(t) = 0$ für $t \neq 0$ und $\delta(t) = \infty$ für $t = 0$, $\int \delta(t)dt = 1$. Dann läßt sich eine spezielle Trajektorie von Z_t durch die Summe

$$Z(t) = \sum_i \delta(t - t_i) \tag{8.54}$$

darstellen, wobei die t_i die zum Poisson-Prozeß gehörenden Zufallspunkte sind. Nach (8.22) ist Z_t ein Prozeß, dessen Trajektorien durch Folgen von Impulsen gegeben sind, die zu den zufälligen Zeiten t_i eintreten. Es gilt

$$E(Z(t)) = \lambda, \quad R(t_1, t_2) = \lambda^2 + \lambda\delta(t_1 - t_2). \tag{8.55}$$

Dies folgt aus (5.2.9) und (8.21), indem man ϵ gegen Null streben läßt.

Definition 8.2.7 *Es sei* $h(t)$ *eine beliebige reelle Funktion, die der Bedingung* $h(t) = 0$ *für* $t < 0$ *genügt. Der stochastische Prozeß* S_t *mit den Trajektorien*

$$s(t) = \sum_{i=0}^{n} h(t - t_i) \tag{8.56}$$

heißt Shot-Noise.

Shot-Noise wurde zuerst im Zusammenhang der Erklärung des durch thermische Bewegungen der Elektronen in Elektronenröhren erzeugten Rauschens beschrieben. Das Eintreffen eines Elektrons zur Zeit t_i (d.h. eines Impulses zur Zufallszeit t_i) erzeugt einen Effekt $h(t-t_i)$, der sich als "Rauschen" äußert. Daß der Shot-Noise-Prozeß auch außerhalb der Elektronik von Interesse ist, wird gleich verdeutlicht werden. Ist der zu den t_i korrespondierende Poisson-Prozeß homogen, so gelten für die Mittelwertsfunktion etc. die folgenden Aussagen (vergl. Papoulis (1965), p. 561)

$$E(s(t)) = \lambda \int_{-\infty}^{\infty} h(t)dt, \quad E(s^2(t)) = \lambda^2 \left[\int_{-\infty}^{\infty} h(t)dt\right]^2 + \lambda \int_{-\infty}^{\infty} h^2(t)\, dt \tag{8.57}$$

und

$$R(\tau) = \lambda^2 \left[\int_{-\infty}^{\infty} h(t)dt\right]^2 + \lambda \int_{-\infty}^{\infty} h(\alpha)h(t+\alpha)d\alpha. \tag{8.58}$$

Die Beziehungen in (8.25) sind auch als *Campbells Theorem* bekannt. Aus (8.26) ergibt sich das Power-Spektrum des Shot-Noise gemäß

$$S(\omega) = 2\pi\lambda^2 H(0)\delta(\omega) + \lambda|H(i\omega)|^2 \tag{8.59}$$

mit $i = \sqrt{-1}$, und H ist die Fourier-Transformierte von h. Hat der Poisson-Prozeß keine konstante Intensität λ, sondern ist er inhomogen, so folgt

$$E(s(t)) = \int_{-\infty}^{\infty} \lambda(\tau)h(t-\tau)d\tau \tag{8.60}$$

und die Autokovarianz ist durch

$$C(t_1, t_2) = \int_{-\infty}^{\infty} \lambda(\tau)h(t_1-\tau)h(t_2-\tau)d\tau \tag{8.61}$$

gegeben.

Definition 8.2.8 *Es sei X_t ein Poisson-Prozeß und es seien ξ_i unabhängige und identisch verteilte zufällige Veränderliche. Weiter sei $n_t = X(t)$, d.h. n_t sei die Anzahl der Ereignisse bis zur Zeit t. Dann ist $U_t = \{U(t)|t \geq 0\}$ mit*

$$U(t) = \sum_{i=1}^{n_t} \xi_i \tag{8.62}$$

ein zusammengesetzter Poisson-Prozeß.

Es läßt sich zeigen (Ross (1983), p. 39, dass die Mittelwertsfunktion eines zusammengesetzten stochastischen Prozesses durch

$$E(X(t)) = \lambda t E(Y) \tag{8.63}$$

und die Varianz durch

$$Var(X(t)) = \lambda t E(Y^2) \tag{8.64}$$

gegeben ist.

Beispiel 8.2.6 Ein Autofahrer fährt auf der Autobahn von A nach B. Während der Fahrt treten zufällig immer wieder unvorhergesehene Situationen auf, die eine Schockwirkung insofern haben, als sie im Autofahrer eine Adrenalinausschüttung bewirken. Es werde angenommen, dass die Schocksituationen als Poisson-Prozeß X_t auftreten, und ξ_i sei die beim i-ten Schock ausgeschüttete Adrenalinmenge. Dann ist

$$U(t) = \sum_{i=1}^{Y(t)} \xi_i$$

die Gesamtmenge von Adrenalin, die aufgrund der Schockwirkungen ausgeschüttet wurde.

$U(t)$ ist eine Trajektorie eines zusammengesetzten Poisson-Prozesses. Es sei noch einmal hervorgehoben, dass $U(t)$ die Gesamtmenge des *ausgeschütteten* Adrenalins ist, nicht aber die Gesamtmenge des zur Zeit t *im Körper befindlichen* Adrenalins. Um diese zu repräsentieren, muß man bedenken, dass das Adrenalin auch wieder abgebaut wird. Es sei t_i ein Zeitpunkt, zu dem ein Schock eintritt, der eine Adrenalinausschütung nach sich zieht. Der Verlauf des Adrenalinspiegels kann u.U. durch eine Funktion $h(t)$ (Impulsantwort) abgebildet werden. Man kann nun die Folge der Schocks durch

$$Z(t) = \sum_i a_i \delta(t - t_i) \tag{8.65}$$

repräsentieren. $a_i\delta(t-t_i)$ ist also ein Schock zur Zeit t_i mit der Intensität a_i. Den tatsächlichen Adrenalingehalt zur Zeit t kann man dann durch

$$S(t) = \sum a_i h(t - t_i) \tag{8.66}$$

darstellen. $S(t)$ ist Shot-Noise charakterisierbar. □

8.2.4 Geburts- und Todesprozesse

Reine Geburtsprozesse

Gegeben sei ein Zählprozeß $X_t = \{X(t), t \geq 0\}$, wobei $X(t)$ die Anzahl der Elemente in einer sich entwickelnden Population zur Zeit t angebe.

Es sollen stochastische Prozesse betrachtet werden, die die Entwicklung bzw. das Wachstum einer Population repräsentieren. Man mag an eine Baktierienkultur denken: nach einer zufällig verteilten Zeit teilt sich ein Bakterium, aus einem werden zwei, und so fort. Aber auch die Ausbreitung von Infektionen oder Gerüchten läßt sich als ein solcher *Wachstumsprozeß* betrachten. Ein besonders einfach strukturierter Prozeß ist der folgende:

Definition 8.2.9 *Es sei $X_t = \{X(t), t \geq 0\}$ ein Zählprozeß, wobei $X(t)$ den Umfang einer Population zur Zeit t repräsentiere; $X(t) = n$, $n \in \mathbb{N}$, bedeutet, dass*

es zur Zeit t gerade n Elemente in der Population gibt. Zu zufälligen Zeitpunkten t_i erhöhe sich $X(t)$ um 1, d.h. $X(t_i) = X(t)+1$, $t_{i-1} \leq t < t_i$. Die Zwischenzeiten $T_i = t_i - t_{i-1}$ seien exponentialverteilt mit dem Parameter λ_i. Dann heißt X_t ein reiner Geburtsprozeß.

Anmerkung: Für $\lambda_i = \lambda$ eine Konstante ist X_t ein Poisson-Prozeß. Der Poisson-Prozeß ist also ein Spezialfall eines reinen Geburtsprozesses.

Insbesondere kann jedes Element der Population stochastisch unabhängig von den anderen ein neues Element erzeugen:

Definition 8.2.10 *Es sei X_t ein reiner Geburtsprozeß, bei dem jedes Element der Population stochastisch unabhängig von den anderen ein neues Element erzeugen kann und wobei der Parameter für die Wartezeit bis zu einer solchen Erzeugung gleich λ ist für alle Elemente. Dann heißt X_t ein* Geburtsprozeß mit linearer Geburtsrate*, oder* Yule- *bzw.* Yule-Furry-Prozeß. *(Yule (1924); Furry (1937))*

Es sei X_t ein Yule-Prozeß und die Population habe den Umfang n erreicht. Wir betrachten die Wahrscheinlichkeit, dass in einem Zeitintervall der Länge h ein neues Element erzeugt wird. Für jedes Element ist die Wahrscheinlichkeit, dass ein neues Element innerhalb der Zeitspanne h gebildet wird, durch $1 - e^{-\lambda h}$ gegeben. Die Wahrscheinlichkeit, dass *kein* neues Element erzeugt wird, ist dann $(e^{-\lambda h})^n = e^{-n\lambda h}$. Dann ist die Wahrscheinlichkeit des Ereignisses $A_0 = \{$mindestens ein Element wird erzeugt$\}$,

$$P(A_0) = 1 - e^{-n\lambda h}.$$

Nun sei $A_h = \{$gerade ein Element wird während der Dauer h gebildet$\}$, und h sei klein in dem Sinne, dass $P(A_0) \approx P(A_h)$, d.h. die Wahrscheinlichkeit, dass 2 oder mehr Elemente erzeugt werden sei, vernachlässigbar klein, und $1 - \exp(-n\lambda h)$ ist in beliebig guter Näherung die Wahrscheinlichkeit, dass gerade ein Element erzeugt wird. Dies bedeutet aber gerade, dass der Parameter für die n-te Wartezeit durch $n\lambda$ gegeben ist. Weiter ist aber auch $\exp(-n\lambda h) \approx 1 - n\lambda h$, und es folgt für das Ereignis $A = \{$ein Element wird gebildet$\}$

$$P(A) \approx 1 - (1 - n\lambda h) = n\lambda h. \tag{8.67}$$

Dieser Ausdruck ist nützlich, wenn man $p_n(t) := P(X(t) = n)$, die Wahrscheinlichkeit, dass die Population zur Zeit t den Umfang n angenommen hat, bestimmen will. Um sie zu bestimmen, betrachtet man $p_n(t + h)$. Dies ist also die Wahrscheinlichkeit, dass die Population zur Zeit $t + h$ den Umfang n hat. Das (zufällige) Ereignis $\{X(t+h) = n\}$ kann nun auf zwei Weisen zustande kommen: entweder es ist $X(t) = n$ und zwischen t und $t + h$ verändert sich der Wert von $X(t)$ nicht, oder es war $X(t) = n - 1$ und und der Zeit zwischen t und $t + h$

erhöht sich der Wert von $n-1$ auf n, so dass ebenfalls $X(t+h)=n$. Die Wahrscheinlichkeit, dass n während der Zeit zwischen t und $t+h$ unverändert bleibt, ist $P(\text{kein Element wird erzeugt}) = 1-P(\text{mindestens ein Element wird erzeugt})$, und nach (8.67) ist dann $P(\text{kein Element wird erzeugt}) = 1-n\lambda h$. Also ist

$$p_n(t+h) = p_n(t)(1-n\lambda h) + p_{n-1}(t)(n-1)\lambda h + o(h) \tag{8.68}$$

Wegen $p_n(t)(1-n\lambda h) = p_n(t) - n\lambda h p_n(t)$ folgt dann

$$p_n(t+h) - p_n(t) = p_{n-1}(t)(n-1)\lambda h - n\lambda h p_n(t) + o(h)$$

und

$$\frac{p_n(t+h)-p_n(t)}{h} = p_{n-1}(t)(n-1)\lambda - n\lambda p_n(t) + o(h). \tag{8.69}$$

Für $h \to 0$ folgt

$$\frac{dp_n(t)}{dt} = p_n'(t) = -n\lambda p_n(t) + (n-1)\lambda p_{n-1}(t), \quad p_n'(0) = 0. \tag{8.70}$$

Dies ist eine Differential-Differenzengleichung für $p_n(t)$, und man verifiziert durch Einsetzen, dass dann für eine anfängliche Größe k der Population

$$p_n(t) = \binom{n-1}{n-k} e^{-k\lambda t}(1-e^{-k\lambda t})^{n-k} \tag{8.71}$$

gelten muß; eine Herleitung von (8.71) aus (8.70) wird im Anhang gegeben[1]. Dies ist der Ausdruck für die *negative Binomialverteilung*. Ist zum Zeitpunkt $t=0$ die Größe der Population gleich 1 (es gibt nur eine infizierte Person), so wird aus (8.71)

$$p_n(t) = e^{-\lambda t}(1-e^{-\lambda t})^{n-1}; \tag{8.72}$$

dies ist die *geometrische Verteilung*. (8.71) und (8.72) liefern die Wahrscheinlichkeit dafür, dass die Population zum Zeitpunkt t den Umfang n hat. Man findet

$$E[X(t)] = E[X(t)-X(0)] + k = ke^{\lambda t} \tag{8.73}$$

$$Var[X(t)] = Var[X(t)-X(0)] = ke^{\lambda t}(e^{\lambda t}-1). \tag{8.74}$$

wobei k wieder der Umfang der Population am Anfang ist.

Beispiel 8.2.7 In der Theorie der Ausbreitung von Infektionen (Epidemiologie) gilt der Yule-Furry-Prozeß als "triviales" Modell (Frauenthal (1980), p. 25; analog hierzu mag man es als triviales Modell für die Ausbreitung von Gerüchten bezeichnen. Um zu sehen, dass das Modell "trivial" ist, betrachte man die Annahmen über die Ausbreitung einer Infektionskrankheit, die auf den Yule-Furry-Prozeß führen:

[1] Anhang hierfür wird noch geliefert!

A1: Gegeben sei eine Population mit konstantem Umfang N, d.h. es gibt keine Geburten, Todesfälle, Zu- oder Abwanderungen.

A2: Eine Person (allgemein: ein Element) aus der Population hat ein Merkmal A (ist mit dem Gerücht infiziert), oder $\bar{A}$ (ist nicht infiziert)

A3: A wird durch Kontakt weitergegeben, wobei es keine latente Phase gibt

A4: Alle Personen (Elemente) der Population sind gleich infizierbar, und alle Infizierten sind gleich ansteckend.

Damit der Ausbreitungsprozeß als reiner Geburtsprozeß betrachtet werden kann, muß darüber hinaus $N = \infty$ gefordert werden, denn es wird bei einem reinen Geburtsprozeß ja keine obere Schranke für die Anzahl der erzeugten Elemente gesetzt. Hat man diese Forderung akzeptiert, so sieht man, dass die Annahmen A1 bis A4 den Annahmen, die zu einem Yule-Furry-Prozeß führen, äquivalent sind, und nach (8.73) erhält man als durschnittliche Zunahme der Infizierten eine Exponentialfunktion. (8.73) ist identisch mit dem Wachstumsgesetz, dass man erhält, wenn man ein *deterministisches Modell* der Infektionsausbreitung formuliert: Ausgehend von den Annahmen A1 bis A4 sei $X(t)$ die Anzahl der Infizierten zur Zeit t, und es sei b die durchschnittliche Anzahl von Kontakten noch Infizierbarer, die zu einem neuen Infizierten pro Zeiteinheit und pro bereits Infiziertem in der Poplation führen. Dann gilt

$$X(t+h) = X(t) + bhX(t),$$

woraus

$$\frac{X(t+h) - X(t)}{h} = bX(t)$$

folgt. Für $h \to 0$ erhält man dann die Differentialgleichung

$$\frac{dX(t)}{dt} = bX(t), \tag{8.75}$$

die bekanntlich die Lösung

$$X(t) = x_o e^{\lambda t} \tag{8.76}$$

hat; dies entspricht (8.73). Dabei ist $x_0 = X(0)$, und b entspricht dem Parameter λ in (8.73). Die Tatsache, dass das deterministische Modell gerade mit der Mittelwertsfunktion des stochastischen Modells (d.h. des Yule-Furry-Prozesses) übereinstimmt, gilt übrigens *nicht notwendig* für deterministische Modelle und ihre stochastischen äquivalente. Als Beispiel werde das *klassische einfache epidemische Modell* (Frauenthal) betrachtet.

Ein unbeschränktes Wachstum der Anzahl der Infizierten ist eine unplausible Inplikation der Annahmen A1 — A4, deshalb soll nun $N < \infty$ gefordert werden. Es sei weiter $A(t)$ die Anzahl der ansteckbaren Elemente zur Zeit t, und β sei die durchschnittliche Anzahl von Kontakten zwischen ansteckbaren und bereits

infizierten, die zu einer neuen Ansteckung pro Zeiteinheit pro bereits infiziertem Element pro ansteckbarem Element führen. Dann ist $N = S(t)+X(t)$, und weiter folgt $dX(t)/dt = \beta S(t)X(t)$, $dS(t)/dt = -\beta S(t)X(t)$. Aber $S(t) = N - X(t)$, so dass

$$\frac{dX(t)}{dt} = \beta X(t)(N - X(t)). \tag{8.77}$$

Dies ist die *logistische Gleichung*, und sie hat die Lösung

$$\begin{aligned} X(t) &= \frac{x_0 \exp(N\beta t)}{1 - x_0(1 - \exp(N\beta t))/N} \\ &= \frac{x_0}{\exp(-N\beta t)(1 - x_0/N) + x_0/N} \end{aligned} \tag{8.78}$$

Für $t \to \infty$ strebt nun $\exp(-N\beta t)$ gegen Null, so dass $X(t)$ gegen N strebt. Das (deterministische) Modell impliziert also, dass am Ende alle Elemente der Population infiziert sind.

Frauenthal (1980), p. 34, diskutiert weiter die stochastische Version des klassischen Modells. $A(t)$ sei wie vorher die Anzahl der Infizierbaren zur Zeit t, und es sei βh die Wahrscheinlichkeit eines Kontaktes zwischen einem bestimmten infizierbarem Element und einem bereits infizierten, der zu einem neuen infizierten Element während des kleinen Intervalles h führt. Es sei weiter

$$p_k(t) = P[X(t) = k \cap A(t) = N - k].$$

Nach (8.62) gilt nun für einen reinen Geburts- (hier: Ansteckungs-) Prozeß $p_n'(t) = p_n'(t) = -n\lambda p_n(t) + (n-1)\lambda p_{n-1}(t)$. Dabei repräsentiert $n\lambda$ den Fall, bei dem keiner der Infizierten die Krankheit weitergibt. Hierfür muß nun $\beta k(N-k)$ eingesetzt werden. Analog dazu hat man für $(n-1)\lambda$ jetzt $\beta(k-1)(n-k+1)$ einzusetzen; dieser Ausdruck beschreibt den Fall, bei dem einer der $k-1$ bereits Infizierten die Krankheit an einen Nichtinfizierten weitergibt. (8.62) nimt dann die Form

$$p_k'(t) = -\beta k(n-k)p_k(t) + \beta(k-1)(N-k+1)p_{k-1}(t) \tag{8.79}$$

an. Diese Differentialgleichung ist nichtlinear im Index k, und es gibt keine Möglichkeit, die Lösung in geschlossener Form anzuschreiben. Es gelingt dementsprechend auch nicht, die Mittelwertsfunktion $E(X(t))$ in geschlossener Form darzustellen. Es läßt sich jedoch zeigen, dass $d^2E(X(t))/dt^2 \to \beta^2(N-3)(N-1)$ für $t \to \infty$, und dies gilt nicht für das deterministische Modell (8.76). Erst für $N \to \infty$ gleichen sich die deterministischen und die stochastischen Versionen einander an. □

Reine Todesprozesse

Beim Yule-Furry-Prozeß werden die Zeiten zwischen den Ereignissen immer kürzer, weil die Anzahl der Elemente, die neue Elemente erzeugen, wächst. Man

kann nun den umgekehrten Fall, bei dem die Anzahl der Elemente mit jedem Ereignis fällt, betrachten. Statt eines Geburts- hat man einen Todesprozeß. Bei einem solchen Prozeß werden dann die Zwischenzeiten immer länger. So kann man einen Prozeß betrachten, der mit n Elementen beginnt und bei dem zunächst ein Element mit einer exponentiellen Wartezeit mit dem Parameter $n\lambda$ "stirbt". Die Wartezeit bis zum nächsten Todesfall ist darnach exponentiell mit dem Parameter $(n-1)\lambda$, etc.

Definition 8.2.11 *Es sei $X_t = \{X(t),\ t \geq 0\}$ ein stochastischer Prozeß mit $X(0) = n$, und zu zufälligen Zeiten t_i verringert sich $X(t)$ um 1. Die Zwischenzeiten seien exponentiell verteilt mit der Rate $\lambda_i = (n-i)\lambda$. Dann heißt X_t* reiner Todesprozeß.

X_t ist kein Zählprozeß im Sinne von Definition 8.2.1, da sich ja der Wert der Trajektorien $X(t)$ mit jedem Ereignis jeweils *verringert*. Man könnte natürlich die Ereignisse auch in diesem Falle zählen und hätte dann einen mit X_t assoziierten Zählprozeß. Die Herleitung der Verteilung $p_k(t)$ geschieht jedenfalls analog zu (8.71), und man erhält

$$P(k;\lambda t) = \binom{n}{k}(e^{-\lambda t})^{n-k}(1-e^{-\lambda t})^n \tag{8.80}$$

für die Wahrscheinlichkeit der Anzahl k der Ereignisse im Intervall $(0,t]$. (8.80) ist die Binomialverteilung.

Beispiel 8.2.8 Das Assoziieren von Begriffen zu einem Oberbegriff läßt sich in erster Näherung durch einen reinen Todesprozeß beschreiben (McGill (1963). So ließen Bousfield und Sedgewick (1944) ihre Vpn soviele amerikanische Städte nennen, wie sie nur konnten, und bestimmten die Zeiten zwischen den Assoziationen. Die Autoren zeigten, dass die Zwischenzeiten zunächst kurz waren, aber immer größer wurden. Insbesondere konnten sie eine Kurve der Form

$$y(t) = c(1-e^{-\lambda t}) \tag{8.81}$$

an die Daten anpassen; dabei ist $y(t)$ die Gesamtzahl von Assoziationen, die bis zur Zeit t erzeugt wurden ($Y_t = \{Y(t),\ t \geq 0\}$ ist hier also ein Zählprozeß), und c ist eine Asymptote. λ ist ein freier Parameter.

Nun ist nach (8.80) $p_k(t)$ durch die Binomialverteilung gegeben, und zwar mit den Parametern n und $p = 1-e^{-\lambda t}$; folglich ist der Erwartungswert $E(k,t) = np = n(1-e^{-\lambda t})$. Dieser Ausdruck ist aber identisch mit dem von Bousfield und Sedgewick angepaßten, d.h. mit (8.81) mit $c = n$. In der Tat muß n die Asymptote sein, denn n ist die Gesamtzahl der gespeicherten Begriffe für den gegebenen Oberbegriffe. (8.81) kann als Mittelwertsfunktion aufgefaßt werden. Die Verbindung von (8.81) mit (8.80) legt also nahe, dass der Suchprozeß, der die assoziierten Begriffe liefert, eine dem Todesprozeß entsprechende Struktur

hat. Dies bedeutet, dass man vermuten kann, dass die Zeitkonstante bei jeder produzierten Assoziation, d.h. bei jedem gefundenen Begriff, reduziert wird, denn dies ist ja die Bedeutung der immer länger werdenden Suchzeiten.

Es muß noch geklärt werden, welche Elemente denn hier "sterben". Dazu kann man annehmen (McGill (1963), p. 345), dass die Suche nach noch nicht produzierten Begriffen einer zufälligen Wanderung durch die Menge (mit dem Umfang n) der gespeicherten Begriffe gleicht. Ist ein Begriff gefunden worden, so ist die Zeit der überprüfung, ob er schon genannt wurde, exponentiell verteilt. Ist der gefundene Begriff schon genannt worden, so geht die Suche weiter. Ist er noch nicht gefunden worden, so wird er genannt und mit einer Marke versehen, um zu kennzeichnen, dass er bereits genannt worden ist (die genannte überprüfung der Begriffe ist also ein Test auf das Vorhandensein einer solchen Marke). Diese Markierung entspricht dem "Sterben", denn das Element ist nun nicht mehr verfügbar.

Wie McGill ausführt ist ein solches Modell nur eine erste Annäherung, denn es kann nicht das tatsächlich beobachtete "Clustering" in den Daten erklären: offenbar gibt es für einen Oberbegriff eine Reihe von Sub-Oberbegriffen. Gelangt die Suche zu einem solchen Sub-Oberbegriff, so werden erst einmal Begriffe aus der entsprechenden Teilmenge genannt, und zwar mit zunächst wieder größerer Zeitkonstante. Weiter werden gelegentlich Abweichungen von der Exponentialkurve insbesondere am Anfang des Assoziationsprozesses beobachtet. Gleichwohl kann das Modell des Todesprozesses als Ausgangsmodell für verfeinerte Versionen gelten. □

Beispiel 8.2.9 McGill liefert ein weiteres Beispiel für die Anwendung des Todesprozesses, nämlich zur Erklärung der Funktionsweise des Photorezeptors, wie sie zuerst von Mueller (1954) vorgeschlagen wurde. Der Muellerschen Theorie zufolge erfährt ein Molekül der photosensitiven Substanz im Rezeptor eine Zustandänderung, wenn es ein Lichtquantum absorbiert. Nach einer zufällig verteilten Zeitdauer führt dieser Zustandwechsel zur Erzeugung eines neuronalen Impulses. Wird der Rezeptor durch einen kurzen Impuls von 1 bis 10 ms Dauer stimuliert, so geht eine Anzahl von Molekülen direkt in den aktivierten Zustand über. Die Anzahl der Lichtquanten, die den Rezeptor erreichen, ist Poisson-verteilt, wird aber der Einfachheit halber als konstant angenommen. Werden nun n Moleküle aktiviert, so beginnen simultan n Prozesse mit zufällig verteilten Dauern, die zur Erzeugung von neuronalen Impulsen führen. Die Prozesse seien stochastisch unabhängig und die Zeitdauern seien exponentialverteilt mit dem Parameter λ. Die Zeitdauer bis zum ersten Impuls ist die kleinste der Zeiten $t_1, \cdots, t_n$, $t = \min(t_1, \cdots, t_n)$. t ist durch

$$P\left(\bigcap_{i=1}^{n}\{t \leq t_i\}\right) = P(t \leq t_1 \cap t \leq t_2 \cdots \cap t \leq t_n)$$

gegeben, und wegen der stochastischen Unabhängigkeit gilt

$$P\left(\bigcap_{i=1}^{n}\{t \leq t_i\}\right) = \prod_{i=1}^{n}(1 - F_i(t)) \tag{8.82}$$

wenn F_i die Verteilungsfunktion von t_i ist. Aber die t_i haben alle die gleiche Verteilungsfunktion F, also gilt

$$P\left(\bigcap_{i=1}^{n}\{t_i \geq t\}\right) = (1 - F(t))^n \tag{8.83}$$

Die Dichte $g_1(t)$ für die erste Zwischenzeit ist dann durch Differentiation von G gegeben, und es folgt

$$g_1(t) = n(1 - F(t))^{n-1} f(t), \tag{8.84}$$

mit $f(t) = dF(t)/dt$. Für $F(t) = 1 - \exp(-\lambda t)$ folgt

$$g_1(t) = n(e^{-\lambda t})^{n-1}\lambda e^{-\lambda t} = n\lambda e^{-n\lambda t}. \tag{8.85}$$

Für die Wartezeit der $(k+1)$-ten Antwort findet man (McGill, p. 347)

$$g_{k+1}(t) = (n-k)\lambda\binom{n}{k}(e^{-\lambda t})^{n-k}(1 - e^{-\lambda t})^k. \tag{8.86}$$

Diese Verteilung ist identisch mit der Verteilung der Zwischenzeiten eines Todesprozesses. Der Prozeß der Erzeugung von neuronalen Impulsen ist strukturell äquivalent einem Suchprozeß, bei dem simultan nach n verschiedenen "Alternativen" gesucht wird; eine Alternative ist die Erzeugung eines Impulses.

In Kapitel 7 wurde die Verteilung von Extremen diskutiert und es wurde gezeigt, dass für hinreichend großes n die Verteilung des Minimums gegen eine Grenzverteilung konvergiert. Insbesondere wurde in Beispiel 7.2.6 (p. 224) die Grenzverteilung für das Minimum exponentialverteilter Variablen hergeleitet; diese Grenzverteilung ist wieder *durch eine Exponentialverteilung gegeben*!

Das Modell stellt wiederum eine starke Vereinfachung dar, illustriert aber den metaphorischen Charakter der Begriffe des "Absterbens" oder der "Geburt", die in der ursprünglichen Formulierung der Definition der Prozesse vorkommen. □

Verallgemeinerung: Geburts- und Todesprozesse

Die reinen Geburts- oder Todesprozesse stellen für viele Anwendungen eine zu starke Vereinfachung dar, deren Nützlichkeit abe darin besteht, einen ersten Ansatz zur Modellierung der zur Diskussion stehenden Prozesse zu liefern. So wird bei der Formulierung des "trivialen" epidemiologischen Modells als einen reinen Geburtsprozesses u.a. nicht in Rechnung gestellt, dass Infizierte auch sterben können oder sich von Kontakten mit Nichtinfizierten fernhalten und damit zur Weitergabe der Infektion nicht mehr zur Verfügung stehen. Man kann nun Geburts-

und Todesprozesse miteinander kombinieren und somit zu realistischeren Modellen gelangen. Dies soll hier nicht weiter verfolgt werden; Darstellungen hierüber findet man in Feller (1968), Ross (1983), Parzen (1962).

8.3 Markov-Ketten

8.3.1 Grundlegende Definitionen

Viele in der Psychologie betrachteten Prozesse können als Folge von Ereignissen

$$S_0, S_1, \cdots, S_n, \cdots$$

angesehen werden. So kann es von Interesse sein, in einer Fabrik die Folge der Unfälle zu analysieren, oder in einer psychiatrischen Klinik die Folge der als "schizophren" eingestuften Patienten, oder die Abfolge bestimmter Klassen von Interaktionen in einer sozialen Gruppe, die Folge der Gewalttaten gegen Ausländer, etc. Der übergang von einem Ereignis zu einem anderen ist dabei zufällig in dem Sinne, als nicht mit Sicherheit gesagt werden kann, welches Ereignis auf ein beeits eingetretenes folgt. Dementsprechend sind die übergangswahrscheinlichkeiten $p(S_k|S_{k-1},\, S_{k-2},\, \cdots,\, S_o)$ von Interesse; sie reflektieren in einem noch zu spezifizierenden Sinne die Struktur des betrachteten Prozesses.

Definition 8.3.1 *Zu einem Zeitpunkt t_k sei der Prozeß im Zustand S_j, und im Zeitpunkt t_{k-1} sei der Prozeß im Zustand S_i. Weiter gelte*

$$P(S_k|S_{k-1}, S_{k-2}, \cdots, S_0) = P(S_j|S_i), \tag{8.87}$$

so dass der übergang nach S_j nur vom Zustand zur Zeit t_{k-1} abhängt. Dann heißt die Folge der Zustände $S_o,\, S_1,\, \cdots,\, S_k,\, \cdots$ Markovkette.

Prozesse mit der Eigenschaft (8.87) nehmen deswegen eine zentrale Rolle bei dem Anwendungen der Theorie der stochastischen Prozesse ein, weil die Analyse solcher Prozesse wesentlich einfacher ist als die von Prozessen, die (8.87) nicht erfüllen. In vielen Fällen ist (8.87) auch nicht nur eine Vereinfachung, sondern durchaus plausibel.

Es sei $p_{ij} = p(S_i|S_j)$. Die p_{ij} können zu einer Matrix zusammengefaßt werden:

$$P = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ p_{j1} & p_{j2} & \cdots & p_{jn} \\ \vdots & \vdots & \cdots & \vdots \end{pmatrix} \tag{8.88}$$

P heißt auch *Matrix der übergangswahrscheinlichkeiten.* Die übergangswahrscheinlichkeiten p_{ij} können dabei eine Funktion der Zeit sein, so dass $p_{ij} = p_{ij}(t)$. Gilt

$p_{ij} = constant$ für alle i, j, so heißt die Markov-Kette *homogen*. Die Anzahl der Zustände kann, muß aber nicht endlich sein.

Beispiel 8.3.1 (Random Walk) Gegeben sei eine Folge von Zuständen

$$\{\cdots - S_3, -S_2, -S_1, S_0, S_1, S_2, S_3, \cdots\}.$$

Es gelte $P(S_{i-1}|S_i) = 1 - p$, $p(S_{i+1}|S_i) = p$, und p sei unabhängig von i, mithin ist die Folge eine Markov-Kette. Offenbar ist $p(S_j|S_i) = 0$ für $j \neq i - 1$ oder $j \neq i + 1$. S_i kann mit dem Wert x_i einer zufälligen Veränderlichen, die ein Merkmal repräsentiert, gleichgesetzt werden. Demnach heißt die Folge *Random Walk*, da sie unter anderem den zufälligen Weg eines Teilchens beschreibt, dass zu bestimmten Zeitpunkten mit der Wahrscheinlichkeit p nach rechts und mit der Wahrscheinlichkeit $1 - p$ nach links gestoßen wird. Wird nicht festgelegt, wann der Random Walk endet, so gibt es unendlich viele Zustände. Für eine endliche Anzahl von möglichen Zuständen ergibt sich insbesondere

$$P = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mn} \end{pmatrix} \tag{8.89}$$

□

Beispiel 8.3.2 (Entscheidungsverhalten: das VTE-Modell (von Vicarious Trial and Error)). Eine Person muß sich zwischen zwei Alternativen A_1 und A_2 entscheiden. Um in der gegebenen Situation zu einer Entscheidung zu gelangen, muß sie verschiedene Aspekte der Situation abwägen. Die Betrachtung eines bestimmten Aspekts erlaubt ihr den übergang zu A_1, falls der Aspekt zu einer bestimmten Menge M_1 gehört, gleichwohl ist der übergang dann nicht sicher, falls der Aspekt nicht genügend Information über die Alternative liefert, sucht sie einen neuen Aspekt. Wenn sie einen neuen Aspekt sucht, sei sie im Zustand S_0. Hat sie einen Aspekt der Menge M_1 gewählt, so ist sie in Zustand S_1. Komplementär zu M_1 gibt es eine Menge M_2; hat sie einen Aspekt aus der Menge M_2 gewählt, so ist die Person im Zustand S_2. Für einen geeigneten Aspekt aus M_2 folgt eine Entscheidung für A_2, sonst kehrt sie zu S_0 zurück. Die übergangswahrscheinlichkeiten seien wie folgt festgelegt:

	S_0	S_1	A_1	A_2
S_0	p_{01}	p_{02}	0	0
S_1	p_{10}	0	0	p_{12}
A_1	0	0	1	0
A_2	0	0	0	1

d.h.

$$P = \begin{pmatrix} p_{01} & p_{02} & 0 & 0 \\ p_{10} & 0 & 0 & p_{12} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Es muß natürlich $p_{01} + p_{02} = 1$, $p_{10} + p_{12} = 1$ gelten. Da $p(A_1|A_1) = p(A_2|A_2) = 1$, bleibt die Person bei ihrer Entscheidung, hat sie sie erst einmal getroffen. Die Zustände A_1 und A_2 heißen demnach *absorbierend.* Die übergangswahrscheinlichkeiten p_{ij} hängen nicht von der Geschichte des Prozesses ab, also ist der Prozeß eine Markov-Kette. Die spezielle Definition der Zustände und er übergangswahrscheinlichkeiten definieren die Struktur des Prozesses. Dieses Entscheidungsmodell ist das VTE-Modell von Bowers (1959, 1962). □

Die Matrix P der übergangswahrscheinlichkeiten bestimmt allerdings den Prozeß noch nicht vollständig. Um den Prozeß vollständig zu charakterisieren müssen noch die Wahrscheinlichkeiten, mit denen der Prozeß in einem der Zustände S_1, $S_2, \cdots$ ist, angegeben werden. Diese werden im *Startvektor* $p_0 = (p_0(S_1), p_0(S_2), \cdots)'$ angegeben. Sind P und der Startvektor p_0 gegeben, können die Wahrscheinlichkeiten dafür, dass der Prozeß nach k Schritten im Zustand S_i, $i = 1, 2, \cdots$ ist, ausgerechnet werden. Für endliches n gilt dann etwa für den Zeitpunkt t_1, wenn der Prozess in t_0 startet, nach dem Satz der totalen Wahrscheinlichkeit:

$$\begin{aligned} p_1(S_i) &= p(S_i|S_1)p_0(S_1) + \cdots + p(S_i|S_n)p_0(S_n) = \\ &= p_{i1}p_0(S_1) + \cdots + p_{in}p_0(S_n). \end{aligned} \tag{8.90}$$

Man sieht, dass nach den Regeln der Matrixmultiplikation der Vektor p_1 kompakt durch

$$\beta_1 = \beta' P \tag{8.91}$$

angegeben werden kann. Die Wahrscheinlichkeit, dass der Prozeß im zweiten Schritt im Zustand S_i ist, ergibt sich analog zu (8.90) und (8.91) zu

$$\beta_2' = \beta_1' P. \tag{8.92}$$

Für p_1 kann man den in (8.91) angegebenen Ausdruck einsetzen und erhält sofort $p_2 = \beta_0 P \cdot P = \beta_0 P^2$. So kann man fortfahren und erhält dann für den Vektor der Zustandswahrscheinlichkeiten nach dem k-ten Schritt

$$p_k = \beta_0^k P^k. \tag{8.93}$$

Die Zustände einer Markov-Kette können wie folgt klassifiziert werden.

Definition 8.3.2 *Ein Zustand heißt* transient*, wenn der Prozeß diesen Zustand wieder verlassen, aber dann nie wieder erreichen kann. Ein Zustand heißt* ergodisch*, wenn er nicht transient ist. Ist ein Zustand* einziger *Zustand einer Menge ergodischer Zustände, so heißt er* absorbierend

Es sei C_k eine Menge von ergodischen Zuständen derart, dass ein übergang von einem Zustand dieser Menge in einen anderen Zustand dieser Menge möglich ist. Dann kann der Prozeß die Menge dieser Zustände nie verlassen. Daraus kann man folgern, dass die Menge der Zustände einer Markov-Kette zerlegt werden kann in eine Menge T von transienten Zuständen und in Teilmengen $C_1, C_2, \cdots$ jeweils ergodischer Zustände.

Definition 8.3.3 *Eine Markov-Kette, deren Zustandsmenge nicht weiter zerlegt werden kann, heißt* irreduzibel.

Eine Markov-Kette, deren Zustandsmenge aus verschiedenen ergodischen Mengen $C_1, C_2, \cdots$ besteht, besteht also eigentlich aus verschiedenen Markov-Ketten, zwischen denen kein übergang möglich ist; die Kette geht immer in den gleichen Zustand über, hat sie ihn ersteinmal erreicht. Ist S_j ein solcher Zustand, so gilt also $p_{jj} = 1$. Dann folgt, dass eine endliche Markov-Kette nie nur transiente Zustände haben kann, denn nach Durchlaufen aller dieser Zustände kann sie in keinen Zustand mehr zurück; dieser Zustand wäre dann aber ein absorbierender und mithin ein ergodischer Zustand.

Es sei eine ergodische Markov-Kette mit der Matrix P von übergangswahrscheinlichkeiten gegeben. Es sei p_n der Vektor der Zustandswahrscheinlichkeiten nach dem n-ten Schritt. Es ist $p_n = Pp_{n-1}$, $p_{n-1} = Pp_{n-2}$, d.h. $p_n = P^2 p_{n-1}$ etc., so dass schließlich $p_n = P^n p_0$, p_0 der Startvektor.

Definition 8.3.4 *Gilt $P^n > 0$ für irgendein $n \in \mathbb{N}$, wie auch immer der Prozeß startet, so heißt die Markov-Kette* regulär. *Ein ergodischer Prozeß heißt* zyklisch, *wenn die Zustände in definiter Ordnung durchlaufen werden.*

Es sei p_0 wieder ein Startvektor für eine reguläre Markov-Kette, und $p_n = P^n p_0$. Es ergibt sich die Frage, ob p_n für $n \to \infty$ gegen einen bestimmten Vektor π konvergiert, für den dann $\pi = P\pi$ gilt (denn dann ist ja p_{n-1} ebenfalls gleich π). Die Bedingung $P\pi = \pi$ besagt aber, dass π ein Eigenvektor von P ist mit dem zugehörigen Eigenwert $\lambda = 1$; eine notwendige Bedingung für die Existenz von π ist also die Existenz eines Eigenwerts $\lambda = 1$.

Beispiel 8.3.3 In einem einfachen Lernexperiment soll die Versuchsperson (Vp) lernen, bestimmte Stimuli entweder in die Kategorie A oder in die Kategorie $\bar{A}$ zu klassifizieren; die Aufgabe der Vp besteht also darin, die Regel heraußufinden, nach der die Stimuli klassifiziert werden sollen. Sie bekommt einen Stimulus dargeboten, dann klassifiziert sie, und danach wird ihr vom Versuchsleiter (Vl) gesagt, welche Kategorie die korrekte war (unabhängig davon, ob die Antwort richtig oder falsch war). Es wird angenommen, dass der Lernprozeß als Markov-Kette dargestellt werden kann; die Matrix der übergangswahrscheinlichkeiten und der Startvektor seien durch

$$P = \begin{bmatrix} 1 & 0 \\ c & 1-c \end{bmatrix}, \quad \beta_0 = \begin{bmatrix} q \\ 1-q \end{bmatrix}$$

gegeben. Es werden also zwei mögliche Zustände angenommen: hat die Vp die Regel gefunden, so ist sie im Zustand S_1, hat sie sie noch nicht gefunden, so ist sie im Zustand S_2. Am Anfang bildet sie eine mögliche Regel, die mit der Wahrscheinlichkeit q korrekt ist. Ist sie in S_1 angelangt, so kann sie den Zustand nicht wieder verlassen, d.h. sie kann, zumindest im Rahmen des Lernexperiments, die Regel nicht wieder vergessen. Der Zustand S_1 ist demnach ein Beispiel für einen *absorbierenden Zustand*. Mit der Wahrscheinlichkeit c geht sie von S_2 nach S_1 über, und es wird angenommen, dass c konstant über die Zeit ist. Man muß sich klarmachen, dass die Vp eine falsche Hypothese über die Regel haben kann und trotzdem eine korrekte Antwort geben kann. In diesem Fall wird sie ihre Hypothese nicht aufgeben, sie bleibt also in S_2. Daraus folgt, dass sie nur nach einer *falschen* Antwort in den Zustand S_1 übergehen kann.

Nach (8.93) kann man nun ausrechnen, mit welcher Wahrscheinlichkeit die Vp nach k Lerndurchgängen (Darbietung eines Stimulus, Antwort und "Reinforcement" des Vl) im Zustand S_i, $i = 1, 2$ ist. Nach dem ersten Durchgang gilt

$$\beta_1 = (q, 1-q) \begin{bmatrix} 1 & 0 \\ c & 1-c \end{bmatrix} = (q + (1-q), (1-q)(1-c)).$$

Für den zweiten Durchgang erhält man den Vektor der Zustandswahrscheinlichkeiten

$$\begin{aligned} \beta_2 &= (q, 1-q) \begin{bmatrix} 1 & 0 \\ c & 1-c \end{bmatrix}^2 \\ &= (q, 1-q) \begin{bmatrix} 1 & 0 \\ c + c(1-c) & (1-c)^2 \end{bmatrix} \\ &= (q + c(2-c) - 2qc + qc^2, (1-q)(1-c)^2), \end{aligned}$$

und für β_3

$$\beta_3 = (q, 1-q) \begin{bmatrix} 1 & 0 \\ c(2-c) + c(1-c)^2, & (1-c)^3 \end{bmatrix}, \quad \text{etc.}$$

Man findet, dass die Wahrscheinlichkeit, *nicht* in S_1 zu sein, nach dem k-ten Schritt durch $(1-c)^k$ gegeben ist, so dass die Wahrscheinlichkeit, in S_1 zu sein, gleich $1 - (1-c)^k$ ist; für $k \to \infty$ folgt also $p(S_1) \to 1$ und damit $p(S_2) \to 0$; also ist der Vektor π mit $P\pi = \pi$ durch $\pi = (1, 0)'$ gegeben. Die Kette ist ergodisch. □

Eine weitere Einführung in diese Modelle findet man in Atkinson, Bower und Crothers (1965)[2] bzw. in den Kapiteln 9 und 10 von Luce, Bush, Galanter (1963)

[2] Atkinson, R.C., Bower, G.H., Crothers, E.J.: An Introduction to Mathematicaal Learning Theory. New York 1965

Beispiel 8.3.4 (2-Personen Interaktion, Atkinson et al (1965), p. 313) Die Interaktion zweier Personen (einer *Dyade*) läßt sich in bestimmten Situationen dadurch charakterisieren, dass die Handlungen der Personen jeweils in eine von zwei Klassen fallen: eine Handlung ist entweder kooperativ oder kompetetiv (nicht kooperativ). Wenn die Person A kooperiert, B aber nicht, so erleidet A einen "Schaden" (worin auch immer der besteht) und B erzielt etwa einen Gewinn, und umgekehrt erleidet B einen Schaden, wenn B kooperiert, A aber nicht; dann hat A einen Gewinn. Kooperieren beide, so haben sie den höchsten Gewinn bzw. den geringsten Verlust, kooperieren beide nicht, so ist ihr Verlust maximal. Die einfachste Lösung für beide besteht darin, stets zu kooperieren. Erzielt man aber den höchsten Gewinn, wenn man selber *nicht* kooperiert, aber der andere kooperiert, so kann es reizvoll sein, eben nicht zu kooperieren. So muß man stets aufs neue entscheiden, ob man kooperieren will oder nicht. Denn nicht zu kooperieren hat den Vorteil, dem anderen keinen Gewinn zukommen zu lassen für den Fall, dass er nicht kooperiert.

Die Dyade kann also in drei verschiedenen Zuständen sein: S_1 — keiner der beiden kooperiert ($\bar{A}\bar{B}$), S_2 — beide kooperieren (AB), S_3 — einer der beiden kooperiert nicht ($\bar{A}B \cup A\bar{B}$). Das Verhalten der Dyade werde nun durch die folgenden Annahmen beschrieben:

A1: Macht eine der beiden Personen A oder B einen Gewinn, so behält sie das Verhalten, das zu diesem Gewinn führte, mit Wahrscheinlichkeit 1 bei.

A2: Macht eine der beiden Personen A oder B keinen Gewinn, macht sie also einen Verlust, so ändert sie das Verhalten, dass zu diesem Verlust führte, mit der Wahrscheinlichkeit θ.

Durch diese Annahmen wird die Folge der Zustände, in der sich die Dyade befindet, als Markov-Kette charakterisiert. Denn das folgende Verhalten hängt immer nur vom Ergebnis des zuletzt gewählten Verhaltens ab, nicht aber von den davor gewählten Verhaltensweisen. Bezeichnet man mit $S_{i,n}$ den Sachverhalt, dass die Dyade im n-ten "Durchgang", d.h. beim n-ten Austausch von Verhaltensweisen, im Zustand S_i ist, so folgen aus den Annahmen A1 und A2 die folgenden übergangswahrscheinlichkeiten: $P(S_{1,n+1}\,|S_{1,n}\,) = (1-\theta)^2$, denn S_1 heißt, dass beide nicht kooperiert haben, und beide ändern - stochastisch unabhängig voneinander — ihren Verhaltenstyp von kompetetiv zu kooperativ bzw bleiben beim kompetetiven Verhaltenstyp, und dies geschieht mit der Wahrscheinlichkeit $(1-\theta)(1-\theta) = (1-\theta)^2$. Weiter ist $p(S_{2,n+1}\,|S_{1,n}\,) = \theta^2$, - den mit Wahrscheinlichkeit θ gehen beide zum kooperativen Veraltenstyp über. Schließlich folgt noch $p(S_{3,n+1}\,|S_{1,n}\,) = 2\theta(1-\theta)$, denn es gilt entweder $\bar{A}\bar{B} \to \bar{A}B$; beide machen einen Verlust, der erste bleibt mit der Wahrscheinlichkeit θ bei seinem Verhaltenstypus, und der andere ändert in von kompetetiv zu kooperativ, und da sie stochastisch unabhängig voneinander handeln, hat dies die Wahrscheinlichkeit $\theta(1-\theta)$. Die andere Möglichkeit ist besteht im übergang $\bar{A}\bar{B} \to A\bar{B}$, der ebenfalls mit der

Wahrscheinlichkeit $(1-\theta)\theta$ erfolgt, und zusammen ergibt sich die Wahrscheinlichkeit $2\theta(1-\theta)$.

Die übergangswahrscheinlichkeiten $p(S_{j,n+1}\,|S_{i,n}\,)$ ergeben sich durch analoge Betrachtungen, so dass man die Matrix

	S_1	S_2	S_3
S_1	$(1-\theta)^2$	θ^2	$2\theta(1-\theta)$
S_2	$\theta^2/4$	$1-\theta+\theta^2/4$	$\theta-\theta^2/2$
S_3	$\theta(1-\theta)$	$\theta(1-\theta)$	$1-\theta+2\theta^2$

bzw.

$$P = \begin{pmatrix} (1-\theta)^2 & \theta^2 & 2\theta(1-\theta) \\ \theta^2/4 & 1-\theta+\theta^2/4 & \theta-\theta^2/2 \\ \theta(1-\theta) & \theta(1-\theta) & 1-\theta+2\theta^2 \end{pmatrix}$$

erhält.

Das Modell definiert eine ergodische Markov-Kette und stellt sicherlich eine drastische Vereinfachung dar, denkt man an die Komplexität menschlichen Handelns. Um so mehr überrascht es, zu sehen, wie gut experimentelle Daten dem Modell entsprechen.

□

Markov-Ketten haben die mathematisch sehr angenehme Eigenschaft der Gedächtnislosigkeit: die Wahrscheinlichkeit, in einen bestimmten Zustand zu gehen, hängt nur von dem Zustand ab, in dem der Prozeß gerade ist. Deswegen lassen sich Markov-Ketten (verhältnismäßig leicht) analysieren. Es ist also nicht verwunderlich, dass sie als Modelle für psychologische Prozesse herangezogen werden. Der klassische Artikel von Miller (1952) liefert eine Reihe wichtiger Einzelheiten über Markov-Ketten, die bei der Modellierung von Nutzen sind. Unter anderem wird gezeigt, wie die zunächst oft psychologisch unplausible Annahme der Gedächtnislosigkeit durch geeignete Definition der Zustände gerechtfertigt werden kann, so dass die mathematischen Vorteile der Markov-Ketten erhalten bleiben. Weitere Details findet man in Kemeney und Snell (1967). Coleman (1964) macht extensiven Gebrauch von Markov-Ketten bei der Modellierung sozialer Prozesse.

Die Zeit ist bisher nicht explizit in die Beschreibung der Markov-Ketten eingegangen. Naürlich ist es plausibel, anzunehmen, dass die Zustandsübergänge in einem bestimmten Zeitraster erfolgen. So kann man z.B. annehmen, dass die übergänge stets nach Ablauf eines Zeitintervalls der Länge T erfolgen; die Werte der p_{ij} können dann von T abhängen. Eine zweite Möglichkeit besteht darin, die p_{ij} explizit als Funktionen der Zeit zu betrachten, $p_{ij} = p_{ij}(t)$. Spezielle Formen dieser Abghängigkeit, die für die Modellierung sozialer Prozesse von Belang sind, diskutieren Singer und Spilerman (1976). Andere Formen der Zeitabhängigkeit werden im folgenden Kapitel behandelt.

8.3.2 Verweildauern und Semi-Markov-Prozesse

Bei der Betrachtung von Markov-Ketten werden die Zeiten, die der Prozeß in einem gegebenen Zustand verbringt - dies sind die *Verweildauern* - nicht explizit betrachtet. Gleichwohl kann es nützlich sein, diese Zeiten zu betrachten. Denn häufig ist man ja an einem Prozeß *in der Zeit* interessiert: man beobachtet die Folge von Zuständen und die Zeiten (*Verweildauern*), die der Prozeß in den einzelnen Zuständen bleibt. Man spricht dann auch von *Markov-Ketten in kontinuierlicher Zeit.*

Beispiel 8.3.5 (Dyadische Interaktion) Man ist an der Interaktion zweier "Spieler" interessiert, die sich in einer Prisoner's Dilemma-Situation befinden. Die Interaktion bestehe aber nicht in einer diskreten Folge solcher Spiele, sondern sei ein kontinuierlicher Verlauf: zu jedem Zeitpunkt soll es möglich sein, sich

- kooperativ (A_1), oder
- kompetetiv (A_2), oder
- neutral (A_3)

zu verhalten; die Menge der möglichen Verhaltensweisen soll also in drei disjunkte Klassen aufteilbar sein. Man hat eine Matrix der Art (8.1), bei der die Zeilen

Tabelle 8.1: übergänge bei Interaktionen

	A_1	A_2	A_3
A_1	p_{11}	p_{12}	p_{13}
A_2	p_{21}	p_{22}	p_{23}
A_3	p_{31}	p_{32}	p_{33}

den Partner I und die Spalten denPartner II repräsentieren; für jeden übergang $\{A_i \to A_j\}$ kann man die entsprechende Wartezeit beobachten. □

In Definition 3.30, p. 84, ist der Begriff der Hazardfunktion eingeführt worden. Auf den Fall von Markov-Ketten kann er wie folgt angewandt werden:

Definition 8.3.5 *Es seien A_i und A_j zwei Zustände einer Markov-Kette, und $\{A_i \to A_j\}$ bezeichne den übergang von A_i nach A_j. Dann heißt*

$$\alpha_{ij} = \lim_{\Delta t \to 0} \frac{P(\{A_i \to A_j\} \in (t, t + \Delta t] | A_i \text{ zur Zeit } t)}{\Delta t} \tag{8.94}$$

die übergangsrate *von A_i nach A_j. Weiter bezeichne $\{A_i^{\uparrow}\}$, dass der Prozeß den Zustand A_i verläßt. Dann heißt*

$$\lambda_i = \lim_{\Delta t \to 0} \frac{P(\{A_i^{\uparrow}\} \in (t, t + \Delta t] | A_i \text{ zur Zeit } t)}{\Delta t} \tag{8.95}$$

die Terminationsrate *des Zustands A_i.*

λ_i ist offenbar die Hazardfunktion für die Wartezeit, die verstreicht, bis der Prozeß den Zustand A_i verläßt. Ein Markov-Prozeß ist aber gedächtnislos, d.h. die Wahrscheinlichkeit, dass ein Markov-Prozeß zur Zeit $t + \Delta t$ in einem Zustand A_i für irgend ein i ist, ist unabhängig von der Geschichte des Prozesses bis zur Zeit t. Auf die Definition von λ_i bezogen heißt dies, dass

$$\lambda_i = konstant$$

Daraus folgt nach Satz 3.3.2 dann, dass Die Verteilung der Verweildauer, d.h. der Wartezeit, die verstreicht, bis der Prozeß den Zustand A_i verläßt, durch die Exponentialverteilung

$$F_i(t) = 1 - e^{-\lambda_i t} \tag{8.96}$$

gegeben ist. Nun verläßt der Prozeß den Zustand A_i, indem er in irgend einen anderen Zustand übergeht, also in A_1 oder A_2 oder A_3 etc., und er kann nicht in zwei Zustände zugleich übergehen. Dementsprechend muß

$$\lambda_i = \sum_j \alpha_{ij} \tag{8.97}$$

gelten.

Aus den Gleichungen (8.94) und (8.95) ergibt sich ein Ausdruck für die übergangswahrscheinlichkeit $p_{ij} = P(A_I \to A_j)$. Es ist ja

$$\frac{\alpha_{ij}}{\lambda_i} = \lim_{\Delta t \to 0} \frac{P(\{A_i \to A_j\} \in (t, t + \Delta t] | A_i \text{ zur Zeit } t)}{P(\{A_i^{\uparrow}\} \in (t, t + \Delta t] | A_i \text{ zur Zeit } t)}$$

Dieser Quotient ist eine bedingte Wahrscheinlichkeit, nämlich die Wahrscheinlichkeit, zum Zeitpunkt t von A_i nach A_j zu wechseln unter der Bedingung, dass der Zustand A_i überhaupt zum Zeitpunkt t verlassen wird. Damit ist dieser Quotient gleich der übergangswahrscheinlichkeit

$$p_{ij}(t) = \frac{\alpha_{ij}}{\lambda_i} \tag{8.98}$$

Von hier Schätzungen für p_{ij}.

8.3.3 Semi-Markov-Prozesse

Es ist möglich, dass die Wechsel der Zustände durch eine Markov-Kette beschrieben werden können, dass aber die Verweildauern nicht exponentialverteilt sind. Prozesse dieser Art heißen *Semi-Markov-Prozesse.*

Literaturverzeichnis

[1] Aitchinson, J. Brown, J.A.C.: The Lognormal distribution. Cambridge 1963

[2] Alexandroff, P.S.: Einführung in die Mengenlehre und die Theorie der reellen Funktionen. Berlin 1956

[3] Atkinson. R.C., Bower, G.H., Crothers, E.J.: An Introduction to Mathematical Learning Theory. New York, 1966

[4] Baumgartner, G., Bornschein, H., Hanitzsch, R., Jung, R., Kornhuber, H.H., Rentschler,I., Schober, H., Thoden, U.: Sehen. Sinnesphysiologie III, München 1978

[5] Bousfield, W.A., Sedgewick, C. (1944) An analysis of sequences of restrictive associative responses. *J. gen. Psychol.* 30, 144 – 165

[6] Coleman, J: Introduction into mathematical sociology. New York 1964

[7] Courant, R.: Vorlesungen über Differential- und Integralrechnung (1961)

[8] Dawes, R.M.: Everyday Irrationality. Boulder, Colorado, 2001

[9] Dinges, H., Rost, H.: Prinzipien der Stochastik. Stuttgart 1982

[10] Feller, W.: An Introduction to Probbility Theory and its Applications, Vol I. New York, 1968

[11] Earman, J.: A primer on determinism. D. Reidel Publishing Company, Dordrecht, 1986

[12] Elstrodt, J.: Maß- und Integrationstheorie. Springer, 2005.

[13] Fisz, M. Wahrscheinlichkeitsrechnung und Mathematische Statistik. Berlin 1962

[14] Furry, W. (1937) On fluctuation phenomena in the passage of high energy electrons through lead. *Phys. Rev.*, 52, 1937

[15] Frauenthal, J.C.: Mathematical Modeling in Epidemiology. Springer-Verlag, Berlin, Heidelberg, New York 1980

[16] Galambos, J.: The asymptotic theory of extreme order statistics. John Wiley & Sons, New York 1978

[17] Gigerenzer, G.:Das Einmaleins der Skepsis. Berlin, 2004.

[18] Gnedenko, B.: The Theory of Probability. Moskau, 1978

[19] Jaynes, E.T.: Probability theory – The logic of science. Cambridge University Press 2003

[20] Jeffreys, H.: Theory of probability, 3rd edition, Oxford 2003

[21] Heuser, H.: Lehrbucgh der Analysis, Teil I. Stuttgart 2001

[22] Kemeny, J.G., Snell, J.L.: Finite MArkov Chains. Toronto, 1967

[23] Keynes, J. M.: A treatise on probability, 1920 – Rough Draft Printers 2008

[24] Khintchine, A.Y.: Mathematical models in the theory of queueing. London 1960

[25] Kolmogorov, A.N.: Grundbegriffe der Wahrscheinlichkeitsrechnung. Berlin, 1933

[26] Leadbetter, M.R., Lindgren, G., Rootzén, H.: Extremes and related properties of random sequences and processes- Springer Verlag, New York, Heidelberg, Berlin 1983

[27] Luce, R.D., Bush, R.R., Galanter, E. Handbook of Mathematical Psychology, Vol. II, New York 1963.

[28] McGill, W. J.: Stochastic Latency Mechanisms. In: Luce, R.D., Bush, R.R., Galanter, E. (Ed.): Handbook of Mathematical Psychology, Vol. I, New York 1963

[29] Miller, G.A. (1952) Finite Markov-Processes in Psychology. *Psychometrika*, 17, 149 – 167. Reprint in Luce, R.D., Bush, R.R., Galanter, E.: Readings in Mathematical Psychology, Vol. I, New York 1963

[30] v. Mises, R.: Wahrscheinlichkeit, Statistik und Wahrheit. Wien, 1936

[31] Monod, J.: Chance and Necessity: Essay on the Natural Philosophy of Modern Biology, 1970

[32] Papoulis, A.: Probability, Random Variables and Stochastic Processes. Tokyo 19654

[33] Parzen, E.: Stochatic Processes. San Francisco 1962

[34] Penrose, R. Shadows of the Mind. London 1994

[35] Plachky, D., Baringhaus, L. Schmitz, N.: Stochastik I. Wiesbaden 1983

[36] Ross, S.M.: Stochastic Processes. New York, 1980

[37] Savage, L.J.: The Foundations of Statistics. New York, 1954

[38] Schnell, E.D. (1960) Samuel Pepys, Isaac Newton and the probability. *American Statistician* 1960, 14, 27-30

[39] Sampath, G. Srinivasan, S.K.: Stochastic Models of Spike Trains of Single Neurons. Lecture Notes in Biomathematics 16, Berlin 1977

[40] Singer, B., Spilerman, S. (1976), The representation of Social Processes by Markov Models. *American Journal of Sociology*, 82 1 – 54

[41] Yule, G. U.: A mathematical theory of evolution, based on the conclusions of Dr. J.C. Willis, F.R.S. *Philosophical Transactions of the Royal Society, London, Series B*, Vol. 213, 1924, 21-87

Index

e-Funktion, 125
Überdeckung
    offene, 38
    endliche, 38
äquivalente Ereignisse, 46
überabzählbar, 32, 72

inverse Stichprobenentnahme, 130
r-Permutationen ohne Wiederholung, 11

Abbildung, 71
Abhängigkeit
    konfigurale, 68
Absolutschwelle, 242
abzählbar, 32, 72, 81
Adrenalin
    ausgeschüttetes, 251
    Gesamtmenge von, 251
    im körper befindliches, 251
Aktionspotential, 238
Algebra, 35, 44
Attraktionsbereich, 210, 214
Auseinandersetzungen
    eheliche, 247

Bayes, 56
bedingte Verteilung, 82
Bernoulli-Variable, 73, 98
    Erwartungswert
        Varianz, 98
Bernoulli-Verteilung, 115
Beta-Verteilung, 180
bias-frei, 101
Binomialkoeffizienten, 14
Binomialverteilung, 115
    negative, 130
    Verallgemeinerung, 118
Borel-Mengen, 40, 44, 80
Borelmengen, 35
Brustkrebsrisiko, 57

Campbells Theorem, 250
Cauchy-Verteilung, 178, 220
conditional failure rate, 59

Deichhöhen, 209
DeMorgansche Regeln, 47
Dichte
    bedingte, 82, 85
    der Maximaverteilung, 207
Dichtefunktion, 51, 78
Differenz von Ereignissen, 46
Dirac-Maß, 40
disjunkte Ereignisse, 47
diskret, 81
Dosiswirkungen, 164
Dyade, 264

Elektronenröhren
    Rauschen in, 250
Epidemiologie, 253
Erneuerungen, 239
Erneuerungsprozeß, 239
Erwartung
    bedingte, 102
erwartungstreu, 99
Erwartungswert, 94
    bedingter, 102
    einer Funktion, 96
    Vektor, 103
Erzeuger, 35
Exponentialfunktion, 125
exponentialverteilt, 149
Exponentialverteilung, 85, 91, 149

k-tes Moment, 150
Grenzverteilung der geometrischen V., 151
Extremwertverteilung, 213
Extremwertverteilungen, 210

failure rate, 84
Fakultät, 11
Faltung, 108
fas sicher (f.s.), 52
Fechners Gesetz, 173
forced choice distribution, 20
Fouriertransformierte, 107
frequentistischer Ansatz, 26
Funktion
charakteristische, 107
erzeugende, 104

Gammaverteilung, 153
Geburtsprozeß
reiner, 252
Gedächtnislosigkeit, 265
Gefangenenproblem, 60
Gerücht, 254
Gesetz der großen Zahlen
starkes, 196
Gesetz der großen Zahlen
schaches, 195
Gleichverteilung, 51, 113, 147
Grenzverteilung, 210
Grenzwert
linksseitiger, 75
rechtsseitiger, 75
Grenzwertsatz
DeMoivre-Laplace, 198
Poisson, 197
zentraler, 201
Grundquote, 58

Hauptachsentransformtion, 186
Hazardfunktion, 59, 84, 241, 247, 266
Hazardrate , 59
HD-Ansatz, 63
hypothetico-deduktiv, 63

i.i.d., 204
Impulsantwort, 251
Infektionen
Ausbreitung von, 253
Infimum, 213
inkrementelle Konfirmation, 62
Intervall
abgeschlossen, 33
halboffenes, 33
offenes, 33
inverse sampling, 130
Isotonie, 36, 50

Klima, eheliches, 247
Kombinationen
mit Wiederholung, 16
ohne Wiederholung, 13
Kombinatorik, 7
Komplementärereignis, 45
Konjunktion, 45
Kontaktpunkt, 34
konträre Ereignisse, 47
Korrelation
Produkt-Moment, 97
Korrelationsmatrix, 103
Korrelationszeit, 236
Kovarianz, 97
Matrix Varianz-, 103
Kovarianzmatriz, 103

Langzeitgedächtnis, 124
Laplace, 26
Lebesgue-Borel-Maß, 40
Lebesgue-Maß, 40
Limes, 75
limiting distribution, 210
Log-Normalverteilung, 172
logistische Verteilung
parametrische Form, 162
logistische Gleichung, 255
logistische Verteilung, 162
Lognormalverteilung, 218

Mächtigkeit des Kontinuums, 32
Mächtigkeit einer Menge, 30

Maß, 36
  σ-endliches, 36
  endliches, 36
  normiertes, 36
  Wahscheinlichkeits-, 36
maßerzeugende Funktion, 50
maßerzeugende Funktionen, 40
Maßraum, 36
Markov-Prozeß
  semi, 267
Markovkette, 259
  homogene, 260
  irreduzible, 262
  reguläre, 262
  zyklische, 262
max-stabil, 213
Maxima expon. vert. z.V., 209
Maxima normalvert. z.V., 206
Maximum von zuf. Veränderlichen, 204
Median, 148
Menge
  abgeschlossene, 33
  kompakte, 33
  offene, 33
Mengen
  vernachlässigbare, 38
messbar, 36
Minima expon. vert. z.V., 210
Minima normalverteilter z.V., 207
Minimum von zuf. Veränderlichen, 204
Modell
  epidemiologisches, 164
Moment, k-tes, 96
Momente, 107
Multinomialkoeffizienten, 19
multinomialverteilt, 119
Multiplets, 8
Multiplikationsregel, 64

Normal-(Gauß-)verteilung, 157
Normalverteilung
  mehrdimensionale, 184
Normierungskonstanten, 211
Notation, polnische, 45
Nullmenge, 39
Nullmengen, 38

Oder, ausschließendes, 47

Panik
  anfälle, 129
  epidemie, 129
  patienten, 130
Pareto-Verteilung, 213, 219
Pascalsches Dreieck, 14
Permutationen, 10
Photorezeptor, 257
Poisson-clusterings, 243
Poisson-Impuls-Prozeß, 249
Poisson-Inkrement-Prozeß, 248
Poisson-Klumpungen, 243
Poisson-Prozeß
  zusammengesetzter, 250
Poisson-Verteilung
  als Limit der neg. Binomialverteilung, 133
Potenzmenge, 31
Präparat
  schmerzlinderndes, 163
prophylaktische Masektomie, 57
Prozeß
  Ansteckungs-, 255
  in stetiger Zeit, 239
  mit stationären Zuwächsen, 239
  mit unabhängigen Zuwächsen, 239
  Poisson, homogener, 240
  Poisson, inhomogener, 244
  Todes-, reiner, 256
  Yule-Furry, 252
Prozess
  autokatalytischer, 177
Punktmaß, 41
Punktmasse, 41
Punktprozeß, 238, 244

Q-Sort-Methode, 20

Radon-Nikodym-Derivierte, 51
Random walk, 260

Randverteilung, 81
Rate, 240
Raum
    topologischer, 34
Rauschen
    additives, 174
    in Elektronenröhren, 250
reinforcement, 263
Rekord, 246
Reliabilität, 83, 215
renewal process, 239
Riemann-Integral, 79
Roulette
    russisches, 123

Sättigung, 164
Satz
    von Bayes, 56, 84
    von Campbell, 250
    von Fisher und Tippett, 213
    von Khintchine, 212
Schockwirkungen, 251
shot-noise, 249
    Powerspektrum, 250
Sigma-Algebra, 35
Skalentransformation, 86
Spike, 238
Spike-Folgen, 241
Stärke von Ankerketten, 225
standardnormalverteilt, 157
stetig, 81
Stichproben
    mit Zurücklegen, 11
    one Zurücklegen, 11
Stichprobenraum, 41
Stieltjes-Integral, 79, 95
Stieltjes-Summe, 79
stochastisch unabhängig, 81
stopping parameter, 130
Streuung, 97
Sub-Additivität, 37
Supremum, 213
symmetrische Differenz von Ereignissen, 47

Taylor-Reihe, 125, 142, 151
Teilkombinationen, 13
Terminationsrate, 267
Todesprozeß
    reiner, 256
Topologie, 34
Trommelrevolver, 123

Uebergangsrate, 267
Unabhängigkeit
    lineare, 186
    stochastische, 64, 97
Unfälle
    alkoholbedingte, 145
Ungleichung
    Bienamé, 191
    Tchebyscheff, 192

Variablen
    latente, 186
Varianz, 97
Vereinigung von Ereignissen, 45
Verteilung
    bedingte, 86
    gemeinsame, 81
    geometrische, 121
    hypergeometrische, 135
    Laplace-, 121
    Mischung von, 248
    Poisson-, 126
    vom gleichen Typ, 213
    zusammengesetzte, 142
        Poisson-, 143
Verteilungen
    stabile, 159, 180
Verteilungsfunktion, 50, 75, 87
    stabile, 155
VTE-Modell, 260

Wachstumsprozeß, 251
Wachstumsprozesse, 164
Wahrnehmungsschwelle
    absolute, 242
Wahrscheinlichkeit
    induzierte, 72

totale, 56
Wahrscheinlichkeitsfunktion
bedingte, 82
Wahrscheinlichkeitsmaß, 71
Wahrscheinlichkeitsraum, 48
Wartezeit
zwischen Unfällen, 242
Weibull-Verteilung, 165, 247
Weibullverteilung, 213

Yule-Furry-Prozeß, 252

Zählmaß, 40
Zählprozeß, 238
Zahl
irrationale, 29
natürliche, 29
rationale, 29
reelle, 29
Ziegenproblem, 62
Ziehen
mit Zurücklegen, 124
mit Zurücklegen, 10, 68
ohne Zurücklegen, 10
zufällige Veränderliche, 71
zufälliger Vektor, 80, 81
zufälliges Ereignis, 45
Zufälligkeit
reine, 148, 152, 240, 241
Zufallspunkte, 248
Zufallsvariable, 71
diskrete, 72
mehrdimensionale, 81
stetige, 72, 87
Zusammenleben
friedliches, 247
Zustand, 259
absorbierender, 261
ergodischer, 261
transienter, 261
Zwischenzeiten, 239
Zwischenzeiten, Verteilung der, 240