

Ulrich Krengel

Einführung in die Wahrscheinlichkeits- theorie und Statistik

Für Studium,
Berufspraxis
und Lehramt

5. Auflage



vieweg studium – Aufbaukurs Mathematik

Ulrich Krengel

Einführung in die Wahrscheinlichkeits- theorie und Statistik

5., neubearbeitete und erweiterte Auflage



Prof. Dr. Ulrich Krengel

Institut für Mathematische Stochastik

Universität Göttingen

Lotzestraße 13

D-37083 Göttingen

Die Deutsche Bibliothek – CIP-Einheitsaufnahme

Ein Titeldatensatz für diese Publikation ist bei

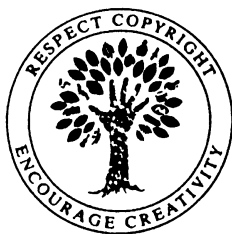
Der Deutschen Bibliothek erhältlich

1. Auflage 1988
- 2., verbesserte Auflage 1990
- 3., erweiterte Auflage 1991
- 4., erweiterte Auflage 1998
- 5., neubearbeitete und erweiterte Auflage 2000

Alle Rechte vorbehalten

© Friedr. Vieweg & Sohn Verlagsgesellschaft mbH, Braunschweig/Wiesbaden, 2000

Der Verlag Vieweg ist ein Unternehmen der Bertelsmann Fachinformation GmbH.



Das Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt. Jede Verwertung außerhalb der engen Grenzen des Urheberrechtsgesetzes ist ohne Zustimmung des Verlags unzulässig und strafbar. Das gilt insbesondere für Vervielfältigungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Verarbeitung in elektronischen Systemen.

<http://www.vieweg.de>

Konzeption und Layout des Umschlags: Ulrike Weigel, www.CorporateDesignGroup.de

Satz des Textes: Erich Berger, Göttingen

Druck und buchbinderische Verarbeitung: Lengericher Handelsdruckerei, Lengerich

Gedruckt auf säurefreiem Papier

Printed in Germany

ISBN 3-528-47259-6

(Paperback)

Vorwort

Stochastik ist die Mathematik des Zufalls. Sie ist von größter Bedeutung für die Berufspraxis der Mathematiker. An vielen Schulen hat sie ihren festen Platz gefunden.

Die beiden Hauptgebiete der Stochastik sind Wahrscheinlichkeitstheorie und Statistik. In der Wahrscheinlichkeitstheorie untersucht man zufällige Prozesse mit festen als bekannt angenommenen steuernden Wahrscheinlichkeiten. Dies ist theoretisch und praktisch von eigenständigem Interesse. Darüber hinaus liefert die Wahrscheinlichkeitstheorie Grundlagen für die Statistik, in der aus beobachteten Daten Schlüsse über unbekannte Wahrscheinlichkeiten und über zweckmäßiges Verhalten gezogen werden sollen.

Stochastische Fragen treten in den unterschiedlichsten Problemkreisen auf. Hier einige Beispiele:

- Was sind gute Strategien bei Glücksspielen und anderen Entscheidungsprozessen unter Unsicherheit?
- Welche Wahrscheinlichkeitsaussagen lassen sich über das Wachstum von Populationen und über die Vererbung von Eigenschaften machen?
- Wie übermittelt man ökonomisch Nachrichten?
- Wie vergleicht man mit vorgegebener Sicherheit die Qualität von Heilmitteln oder Produktionsverfahren?
- Was lässt sich über die Genauigkeit von Messungen aussagen?

Dies sind Fragen, die sich nicht ohne Zusatzüberlegungen nur durch den Beweis mathematischer Sätze beantworten lassen. Ein wesentlicher Teil der Schwierigkeit besteht bereits darin, die passenden mathematischen Begriffe zu entwickeln, die es erlauben, diese „realen“ Fragen angemessen mathematisch auszudrücken. Die für Berufspraxis und Schule gleichermaßen wichtige Umsetzung von realen Problemen in eine adäquate theoretische Form kann man wohl nirgends besser üben als in der Stochastik. Die Übungsaufgaben, die oft von der „eingekleideten“ Art sind, sind dabei äußerst wichtig. Der Leser sollte so viele wie möglich lösen.

Ich habe versucht, ein wenig von der Faszination zu vermitteln, die Stochastik ausüben kann. Dies war mir wichtiger als eine möglichst vollständige Abhandlung der praktisch gebräuchlichen Verfahren. Ist das Interesse geweckt, kann ja der Leser weitere Literatur heranziehen. Immerhin wird aber ein gewisser Fundus der Methodenlehre vermittelt, und ich denke, dass der Leser, der hier die Grundideen verstanden hat, sich schnell in systematischere Darstellungen und Handbücher hineinfinden wird.

Das Buch wendet sich an Studenten der Mathematik, der Physik und der Informatik vom dritten Semester an. Es setzt nur Grundkenntnisse aus der Analysis und der linearen Algebra voraus. Nur in einigen späteren Abschnitten würde man eigentlich ein wenig Maßtheorie brauchen. Die Aussagen lassen sich aber auch ohne solche weiter gehenden Vorkenntnisse verstehen, wenn man bereit ist, auf einzelne Beweise (vor allem von Existenzsätzen) zu verzichten. Diese sind in vertiefenden Vorlesungen leicht nachzuholen.

Das Buch enthält mehr Stoff als man bei angemessenem Tempo in einer vierstündigen Vorlesung vermitteln kann. Dies gibt Wahlmöglichkeiten. Die relativ zahlreichen mit einem Stern versehenen Abschnitte, Sätze und Beispiele und die Anhänge können am leichtesten weggelassen werden. Jedenfalls werden sie später nicht unbedingt benötigt.

Allerdings sind darunter viele Rosinen, so dass vieles dafür spricht, lieber einen Teil des Kuchens nicht zu essen.

Vieles aus den ersten Paragraphen ist Schulstoff. Weil Anfänger mit der mathematischen Modellierung realer Experimente oft Schwierigkeiten haben, scheint mir eine ausführliche Darstellung nicht nur für die zukünftigen Lehrer sinnvoll.

Man kann im Prinzip den gesamten Statistikteil auf eine spätere Lehrveranstaltung verschieben, aber Mathematikstudenten mit anderen Studienschwerpunkten und Physikstudenten fehlt oft die Zeit, eine solche zu besuchen.

Wie bei Lehrbüchern üblich habe ich die Quellen in der Regel nicht genannt. Es gibt aber historische Hinweise und Hinweise zum Weiterlesen.

Ich möchte den vielen Mitarbeitern und Freunden herzlich danken, die bei der Entstehung dieses Buches geholfen haben. Petra Küster hat schon bei der Ausarbeitung des Skripts mitgewirkt, das als Grundlage diente. Aus Vorlesungsnotizen von Götz Kersting habe ich manche Anregung geschöpft. Erich Berger, Wolfgang Stadje, Götz Kersting, Uwe Rösler, Hans-Jürgen Döring, Ulrich Wacker, Catherine Pallenberg, Norbert Neumann, Herold Dehling und Heinrich Hering haben Teile des Manuskripts gelesen, und ihre Vorschläge haben zu wesentlichen Verbesserungen geführt. Michael Krawczak hat das schöne Titelbild beigetragen. Das Manuskript haben Frau Schrörs, Frau Zimmer, Frau Graupner, Frau Giesecking und Frau Steffen sehr einwandfrei getippt. Dem Vieweg-Verlag, insb. Frau Schmickler-Hirzebruch, danke ich für die gute Zusammenarbeit. Meiner Frau danke ich für ihr Verständnis dafür, dass ich oft selbst sonntags so schwer vom Schreibtisch wegzukriegen war.

Die fünfte Auflage enthält u.a. einen zusätzlichen Paragraphen über Laufzeitanalysen für rekursive Algorithmen. Ich danke Herrn Uwe Rösler und Herrn Ludger Rüschendorf für wertvolle Hinweise zu diesem aktuellen Thema. Der Abschnitt über nichtparametrische Tests wurde deutlich erweitert.

Herrn Erich Berger danke ich für die sorgfältige Herstellung der neuen Druckvorlage mit LATEX und für unzählige Verbesserungsvorschläge, die ich gerne aufgegriffen habe.

Ich widme dieses Buch meinem Lehrer Konrad Jacobs, der mein Interesse an Stochastik geweckt hat und dem ich viel verdanke.

Göttingen, im Oktober 1999.

Ulrich Krengel

Inhaltsverzeichnis

<i>Kapitel I</i>	Diskrete Wahrscheinlichkeitsräume	1
§ 1	Modelle für Zufallsexperimente, Abzählmethoden	1
1.1	Endliche Wahrscheinlichkeitsräume	2
1.2	Einfache Urnenmodelle	6
1.3	Anwendungsbeispiele	10
1.4	Die hypergeometrische Verteilung	12
1.5	Vereinigungen von Ereignissen	12
1.6	Multinomialkoeffizienten	14
1.7	Runs	14
1.8	Einfache Identitäten für Binomialkoeffizienten	15
	Anhang	17
	Aufgaben	19
§ 2	Bedingte Wahrscheinlichkeit und Unabhängigkeit	21
2.1	Definition und Eigenschaften bedingter Wahrscheinlichkeiten . . .	21
2.2	Unabhängigkeit	25
2.3	Produktexperimente	27
2.4	Einige Verteilungen für Produktexperimente	29
2.5	Diskrete Wahrscheinlichkeitsräume	31
2.6	Konstruktion von Wahrscheinlichkeitsräumen aus bedingten Wahrscheinlichkeiten	32
2.7	Austauschbare Verteilungen	34
2.8	Genetische Modelle	35
2.9	Bedingte Wahrscheinlichkeit und Scheinkorrelation	37
	Anmerkungen	39
	Aufgaben	40
§ 3	Zufallsvariable, Erwartungswert, Varianz	42
3.1	Verteilungen von Zufallsvariablen	42
3.2	Unabhängigkeit	45
3.3	Erwartungswerte	46
3.4	Das Rechnen mit Indikatorfunktionen	49
3.5	Varianz und Kovarianz	52
3.6	Das schwache Gesetz der großen Zahlen	56
	Aufgaben	58
§ 4	Grundbegriffe der Schätztheorie	60
4.1	Der allgemeine Rahmen von Schätzproblemen	61
4.2	Maximum-Likelihood-Schätzer	62
4.3	Erwartungstreue	63
4.4	Der mittlere quadratische Fehler	65

4.5	Die Informationsungleichung	66
4.6	Konsistenz	68
4.7	Konfidenzintervalle	69
	Aufgaben	74
§ 5	Approximationen der Binomialverteilung	76
5.1	Approximation von $n!$ und $b_{n,p}(k)$	76
5.2	Der Satz von de Moivre-Laplace	80
5.3	Anwendungen	83
5.4	Die Poisson-Approximation	85
	Anhang	89
	Aufgaben	90
§ 6	Tests	92
6.1	Beispiel der „tea tasting Lady“	92
6.2	Grundbegriffe der Testtheorie	94
6.3	Mehr zur „tea tasting Lady“	95
6.4	Ein verfeinertes Modell für den Tee-Test	97
6.5	Beispiel des Testens der Existenz von außersinnlicher Wahrnehmung	99
6.6	Eine Erweiterung des Testbegriffs: Randomisierte Tests	100
6.7	Tests einfacher Hypothesen gegen einfache Alternativen	101
6.8	Anwendung auf zusammengesetzte Alternativen	103
6.9	Allgemeine Hinweise zur Testtheorie	103
6.10	p -Werte	104
	Aufgaben	105
§ 7	Erzeugende Funktionen	107
	Verzweigungsprozesse	111
	Aufgaben	113
§ 8	Entropie und Codierung	114
8.1	Der Quellen-Codierungssatz	114
8.2	Anwendung auf mehrstufige Zufallsexperimente	117
	Aufgaben	118
§ 9	Laufzeitanalysen von rekursiven Algorithmen	120
	Aufgaben	126
Kapitel II Allgemeine Modelle		128
§ 10	Wahrscheinlichkeitsmaße mit Dichten	128
10.1	σ -Algebren und allgemeine Wahrscheinlichkeitsmaße	128
10.2	Beispiele von Verteilungen mit Dichten	131
	Anhang	136
	Aufgaben	138

§ 11	Zufallsvariable und ihre Momente	140
11.1	Messbare Funktionen	140
11.2	Verteilungen von Zufallsvariablen	142
11.3	Unabhängigkeit	143
11.4	Erwartungswerte	145
11.5	Mehrdimensionale Dichtetransformation und Normalverteilung . .	147
	Aufgaben	151
§ 12	Grenzwertsätze	153
12.1	Das starke Gesetz der großen Zahlen	153
12.2	Normale Zahlen	157
12.3	Der Zentrale Grenzwertsatz	158
	Anhang	162
	Aufgaben	163
§ 13	Schätzverfahren und Fehlerrechnung	164
13.1	Maximum-Likelihood-Schätzungen bei Dichten	164
13.2	Konfidenzintervalle	166
13.3	Das Fehlerfortpflanzungsgesetz	167
13.4	Die Methode der kleinsten Quadrate	168
13.5	Median, Ausreißer und Robuste Schätzer	170
	Anhang	172
	Aufgaben	174
§ 14	Einige wichtige Testverfahren	175
14.1	Der t -Test	175
14.2	Einfache Varianzanalyse	180
14.3	χ^2 -Tests	182
14.4	Nichtparametrische Tests	187
	Anhang	192
	Aufgaben	194

Kapitel III **Markowsche Ketten**

195

§ 15	Die markowsche Eigenschaft	195
15.1	Definition und Beispiele	195
15.2	Einfache Folgerungen aus der markowschen Eigenschaft	197
15.3	Stationäre Übergangswahrscheinlichkeiten	198
15.4	Absorptionswahrscheinlichkeiten	200
15.5	Absorptionsverteilungen	201
	Aufgaben	203
§ 16	Das Verhalten markowscher Ketten in langen Zeiträumen	205
16.1	Ketten mit endlich vielen Zuständen	205
16.2	Kommunizierende Zustände und Periodizität	208

16.3	Rekurrenz und Transienz	210
	Anhang	215
	Aufgaben	216
§ 17	Der Erneuerungssatz	218
17.1	Die Erneuerungsgleichung	218
17.2	Anwendung auf Übergangswahrscheinlichkeiten	221
17.3	Bestimmung der m_{ii}	223
	Aufgaben	226
§ 18	Der Poisson-Prozess	227
18.1	Charakterisierung des Poisson-Prozesses	227
18.2	Sprungzeiten beim Poisson-Prozess	230
	Aufgaben	232
Hinweise zum Weiterlesen		234
Lösungen der mit (L) gekennzeichneten Aufgaben		236
Literaturverzeichnis		243
Tabellen		247
Symbolverzeichnis		252
Namen- und Sachwortverzeichnis		253

Kapitel I Diskrete Wahrscheinlichkeitsräume

Zunächst sollen die wichtigsten Grundbegriffe für solche Zufallsexperimente entwickelt werden, die nur endlich viele oder abzählbar viele mögliche Versuchsausgänge haben. Für sie benötigt man weniger theoretischen Hintergrund.

§ 1 Modelle für Zufallsexperimente, Abzählmethoden

Ziel der Wahrscheinlichkeitstheorie ist die Analyse der Gesetzmäßigkeiten, die bei der Beschreibung so genannter „Zufallsexperimente“ eine Rolle spielen. Darunter verstehen wir Experimente, deren Ausgänge nicht durch logische oder andere Gründe durch die Versuchsbedingungen determiniert sind. Wenigstens gedanklich sollten die Experimente unter den gleichen Bedingungen wiederholbar sein, und zwar so, dass der Versuchsausgang bei unabhängig angestellten Wiederholungen nicht notwendig stets der gleiche ist, sondern nur statistischen Regelmäßigkeiten folgt.

Beispiele

- Würfelexperimente
- Blindes zufälliges Ziehen von Kugeln oder Losen aus einer Urne
- Kartenspiele

Die gleichen Gesetzmäßigkeiten treten auch bei „Experimenten“ ganz anderen Typs auf, etwa

- Geburten (Junge oder Mädchen)
- Ermittlung der Anzahl der Ausschussexemplare in der Tagesproduktion einer Maschine
- Unfallstatistiken
- Registrierung von Lebensdauern
- Anzahl der Ausschläge eines Geiger-Zählers zur Messung von radioaktiver Strahlung.

Dagegen ist die Frage, wie „wahrscheinlich“ es ist, dass auf dem Mars einmal Leben existiert hat, von ganz anderer Art. Entweder ist die richtige Antwort ja oder nein, unabhängig davon, ob wir sie kennen oder nicht.

Um nun Gesetzmäßigkeiten in Zufallsexperimenten mathematisch untersuchen zu können, muss man zunächst *mathematische Modelle* dafür bilden, innerhalb derer sie sich rein mathematisch beschreiben und „definieren“ lassen. Diese Modellbildung ist ein Vorgang, der auch bei anderen mathematischen Disziplinen auftaucht, etwa in der Geometrie. Zum Beispiel ist eine Kugel für viele Zwecke ein passendes Modell für die Erde. Will man etwa eine Kirchturmhöhe aus einer Längenmessung und aus Winkelmessungen bestimmen, so ist ein Dreieck ein Modell.

Der Übergang von der Wirklichkeit zum Modell ist nie rein logisch begründbar. Er setzt in starkem Maß *Erfahrung* über die Natur des Experiments voraus. Das ist keine Besonderheit der Modelle für Zufallsexperimente. So genügt das ebene Modell der Erdoberfläche vollauf, wenn man eine Landkarte des Landkreises Göttingen herstellen will. Für feine geophysikalische Betrachtungen ist selbst das Modell der Erdkugel zu grob und man betrachtet im feineren Modell Abplattungen. Wir sehen daran auch gleich, dass die Wahl des Modells von der Zielsetzung mitbestimmt wird.

Wir halten also fest, dass es keinen *prinzipiellen* Unterschied zwischen den Rollen, die Modelle in Geometrie und Stochastik spielen, gibt. *Praktisch* scheint es aber Unterschiede zu geben, weil das passende Modell für Zufallsexperimente oft weniger offensichtlich ist. Man täuscht sich leicht darüber, ob ein Versuch wirklich hinreichend viele Symmetrien enthält, um die Annahme zu rechtfertigen, alle Versuchsausgänge seien gleichwahrscheinlich. Häufig lässt sich die Angemessenheit eines Modells nur empirisch prüfen, und das ist stets mit Unsicherheiten behaftet, die mit dem zufälligen Ausgang der Experimente zusammenhängen. Häufiger als in anderen Gebieten wird man schon aus Gründen der mathematischen Praktikabilität bewusst ein relativ grobes Modell in Kauf nehmen müssen. Gelegentlich wird man sogar ohne genügend Erfahrung über die Natur des Experiments ad hoc Modelle entwerfen, um gewisse Phänomene überhaupt erst einmal einer Rechnung zugänglich zu machen, und um zu vergleichen, welche qualitativen Auswirkungen verschiedene Modellannahmen haben.

Zeitweise haben Mathematiker sogar geglaubt, es läge im Wesen der zufälligen Erscheinungen, dass sie sich eben nicht mathematisieren ließen. Jedenfalls hat es — im Gegensatz z.B. zur Geometrie — bis in dieses Jahrhundert hinein gedauert, bis man eine gesicherte axiomatische Grundlegung gegeben hat. Andererseits macht gerade dieses Phänomen, dass man über Zufallsereignisse mathematisch rigorose Resultate beweisen kann, einen Reiz des Gebietes aus.

Ist man sich der Schwierigkeit der Modellbildung in der Stochastik bewusst, so wird man die Anwendbarkeit der abgeleiteten Resultate auf reale Probleme auch nachträglich noch überprüfen müssen. Eine Art Rückkopplung kann hilfreich sein. Wenn die abgeleiteten Resultate nicht gut mit der Erfahrung übereinstimmen, wird das Modell revidiert werden müssen.

Die Frage der Modellbildung wird für uns ein wiederkehrendes Leitthema sein, das mit der Entwicklung der mathematischen Theorie in Wechselwirkung steht.

1.1 Endliche Wahrscheinlichkeitsräume

Beginnen wir mit einer kleinen Aufgabe: Es sei die Wahrscheinlichkeit dafür zu bestimmen, dass die Summe der bei zwei Würfeln eines Würfels erhaltenen Augenzahlen mindestens 10 ist.

Wir können das Ergebnis des aus zwei Würfeln bestehenden Zufallsexperiments durch das Paar (i, k) der beobachteten Augenzahlen beschreiben. Z.B. bedeutet $(4, 1)$, dass der

erste Wurf eine 4 und der zweite eine 1 ist. Die Menge $\Omega = \{(i, k) : 1 \leq i, k \leq 6\}$ aller möglichen Ergebnisse hat 36 Elemente. Aus Symmetriegründen ist es nahe liegend, sie als gleichwahrscheinlich anzusehen. Jedes $(i, k) \in \Omega$ hat also die Wahrscheinlichkeit $1/36$. Die Menge der Ergebnisse, für die die Summe $i + k$ der Augenzahlen mindestens 10 ist, ist

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

Da A sechs Elemente hat, ist die gesuchte Wahrscheinlichkeit also $6/36 = 1/6$.

Wir wollen auch Experimente betrachten, für die nicht alle möglichen Ergebnisse gleichwahrscheinlich sind. Werfen wir eine Reißzwecke auf den Fußboden, so landet sie entweder mit der Spitze nach oben oder mit der Spitze schräg nach unten. Bezeichnen wir das erste dieser möglichen Ergebnisse mit o und das zweite mit u , so ist $\Omega = \{o, u\}$ eine Beschreibung der Ergebnismenge. Die Annahme, o und u seien gleichwahrscheinlich, wäre sehr unplausibel. Man könnte versuchen, die Wahrscheinlichkeit p von o durch Ausführung vieler Würfe zu schätzen.

Allgemein ordnen wir einem Zufallsexperiment mit endlich vielen möglichen Ausgängen eine endliche nichtleere Menge Ω zu, deren Elemente ω die Versuchsausgänge bezeichnen. Sie heißen **Ergebnisse** oder oft auch **Stichproben**, **Realisierungen** oder **Elementarereignisse**. Ω heißt **Ergebnismenge** oder **Stichprobenraum**, gelegentlich auch Grundraum oder Ergebnisraum.

Die Teilmengen von Ω sind die **Ereignisse**, die in unserem Modell in Betracht gezogen werden. Genauer: Wir identifizieren $A \subset \Omega$ mit dem Ereignis, dass ein $\omega \in A$ der beobachtete Versuchsausgang ist. Diese Konvention gestattet es, mengentheoretische Notationen einzusetzen. So bedeutet $A \cap B$ das Ereignis, dass sich A und B ereignen, denn $\omega \in A \cap B$ besagt, dass ω in A und in B liegt. Offenbar ist $A \cup B$ das Ereignis, dass sich A oder B ereignet. (Das Wort „oder“ ist immer im nicht ausschließenden Sinn zu verstehen.) Das Komplement A^c von A in Ω bezeichnet das Ereignis, dass A nicht geschieht. Ereignisse A und B heißen **unvereinbar**, wenn die Mengen A und B disjunkt sind, d.h. wenn ihr Durchschnitt $A \cap B$ die leere Menge $\emptyset$ ist. $\emptyset$ heißt auch das **unmögliche Ereignis**. Ω heißt das **sichere Ereignis**.

Nun müssen den Ereignissen noch Wahrscheinlichkeiten zugeordnet werden. Die Menge der Ereignisse ist mengentheoretisch die Potenzmenge $\mathcal{P}(\Omega)$, d.h. die Menge aller Teilmengen von Ω . Eine Abbildung P von $\mathcal{P}(\Omega)$ in $[0, 1]$ heißt **Wahrscheinlichkeitsverteilung** oder **Wahrscheinlichkeitsmaß**, wenn sie die folgenden Eigenschaften hat:

$$P(\Omega) = 1; \tag{1.1}$$

$$P(A) \geq 0 \quad \text{für alle } A; \tag{1.2}$$

$$P(A \cup B) = P(A) + P(B) \quad \text{für alle disjunkten } A, B. \tag{1.3}$$

Ist (1.1) erfüllt, so nennt man P **normiert**. Die letzte der obigen drei Eigenschaften heißt **Additivität** von P . $P(A)$ heißt **Wahrscheinlichkeit** von A . Das Paar (Ω, P) heißt der dem Experiment zugeordnete **Wahrscheinlichkeitsraum**. Er dient uns als Modell für das reale Experiment. Die Wahl von Ω ist oft nahe liegend. Die „richtige“ Wahl von P ist nur in wenigen Fällen klar und wird uns noch viel beschäftigen. Die oben genannten „Axiome“ (1.1)–(1.3) stecken nur einen Rahmen ab.

Indem wir Wahrscheinlichkeiten nur im Modell definiert haben, sind wir der Frage, was Wahrscheinlichkeiten realer Ereignisse sind, aus dem Weg gegangen. Man versteht wohl allgemein die Wahrscheinlichkeit eines realen Ereignisses als Maß für die Sicherheit oder Häufigkeit, mit der es bei wiederholter Ausführung des Experiments auftreten wird.

Sei etwa A beim Würfeln mit einem Würfel das Ereignis, dass die geworfene Augenzahl höchstens 4 ist. Wirft man den Würfel n -mal, so unterscheiden wir zwischen der **absoluten Häufigkeit** $k_n(A)$, also der Zahl der Würfe unter diesen n Würfeln, bei denen A auftritt, und der **relativen Häufigkeit** $h_n(A) = k_n(A)/n$. Empirisch beobachtet man, dass sich die $h_n(A)$ für sehr große n einem Grenzwert $P(A)$ annähern. (Konvergenz lässt sich natürlich empirisch nie prüfen, da wir dazu eine unendliche Folge von Würfeln durchführen müssten.) Nehmen wir an, dass für alle Ereignisse A Konvergenz von $h_n(A)$ gegen $P(A)$ vorliegt, so überlegt man sich leicht, dass P die Eigenschaften (1.1)–(1.3) haben muss.

Man hat daher Versuche unternommen, Wahrscheinlichkeiten mit Hilfe von relativen Häufigkeiten zu definieren. Aber dies führt zu großen Schwierigkeiten. Relative Häufigkeiten hängen von der konkret beobachteten Folge von Würfeln ab und sind deshalb zufällig. Begrifflich sind damit relative Häufigkeiten sehr verschieden von Wahrscheinlichkeiten. Wir werden später, in § 3, umgekehrt einen Zusammenhang von Wahrscheinlichkeiten und relativen Häufigkeiten herstellen, indem wir ausgehend von dem axiomatischen Modell Konvergenz von relativen Häufigkeiten in geeignetem Sinn beweisen. Dafür ist es aber jetzt noch zu früh.

Nun zurück zu unserem mathematischen Modell. Wir ersparen uns den ziemlich trivialen Beweis der folgenden

Eigenschaften von P : Für $A, B, A_i \in \mathcal{P}(\Omega)$ gilt

$$P(A^c) = 1 - P(A), \quad \text{speziell} \quad P(\emptyset) = 0; \quad (1.4)$$

$$A \subset B \Rightarrow P(A) \leq P(B) \quad (\Rightarrow \text{bezeichnet die logische Implikation}); \quad (1.5)$$

$$P(A \setminus B) = P(A) - P(A \cap B) \quad (\text{mit } A \setminus B = A \cap B^c); \quad (1.6)$$

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i), \quad \text{falls } A_1, \dots, A_n \text{ paarweise disjunkt sind}; \quad (1.7)$$

$$P\left(\bigcup_{i=1}^n A_i\right) \leq \sum_{i=1}^n P(A_i) \quad \text{für beliebige } A_1, \dots, A_n; \quad (1.8)$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \quad (1.9)$$

Aus (1.7) folgt

$$P(A) = \sum_{\omega \in A} P(\{\omega\}). \quad (1.10)$$

Die Wahrscheinlichkeit eines Ereignisses A ist die Summe der Wahrscheinlichkeiten der Ergebnisse, bei denen A eintritt. P ist also durch die Werte aller $P(\{\omega\})$ mit $\omega \in \Omega$ bestimmt. Wir schreiben auch $P(\omega)$ statt $P(\{\omega\})$. Die Abbildung $\omega \mapsto P(\omega)$ heißt **Wahrscheinlichkeitsfunktion**. Die Wahrscheinlichkeitsverteilung kann also durch Angabe der zugehörigen Wahrscheinlichkeitsfunktion beschrieben werden. Offenbar gilt für eine Wahrscheinlichkeitsfunktion stets

$$P(\omega) \geq 0 \text{ für alle } \omega \in \Omega, \text{ und } \sum_{\omega \in \Omega} P(\omega) = 1. \quad (1.11)$$

Umgekehrt definiert jede Abbildung $\omega \mapsto P(\omega)$ mit diesen Eigenschaften gemäß (1.10) ein eindeutig bestimmtes Wahrscheinlichkeitsmaß. Eine Wahrscheinlichkeitsfunktion kann daher auch als Abbildung $\omega \mapsto P(\omega)$ mit den Eigenschaften (1.11) definiert werden. Auch der Name **Wahrscheinlichkeitsvektor** ist gebräuchlich. In vielen Fällen ist nämlich Ω von der Form $\{1, 2, \dots, n\}$, und P durch einen Vektor $(p_1, \dots, p_n)$ mit $p_i \geq 0$ und $\sum_{i=1}^n p_i = 1$ gegeben, indem man $P(i) = p_i$ setzt.

Für ein Experiment gibt es im Allgemeinen mehrere mögliche Modelle. Die Auswahl des Modells richtet sich nach Kriterien wie Einfachheit, Bestimmbarkeit der Wahrscheinlichkeiten, usw. In dem eingangs gegebenen Beispiel mit den zwei Würfeln eines Würfels könnte man z.B. auch daran denken, die Menge $\{2, 3, \dots, 12\}$ als Ergebnismenge zu wählen, denn die Zahlen $2, \dots, 12$ sind die möglichen Augensummen, und nur nach diesen ist gefragt. Allerdings wären dann die zugehörigen Wahrscheinlichkeiten $P(2), \dots, P(12)$ nicht alle gleich, und die einfachste Art sie zu bestimmen wäre wohl ein Rückgriff auf das oben verwendete Modell.

Wie findet man nun die Wahrscheinlichkeiten $P(\omega)$? Dies ist von Beispiel zu Beispiel verschieden. Wir werden drei verschiedene Ansätze kennen lernen:

(1) *Kombinatorische Bestimmung* von $P(\omega)$ durch Abzählen von Ω . Dies ist möglich, wenn man z.B. auf Grund von Symmetrie-Erwägungen davon ausgehen kann, dass alle $\omega \in \Omega$ gleichwahrscheinlich sind. P heißt dann die **Gleichverteilung** auf Ω , und (Ω, P) wird in diesem Fall als **laplacescher Wahrscheinlichkeitsraum** bezeichnet. Bezeichnet $\text{card}(\Omega)$ die Anzahl der Elemente von Ω , so ist in diesem Fall

$$P(\omega) = \frac{1}{\text{card}(\Omega)}$$

für alle $\omega \in \Omega$, und

$$P(A) = \frac{\text{card}(A)}{\text{card}(\Omega)} \quad (1.12)$$

die **Laplace-Wahrscheinlichkeit** von A . Das zugehörige Zufallsexperiment wird **Laplace-Experiment** genannt.

Wenn man davon spricht, dass ein Element einer Menge Ω *zufällig* ausgewählt wird, meint man — sofern nichts anderes gesagt wird — dass alle ω gleichwahrscheinlich sein sollen.

(2) *Statistische Schätzung*. Z.B. findet man Schätzungen von Lebensdauern durch Sterbetafeln. Die Wahrscheinlichkeiten der Geburten von Jungen bzw. Mädchen und die Wahrscheinlichkeit, im zweiten Jahr nach Erwerb des Führerscheins einen Unfall zu verursachen, kann man statistisch schätzen. Zur Diskussion dieser Verfahren werden wir theoretische Überlegungen mit heranziehen müssen.

(3) *Logische Überlegung und Kombination statistischer Methoden mit mathematischen Ableitungen*. Oft lässt sich durch eine Überlegung die Form von Wahrscheinlichkeitsverteilungen bis auf unbekannte Parameter bestimmen. Diese müssen dann noch geschätzt werden.

Am einfachsten sind die Laplace-Experimente. Daher beginnen wir damit. Ein historisch wichtiges Beispiel soll aber vorher zeigen, dass die Gleichwahrscheinlichkeit manchmal erst durch passende Definition der Ergebnisse herbeigeführt werden muss.

Beispiel 1.1 (*Spielabbruch*) Zwei Spieler spielen eine Folge von Spielen, wobei jedes Mal jeder der beiden mit Wahrscheinlichkeit $1/2$ gewinnt, unabhängig davon wie die anderen Spiele ausgehen. Beide bringen den gleichen Einsatz ein und vereinbaren, dass derjenige die Gesamtsumme erhält, der als Erster 10 Spiele gewonnen hat. Nach 15 Spielen habe Spieler A 8 Spiele gewonnen und Spieler B 7 Spiele. Durch ein unvorhergesehenes Ereignis werden die beiden gezwungen, ihre Spielfolge abzubrechen. Wie ist das Geld aufzuteilen?

Da in den Spielregeln der Spielabbruch nicht vorgesehen war, könnte wohl Spieler B verlangen, den vollen Einsatz zurückzuerhalten oder das Spiel später zu beenden. Nehmen wir aber an, man einigt sich darauf, den Einsatz entsprechend den Wahrscheinlichkeiten zu verteilen, mit denen das Weiterspielen zum Erfolg führen würde. Wie groß sind diese?

Ansatz 1: Spätestens nach vier weiteren Spielen wäre alles entschieden. Schreiben wir z.B. $ABBA$ für den Fall, dass Spieler A das erste und letzte dieser vier Spiele gewinnt und Spieler B die beiden anderen, so wären die folgenden Spielverläufe (Ergebnisse) möglich:

$AAAA$	$AABB$	$ABBB$
$AAAB$	$ABAB$	$BABB$
$AABA$	$ABBA$	$BBAB$
$ABAA$	$BAAB$	$BBBA$
$BAAA$	$BABA$	$BBBB$
	$BBAA$	

Da bei den 11 Spielverläufen der ersten beiden Spalten der Spieler A gewinnt, und bei den 5 der letzten Spalte der Spieler B , und da aus Symmetriegründen die obigen Spielverläufe gleichwahrscheinlich sind, müsste Spieler A $11/16$ des Geldes erhalten und Spieler B $5/16$. So argumentierte Pascal 1654 in einem Brief an Fermat.

Ansatz 2: Man kann einwenden, dass diese Spielverläufe gar nicht eintreten, da meist schon vorher ein Spieler genug Siege beisammen hat. Die wirklich möglichen Fortsetzungen wären

AA	$ABBB$
ABA	BBB
$ABBA$	$BABB$
BAA	$BBAB$
$BBAA$	
$BABA$	

Hier gewinnt A in den 6 Spielen der ersten Spalte, und B in den 4 Spielen der zweiten Spalte. Müsste A also $6/10$ des Geldes erhalten? Das wäre richtig, wenn hier alle Spielverläufe gleichwahrscheinlich wären. Es ist aber $P(\omega) = 1/4$ für $\omega = AA$ und $P(\omega') = 1/16$ für $\omega' = BBAB$. Wenn man die unterschiedlichen Wahrscheinlichkeiten berücksichtigt, kommt man auf die gleiche Lösung wie oben. Durch die künstliche Verlängerung der Folgen ist es in Ansatz 1 gelungen, den gleichwahrscheinlichen Fall herbeizuführen.

1.2 Einfache Urnenmodelle

Zahlreiche Abzählprobleme lassen sich auf kombinatorische Formeln zurückführen, die wir am Beispiel von mehrfachen Ziehungen aus einer Urne kennen lernen wollen. In einer Urne seien N Kugeln, die wir uns mit $1, 2, \dots, N$ nummeriert denken. Sukzessive werden insgesamt n Kugeln zufällig gezogen.

Wir sprechen von einer Stichprobe **mit Zurücklegen**, wenn Kugeln mehrfach gezogen werden dürfen, also nach jeder Ziehung die gezogene Kugel zurückgelegt wird und neu geschüttelt wird bevor die nächste Kugel gezogen wird. Kann jede Kugel nur einmal auftreten, so spricht man von einer **Stichprobe ohne Rücklegen**. (Statt „mit Rücklegen“ und „ohne Rücklegen“ sagt man auch „mit Wiederholung“ und „ohne Wiederholung“, obwohl es zu Mehrfachziehungen irgendwelcher Kugeln natürlich beim Rücklegen nicht kommen *muss*.)

Man kann das Ergebnis der Folge der Ziehungen dadurch beschreiben, dass man das n -Tupel $(\omega_1, \omega_2, \dots, \omega_n)$ angibt, in dem ω_i die Nummer der bei der i -ten Ziehung gezogenen Kugel ist. Hierbei kommt es also auf die Reihenfolge des Erscheinens der einzelnen Kugeln an. Wir sprechen dann von einer **Stichprobe in Reihenfolge**. $(1, 3, 4, 1)$ und $(1, 1, 3, 4)$ werden dann als verschiedene Ergebnisse betrachtet. Soll dagegen nur unterschieden werden wie oft die einzelnen Kugeln in der Stichprobe erscheinen, so spricht man von einer **Stichprobe ohne Reihenfolge**. Die beiden eben genannten Folgen wären dann das gleiche Ergebnis $\{1, 1, 3, 4\} = \{1, 3, 4, 1\}$. Wir verwenden dann geschweifte Klammern wie bei Mengen. Hier kommt es aber auf die Vielfachheit der Elemente an: $\{1, 3, 4\}$ ist von $\{1, 1, 3, 4\}$ verschieden.

[Wieder sind auch andere Bezeichnungen in Gebrauch: Stichproben in Reihenfolge werden auch *geordnete Stichproben* genannt. Stichproben in Reihenfolge ohne Rücklegen heißen auch *Permutationen* von n aus N Elementen, und Stichproben ohne Reihenfolge ohne Rücklegen auch *Kombinationen* von n aus N Elementen.]

Durch die eben genannten Unterscheidungen ergeben sich vier verschiedene Stichprobenräume, deren Elemente wir nun zählen wollen. Sei $\mathbb{A} = \{1, 2, \dots, N\}$.

I) Stichproben in Reihenfolge mit Zurücklegen: Als Stichprobenraum kann

$$\Omega_I := \{\omega = (\omega_1, \dots, \omega_n) : \omega_i \in \mathbb{A} \text{ für } i = 1, \dots, n\} = \mathbb{A}^n$$

gewählt werden. Offenbar ist $\text{card}(\Omega_I) = N^n$.

II) Stichproben in Reihenfolge ohne Zurücklegen: Als Stichprobenraum dient

$$\Omega_{II} := \{\omega = (\omega_1, \dots, \omega_n) : \omega_i \in \mathbb{A}, \omega_i \neq \omega_j \text{ für } i \neq j \ (1 \leq i, j \leq n)\}.$$

Um die Anzahl der Elemente von Ω_{II} zu bestimmen, bedienen wir uns eines Abzählprinzips, das gleich etwas allgemeiner formuliert wird als wir es hier brauchen.

Abzählprinzip: Sei Ω eine Menge von n -Tupeln $\omega = (\omega_1, \omega_2, \dots, \omega_n)$, die wir als Ergebnisse eines aus n Telexperimenten (*Stufen*) bestehenden Experiments auffassen, wobei ω_i das Ergebnis der i -ten Stufe ist. Für das erste Telexperiment gebe es k_1 mögliche Ausgänge. Für jedes i sei die Zahl k_i der möglichen Ausgänge des i -ten Telexperiments unabhängig davon wie die früheren Telexperimente ausgegangen sind. Dann ist

$$\text{card}(\Omega) = k_1 \cdot k_2 \cdot \dots \cdot k_n.$$

Beweis: Simple Induktion. □

Beim Ziehen von Kugeln ohne Zurücklegen ist $k_1 = N$, da zuerst noch alle Kugeln in der Urne sind. Nach der ν -ten Ziehung liegt $(\omega_1, \omega_2, \dots, \omega_\nu)$ fest, und für jede Wahl von $(\omega_1, \omega_2, \dots, \omega_\nu)$ gibt es noch $k_{\nu+1} = N - \nu$ Kugeln zur Auswahl für die $(\nu + 1)$ -te Ziehung. Welche dies sind, hängt von $(\omega_1, \omega_2, \dots, \omega_\nu)$ ab, aber nicht ihre Zahl. Es folgt

$$\text{card}(\Omega_{II}) = N \cdot (N - 1) \cdot \dots \cdot (N - n + 1).$$

Das aus n Faktoren bestehende Produkt auf der rechten Seite wird meist abgekürzt durch die Notation

$$(N)_n := N(N-1) \cdot \dots \cdot (N-n+1). \quad (1.13)$$

Man kann nicht mehr Kugeln ohne Rücklegen ziehen als in der Urne enthalten sind. In der Definition von $(N)_n$ kommt dies auch zum Ausdruck. Für $n > N$ ist $(N)_n = 0$.

Spezialfall $N = n$: Dann sind am Schluss alle Kugeln gezogen und Ω_{II} besteht gerade aus der Menge der **Permutationen** von $\{1, \dots, N\}$. Mit der üblichen Notation $N! = 1 \cdot 2 \cdot 3 \cdot \dots \cdot N$ ist $(N)_N = N!$, und wir erhalten den bekannten Satz, dass $N!$ die Zahl der Permutationen einer Menge der Mächtigkeit N ist, als Spezialfall wieder.

III) Stichproben ohne Reihenfolge ohne Rücklegen: Jede solche Stichprobe ist dadurch beschrieben, dass man angibt, welche Kugeln darin vorkommen. Der zugehörige Stichprobenraum Ω_{III} ist demnach gerade die Menge der Teilmengen der Mächtigkeit n aus $\mathbb{A} = \{1, \dots, N\}$, also

$$\Omega_{III} = \{ \{ \omega_1, \omega_2, \dots, \omega_n \} : \omega_i \in \mathbb{A}, \omega_i \neq \omega_j \text{ für } i \neq j \}.$$

Wir können Ω_{III} auch beschreiben, indem wir eine Äquivalenzrelation in Ω_{II} einführen:

$$(\omega_1, \omega_2, \dots, \omega_n) \sim (\omega'_1, \omega'_2, \dots, \omega'_n)$$

gelte genau dann, wenn es eine Permutation π von $\{1, \dots, n\}$ gibt mit $\omega'_i = \omega_{\pi i}$ für $i = 1, \dots, n$. Dabei sei πi das Bild von i unter der Permutation π . Die Elemente von Ω_{III} sind dann die Äquivalenzklassen. Jede Äquivalenzklasse kann z.B. durch den Repräsentanten $(\omega_1, \omega_2, \dots, \omega_n)$ mit $\omega_1 < \omega_2 < \dots < \omega_n$ beschrieben werden.

Da jede der Äquivalenzklassen $n!$ Elemente hat, ist $\text{card}(\Omega_{III}) = n! \cdot \text{card}(\Omega_{II})$. Daraus folgt

$$\text{card}(\Omega_{III}) = \frac{(N)_n}{n!}.$$

Wir schreiben für die rechte Seite auch $\binom{N}{n}$ (gelesen: N über n). Schreibt man $(N)_n$ in der Form $N!/(N-n)!$, so ergibt sich

$$\binom{N}{n} = \frac{N!}{n!(N-n)!} = \binom{N}{N-n} \quad (1 \leq n \leq N). \quad (1.14)$$

Im Fall $n > N$ ist wegen $(N)_n = 0$ auch $\binom{N}{n} = 0$. $\binom{N}{n}$ ist die Anzahl der n -elementigen Teilmengen einer Menge der Mächtigkeit N . Da jede Menge genau eine 0-elementige Teilmenge hat, nämlich die leere Menge $\emptyset$, setzen wir $\binom{N}{0} = 1$. Indem wir noch $0! = 1$ setzen, erreichen wir, dass (1.14) auch für $n = 0$ gilt.

[Eine französische Fabel berichtet von einem Schäfer, der seine Schafe besonders schnell zählt, indem er deren Beine zählt und durch 4 teilt. Nennt man zwei Beine äquivalent, wenn sie zum gleichen Schaf gehören, so entspricht jedem Schaf eine Äquivalenzklasse. Wir haben also das „Prinzip des Schäfers“ angewandt, um unsere Äquivalenzklassen zu zählen.]

IV) Stichproben ohne Reihenfolge mit Rücklegen: Als Stichprobenraum Ω_{IV} können wir die Menge der Äquivalenzklassen unter der obigen Äquivalenzrelation in Ω_I nehmen. Indem wir aus jeder Äquivalenzklasse den Repräsentanten mit $\omega_1 \leq \omega_2 \leq \dots \leq \omega_n$ auswählen, sehen wir, dass sich Ω_{IV} auch als die Menge

$$\{(\omega_1, \omega_2, \dots, \omega_n) \in \mathbb{A}^n : \omega_1 \leq \omega_2 \leq \dots \leq \omega_n\}$$

beschreiben lässt. Da nun die Äquivalenzklassen nicht gleich mächtig sind, ist das Prinzip des Schäfers nicht anwendbar.

Wir ordnen den Elementen $\omega = (\omega_1, \dots, \omega_n)$ von Ω_{IV} die Folgen $(\omega'_1, \dots, \omega'_n)$ mit $\omega'_i = \omega_i + i - 1$ zu. Durch diese Zuordnung wird Ω_{IV} bijektiv auf die Menge

$$\Omega'_{III} = \{(\omega'_1, \dots, \omega'_n) \in \mathbb{B}^n : \omega'_1 < \omega'_2 < \dots < \omega'_n\}$$

mit $\mathbb{B} = \{1, 2, \dots, N + n - 1\}$ abgebildet. Nach Fall III folgt

$$\text{card}(\Omega_{IV}) = \text{card}(\Omega'_{III}) = \binom{N + n - 1}{n}.$$

Eine alternative Interpretation: Bevor wir Anwendungen diskutieren, wollen wir sehen, dass wir gleich noch vier weitere Abzählprobleme mitgelöst haben, indem wir nur die Stichproben anders interpretieren:

Wir fragen nach der Anzahl der Möglichkeiten n Murmeln auf N **Zellen** genannte Plätze zu verteilen. Sind die Murmeln **unterscheidbar**, so ist eine Verteilung dadurch beschrieben, dass man für jedes i mit $1 \leq i \leq n$ die Nummer ω_i der Zelle angibt, in die man die i -te Murmel platziert hat. Eine Verteilung ist also beschrieben durch ein n -Tupel $\omega = (\omega_1, \dots, \omega_n)$ mit $1 \leq \omega_i \leq N$. Die Menge der Ergebnisse (Verteilungen) ist also wieder Ω_I , aber jetzt mit der Uminterpretation

Nummer der Ziehung $\sim$ Nummer der Murmel,
 Nummer der Kugel $\sim$ Nummer der Zelle.

Sind die Murmeln **ununterscheidbar**, so können wir zwischen Permutationen von $(\omega_1, \dots, \omega_n)$ nicht mehr unterscheiden. Diese werden also identifiziert. Ununterscheidbare Murmeln entsprechen Stichproben ohne Reihenfolge, während unterscheidbare Murmeln Stichproben in Reihenfolge entsprechen.

Stichproben vom Umfang n aus $\{1, \dots, N\}$	mit Rücklegen	ohne Rücklegen	
in Reihenfolge	I N^n	II $(N)_n$	unterscheidbare Murmeln
ohne Reihenfolge	IV $\binom{n + N - 1}{n}$	III $\binom{N}{n}$	ununterscheid- bare Murmeln
	mit Mehrfachbe- setzung	ohne Mehrfach- besetzung	Verteilungen von n Murmeln auf N Zellen

Tabelle 1.1

Ziehen ohne Rücklegen bedeutet, dass die Nummer einer Kugel in der Stichprobe nicht mehrfach auftreten darf. Dies bedeutet jetzt, dass jede Zelle nur einmal auftreten darf,

dass man also in jede Zelle maximal eine Murmel legen darf. Solche Verteilungen nennen wir Verteilungen **ohne Mehrfachbesetzung**. Verteilungen, bei denen jede Zelle beliebig viele Murmeln enthalten darf, nennen wir Verteilungen **mit Mehrfachbesetzung**.

Unsere Überlegungen sind in Tabelle 1.1 zusammengefasst.

1.3 Anwendungsbeispiele

(1) (4 Würfel) Es werden vier völlig gleich aussehende Würfel gleichzeitig geworfen. Welches ist die Wahrscheinlichkeit p dafür, dass die vier erscheinenden Augenzahlen verschieden sind?

Man könnte versucht sein, folgendermaßen zu argumentieren: Jedes Ergebnis ist dadurch beschrieben, dass man sagt, wie oft die 1, die 2, usw. auftritt. Also ist Ω_{IV} mit $N = 6$ und $n = 4$ ein passender Stichprobenraum. Die Menge der Ergebnisse, bei denen die vier Augenzahlen verschieden sind, ist dann Ω_{II} . Also ist

$$p = \binom{6}{4} / \binom{4+6-1}{4} = \frac{5}{42}.$$

Falsch daran ist das „also“, denn damit wird Gleichverteilung unterstellt. Aber wären die Würfel verschieden gefärbt, oder würden sie nacheinander geworfen, so wären sie unterscheidbar, ohne dass sich an p etwas ändern dürfte. Dem Ergebnis $\{1, 1, 1, 1\}$ in Ω_{IV} (vier Einsen zu werfen) entspricht im unterscheidbaren Fall nur das Ergebnis $(1, 1, 1, 1)$, aber dem Ergebnis $\{1, 2, 3, 4\}$, dass die Augenzahlen 1, 2, 3, 4 auftreten, entsprechen $4! = 24$ Ergebnisse im unterscheidbaren Fall, in dem Ω_I der angemessene Stichprobenraum ist. Die richtige Lösung ist daher

$$p = \frac{\text{card}(\Omega_{II})}{\text{card}(\Omega_I)} = \frac{(6)_4}{6^4} = \frac{5}{18}.$$

(2) (Geburtstagszwillinge)* Wie groß ist die Wahrscheinlichkeit p , dass mindestens zwei von den $n = 25$ Schülern einer Klasse am gleichen Tag Geburtstag haben?

Als Stichprobenraum können wir Ω_I mit $n = 25$ und $N = 365$ wählen. Das Ergebnis $(\omega_1, \dots, \omega_{25})$ bedeutet, dass Schüler Nummer 1 am ω_1 -ten Tag des Jahres Geburtstag hat, Schüler 2 am ω_2 -ten Tag, usw. Das interessierende Ereignis ist das Komplement des Ereignisses Ω_{II} , dass alle Schüler an verschiedenen Tagen Geburtstag haben. Also ist $p = 1 - P(\Omega_{II})$. Unter der Annahme der Gleichverteilung auf Ω_I ist

$$P(\Omega_{II}) = \frac{\text{card}(\Omega_{II})}{\text{card}(\Omega_I)} = \frac{(N)_n}{N^n} = 1 \cdot \left(1 - \frac{1}{N}\right) \cdot \left(1 - \frac{2}{N}\right) \cdot \dots \cdot \left(1 - \frac{n-1}{N}\right).$$

Wir begnügen uns mit einer Approximation. Ist k^3 viel kleiner als N^2 , so ist $1 - k/N \approx \exp(-k/N)$, wobei wir die Notation $\exp(a) = e^a$ benutzen und wie üblich $\approx$ für „ungefähr gleich“ schreiben. Also ist

$$P(\Omega_{II}) \approx \exp\left(-\sum_{k=1}^{n-1} \frac{k}{N}\right) = \exp\left(-\frac{n(n-1)}{2N}\right) = \exp\left(-\frac{600}{730}\right) \approx 0,44.$$

Damit ist $p \approx 0,56$. (Ohne die Approximation ergäbe sich $p = 0,568$.)

(3) (*Zahlenlotto*) Es werden $n = 6$ Kugeln aus $N = 49$ Kugeln ohne Rücklegen gezogen. Notieren wir die Kugeln in der Reihenfolge der Ziehungen, so ist Ω_{II} die Ergebnismenge; notiert man nur die Teilmenge der gezogenen Kugeln, so muss man mit Ω_{III} arbeiten.

Im Gegensatz zum ersten Beispiel führen nun beide Beschreibungen bei Annahme der Gleichverteilung zu den gleichen Wahrscheinlichkeiten für die in den beiden Modellen einander entsprechenden Ereignisse, denn da nun alle ω_i verschieden sind, entsprechen jeder Teilmenge $\{\omega_1, \dots, \omega_6\}$ genau $6!$ geordnete 6-Tupel. Der Gleichverteilung auf Ω_{II} entspricht also die Gleichverteilung auf Ω_{III} . Natürlich lassen sich in Ω_{II} auch Ereignisse beschreiben wie „die dritte gezogene Kugel ist eine 5“, die in Ω_{III} nicht definiert sind. Interessiert man sich aber nur für Ereignisse wie „unter den 6 geratenen Zahlen $\{2, 11, 17, 32, 40, 48\}$ sind mindestens drei Richtige“, so hat man die freie Wahl zwischen Ω_{II} und Ω_{III} . Wir wollen Ω_{III} benutzen. Die Wahrscheinlichkeit jedes Ergebnisses ist dann

$$P(\omega) = \frac{1}{\binom{49}{6}} = \frac{1}{13\,983\,816} \approx 7,1511 \cdot 10^{-8}.$$

Dies ist auch die Wahrscheinlichkeit p_6 für „6 Richtige“, denn dazu muss das gezogene ω genau dem geratenen gleich sein.

Wir wollen nun $p_4 = P(\text{genau 4 Richtige})$ bestimmen: Sei $\bar{\omega} = \{\bar{\omega}_1, \bar{\omega}_2, \dots, \bar{\omega}_6\}$ die Menge der geratenen Zahlen. Dann lässt sich das Ereignis „genau 4 Richtige“ in Ω_{III} beschreiben durch

$$A_4 = \{\omega \in \Omega_{III} : \text{card}(\bar{\omega} \cap \omega) = 4\}.$$

Wir können eine zu A_4 gehörige Menge ω von Kugeln dadurch beschreiben, dass wir zuerst 4 Kugeln von $\bar{\omega}$ und dann 2 Kugeln von $\{1, 2, \dots, 49\} \setminus \bar{\omega}$ festlegen. Da $\bar{\omega}$ aus 6 Kugelnummern besteht, gibt es für die erste Wahl $\binom{6}{4}$ Möglichkeiten, für die zweite gibt es $\binom{43}{2}$ Möglichkeiten. Jede Wahl der 4 Kugeln aus $\bar{\omega}$ lässt sich mit jeder Wahl von 2 Kugeln aus $\{1, \dots, 49\} \setminus \bar{\omega}$ zu einem Element $\omega \in A_4$ zusammensetzen. Also ist $\text{card}(A_4) = \binom{6}{4} \cdot \binom{43}{2}$, und

$$p_4 = P(A_4) = \frac{\binom{6}{4} \binom{43}{2}}{\binom{49}{6}}.$$

Ebenso bestimmt man z.B. $p_5 = P(\text{genau 5 Richtige})$. Die Wahrscheinlichkeit für mindestens 4 Richtige wäre $p_6 + p_5 + p_4$.

Aus Symmetriegründen ist klar, dass die gesuchten Wahrscheinlichkeiten nicht von $\bar{\omega}$ abhängen. Das wäre nur anders, wenn der Zufallsmechanismus keine Gleichverteilung herstellen würde. Trotzdem heißt dies nicht, dass man nicht durch geschickte Wahl von $\bar{\omega}$ seine Gewinnchancen verbessern kann. Die Höhe der Auszahlung hängt nämlich davon ab, was die übrigen Lottospieler tun. Da sie gewisse Zahlenkombinationen bevorzugen, kann man mit einer höheren Auszahlung rechnen, wenn man solche $\bar{\omega}$ angibt, die die übrigen Lottospieler vernachlässigen. Welche das sind, kann man untersuchen, indem man von früheren Auslosungen die Gewinnhöhen mit den ausgelosten Zahlen vergleicht. Z.B. werden (bisher) die Zahlen der ersten Spalte und die über 31 etwas seltener getippt. Allerdings kommen nur 50 % der Einzahlungen zur Auszahlung. Der typische Spieler erhält also im Mittel nur 50 Pf von jeder eingesetzten Mark zurück. Einer meiner Diplomanden, Herr G. Hoffmann, zeigte, dass man von jeder eingesetzten Mark im Mittel 67 Pf zurückerhalten kann, wenn man aus den seltener getippten Zahlen seinen Tipp zufällig auswählt.

(4) (*Spatzen*) Auf wie viele Arten können sich zwei nicht unterscheidbare Spatzen auf 4 Telegraphenleitungen verteilen? Nach Fall IV mit $N = 4$ und $n = 2$ erhält man die Antwort $\binom{4+2-1}{2} = 10$. Man prüfe dies durch Auflistung aller Verteilungen.

1.4 Die hypergeometrische Verteilung

Mit dem Lottospiel haben wir schon einen Spezialfall einer wichtigen Verteilung kennen gelernt, die auch in vielen anderen Situationen auftritt. Um eine allgemeine Formel abzuleiten, betrachten wir eine Urne, welche S schwarze und W weiße Kugeln enthält, insgesamt also $N = S + W$ Kugeln. Es werden $n \leq S + W$ Kugeln ohne Rücklegen gezogen. Die Wahrscheinlichkeit dafür, dass die Stichprobe genau s schwarze und $n - s = w$ weiße Kugeln enthält, ist

$$h(s; n, N, S) := \frac{\binom{S}{s} \binom{W}{w}}{\binom{S+W}{n}} \quad (0 \leq s \leq n). \quad (1.15)$$

Zur Ableitung argumentieren wir wie oben. Wir denken uns die Kugeln mit $1, 2, \dots, N$ nummeriert, wobei etwa die Kugeln $1, \dots, S$ schwarz seien. Es gibt $\binom{S+W}{n}$ Ereignisse in Ω_{III} . Es gibt $\binom{S}{s}$ Möglichkeiten, s schwarze Kugeln aus S vorhandenen schwarzen Kugeln ohne Zurücklegen zu ziehen, da es $\binom{S}{s}$ Teilmengen der Mächtigkeit s von $\{1, \dots, S\}$ gibt. Entsprechend gibt es $\binom{W}{w}$ Möglichkeiten, w weiße Kugeln aus einem Reservoir von W weißen Kugeln ohne Rücklegen zu ziehen. Jede Kombination einer dieser Mengen von s schwarzen Kugeln mit einer dieser Mengen von w weißen Kugeln entspricht genau einem Ergebnis ω aus dem Ereignis „die Stichprobe enthält genau s schwarze Kugeln“.

Beispiel 1.2 (Skatspiel) Beim Skat erhält jeder der drei Spieler zehn Karten aus einem Pack mit 32 Karten, und zwei Karten (der *Skat*) werden zunächst beiseite gelegt. Es gibt 4 Asse. Wie groß ist die Wahrscheinlichkeit, dass Spieler A genau 3 Asse erhält? Dies entspricht dem obigen Experiment mit $S = 4$, $W = 28$, $n = 10$ und $s = 3$. Die Asse entsprechen den schwarzen Kugeln, die anderen Karten den weißen. Also ist die gesuchte Wahrscheinlichkeit

$$\frac{\binom{4}{3} \binom{28}{7}}{\binom{32}{10}} = \frac{66}{899}.$$

Die Wahrscheinlichkeit, dass mindestens einer der drei Spieler 3 Asse erhält, ist dreimal so groß, da die Ereignisse, dass Spieler A , B bzw. C drei Asse erhält, unvereinbar sind.

1.5 Vereinigungen von Ereignissen

Es ist manchmal schwierig, die Ergebnisse direkt abzuzählen, die zu einem Ereignis gehören, nach dem gefragt ist, aber die eingangs angegebenen Eigenschaften von P können weiterhelfen. Nicht selten ist z.B. $P(A^c)$ leichter zu ermitteln als $P(A)$. Ziemlich typisch ist es, dass die Wahrscheinlichkeit von Durchschnitten leichter zu bestimmen ist als die von Vereinigungen. Man benutzt dann die Formel (1.9) oder entsprechende Formeln für Vereinigungen von mehr als zwei Ereignissen. Z.B. rechnet man durch doppelte Anwendung von (1.9) auf $(A_1 \cup A_2) \cup A_3$ leicht nach, dass

$$\begin{aligned} P(A_1 \cup A_2 \cup A_3) &= [P(A_1) + P(A_2) + P(A_3)] \\ &\quad - [P(A_1 \cap A_2) + P(A_1 \cap A_3) + P(A_2 \cap A_3)] \\ &\quad + P(A_1 \cap A_2 \cap A_3) \end{aligned} \quad (1.16)$$

ist. Dies lässt sich noch verallgemeinern. Sind $A_1, A_2, \dots, A_m$ m beliebige Ereignisse, so sei

$$S_k = \sum_{\{i_1, i_2, \dots, i_k\} \subset \{1, \dots, m\}} P(A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_k}) \quad (1.17)$$

die Summe der Wahrscheinlichkeiten von allen Durchschnitten, die aus genau k der m Ereignisse gebildet werden können. Z.B. ist in (1.16) die erste eckige Klammer S_1 , die zweite S_2 und der letzte Term S_3 . Induktiv lässt sich für alle m zeigen

$$P\left(\bigcup_{i=1}^m A_i\right) = S_1 - S_2 + S_3 - \dots + (-1)^{m+1} S_m. \quad (1.18)$$

Da wir in § 3 ohnehin eine allgemeinere Formel ableiten, lassen wir den Beweis hier weg.

Beispiel 1.3* (*2 Asse beim Skat*) Wie groß ist beim Skat die Wahrscheinlichkeit p , dass mindestens einer der drei Spieler genau 2 Asse erhält?

Sei A_i das Ereignis, dass der i -te der drei Spieler genau zwei Asse erhält. Die gesuchte Wahrscheinlichkeit ist $p = P(A_1 \cup A_2 \cup A_3)$. Wie im obigen Beispiel zeigt man

$$P(A_i) = \binom{4}{2} \binom{28}{8} / \binom{32}{10} = \frac{2079}{7192} \approx 0,289.$$

$P(A_1 \cap A_2)$ lässt sich am schnellsten mit Hilfe der erst im nächsten Paragraphen diskutierten bedingten Wahrscheinlichkeiten berechnen, aber auch die folgende Rechnung enthält für kombinatorisches Abzählen typische und daher lehrreiche Elemente:

Wir denken uns die Karten mit 1, 2, ..., 32 nummeriert, und zwar so dass 1, 2, 3, 4 die Asse sind, und wählen als Stichprobenraum Ω die Menge der Permutationen $\omega = (\omega_1, \dots, \omega_{32})$ der Karten. $(\omega_1, \dots, \omega_{10})$ seien die zehn Karten des ersten Spielers in der Reihenfolge des Aufnehmens und $(\omega_{11}, \dots, \omega_{20})$ die des zweiten Spielers. Wie viele Elemente hat $A_1 \cap A_2$? Man kann auf $\binom{10}{2}$ Arten die Zeitpunkte $i_1 < i_2$ bestimmen, an denen Spieler 1 die Asse aufnimmt, und analog die Zeitpunkte $i_3 < i_4$ für Spieler 2 mit $i_3, i_4 \in \{11, \dots, 20\}$. Zum Zeitpunkt i_1 gibt es 4 Wahlmöglichkeiten für das Ass, zum Zeitpunkt i_2 noch 3, zum Zeitpunkt i_3 noch 2 und zum Zeitpunkt i_4 nur eine. Die übrigen Koordinaten wählt man aus $\{5, \dots, 32\}$ nach dem Schema der Stichprobe in Reihenfolge ohne Rücklegen. Es folgt

$$\text{card}(A_1 \cap A_2) = \binom{10}{2}^2 \cdot 4! \cdot 28!$$

und

$$P(A_1 \cap A_2) = \frac{\text{card}(A_1 \cap A_2)}{32!} = \frac{405}{7192} \approx 0,056.$$

Aus Symmetriegründen ist $P(A_1 \cap A_3) = P(A_2 \cap A_3)$. Da nicht jeder der drei Spieler zwei Asse haben kann, ist $P(A_1 \cap A_2 \cap A_3) = 0$. Mit (1.16) folgt $p \approx 3 \cdot (0,289 - 0,056) = \underline{\underline{0,699}}$.

[Der Leser lasse sich nicht dadurch verwirren, dass wir bei der Ableitung der hypergeometrischen Verteilung mit einem anderen Stichprobenraum gearbeitet haben. Häufig lassen sich Rechnungen vereinfachen, wenn man Wahrscheinlichkeiten, die das gleiche reale Ereignis in zwei verschiedenen Modellen hat, gleichsetzt. Natürlich muss gewährleistet sein, dass das Resultat davon unbeeinflusst bleibt, die Modelle also in diesem Sinn äquivalent sind.]

Die berechneten Wahrscheinlichkeiten beziehen sich natürlich auf die Kartenverteilungen nach dem Austeilen aber vor dem Aufnehmen des Skats. Die Verteilungen nach dem Aufnehmen des Skats hängen von der Spielweise ab.

1.6 Multinomialkoeffizienten

Die Binomialkoeffizienten $\binom{n}{k}$ beschreiben auf wie viele Arten man eine Menge von n nummerierten Kugeln derart in zwei Gruppen teilen kann, dass die erste Gruppe k Kugeln enthält. Wie viele Möglichkeiten gibt es nun, die Zahlen $1, \dots, n$ so in r Gruppen zu teilen, dass die erste Gruppe k_1 Elemente hat, die zweite Gruppe k_2 Elemente usw.? Offenbar muss man $k_1 + k_2 + \dots + k_r = n$ voraussetzen. Man kann zuerst auf $\binom{n}{k_1}$ Arten die erste Gruppe auswählen, dann auf $\binom{n-k_1}{k_2}$ Arten die zweite Gruppe, usw., schließlich auf $\binom{n-k_1-k_2-\dots-k_{r-1}}{k_r}$ Arten die r -te Gruppe. Multipliziert man diese Anzahlen miteinander, so ergibt sich nach leichter Rechnung die gesuchte Zahl

$$\frac{n!}{k_1! k_2! \dots k_r!}.$$

Wir bezeichnen sie mit

$$\binom{n}{k_1, k_2, \dots, k_r}.$$

Die so bestimmten Zahlen heißen **Multinomialkoeffizienten**.

Warnung: Die Gruppen, in die wir die Kugeln eingeteilt haben, sind nummeriert. Wir haben also nicht die Anzahl der Einteilungen in Gruppen der Größe $k_1, \dots, k_r$ unabhängig von deren Bedeutung oder Reihenfolge studiert, sondern in benannte Gruppen. Wir wollen den Unterschied an einem Beispiel erläutern.

Beispiel 1.4 (Mannschaftseinteilung) 26 Schulkinder haben einen Fußball, vier Tennisschläger, einen Fußballplatz und einen Tennisplatz zur Verfügung. Die Zahl der Einteilungen in zwei Fußballmannschaften A, B mit je 11 Spielern und zwei Tennisteamen C, D mit je 2 Spielern (für ein „Doppel“) ist $26!/(2! 2! 11! 11!)$. Dabei sind die Mannschaften benannt. Z.B. sind A und C die Mannschaften, die bei Spielbeginn mit der Sonne im Rücken spielen. Die Gruppeneinteilungen

$$A = \{1, 2, \dots, 11\}, \quad B = \{12, \dots, 22\}, \quad C = \{23, 24\}, \quad D = \{25, 26\}$$

und

$$A' = \{12, \dots, 22\}, \quad B' = \{1, \dots, 11\}, \quad C' = \{23, 24\}, \quad D' = \{25, 26\}$$

würden als verschieden gezählt, obwohl bei beiden Einteilungen die gleichen Teams gegeneinander spielen.

Interessiert man sich für die Zahl der Einteilungen in zwei Teams der Größe 11 und zwei Teams der Größe 2, so wäre unsere obige Lösung noch durch 4 zu teilen, weil immer vier „benannte“ Gruppeneinteilungen die gleiche Einteilung in Mannschaften ergeben.

1.7 Runs*

In einem Hörsaal sitzen in der ersten Reihe sieben Damen und neun Herren nebeneinander, und zwar in der Reihenfolge

$$DHDDDDHHHHHHDHHD.$$

Deutet diese Anordnung auf eine Tendenz hin, dass Nachbarn zum gleichen Geschlecht gehören? In einer fortlaufenden Folge von Symbolen nennen wir jede maximale Teilfolge von einander benachbarten gleichen Symbolen einen **run**. Die obige Folge beginnt also mit einem D -Run der Länge 1, auf den ein H -Run der Länge 1 und dann ein D -Run der Länge 4 folgt, usw. Insgesamt gibt es 7 Runs, von denen vier D -Runs und drei H -Runs sind. Wenn Symbole der gleichen Art häufiger benachbart sind, als es dem Zufall entspricht, wird es relativ wenige Runs geben, bei einer Tendenz zur Abwechslung viele. Für Runs interessiert man sich in vielen Situationen. Lange Runs von defekten Produkten deuten auf Fehler in der Produktionsanlage, lange Runs von erkrankten Pflanzen in einer Reihe auf Ansteckung.

Wie groß ist nun bei rein zufälliger Anordnung von 7 D 's und 9 H 's die Wahrscheinlichkeit p , wie oben 4 D -Runs und 3 H -Runs anzutreffen? Wir benötigen das

Lemma 1.5 *Die Zahl der Möglichkeiten, n ununterscheidbare Murmeln auf N Zellen so zu verteilen, dass keine Zelle leer bleibt, ist $\binom{n-1}{N-1}$.*

Beweis: Man legt zuerst in jede der N Zellen eine Murmel. Dann sind noch $n' = n - N$ Murmeln zu verteilen. Nach Modell Ω_{IV} ist die Anzahl der Verteilungen dieser ununterscheidbaren Murmeln auf die N Zellen

$$\binom{N + n' - 1}{n'} = \binom{N + n' - 1}{N - 1} = \binom{n - 1}{N - 1}. \quad \square$$

Nun zu unserer Frage über die Runs. Für eine Folge von d D 's und h H 's gibt es $\binom{d+h}{d}$ Anordnungen, wenn wir sowohl die D 's wie die H 's untereinander als nicht unterscheidbar betrachten. Wie viele davon bestehen aus, sagen wir, r D -Runs und s H -Runs? Da D -Runs und H -Runs abwechseln, ist $r - 1 \leq s \leq r + 1$. Nach dem Lemma gibt es $\binom{d-1}{r-1}$ Möglichkeiten, die Längen $k_1, k_2, \dots, k_r \geq 1$ der D -Runs festzulegen, und $\binom{h-1}{s-1}$ Möglichkeiten die Längen der H -Runs festzulegen, also $\binom{d-1}{r-1} \binom{h-1}{s-1}$ Möglichkeiten, die Längen aller Runs festzulegen. Ist $r = s + 1$, so muss der erste Run ein D -Run sein, im Fall $r = s - 1$ ein H -Run. Dann liegt also nach Wahl der Längen aller Runs die ganze Folge fest. In diesen Fällen ist also die gesuchte Wahrscheinlichkeit

$$\binom{d-1}{r-1} \binom{h-1}{s-1} / \binom{d+h}{d}.$$

Im Fall $r = s$ muss dies noch mit 2 multipliziert werden, da man noch die Wahl hat, die Folge mit einem D oder einem H zu beginnen.

In unserem Beispiel war $d = 7$, $h = 9$, $r = 4$, $s = 3$. Die Wahrscheinlichkeit p ist $\binom{6}{3} \binom{8}{2} / \binom{16}{7} = \frac{7}{141}$.

Um zu beurteilen, ob die beobachtete Zahl von 7 Runs ungewöhnlich klein ist, würde man nach der Wahrscheinlichkeit fragen, höchstens 7 Runs zu beobachten. Diese lässt sich nun leicht berechnen. In der Praxis benutzt man natürlich Tabellen.

1.8 Einfache Identitäten für Binomialkoeffizienten

Die Binomialkoeffizienten $\binom{n}{k}$ treten so oft auf, dass es lohnt, sich ein paar Rechenregeln zu merken. Wir beginnen mit dem Satz, nach dem die Koeffizienten benannt sind.

Satz 1.6 (Binomischer Lehrsatz)

$$(x + y)^n = \sum_{k=0}^n \binom{n}{k} x^k y^{n-k}. \quad (1.19)$$

Beweis: Schreibt man $(x + y)^n = (x_1 + y_1)(x_2 + y_2) \dots (x_n + y_n)$ mit $x_i = x$ und $y_i = y$, so tritt das Glied $x^k y^{n-k}$ beim Ausmultiplizieren immer dann auf, wenn in k Klammern der Faktor x_i und in $n - k$ Klammern der Faktor y_i gewählt wird, also in $\binom{n}{k}$ Fällen. $\square$

Folgerung 1.7

$$(i) \quad \sum_{k=0}^n \binom{n}{k} = 2^n; \quad (ii) \quad \sum_{k=0}^n (-1)^k \binom{n}{k} = 0; \quad (iii) \quad \sum_{k=0}^n k \binom{n}{k} = n \cdot 2^{n-1}.$$

Beweis: (i) Setze in (1.19) $x = 1, y = 1$. (ii) Setze $x = -1, y = 1$. (iii) Leite (1.19) nach x ab und setze dann $x = y = 1$. $\square$

Nützlich sind auch die für $0 \leq k \leq n$ gültigen Relationen

$$\binom{n}{k} = \binom{n}{n-k} \quad (1.20)$$

und

$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}. \quad (1.21)$$

(1.20) folgt direkt aus der Definition; (1.21) ergibt sich aus der Beobachtung, dass es $\binom{n}{k}$ Teilmengen der Mächtigkeit k von $\{1, \dots, n\}$ gibt. Dies sind $\binom{n-1}{k-1}$ Teilmengen, die das Element 1 enthalten und $\binom{n-1}{k}$ Teilmengen, die es nicht enthalten. Aus (1.21) folgt, dass in **Pascals Dreieck** (Bild 1.1) die $(n + 1)$ -te Reihe die Binomialkoeffizienten $\binom{n}{k}$

$$\begin{array}{ccccccc}
 & & & & 1 & & & & \\
 & & & & 1 & & 1 & & \\
 & & & 1 & & 2 & & 1 & \\
 & & 1 & & 3 & & 3 & & 1 \\
 & 1 & & 4 & & 6 & & 4 & & 1 \\
 1 & & 5 & & 10 & & 10 & & 5 & & 1
 \end{array}$$

Bild 1.1 Pascals Dreieck

($k = 0, \dots, n$), enthält. Aus dem Dreieck wiederum entnimmt man, dass die Binomialkoeffizienten bis zur Mitte hin ansteigen, d.h.

$$k + 1 \leq \frac{n}{2} \quad \text{impliziert} \quad \binom{n}{k} < \binom{n}{k+1}. \quad (1.22)$$

Aus unseren Abzählüberlegungen ergibt sich auch noch leicht:

Folgerung 1.8

$$\sum_{s=0}^S \binom{S}{s} \binom{W}{n-s} = \binom{S+W}{n} \quad (n \leq S+W).$$

Beweis: Die Summe der Wahrscheinlichkeiten $h(s; n, W + S, S)$ in (1.15) ist 1. $\square$

Anhang*

Historische Anmerkungen: Die Wahrscheinlichkeiten der verschiedenen Augensummen bei einem Wurf mit drei Würfeln wurden schon im 13. Jh. in dem Epos „De Vetula“ richtig bestimmt, als dessen Autor man Richard de Fournival (1201–1260), den Kanzler der Kathedrale von Amiens, vermutet. Auf einen (evtl. mittelbaren) islamischen Einfluss deuten darin vorkommende arabische Fachtermini. Geronimo Cardano (1501–1576) schrieb um 1564 ein Manuskript über Glücksspiele „De ludo aleae“, das aber erst 1663 erschien. Darin ist bereits das Prinzip des fairen Spiels formuliert. Als Beginn der wissenschaftlichen Beschäftigung mit Wahrscheinlichkeiten betrachtet man den Briefwechsel von Blaise Pascal (1623–1662) und Pierre Fermat (1601–1665) aus dem Jahre 1654. Dieser war angeregt durch Fragen des Schriftstellers und Glücksspielers Chevalier de Méré über Gewinnaussichten in einigen Spielsituationen, u.a. auch bei einem Spielabbruch. (Es gibt übrigens eine interessante fiktive Fortsetzung des Briefwechsels über die Grundlagen der Wahrscheinlichkeitsrechnung in dem Buch von A. Rényi: „Briefe über die Wahrscheinlichkeit“.)

Christiaan Huygens (1629–1695) hörte von dem Briefwechsel von Pascal und Fermat, hatte aber keinen Zugang dazu. Er ging über Pascal und Fermat hinaus, indem er den Begriff des Erwartungswerts einführte. Sein „Tractatus de Ratiociniis in Ludo Aleae“ (1657) war sehr einflussreich.

Jakob Bernoulli (1654–1705) stellte den Zusammenhang von Wahrscheinlichkeiten und relativen Häufigkeiten durch das (in §3 diskutierte) „Gesetz der großen Zahlen“ her. Sehr wahrscheinlich fand er den Beweis schon zwischen 1687 und 1689, aber dieser wurde erst 1713 in seinem nicht ganz vollendeten posthum herausgegebenen Buch „Ars conjectandi“ veröffentlicht; vgl. A. P. Yushkevich (1986). Pierre Simon Marquis de Laplace sah 1812 in seinem berühmten Buch „Théorie analytique des probabilités“ als Ziel der Wahrscheinlichkeitstheorie die Rückführung auf gleichwahrscheinliche Fälle an. Für ihn waren dies Fälle, über die wir gleich wenig wissen; er glaubte, dass der Zufall in einer in seiner Sicht deterministischen Welt nur ins Spiel kommt, weil wir etwas nicht wissen.

Für die gegenteilige Ansicht, dass es echt nichtdeterministische Vorgänge gibt, lässt sich der Atomzerfall anführen. Zwar ist es theoretisch denkbar, dass bisher unbekannte Variable den Zeitpunkt des Zerfalls von Atomen bestimmen, aber es gibt keinerlei Anhaltspunkte für die Existenz solcher Variablen. Die Quantenmechanik enthält den Begriff Wahrscheinlichkeit als konstitutives Element. Es gibt keine deterministische Hintergrundtheorie zur Quantentheorie, und wenn es eine gäbe, so müsste sie Eigenschaften haben, die mit anderen als gesichert geltenden Ergebnissen der Physik in Widerspruch stehen. Für die mathematische Theorie ist es unwesentlich, ob das deterministische Weltbild richtig oder falsch ist.

Auf die Entwicklung des axiomatischen Ansatzes kommen wir in §10 zurück.

D'Alembert stellte 1761 die Vermutung auf, lange Runs träten viel seltener auf als nach den Regeln der Wahrscheinlichkeitstheorie zu erwarten und leitete daraus Zweifel an der Anwendbarkeit der Wahrscheinlichkeitstheorie auf die reale Welt ab. Als Marbe dies zu Beginn dieses Jh. experimentell zu untermauern versuchte, widerlegten ihn H. Bruns und L. v. Bortkiewicz durch Herleitung des Erwartungswerts und der Varianz der Zahl der Runs (damals Iterationen genannt) gegebener Länge. Später erlangten Runs in der Testtheorie Bedeutung.

Anwendung der Urnenmodelle auf die statistische Physik: In der statistischen Physik sieht man von der Beschreibung der Bewegung von Teilchen nach den Gesetzen der newtonschen Mechanik ab. Die Teilchenzahlen sind zu groß, als dass man noch eine Chance hätte, die zugehörigen Differenzialgleichungssysteme zu lösen oder die Anfangsbedingungen zu bestimmen.

Stattdessen macht man einen stochastischen Ansatz. Für jedes Teilchen wird der Zustand in jedem Zeitpunkt durch die drei Ortskoordinaten und die drei Geschwindigkeitskoordinaten beschrieben. Für ein abgeschlossenes System können wir annehmen, dass die Menge der möglichen Zustände, der Phasenraum, eine beschränkte Teilmenge des $\mathbb{R}^6$ ist. Man teilt ihn in Zellen genannte gleich große 6-dimensionale Quader, die so klein sind, dass die interessierenden Größen, etwa die Energie, innerhalb jeder Zelle als konstant angesehen werden können. Es gebe insgesamt N Zellen.

Interessiert man sich für die Energie der Teilchen, so teilt man die Zellen in Gruppen mit jeweils gleichem Energieniveau ein. Es möge N_j Zellen vom Energieniveau E_j geben. Sind n Teilchen vorhanden, so sei der **Mikrozustand** des Systems das n -Tupel $\omega = (\omega_1, \dots, \omega_n)$, in dem ω_i die Nummer der Zelle angibt, in der sich das i -te Teilchen gerade aufhält. Die **Besetzungszahl** $X_j(\omega)$ des j -ten Energieniveaus ist die Zahl der Teilchen in Zellen vom Energieniveau E_j . Man interessiert sich für die Wahrscheinlichkeit des Ereignisses

$$A(k_1, \dots, k_d) = \{\omega : X_1(\omega) = k_1, \dots, X_d(\omega) = k_d\},$$

dass sich für $j = 1, \dots, d$ gerade k_j Teilchen in Zellen vom Energieniveau E_j aufhalten. Natürlich muss $k_1 + k_2 + \dots + k_d = n$ sein. A priori hat man keinerlei Kenntnis darüber, welche Wahrscheinlichkeitsverteilung auf Ω angemessen ist.

Beispiel: In Wirklichkeit sollen N und n groß sein, aber zur Illustration verwenden wir kleine Werte. Der Phasenraum sei in 6 Zellen $1, \dots, 6$ unterteilt. Die Zellen $1, 2, 3$ mögen zum Energieniveau E_1 gehören, die Zellen $4, 5, 6$ zum Energieniveau E_2 (Bild 1.2). Der Mikrozustand $\omega = (2, 6, 3, 6, 4)$ beschreibt die Situation, in der Teilchen 1 in Zelle 2 ist, Teilchen 2 in Zelle 6, usw. Das Energieniveau E_1 ist doppelt besetzt, nämlich durch die Teilchen 1 und 3. Also ist $X_1(\omega) = 2$. Analog ist hier $X_2(\omega) = 3$.

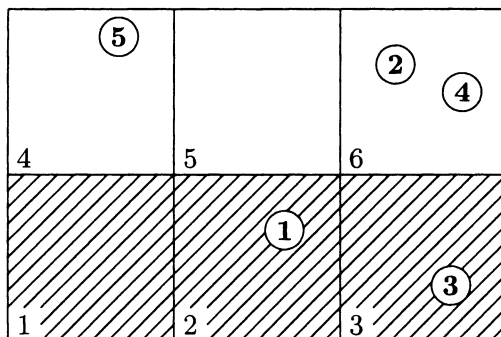


Bild 1.2 Besetzung von Zellen des Phasenraums

Maxwell-Boltzmann-Statistik: Können beliebig viele der Teilchen in jeder Zelle auftreten und sind die Teilchen unterscheidbar, so ist $\Omega_I = \{1, \dots, N\}^n$ die Menge der Mikrozustände. Betrachtet man jedes Element aus Ω_I als gleichwahrscheinlich, so spricht man von der Maxwell-Boltzmann-Statistik. Für jedes ω ist dann $P(\omega) = N^{-n}$. Es gibt $n!/(k_1! k_2! \dots k_d!)$ mögliche Aufteilungen von $\{1, \dots, n\}$ auf die verschiedenen Energieniveaus, die die Eigenschaft haben, dass k_j Teilchen in Zellen vom Energieniveau E_j sind. Für jede dieser Aufteilungen gibt es $N_j^{k_j}$ mögliche Verteilungen der k_j Teilchen vom Niveau E_j auf die N_j Zellen zu diesem Niveau, und bei fester Verteilung auf die Niveaus kann man all diese Verteilungen frei miteinander kombinieren. Es folgt also

$$P(A(k_1, \dots, k_d)) = \frac{1}{N^n} \frac{n!}{k_1! k_2! \dots k_d!} N_1^{k_1} \dots N_d^{k_d}.$$

Man kann Grenzübergänge mit wachsender Teilchenzahl vollziehen. Bei geeigneter Einteilung in Niveaus kann man auch zu empirisch überprüfbar Resultaten gelangen. Spricht man etwa vom Niveau E_1 , wenn die Energie eines in dieser Zelle befindlichen Teilchens $\leq \alpha$ ist, und sonst vom Niveau E_2 , und zählt man ab, wie viele Zellen vom Niveau E_1 sind, so lässt sich eine Energieverteilung ableiten. Für „klassische“ Teilchen, etwa für Moleküle, stimmt die so theoretisch aus einem an sich willkürlichen Ansatz abgeleitete Energieverteilung gut mit Messergebnissen überein. So betrachtet man die Annahme der Gleichverteilung als nachträglich gerechtfertigt. (Einzelheiten finden sich in Lehrbüchern über theoretische Physik.)

Fermi-Dirac-Statistik: Für Elementarteilchen stimmen die Resultate, die man aus der Annahme der Gleichverteilung auf Ω_I ableiten kann, nicht mit den Beobachtungen überein. Für Elementarteilchen mit halbzahligen Spin, sog. Fermionen, zu denen die Elektronen, Protonen und Neutronen gehören, kommt man aber mit einer Gleichverteilungsannahme auf $\Omega_{II} = \{\omega \in \Omega_I : \omega_i \neq \omega_j \forall i \neq j\}$ zu guten Übereinstimmungen. Diese Annahme lässt sich dann so interpretieren, dass für Fermionen das so genannte **Pauli-Verbot** gilt: Es ist verboten, dass sich in einer Zelle zugleich zwei Teilchen aufhalten.

Statt Ω_{II} können wir ebenso gut Ω_{III} benutzen, denn jedes $\omega \in \Omega_{III}$ lässt sich auf $n!$ Arten permutieren, ohne dass sich etwas an den Besetzungszahlen ändert. (Die Besetzungszahlen $X_j(\omega)$ sind daher auch auf den Äquivalenzklassen wohldefiniert.) Benutzt man Ω_{III} als Ergebnisraum, so ist ein „Mikrozustand“ $\omega \in \Omega_{III}$ beschreibbar, indem man angibt, welche der N Zellen eines der Teilchen enthalten. (Die übliche Sprechweise ist dann „Zustand“ statt Zelle und „besetzter Zustand“ für eine Zelle, die ein Teilchen enthält.) Es gibt dann $\binom{N}{n}$ Mikrozustände, und ein $\omega \in A(k_1, \dots, k_d)$ kann dadurch festgelegt werden, dass man für jedes j angibt, welche k_j der N_j Zustände vom Niveau E_j besetzt sein sollen. Also ist dann

$$P(A(k_1, \dots, k_d)) = \binom{N}{n}^{-1} \binom{N_1}{k_1} \dots \binom{N_d}{k_d}.$$

Bose-Einstein-Statistik: Für Bosonen, d.h. Elementarteilchen mit ganzzahligem Spin, zu denen Photonen und Mesonen gehören, kommt man zu Ergebnissen in Übereinstimmung mit physikalischen Beobachtungen, wenn man die Gleichverteilung auf Ω_{IV} annimmt. Man überlegt sich analog wie oben, dass dann

$$P(A(k_1, \dots, k_d)) = \binom{N+n-1}{n}^{-1} \binom{N_1+k_1-1}{k_1} \dots \binom{N_d+k_d-1}{k_d}$$

ist. Diese Verteilung ist äußerst bemerkenswert, weil kein plausibler Zufallsmechanismus zu ihr führt. (Das soll nicht heißen, dass es keinen Zufallsmechanismus gibt, der diese Verteilung zur Folge hat.)

Aufgaben

Aufgaben, deren Lösungen am Ende des Buches skizziert sind, sind durch (L) gekennzeichnet.

In allen Aufgaben gebe man ein Modell für das beschriebene Experiment an. Die auftretenden Ereignisse sind mathematisch zu beschreiben.

1. Aus einer Urne mit 3 roten und 4 schwarzen Kugeln und aus einer Urne mit 2 roten, 2 weißen und 3 schwarzen Kugeln wird je eine Kugel gezogen. Wie groß ist die Wahrscheinlichkeit, dass die gezogenen Kugeln die gleiche Farbe haben?
2. Ein Würfel wird 7 mal geworfen. Wie groß ist die Wahrscheinlichkeit, dass jede der Ziffern 1, ..., 6 unter den Wurfresultaten vorkommt?
3. Unter 32 Karten befinden sich 4 Asse. Die Karten werden gemischt und nacheinander aufgedeckt. Wie groß ist die Wahrscheinlichkeit, dass die neunte aufgedeckte Karte das zweite aufgedeckte Ass ist? (L)
4. Die Ecken eines Würfels sind gleichmäßig schräg abgeschliffen worden, so dass der Würfel auch auf jeder dieser Ecken liegen bleiben kann. Allerdings ist die Wahrscheinlichkeit jeder Ecke nur $1/4$ so groß wie die jeder Seite. Wie groß ist die Wahrscheinlichkeit einer Sechse?
5. Für einen gefälschten Würfel ist $P(\{1, 3, 5\}) = P(\{1, 2, 3\}) = P(\{2, 4, 5\})$. Man bestimme $P(2), \dots, P(6)$ in Abhängigkeit von $P(1)$.
6. Durch Berechnung der jeweiligen Wahrscheinlichkeiten nehme man Stellung zum folgenden Argument:
Beim dreimaligen Würfeln sind die Ereignisse „die Augensumme ist 11“ und „die Augensumme ist 12“ gleichwahrscheinlich, da beide Summen auf sechs Arten dargestellt werden können.
($11 = 6 + 4 + 1 = 6 + 3 + 2 = 5 + 5 + 1 = 5 + 4 + 2 = 5 + 3 + 3 = 4 + 4 + 3$;
 $12 = 6 + 5 + 1 = 6 + 4 + 2 = 6 + 3 + 3 = 5 + 5 + 2 = 5 + 4 + 3 = 4 + 4 + 4$.)
7. Man berechne die Wahrscheinlichkeit, dass beim viermaligen Werfen eines Würfels
a) das Maximum der erhaltenen Augenzahlen gleich 4 ist,
b) das Minimum der erhaltenen Augenzahlen kleiner oder gleich 4 ist.
8. In einer Lotterie wurde eine siebenstellige Gewinnzahl auf die folgende Weise ermittelt: In einer Trommel kommen die Ziffern 0 bis 9 je sieben mal vor. Die sieben Ziffern der Gewinnzahl werden nacheinander ohne Rücklegen gezogen. Wie groß sind die Wahrscheinlichkeiten der Gewinnzahlen 8888888, 1234567, 4491101?
9. Man beweise induktiv die Formel (1.18).
10. Ein Prüfer hat 18 Standardfragen, von denen er in jeder Prüfung 6 zufällig auswählt. Ein Kandidat kennt die Antworten von 10 Fragen. Wie groß ist die Wahrscheinlichkeit, dass er die Prüfung besteht, wenn er dazu mindestens drei Fragen richtig beantworten muss?
11. Eine Gruppe von n Personen, darunter A und B , setzt sich zufällig an einen runden Tisch. Wie groß ist die Wahrscheinlichkeit, dass genau k Personen zwischen A und B sitzen? (L)
12. Wie viel Rosinen muss man in 500 g Teig tun, damit ein 50-g-Brötchen mit Wahrscheinlichkeit 99 % mindestens eine Rosine enthält?
13. Drei Kugeln werden in 6 Schachteln eingeordnet, welche von 1 bis 6 nummeriert seien. Jeder Verteilung wird eine Zahl zugeordnet, indem man die Schachtelnummer mit den Besetzungszahlen multipliziert und die Resultate dann aufaddiert. Berechnen Sie die Wahrscheinlichkeit, dass die Zahl 9 bzw. 10 ist, wenn die Einordnung a) der Maxwell-Boltzmann, b) der Fermi-Dirac, c) der Bose-Einstein-Statistik folgt.

14. Beim Schachspiel kann ein Turm nur horizontal und vertikal schlagen. Wir nehmen nun den allgemeineren Fall an, dass das Spielbrett aus $n \times n$ Feldern besteht.
- (A) - Wie viele Möglichkeiten gibt es, n einander gleiche Türme auf dieses Brett zu stellen, so dass keiner den anderen bedroht?
- (B) - Bezeichnet A_n die gesuchte Zahl, so könnte man wie folgt argumentieren: für *einen* Turm hat man n^2 Möglichkeiten, ihn zu platzieren; dieser bedroht dann eine Reihe und eine Spalte. Das Problem reduziert sich daher auf ein $(n-1) \times (n-1)$ -Brett mit $n-1$ Türmen, so dass $A_n = n^2 A_{n-1}$ ist. Dies bedeutet aber $A_n = n^2 (n-1)^2 (n-2)^2 \dots 2^2 1^2 = (n!)^2$. Warum ist dieses *nicht* die gesuchte Lösung von (A)? (L)
15. Wie viele verschiedene 5-stellige Zahlen kann man durch Nebeneinanderlegen von 5 von 6 Kärtchen bilden, auf denen die Ziffern 1, 1, 2, 2, 2, 3 stehen? (L)

§ 2 Bedingte Wahrscheinlichkeit und Unabhängigkeit

2.1 Definition und Eigenschaften bedingter Wahrscheinlichkeiten

Häufig steht, bevor das Ergebnis eines Zufallsexperiments bekannt ist, schon die Information zur Verfügung, dass das Ergebnis zu einer bestimmten Teilmenge des Stichprobenraums gehört. Z.B. sieht ein Spieler beim Skat seine eigenen zehn Karten. Interessiert sich Spieler 1 für die Wahrscheinlichkeit des Ereignisses A , dass Spieler 2 zwei Asse hat, so wird er zunächst seine eigenen Asse zählen. Hat er selbst drei oder vier Asse, so ist für ihn die Wahrscheinlichkeit des Ereignisses A natürlich 0, hat er maximal zwei Asse, so ist sie positiv.

Ein anderes Beispiel: Wenn ein 30-jähriger Mann eine Lebensversicherung abschließt, wird man bei der Berechnung der Wahrscheinlichkeit, dass er im l -ten Lebensjahr stirbt, berücksichtigen müssen, dass er das Alter 30 bereits erreicht hat.

Für ein Laplace-Experiment ist der Ansatz für die Definition bedingter Wahrscheinlichkeiten sehr nahe liegend. Waren ursprünglich alle Ergebnisse $\omega \in \Omega$ gleichwahrscheinlich, und erhält man nur die Information, dass $\omega \in B$ liegt, so ordnen wir den Ergebnissen aus B^c die bedingte Wahrscheinlichkeit 0 zu, und betrachten die Ergebnisse aus B als gleichwahrscheinlich unter der bedingten Wahrscheinlichkeit. Dies bedeutet, dass für jedes $A \subset \Omega$ die bedingte Wahrscheinlichkeit von A bei gegebenem B den Wert

$$P(A | B) = \frac{\text{card}(A \cap B)}{\text{card}(B)} \quad (2.1)$$

erhält. Aus $P(A \cap B) = \text{card}(A \cap B)/\text{card}(\Omega)$ und $P(B) = \text{card}(B)/\text{card}(\Omega)$ ergibt sich in diesem Fall

$$P(A | B) = \frac{P(A \cap B)}{P(B)}. \quad (2.2)$$

Wir definieren nun auch für beliebige Wahrscheinlichkeitsräume (Ω, P) und für beliebige Ereignisse B mit $P(B) > 0$ die **bedingte Wahrscheinlichkeit** $P(A | B)$ von A bei gegebenem B durch (2.2).

In § 1 hatten wir — zunächst noch relativ vage und ohne Beweis — einen Zusammenhang von Wahrscheinlichkeiten und relativen Häufigkeiten angeführt. Sei wieder $k_n(A)$ die absolute Häufigkeit, mit der sich Ereignis A in n unabhängigen Wiederholungen des durch (Ω, P) modellierten Experiments ereignet, und $h_n(A) = k_n(A)/n$ die relative Häufigkeit. Für große n sollte mit großer Wahrscheinlichkeit $h_n(A)$ ungefähr $P(A)$ sein. Die absolute Häufigkeit des Ereignisses A unter denjenigen von den ersten n Experimenten, in denen sich B ereignet, ist $k_n(A \cap B)$. Also ist die relative Häufigkeit von A unter diesen Experimenten $k_n(A \cap B)/k_n(B) = h_n(A \cap B)/h_n(B) \approx P(A \cap B)/P(B) = P(A | B)$. $P(A | B)$ ist also ein Maß für die langfristige relative Häufigkeit, mit der sich A ereignet, wenn man nur die Teilfolge der Experimente betrachtet, in denen B eintritt.

Da die zu B^c gehörigen Ergebnisse ω die bedingte Wahrscheinlichkeit $P(\{\omega\} | B) = 0$ haben, liegt die Idee nahe, für die bedingte Wahrscheinlichkeit nur den kleineren Stichprobenraum B statt Ω zu verwenden. Dies ist aber nicht zweckmäßig, zumal wir später mehrere Bedingungen nebeneinander betrachten werden.

Nun aber zunächst ein paar

Beispiele: (1) (*Lebensdauer*) Nehmen wir einmal der Einfachheit halber an, die Wahrscheinlichkeit p_k , dass ein zufällig ausgewählter männlicher Bundesbürger genau im k -ten Lebensjahr stirbt, sei unabhängig von seinem Geburtsjahr. Wir vernachlässigen also die durch Einflüsse von Kriegen und durch Änderungen der Lebensgewohnheiten und der ärztlichen Versorgung bewirkten Schwankungen. p_k sei durch die Auswertung von Sterbetafeln bekannt. Die Wahrscheinlichkeit, dass das k -te Lebensjahr erreicht wird, ist $s_k := p_k + p_{k+1} + p_{k+2} + \dots$. Die bedingte Wahrscheinlichkeit, im l -ten Lebensjahr zu sterben, unter der Bedingung, dass das k -te Lebensjahr gerade erreicht ist, ist 0 für $l < k$ und p_l/s_k für $l \geq k$.

(2) (*Ziehen von Kugeln ohne Rücklegen*) Aus einer Urne, die zwei weiße und drei schwarze Kugeln enthält, werden nacheinander ohne Rücklegen zwei Kugeln gezogen. Die Wahrscheinlichkeit des Ereignisses A , dass die zweite gezogene Kugel schwarz ist, gegeben das Ereignis B , dass die erste gezogene Kugel weiß ist, sollte $3/4$ sein, denn, wenn B geschieht, sind nach der ersten Ziehung noch drei schwarze Kugeln in der Urne, und eine weiße. Kommt dieser Wert auch bei Anwendung der Definitionsgleichung (2.2) heraus?

Es handelt sich um ein Laplace-Experiment. Geben wir den weißen Kugeln die Nummern 1, 2, und den schwarzen die Nummern 3, 4, 5, so haben die interessierenden Ereignisse die Form

$$\begin{aligned} A \cap B &= \{(1, 3), (1, 4), (1, 5), (2, 3), (2, 4), (2, 5)\}, \\ B &= \{(1, 2), (1, 3), (1, 4), (1, 5), (2, 1), (2, 3), (2, 4), (2, 5)\}. \end{aligned}$$

Also ist $P(A|B) = \text{card}(A \cap B)/\text{card}(B) = 6/8 = 3/4$, wie erwartet.

In diesem Beispiel betraf die Bedingung B ein Ereignis, das zeitlich vor A beobachtet werden konnte. In der Definition (2.2) spielt jedoch die zeitliche Reihenfolge keine Rolle. Wir können z.B. ebenso gut nach der Wahrscheinlichkeit $P(B|A)$ fragen, dass die erste Kugel weiß ist, gegeben die zweite Kugel ist schwarz. (Es ist $P(B|A) = \text{card}(A \cap B)/\text{card}(A) = 6/12 = 1/2$.)

In der Praxis wird meist nicht $P(A|B)$ aus $P(B)$ und $P(A \cap B)$ berechnet, sondern umgekehrt $P(A \cap B)$ aus $P(B)$ und $P(A|B)$. Man verwendet dazu die Definitionsgleichung (2.2) in der umgeschriebenen Form

$$P(A \cap B) = P(B) P(A|B). \quad (2.3)$$

Im obigen Beispiel würde man etwa schließen: Die Wahrscheinlichkeit $P(B)$ ist $2/5$, da zu Beginn zwei von fünf Kugeln weiß sind. Nachdem B eingetreten ist, sind drei von vier Kugeln schwarz. Also ist $P(A|B) = 3/4$ und somit $P(A \cap B) = \frac{2}{5} \cdot \frac{3}{4} = \frac{3}{10}$.

Genau genommen hat man bei dieser Rechnung gar nicht den Stichprobenraum für zwei Ziehungen ohne Rücklegen benutzt, sondern Modelle für die beiden Teilerperimente. Stillschweigend wird also angenommen, dass $P(B)$ und $P(A|B)$ im Gesamtexperiment den gleichen Wert hätten, wie man ihn den Teilerperimenten entnimmt. Das ist sicher vernünftig, wenn die Modelle adäquat sind. Wir gehen später noch etwas darauf ein.

Die Formel (2.3) lässt sich induktiv verallgemeinern:

Sind $A_1, A_2, \dots, A_k$ Ereignisse mit $P(A_1 \cap A_2 \cap \dots \cap A_{k-1}) > 0$, so ist

$$\begin{aligned} P(A_1 \cap A_2 \cap \dots \cap A_k) &= P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1 \cap A_2) \\ &\quad \cdot \dots \cdot P(A_k|A_1 \cap A_2 \cap \dots \cap A_{k-1}). \end{aligned} \quad (2.4)$$

Beweis: Simple Induktion. (Aus $P(A_1 \cap A_2 \cap \dots \cap A_{k-1}) > 0$ folgt $P(A_1 \cap A_2 \cap \dots \cap A_j) > 0$ für alle j mit $1 \leq j \leq k-1$, so dass die auftretenden bedingten Wahrscheinlichkeiten definiert sind.) $\square$

Beispiel 2.1 Beim Skat bestimme man die Wahrscheinlichkeit dafür, dass die drei Spieler je genau ein Ass haben.

Aus Symmetriegründen können wir annehmen, dass Spieler 1 die ersten 10 ausgeteilten Karten erhält, Spieler 2 die nächsten 10, dann Spieler 3 zehn, und die letzten 2 in den Skat kommen. Sei A_i das Ereignis, dass Spieler i genau ein Ass erhält. Dann ist die gesuchte Wahrscheinlichkeit

$$P(A_1 \cap A_2 \cap A_3) = P(A_1) \cdot P(A_2 | A_1) \cdot P(A_3 | A_1 \cap A_2).$$

Es ist $P(A_1) = \binom{4}{1} \binom{28}{9} / \binom{32}{10}$ und $P(A_2 | A_1) = \binom{3}{1} \binom{19}{9} / \binom{22}{10}$, denn nachdem Spieler 1 zehn Karten und darunter genau ein Ass erhalten hat, sind noch 3 Asse unter den 22 verbliebenen Karten. Analog ist $P(A_3 | A_1 \cap A_2) = \binom{2}{1} \binom{10}{9} / \binom{12}{10}$.

[Man überzeuge sich, dass diese Methode für die Wahrscheinlichkeit, dass Spieler 1 und Spieler 2 je zwei Asse haben, den gleichen Wert wie die kompliziertere Überlegung in Beispiel 1.3 liefert.]

Die folgenden Eigenschaften bedingter Wahrscheinlichkeiten sind trivial aber nützlich.

Satz 2.2 (i) Sei $P(B) > 0$. Durch $P_B(A) := P(A | B)$ ist ein Wahrscheinlichkeitsmaß auf Ω definiert. Ist $A \subset B^c$ oder $P(A) = 0$, so ist $P(A | B) = 0$.

(ii) (Formel von der totalen Wahrscheinlichkeit) $\{B_1, B_2, \dots\}$ heißt Zerlegung von Ω , wenn die B_i disjunkt sind, und ihre Vereinigung Ω ist. Für jede Zerlegung und jedes Ereignis A gilt

$$P(A) = \sum_k P(B_k) P(A | B_k).$$

(Ist $P(B_k) = 0$, so ist $P(A | B_k)$ nicht definiert; das Produkt $P(B_k) P(A | B_k)$ kann dann = 0 gesetzt werden.)

(iii) (Formel von Bayes) Ist $P(A) > 0$, und gelten die Voraussetzungen von (ii), so ist für alle i

$$P(B_i | A) = \frac{P(B_i) P(A | B_i)}{\sum_k P(B_k) P(A | B_k)}.$$

(iv) Ist C die disjunkte Vereinigung von Ereignissen $C_1, C_2, \dots$ mit $P(C_i) > 0$, und sind die $P(A | C_i)$ alle gleich, so ist $P(A | C) = P(A | C_1)$.

Beweis: (i) Die Eigenschaften $P_B(A) \geq 0$ und $P_B(\Omega) = 1$ sind klar. Ist A disjunkte Vereinigung von Ereignissen $A_1, A_2, \dots$ so ist $A \cap B$ die disjunkte Vereinigung der $A_i \cap B$. Daraus folgt die Additivität.

(ii) folgt, weil A die disjunkte Vereinigung der $A \cap B_k$ ist, aus der Additivität von P und (2.3).

(iii) ergibt sich aus (ii) und $P(B_i \cap A) = P(B_i) P(A | B_i)$.

(iv) Es ist $P(A | C_1) P(C) = P(A | C_1) \sum P(C_i) = \sum P(A | C_i) P(C_i) = \sum P(A \cap C_i) = P(A \cap C) = P(A | C) P(C)$. Nach Division durch $P(C)$ folgt die Behauptung. $\square$

Beispiele: (1) (*Ziehen von Kugeln der gleichen Farbe*)* Wie groß ist für die obige Urne mit zwei weißen und drei schwarzen Kugeln die Wahrscheinlichkeit des Ereignisses C , dass beide (ohne Rücklegen) gezogenen Kugeln die gleiche Farbe haben? Sei B_1 das Ereignis, dass die erste gezogene Kugel weiß ist, und B_2 das Ereignis, dass sie schwarz ist. Offenbar ist $P(C|B_1) = 1/4$, denn nachdem eine weiße Kugel gezogen ist, ist nur noch eine von vier Kugeln weiß. Analog überlegt man sich $P(C|B_2) = 2/4$. Nach der Formel von der totalen Wahrscheinlichkeit ist

$$P(C) = P(B_1)P(C|B_1) + P(B_2)P(C|B_2) = \frac{2}{5} \cdot \frac{1}{4} + \frac{3}{5} \cdot \frac{2}{4} = \frac{2}{5}.$$

(2) (*Test für eine seltene Krankheit*) Eine Krankheit kommt bei ca. 0,5 % der Bevölkerung vor. Ein Test zur Auffindung der Krankheit führt bei 99 % der Kranken zu einer Reaktion, aber auch bei 2 % der Gesunden. Wie groß ist die Wahrscheinlichkeit, dass eine Person, bei der die Reaktion eintritt, die Krankheit wirklich hat?

Zur Lösung denken wir uns die Bevölkerung mit $\{1, \dots, N\}$ nummeriert. Ist B_1 die Menge der Kranken und B_2 die der Gesunden, so soll also $\text{card}(B_1) \approx 0,005 N$, $\text{card}(B_2) \approx 0,995 N$ sein. Ist A die Teilmenge der Personen, bei denen ein Test im Fall der Anwendung zur Reaktion führt, so soll ferner $\text{card}(A \cap B_1) \approx 0,99 \text{ card}(B_1)$ und $\text{card}(A \cap B_2) \approx 0,02 \text{ card}(B_2)$ gelten. Bei zufälliger Auswahl einer Person ist jeder Person die Wahrscheinlichkeit $1/N$ zugeordnet. Also setzen wir

$$\begin{aligned} P(B_1) &= 0,005, & P(B_2) &= 0,995, \\ P(A \cap B_1) &= 0,99 \cdot 0,005 & \text{und} & P(A \cap B_2) = 0,02 \cdot 0,995. \end{aligned}$$

Nach der Formel von Bayes folgt

$$P(B_1|A) = \frac{P(A \cap B_1)}{P(A)} = \frac{0,99 \cdot 0,005}{0,99 \cdot 0,005 + 0,02 \cdot 0,995} = \frac{495}{2485} \approx 0,2.$$

Von allen Personen, an denen die Reaktion beobachtet wird, sind also nur 20 % krank. Sie sollten sich weiteruntersuchen lassen, aber nicht über Gebühr beunruhigt sein.

Achtung: Nicht jede Information, die wir über ein Zufallsexperiment erhalten, eignet sich zur Bestimmung von bedingten Wahrscheinlichkeiten nach dem obigen Schema. Angenommen Herr Hans Schlau zieht aus einer Urne mit 4 weißen und 4 schwarzen Kugeln ohne Rücklegen 4 Kugeln und sagt wahrheitsgemäß, dass unter den ersten 3 gezogenen Kugeln mindestens eine schwarze war. Ist nach dieser Information die Wahrscheinlichkeit, dass die vierte gezogene Kugel weiß ist, größer als $1/2$? Könnte man also darauf vorteilhaft wetten?

Ist B das von Hans genannte Ereignis und A das Ereignis, dass die vierte gezogene Kugel weiß ist, so ist, wie man leicht prüft, $P(A|B) > 1/2$. Aber dies beantwortet die gestellte Frage nur, wenn feststeht, dass Hans Auskunft geben *musste*, ob B eingetreten war, und dass es ihm nicht frei stand, welche wahre Teilinformation er gab. Dann wäre es nämlich denkbar, dass er die Information „ B ist eingetreten“ nur äußerte, wenn genau eine der drei ersten gezogenen Kugeln schwarz war, und dass er z.B. bei genau zwei schwarzen unter den ersten drei Kugeln gesagt hätte: „Es ist mindestens eine weiße darunter.“ Wir sehen, dass der gemachte Ansatz für bedingte Wahrscheinlichkeiten nur sinnvoll ist, wenn die Information, ob B eintritt, in *jedem* Fall mitgeteilt wird.

2.2 Unabhängigkeit

Zieht man aus der Urne mit den 2 weißen und 3 schwarzen Kugeln zwei Kugeln *nun mit Rücklegen*, so hat der Stichprobenraum Ω die 25 Elemente (i, k) mit $1 \leq i, k \leq 5$. Das Ereignis A , dass die zweite Kugel schwarz ist, hat nun die 15 Elemente (i, k) mit $3 \leq k \leq 5$ und das Ereignis B , dass die erste Kugel weiß ist, besteht aus den 10 Elementen (i, k) mit $1 \leq i \leq 2$. Es ist

$$P(A) = \frac{3}{5}, \quad P(B) = \frac{2}{5} \quad \text{und} \quad P(A \cap B) = \frac{6}{25}.$$

Daraus folgt $P(A|B) = (6/25) \cdot (5/2) = 3/5$, und somit $P(A|B) = P(A)$. Das ist auch zu erwarten, da im Fall des Rücklegens die erste Ziehung „keinerlei Einfluss“ auf die zweite hat. Für die mathematische Formulierung dieses fehlenden Einflusses müssen wir von Ursache-Wirkung-Vorstellungen absehen und nur den für die Berechnung von Wahrscheinlichkeiten wesentlichen Aspekt betrachten. Dann legt es das Beispiel nahe, A von B unabhängig zu nennen, wenn $P(A) = P(A|B)$ gilt. Da $P(A|B)$ nur für Ereignisse B mit $P(B) > 0$ definiert ist, hat es sich als zweckmäßig erwiesen, zu definieren:

*Zwei Ereignisse A, B heißen **unabhängig**, wenn $P(A \cap B) = P(A)P(B)$ ist.*

Ist $P(B) > 0$, so ist dies äquivalent zu $P(A) = P(A|B)$.

Unabhängigkeit von A und B drückt aus, dass A und B *wahrscheinlichkeitstheoretisch* in dem Sinn keinerlei Einfluss aufeinander haben, dass die Information, „ B geschieht“ nichts an der Wahrscheinlichkeit von A ändert. Dies muss man von realer Beeinflussung unterscheiden. So ist z.B. im obigen Urnenexperiment, selbst wenn es ohne Rücklegen ausgeführt wird, B *real* unbeeinflusst von A , denn B betrifft die erste Ziehung und A die zweite. Dennoch war beim Ziehen ohne Rücklegen $P(B|A) = 1/2$ und $P(B) = 2/5$, d.h. $P(B) \neq P(B|A)$ und damit B nicht unabhängig von A . Unabhängigkeit ist ein in A und B *symmetrischer Begriff*: Ist A unabhängig von B , so auch B von A . Bei realer Beeinflussung ist dies sicher nicht der Fall. Insofern wäre „fehlender Zusammenhang“ schon eher die Entsprechung von Unabhängigkeit als „fehlender Einfluss“.

Andererseits kann Unabhängigkeit von zwei Ereignissen A, B selbst dann vorliegen, wenn real das Eintreten von A davon abhängt, ob B geschieht. Als Beispiel betrachten wir ein Experiment, das aus zwei Würfeln eines perfekten Würfels besteht. Sei A das Ereignis, dass die Summe der beiden geworfenen Augenzahlen gerade ist, und B das Ereignis, dass die zweite geworfene Augenzahl gerade ist. Dann ist $P(A) = P(B) = 1/2$ und $P(A \cap B) = 1/4$. Also sind A und B unabhängig, obwohl B mitbestimmt, ob A eintritt. Dies drückt sich hier auch *wahrscheinlichkeitstheoretisch* aus, sobald wir nur einen nicht perfekten Würfel nehmen. Ist z.B. die Wahrscheinlichkeit für eine gerade Augenzahl $2/5$ statt $1/2$, so erhält man, wenn man den ersten Wurf als unabhängig vom zweiten ansetzt, $P(B) = \frac{2}{5}$, $P(A) = \left(\frac{2}{5}\right)^2 + \left(\frac{3}{5}\right)^2 = \frac{13}{25}$ und $P(A \cap B) = \left(\frac{2}{5}\right)^2 \neq P(A) \cdot P(B)$. Beim nicht perfekten Würfel sind also A und B abhängig.

Ein beliebter Anfängerfehler ist die Verwechslung von Unabhängigkeit und Disjunktheit. Sind zwei Ereignisse A, B disjunkt, so ist $P(A \cap B) = 0$. Disjunkte (d.h. unvereinbare) Ereignisse sind also nur dann unabhängig, wenn mindestens eines von beiden die Wahrscheinlichkeit 0 hat.

Wir benötigen den Unabhängigkeitsbegriff auch für mehr als zwei Ereignisse.

Definition 2.3 Wir sagen, die **Produktformel** gelte für eine endliche Familie $\{A_i, i \in J\}$ von Ereignissen, wenn die Identität

$$P\left(\bigcap_{i \in J} A_i\right) = \prod_{i \in J} P(A_i) \quad (2.5)$$

erfüllt ist. Eine Familie $\{A_i, i \in I\}$ von Ereignissen heißt **unabhängig**, wenn für jede endliche Teilfamilie die Produktformel gilt, also (2.5) für alle (nichtleeren) endlichen Teilmengen $J \subset I$ erfüllt ist. (Oft sagt man statt „unabhängig“ auch ausführlicher „stochastisch unabhängig“. Verkürzt ist die Sprechweise „die A_i sind unabhängig“ üblich.)

Man beachte, dass die Gültigkeit der Produktformel (2.5) für eine endliche Familie $A_1, \dots, A_m$ (also für $J = \{1, \dots, m\}$) noch nicht bedeutet, dass $A_1, \dots, A_m$ unabhängig sind, denn die Produktformel braucht dann noch nicht für alle Teilfamilien zu gelten. Z.B. besagt Unabhängigkeit von drei Ereignissen A, B, C nicht nur, dass $P(A \cap B \cap C) = P(A)P(B)P(C)$ gilt, sondern außerdem, dass $P(A \cap B) = P(A)P(B)$, $P(A \cap C) = P(A)P(C)$ und $P(B \cap C) = P(B)P(C)$ gilt. Im folgenden Beispiel gilt die Produktformel für A, B, C , aber nicht für A, B .

Beispiel 2.4* Eine Münze wird dreimal geworfen. Steht K für „Kopf“ und Z für „Zahl“, so sei $\Omega = \{K, Z\}^3$ und $P(\{\omega\}) = 1/8$ für alle ω . Ist

$$A = \{(K, K, K), (K, K, Z), (K, Z, K), (Z, K, K)\}$$

das Ereignis, dass mindestens zweimal Kopf kommt, B das Ereignis, dass beim ersten Wurf Kopf kommt, und C das Ereignis, dass beim zweiten und dritten Wurf die gleiche Seite der Münze oben liegt, so ist $P(A) = P(B) = P(C) = 1/2$ und $P(A \cap B \cap C) = 1/8$, aber $P(A \cap B) = 3/8 \neq P(A)P(B)$.

Eine Familie $\{A_i, i \in I\}$ von Ereignissen heißt **paarweise unabhängig**, wenn für alle $i \neq j$ die Ereignisse A_i, A_j unabhängig sind. Unabhängigkeit impliziert natürlich paarweise Unabhängigkeit, aber nicht umgekehrt. Beim zweifachen Münzwurf sind z.B. die Ereignisse $A_1 = \{1. \text{ Wurf Kopf}\}$, $A_2 = \{2. \text{ Wurf Kopf}\}$, $A_3 = \{1. \text{ und } 2. \text{ Wurf gleich}\}$ paarweise unabhängig aber nicht unabhängig.

Der Begriff der paarweisen Unabhängigkeit ist unwichtig. Er wurde hier nur erwähnt, um den Begriff der Unabhängigkeit deutlicher zu machen.

Satz 2.5 (i) *Jede Teilfamilie einer unabhängigen Familie von Ereignissen ist unabhängig. Eine Familie ist genau dann unabhängig, wenn jede endliche Teilfamilie unabhängig ist.*

(ii) *Ist $\{A_i, i \in I\}$ eine Familie von unabhängigen Ereignissen, k ein nicht zu I gehöriger Index, und $P(A_k) = 0$ oder $P(A_k) = 1$, so ist auch $\{A_i, i \in I \cup \{k\}\}$ unabhängig.*

(iii) *Ist $\{A_i, i \in I\}$ unabhängig und für jedes i B_i eines der Ereignisse $A_i, A_i^c, \emptyset, \Omega$, so ist $\{B_i, i \in I\}$ unabhängig.*

(iv) *Ist $I = \{1, \dots, n\}$ endlich, so ist $\{A_i, i \in I\}$ genau dann unabhängig, wenn für jede Wahl von $B_i \in \{A_i, A_i^c\}$ die Produktformel für $B_1, \dots, B_n$ gilt.*

Beweis: (i) folgt direkt aus der Definition der Unabhängigkeit.

(ii) Wir müssen (2.5) für endliche Mengen $J \subset I \cup \{k\}$ mit $k \in J$ prüfen. Ist $P(A_k) = 0$, so steht auf beiden Seiten 0. Ist $P(A_k) = 1$, so ist $P(A_k \cap A) = P(A)$ für alle Ereignisse A . Auf beiden Seiten von (2.5) steht dann der gleiche Wert, den man erhält, wenn J um $\{k\}$ verkleinert wird.

(iii) Wegen (ii) brauchen wir nur den Fall $B_i \in \{A_i, A_i^c\}$ zu betrachten. Durch Induktion über m beweisen wir: Ist $J \subset I$ endlich und $\text{card}(\{j \in J : B_j \neq A_j\}) \leq m$, so gilt die Produktformel für $\{B_j, j \in J\}$.

Ist $m = 0$, so sind alle $B_j = A_j$ und (2.5) besagt, dass die Produktformel für die B_j gilt. Sei nun die Induktionsannahme für m bewiesen und J' eine endliche Teilmenge von I mit $\text{card}\{j \in J' : B_j \neq A_j\} = m + 1$. Da es auf die Nummerierung nicht ankommt, sei o.B.d.A. $J' = \{1, \dots, N\}$ mit $N \geq m + 1$ und $B_1 = A_1^c$. Wenden wir nun die Induktionsannahme auf $A_1, B_2, \dots, B_N$ und auf $B_2, \dots, B_N$ an, so folgt

$$\begin{aligned} P\left(\bigcap_{j=1}^N B_j\right) &= P\left(\bigcap_{j=2}^N B_j\right) - P\left(A_1 \cap \bigcap_{j=2}^N B_j\right) \\ &= \prod_{j=2}^N P(B_j) - P(A_1) \cdot \prod_{j=2}^N P(B_j) = \prod_{j=1}^N P(B_j). \end{aligned}$$

Also gilt die Produktformel für $\{B_j, j \in J'\}$.

(iv) Dass die Gültigkeit der Produktformel für jede Wahl der B_i notwendig ist, haben wir gerade gezeigt. Jetzt schließen wir umgekehrt. Addieren wir die Produktformeln für $B_1, B_2, \dots, B_n$ und für $B_1^c, B_2, \dots, B_n$, so folgt

$$P(B_2 \cap \dots \cap B_n) = \prod_{i=2}^n P(B_i).$$

So erhält man die Gleichung auch für Durchschnitte von $n - 1$ Mengen, dann für solche aus $n - 2$ Mengen, usw. $\square$

2.3 Produktexperimente

Wir nehmen an, wir kennen schon Modelle $(\Omega_1, P_1), (\Omega_2, P_2), \dots, (\Omega_n, P_n)$ für gewisse Zufallsexperimente und wollen nun ein Modell für das Experiment konstruieren, welches in der unabhängigen Hintereinanderausführung dieser Telexperimente besteht. Z.B. könnte jedes $\Omega_i = \{1, 2, \dots, 6\}$ und jedes P_i die Gleichverteilung sein, also (Ω_i, P_i) ein Modell für das einfache Würfeln, und wir wollen ein Modell für das n -fache Würfeln bilden. Wir setzen aber nicht voraus, dass die Telexperimente gleich sind, oder dass ein P_i die Gleichverteilung ist. Es liegt nahe, als Ergebnisraum das kartesische Produkt

$$\Omega = \prod_{i=1}^n \Omega_i = \{\omega = (\omega_1, \omega_2, \dots, \omega_n) : \omega_i \in \Omega_i \ (i = 1, \dots, n)\} \quad (2.6)$$

zu wählen, und als Wahrscheinlichkeitsfunktion

$$P(\omega) = \prod_{i=1}^n P_i(\omega_i). \quad (2.7)$$

Man bezeichnet (Ω, P) als **Produkt der Wahrscheinlichkeitsräume** (Ω_i, P_i) , und schreibt

$$\Omega = \Omega_1 \times \dots \times \Omega_n, \quad P = P_1 \times \dots \times P_n = \prod_{i=1}^n P_i.$$

Man prüft leicht nach, dass durch (2.7) eine Wahrscheinlichkeitsfunktion gegeben ist. Ist $X_i(\omega)$ die i -te Koordinate von $\omega = (\omega_1, \omega_2, \dots, \omega_n)$, so ist in Ω das Ereignis, dass sich im i -ten Telexperiment $A_i \subset \Omega_i$ ereignet, durch $\{\omega \in \Omega : X_i(\omega) \in A_i\}$ beschrieben. Abkürzend schreiben wir dafür auch $\{X_i \in A_i\}$. Das kartesische Produkt

$$\prod_{i=1}^n A_i = A_1 \times A_2 \times \dots \times A_n = \bigcap_{i=1}^n \{X_i \in A_i\}$$

beschreibt das Ereignis, dass sich für alle $i = 1, 2, \dots, n$ im i -ten Telexperiment A_i ereignet. Unter P ist die Wahrscheinlichkeit dafür

$$\begin{aligned} P\left(\bigcap_{i=1}^n \{X_i \in A_i\}\right) &= \sum_{\omega \in A_1 \times \dots \times A_n} P(\omega) = \sum_{\omega \in A_1} \dots \sum_{\omega \in A_n} P_1(\omega_1) \cdot \dots \cdot P_n(\omega_n) \\ &= \left(\sum_{\omega \in A_1} P_1(\omega_1)\right) \cdot \dots \cdot \left(\sum_{\omega \in A_n} P_n(\omega)\right) = \prod_{i=1}^n P_i(A_i). \end{aligned}$$

Hält man ein beliebiges k mit $1 \leq k \leq n$ fest und setzt für alle $j \neq k$ $A_j = \Omega_j$, so folgt speziell $\{X_j \in A_j\} = \Omega$ für $j \neq k$. Dann ist

$$\bigcap_{i=1}^n \{X_i \in A_i\} = \{X_k \in A_k\},$$

und aus der eben ausgeführten Rechnung ergibt sich $P(\{X_k \in A_k\}) = P_k(A_k)$. Dies entspricht der selbstverständlichen Forderung an (Ω, P) , dass die Wahrscheinlichkeit dafür, dass sich im k -ten Telexperiment A_k ereignet, mit der Wahrscheinlichkeit übereinstimmen soll, die für A_k im k -ten Teilmodell (Ω_k, P_k) vorgegeben ist.

Zugleich folgt aus der Rechnung nun

$$P\left(\bigcap_{i=1}^n \{X_i \in A_i\}\right) = \prod_{i=1}^n P(\{X_i \in A_i\}).$$

Da hierin beliebig viele $A_i = \Omega_i$ gesetzt werden dürfen (alle Rechnungen galten für beliebige A_i), gilt auch für alle $J \subset \{1, \dots, n\}$ die Produktformel für $\{\{X_i \in A_i\}, i \in J\}$.

Das Modell hat also wirklich die geforderte Eigenschaft, dass darin Ereignisse, die etwas über die Ausgänge verschiedener Telexperimente aussagen, unabhängig sind. Man überlegt sich nun auch leicht, dass *nur* der Ansatz (2.7) die Forderungen erfüllt, dass für alle A_j $P(\{X_j \in A_j\}) = P_j(A_j)$ ist und die $\{X_j \in A_j\}$ unabhängig sind.

Wir wollen die oben nachgewiesene Unabhängigkeit von Ereignissen, die auf verschiedenen Telexperimenten basieren, noch etwas verallgemeinern: Ist z.B. A das Ereignis, dass unter den ersten fünf von zehn Würfeln eines Würfels keine 6 vorkommt, und B das Ereignis, dass unter den letzten 5 Würfeln 4 Sechsen vorkommen, so sollten A und B unabhängig sein. Gewöhnlich wird dies aus der Intuition heraus auch ohne Beweis benutzt, aber korrekter ist es sicher, dies auch zu beweisen.

Ist $I = \{i_1, \dots, i_d\} \subset \{1, \dots, n\}$, so sei im obigen Produktmodell

$$\Omega_I = \Omega_{i_1} \times \Omega_{i_2} \times \dots \times \Omega_{i_d}$$

und X_I die Abbildung

$$\omega \mapsto X_I(\omega) = (X_{i_1}(\omega), \dots, X_{i_d}(\omega)) \in \Omega_I.$$

Wir sagen, dass ein Ereignis $A \subset \Omega$ nur von den Koordinaten in I abhängt, wenn es ein $A_I \subset \Omega_I$ mit $A = \{\omega \in \Omega : X_I(\omega) \in A_I\}$ gibt. (Man überlege sich, dass im eben gegebenen Würfelbeispiel das Ereignis A nur von den Koordinaten mit Index in $\{1, \dots, 5\}$ und B nur von den Koordinaten mit Index in $\{6, \dots, 10\}$ abhängt. Man sagt auch, A hänge von den Zeitpunkten $1, \dots, 5$ ab.)

Es gilt:

Satz 2.6 *Ist im obigen Produktmodell $\{1, \dots, n\}$ die disjunkte Vereinigung von Indexmengen $I_1, \dots, I_p \neq \emptyset$, und hängt für $\nu = 1, \dots, p$ das Ereignis $A^{(\nu)}$ nur von den Koordinaten in I_ν ab, so sind $A^{(1)}, \dots, A^{(p)}$ unabhängig.*

Beweis: Durch Umnummerierung der Koordinaten erreichen wir $I_1 = \{1, \dots, d_1\}$, $I_2 = \{d_1 + 1, \dots, d_1 + d_2\}$, Jedes $\omega = (\omega_1, \dots, \omega_n)$ können wir auch als Folge $(\omega(1), \dots, \omega(p))$ mit $\omega(\nu) \in \Omega_{I_\nu}$ auffassen, d.h. es ist $\Omega = \prod_{\nu=1}^p \Omega_{I_\nu}$. Die Variablen $X_{I_1}, \dots, X_{I_p}$ können wir als die p Koordinaten in dieser Produktraum-Darstellung ansehen. Setzt man $P^{(\nu)} = \prod_{i \in I_\nu} P_i$, so ist auch $P = \prod_{\nu=1}^p P^{(\nu)}$. Da $A^{(\nu)}$ nur von den Koordinaten in I_ν abhängt, gibt es eine Menge $A_{I_\nu} \subset \Omega_{I_\nu}$ mit $A^{(\nu)} = \{\omega = (\omega(1), \dots, \omega(p)) : X_{I_\nu}(\omega) \in A_{I_\nu}\}$. Damit ist die Behauptung auf den oben diskutierten Fall der Unabhängigkeit der Mengen $\{X_i \in A_i\}$ reduziert. $\square$

2.4 Einige Verteilungen für Produktexperimente

Es gibt in der Wahrscheinlichkeitstheorie einige Verteilungen, die immer wieder in Theorie und Praxis auftreten. Ein paar davon wollen wir nun kennen lernen.

Die Binomialverteilung: Wir betrachten ein Experiment, welches in der unabhängigen n -fachen Wiederholung eines Einzelexperiments mit nur zwei verschiedenen möglichen Ausgängen besteht.

Bezeichnen wir die beiden Ausgänge mit 0 und 1, so ist also $\Omega_i = \{0, 1\}$ der Stichprobenraum für das i -te Teilexperiment. Da die Teilexperimente Wiederholungen des gleichen Experiments sind, sollen in allen Teilexperimenten die gleichen Wahrscheinlichkeiten P_i auftreten. $p = P_i(1)$ soll also nicht von i abhängen. Natürlich ist dann $P_i(0) = 1 - p$. Im Produktmodell ist

$$\Omega = \{0, 1\}^n = \{(\omega_1, \dots, \omega_n) : \omega_i \in \{0, 1\}, 1 \leq i \leq n\}$$

und $P(\omega) = p^k (1 - p)^{n-k}$, wenn k die Anzahl der Einsen in $\omega = (\omega_1, \dots, \omega_n)$ ist. Oft nennt man p **Erfolgswahrscheinlichkeit** und spricht von einem Erfolg im i -ten Teilexperiment, wenn $\omega_i = 1$ ist. Ein Experiment dieser Form nennt man **Bernoulli-Experiment** und P heißt **Bernoulli-Verteilung**. Man spricht auch von n Bernoulli-Experimenten mit Erfolgswahrscheinlichkeit p . Das Ereignis, dass insgesamt genau k Erfolge auftreten, wird durch

$$E_k = \left\{ \omega \in \Omega : \sum_{i=1}^n \omega_i = k \right\}$$

beschrieben. Die Zahl der Elemente von E_k ist gleich der Zahl der Möglichkeiten, die k Zeitpunkte in $\{1, \dots, n\}$ festzulegen, an denen die Erfolge auftreten sollen, also $\binom{n}{k}$. Es folgt

$$P(E_k) = \binom{n}{k} p^k (1-p)^{n-k} \quad (0 \leq k \leq n). \quad (2.8)$$

Die rechts auftretenden Terme

$$b_{n,p}(k) = \binom{n}{k} p^k (1-p)^{n-k} \quad (2.9)$$

sind ≥ 0 und haben die Summe 1, da die E_k eine disjunkte Zerlegung von Ω bilden. Sie bestimmen also eine Wahrscheinlichkeitsverteilung auf $\{0, 1, \dots, n\}$. Man nennt sie **Binomialverteilung** mit Parametern n und p oder $b_{n,p}$ -Verteilung. Zur Abkürzung setzt man oft $q = 1 - p$.

Beispiele: (1) (10-mal Würfeln) Ein Würfelexperiment hat natürlich nicht zwei sondern sechs mögliche Ausgänge. Fragt man aber z.B. nach der Wahrscheinlichkeit bei 10 Würfeln 3 Sechsen zu erhalten, so können wir die geworfenen Sechsen als Erfolge betrachten und die übrigen Würfe als Misserfolge. Die Erfolgswahrscheinlichkeit p ist $1/6$. Also ist die gesuchte Wahrscheinlichkeit $\binom{10}{3} \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^7$.

(2) Ist $p = 0,51$ die Wahrscheinlichkeit für die Geburt eines Jungen, und sind aufeinander folgende Geburten unabhängig, so ist die Wahrscheinlichkeit dafür, dass in einer Familie mit 4 Kindern 2 Jungen und 2 Mädchen vorkommen, $\binom{4}{2} p^2 (1-p)^2 \approx 0,374$.

Die Multinomialverteilung: Hier hat man wieder n unabhängige identische, d.h. untereinander gleiche, Teilversuche, aber jeder Teilversuch hat nun allgemeiner r verschiedene mögliche Ausgänge. Die Teilerperimente sind also beschreibbar durch $P_i(j) = p_j$ ($j = 1, \dots, r$), wobei der Wahrscheinlichkeitsvektor $(p_1, \dots, p_r)$ beliebig vorgegeben ist. Die Wahrscheinlichkeit, in den n Teilversuchen insgesamt k_1 -mal das Ergebnis 1 k_2 -mal das Ergebnis 2, $\dots$, und k_r -mal das Ergebnis r zu erhalten, ist

$$\frac{n!}{k_1! k_2! \dots k_r!} p_1^{k_1} \dots p_r^{k_r},$$

wenn die k_i ganze Zahlen ≥ 0 mit $k_1 + k_2 + \dots + k_r = n$ sind, und natürlich $= 0$ sonst.

Der Beweis ist der gleiche wie im Fall der Binomialverteilung. An die Stelle der Binomialkoeffizienten $\binom{n}{k}$ treten die Multinomialkoeffizienten.

Die geometrische Verteilung: Die Wahrscheinlichkeit, erst im neunten Wurf mit einem korrekten Würfel die erste 6 zu würfeln, ist offenbar $\left(\frac{5}{6}\right)^8 \cdot \frac{1}{6}$, denn dies Ereignis tritt genau dann ein, wenn die ersten 8 Würfe keine 6 ergeben, aber der neunte. Allgemein ist in einem Bernoulli-Experiment mit Erfolgswahrscheinlichkeit $p > 0$ die Wahrscheinlichkeit, den ersten Erfolg genau im k -ten Teilversuch zu haben,

$$p(1-p)^{k-1} = p q^{k-1} \quad (k = 1, 2, \dots).$$

Im Produktraum $\{0, 1\}^n$ ist das gesuchte Ereignis natürlich nur beschreibbar, wenn $n \geq k$ ist. Aber wir können auch k als Ergebnis eines Experiments auffassen, das darin besteht, zu beobachten, in welchem Teilerperiment einer Folge von Bernoulli-Versuchen der erste Erfolg eintritt. Dann ist $\Omega = \{1, 2, \dots\}$ die zugehörige Ergebnismenge und $P(k) = p q^{k-1}$ die Wahrscheinlichkeitsfunktion. Es ist $P(k) \geq 0$ und $\sum P(k) = 1$. Der einzige Unterschied zu den bisher benutzten Modellen besteht darin, dass Ω nun abzählbar unendlich ist. Die durch die Wahrscheinlichkeitsfunktion $P(k)$ auf Ω definierte Verteilung heißt **geometrische Verteilung**. (Manche Autoren nennen die Verteilung $p q^k$ auf $\{0, 1, \dots\}$ geometrische Verteilung.)

Die negative Binomialverteilung: Dies ist eine Verallgemeinerung der geometrischen Verteilung. Sei $f(k; r, p)$ die Wahrscheinlichkeit dafür, dass bei $n \geq r + k$ Bernoulli-Experimenten mit Erfolgswahrscheinlichkeit $p > 0$ genau k Misserfolge dem r -ten Erfolg vorangehen. Ist $n = r + k$, so ist das beschriebene Ereignis die Menge der Folgen $\omega = (\omega_1, \omega_2, \dots, \omega_n) \in \{0, 1\}^n$, welche genau r Einsen enthalten und mit einer 1 enden. Es gibt genau $\binom{k+r-1}{k}$ solche Folgen, denn dies ist die Zahl der Möglichkeiten, die Nullen auf die Stellen $1, 2, \dots, r + k - 1$ zu verteilen. Jedes solche ω hat Wahrscheinlichkeit $p^r q^k$. Also ist

$$f(k; r, p) = \binom{k+r-1}{k} p^r q^k. \quad (2.10)$$

Man überlegt sich leicht, dass man mit mehr als $r+k$ Bernoulli-Experimenten den gleichen Wert erhält. Die durch (2.13) bei festem r für $k = 0, 1, 2, \dots$ definierte Verteilung auf $\{0, 1, 2, \dots\}$ nennt man die **negative Binomialverteilung** (oder **Pascal-Verteilung**) mit Parametern $r, p > 0$.

Der Grund dafür liegt in der Möglichkeit, (2.13) so umzuschreiben, dass formal eine Ähnlichkeit zur Binomialverteilung entsteht. Setzt man nämlich für beliebige, nicht mehr notwendig positive ganze Zahlen m

$$\binom{m}{0} = 1 \quad \text{und} \quad \binom{m}{k} = \frac{m \cdot (m-1) \cdot \dots \cdot (m-k+1)}{k!} \quad \text{für } k \geq 1,$$

so lässt sich für $r \geq 1$ und $k \geq 1$ leicht nachrechnen, dass

$$f(k, r, p) = \binom{-r}{k} (-1)^k p^r q^k$$

ist, und dies ist auch für $k = 0$ richtig.

2.5 Diskrete Wahrscheinlichkeitsräume

Wir haben damit Beispiele für Verteilungen auf Ergebnismengen Ω erhalten, die nicht mehr endlich, aber noch abzählbar unendlich sind. In Verallgemeinerung der endlichen Wahrscheinlichkeitsräume nennen wir ein Paar (Ω, P) einen **diskreten Wahrscheinlichkeitsraum**, wenn Ω eine nichtleere höchstens abzählbar unendliche Menge ist und P eine Abbildung, die jedem $A \subset \Omega$ einen Wert $P(A)$ so zuordnet, dass die folgenden drei Beziehungen erfüllt sind:

$$P(\Omega) = 1. \quad (2.11)$$

$$P(A) \geq 0 \quad \text{für alle } A \subset \Omega. \quad (2.12)$$

Ist $A_1, A_2, \dots$ eine Folge disjunkter Ereignisse in Ω , so ist

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i). \quad (2.13)$$

Wenn (2.13) für beliebige disjunkte Ereignisse A_i gilt, nennen wir P **σ -additiv**. Dabei nennen wir immer noch alle Teilmengen von Ω Ereignisse. In endlichen Wahrscheinlichkeitsräumen folgt σ -Additivität aus der dort geforderten schwächeren Additivität, denn in der disjunkten Folge $A_1, A_2, \dots$ müssen dann alle A_i bis auf endlich viele leer sein. In abzählbaren Wahrscheinlichkeitsräumen ist σ -Additivität eine echt stärkere Forderung. Wir gehen auf diese Problematik kurz im Anhang zu diesem Paragraphen ein.

Ist $B_1 \subset B_2 \subset \dots$ eine wachsende Folge von Ereignissen und B ihre Vereinigung, so gilt $P(B_k) \rightarrow P(B)$, denn B ist die abzählbare disjunkte Vereinigung der Mengen $A_1 = B_1$, $A_2 = B_2 \setminus B_1$, $A_3 = B_3 \setminus B_2$, $\dots$, und B_k ist die endliche disjunkte Vereinigung der Mengen $A_1, \dots, A_k$. Die Behauptung folgt also aus $\sum_{i=1}^k P(A_i) \rightarrow \sum_{i=1}^{\infty} P(A_i)$. Ist $C_1 \supset C_2 \supset \dots$ eine fallende Folge von Ereignissen und C ihr Durchschnitt, so gilt $P(C_k) \rightarrow P(C)$. Dies folgt, indem man die Komplemente $B_k = C_k^c$ und $B = C^c$ bildet, wegen $P(C_k) = 1 - P(B_k)$ und $P(C) = 1 - P(B)$. Man spricht von der **Stetigkeit** von P für monotone Folgen von Ereignissen.

Die bisher bewiesenen Sätze über Eigenschaften von Wahrscheinlichkeitsräumen, bedingte Wahrscheinlichkeiten, Unabhängigkeit, usw. gelten auch für allgemeine diskrete Wahrscheinlichkeitsräume, wie man sich leicht überzeugt.

2.6 Konstruktion von Wahrscheinlichkeitsräumen aus bedingten Wahrscheinlichkeiten

Vielfach lassen sich aus der Art der Konstruktion eines Zufallsexperiments Schlüsse darüber ziehen, welche bedingten Wahrscheinlichkeiten für gewisse Ereignisse sich in einem realistischen Modell ergeben müssten. Dann kann es zweckmäßig sein, das Modell mit Hilfe dieser der experimentellen Anordnung entnommenen bedingten Wahrscheinlichkeiten zu definieren. Ein Beispiel ist das **Urnenmodell von Pólya**:

Eine Urne enthält zu Beginn W weiße und S schwarze Kugeln. Es wird eine Kugel entnommen. Sie wird zurückgelegt, und für ein vorgegebenes $c \in \mathbb{Z}$ werden c Kugeln der gezogenen Farbe zusätzlich in die Urne gelegt. (Ist c negativ, so werden $|c|$ Kugeln entnommen. Man muss dann annehmen, dass W und S hinreichend groß sind.) Nun wird erneut eine Kugel gezogen. Sie wird wieder zurückgelegt und c Kugeln der gezogenen Farbe werden zugefügt, bevor neu gezogen wird. (Vor jeder Ziehung wird neu geschüttelt.)

Im Fall $c = 0$ ergibt sich eine Folge von Ziehungen mit Rücklegen und im Fall $c = -1$ ohne Rücklegen. Im Fall $c > 0$ wird bei nachfolgenden Ziehungen eine Farbe umso wahrscheinlicher, je öfter sie schon gezogen worden ist. Pólya wollte damit ein Modell für die zufällige Ausbreitung ansteckender Krankheiten vorschlagen. Repräsentiert etwa eine gezogene schwarze Kugel einen beobachteten Krankheitsfall, so ist im nächsten Zeitpunkt wegen der Gefahr der Ansteckung die Wahrscheinlichkeit der Beobachtung eines weiteren Krankheitsfalls erhöht.

Das Ergebnis von n Ziehungen können wir durch eine Folge $\omega = (\omega_1, \dots, \omega_n) \in \Omega = \{w, s\}^n$ beschreiben, in der $\omega_i = w$ sein soll, wenn die i -te gezogene Kugel weiß und $= s$, wenn sie schwarz ist. Um z.B. für $\omega = (w, s, w, w)$ die Wahrscheinlichkeit $P(\omega)$ zu bestimmen, können wir wie folgt vorgehen:

Sei $A_1 = \{\omega : \omega_1 = w\}$, $A_2 = \{\omega : \omega_2 = s\}$, $A_3 = \{\omega : \omega_3 = w\}$, $A_4 = \{\omega : \omega_4 = w\}$. Dann ist $\{\omega\} = A_1 \cap A_2 \cap A_3 \cap A_4$ und

$$P(\omega) = P(A_1) \cdot P(A_2 | A_1) \cdot P(A_3 | A_1 \cap A_2) \cdot P(A_4 | A_1 \cap A_2 \cap A_3).$$

Offenbar muss bei einem realistischen Ansatz $P(A_1) = W/(W + S)$ sein. Ist die erste Kugel weiß, so sind bei der zweiten Ziehung $W + c$ weiße Kugeln und S schwarze in der Urne. Also muss $P(A_2 | A_1) = S/(W + c + S)$ sein. Wir erhalten so

$$P((w, s, w, w)) = \frac{W}{W + S} \cdot \frac{S}{W + c + S} \cdot \frac{W + c}{W + 2c + S} \cdot \frac{W + 2c}{W + 3c + S}.$$

Nach genau dem gleichen Schema lässt sich die gesamte Wahrscheinlichkeitsfunktion bestimmen. Bei diesem Experiment gibt es im Prinzip auch die Möglichkeit der Modellierung mit einem laplaceschen Wahrscheinlichkeitsraum, aber ein solches Modell wäre äußerst kompliziert.

Das oben benutzte Schema zur Definition einer Wahrscheinlichkeitsfunktion mit Hilfe von bedingten Wahrscheinlichkeiten lässt sich folgendermaßen allgemein fassen:

Ein Experiment bestehe aus n Teilversuchen (**Stufen**). Ω_i sei Ergebnismenge für den i -ten Teilversuch. $P_1(\omega_1)$ sei die Wahrscheinlichkeit dafür, dass im ersten Teilversuch ω_1 beobachtet wird. Für $1 < i \leq n$ sei $P_i(\omega_i | \omega_1, \dots, \omega_{i-1})$ die bedingte Wahrscheinlichkeit dafür, dass ω_i in der i -ten Stufe beobachtet wird, wenn $\omega_1, \dots, \omega_{i-1}$ die Ergebnisse der Stufen $1, \dots, i-1$ sind. Dann wird das Gesamtexperiment beschrieben durch

$$\Omega = \Omega_1 \times \dots \times \Omega_n = \{\omega = (\omega_1, \dots, \omega_n) : \omega_i \in \Omega_i \ (1 \leq i \leq n)\} \quad (2.14)$$

und

$$P(\omega) = P_1(\omega_1) \cdot P_2(\omega_2 | \omega_1) \cdot \dots \cdot P_n(\omega_n | \omega_1, \dots, \omega_{n-1}). \quad (2.15)$$

Sei wieder $X_i((\omega_1, \dots, \omega_n)) = \omega_i$. Das Ereignis, dass im i -ten Teilexperiment $\eta_i \in \Omega_i$ beobachtet wird, wird dann in Ω durch $\{X_i = \eta_i\}$ beschrieben. Der folgende Satz zeigt, dass der Ansatz (2.15) gerade die gewünschten Eigenschaften hat:

Satz 2.7 (i) *Durch (2.15) ist eine Wahrscheinlichkeitsfunktion auf Ω definiert.*

(ii) *Für alle $\eta_i \in \Omega_i$ ($i = 1, \dots, n$) ist*

$$P(\{X_1 = \eta_1\}) = P_1(\eta_1) \quad (2.16)$$

und

$$P(\{X_i = \eta_i\} | \{X_1 = \eta_1\} \cap \dots \cap \{X_{i-1} = \eta_{i-1}\}) = P_i(\eta_i | \eta_1, \dots, \eta_{i-1}). \quad (2.17)$$

(iii) *Der Ansatz (2.15) ist der einzige mit den Eigenschaften (i) und (ii).*

Beweis: Zum Beweis von (2.16) und (2.17) berechnen wir

$$\begin{aligned} & P(\{X_1 = \eta_1\} \cap \dots \cap \{X_i = \eta_i\}) \\ &= \sum_{(\omega_{i+1}, \dots, \omega_n)} P_1(\eta_1) \cdot P_2(\eta_2 | \eta_1) \cdot \dots \cdot P_i(\eta_i | \eta_1, \dots, \eta_{i-1}) \cdot \\ & \quad \cdot P_{i+1}(\omega_{i+1} | \eta_1, \dots, \eta_i) \cdot \\ & \quad \cdot P_{i+2}(\omega_{i+2} | \eta_1, \dots, \eta_i, \omega_{i+1}) \cdot \dots \cdot P_n(\omega_n | \eta_1, \dots, \eta_i, \omega_{i+1}, \dots, \omega_{n-1}). \end{aligned}$$

Bei der Summation über $\omega_n \in \Omega_n$ bei festem $\eta_1, \dots, \eta_i, \dots, \omega_{n-1}$ ist die Summe der $P_n(\omega_n | \dots) = 1$, so dass wir den letzten Faktor und die Summation über ω_n streichen können. Dann fällt der vorletzte Faktor durch Summation über ω_{n-1} weg, usw. Schließlich erhalten wir

$$= P_1(\eta_1) \cdot \dots \cdot P_i(\eta_i | \eta_1, \dots, \eta_{i-1}).$$

Im Fall $i = 1$ steht nur der erste Faktor da, so dass (2.16) gilt. Im Fall $i > 1$ erhalten wir (2.17) durch Einsetzen in die Definitionsgleichung der bedingten Wahrscheinlichkeit. Summiert man über alle $(\omega_1, \dots, \omega_n)$, so folgt ebenso $\sum P(\omega) = 1$. $P(\omega) \geq 0$ ist trivial. Dass (2.15) der einzige Ansatz mit (2.16) und (2.17) ist, folgt aus (2.4). $\square$

Die Möglichkeit zu vorgegebenen Werten von $P_1(\omega_1)$ und $P_i(\omega_i | \omega_1, \dots, \omega_{i-1})$ ein Modell mit den Eigenschaften (2.16) und (2.17) zu konstruieren, kann man als Rechtfertigung für Rechnungen mit bedingten Wahrscheinlichkeiten ansehen, die unmittelbar der Versuchsanordnung entnommen werden. Jedenfalls kann man nun sicher sein, dass ein Modell existiert, in dem dies wirklich bedingte Wahrscheinlichkeiten im Sinn der Definition (2.2) sind, und es gibt keinen Grund, ein mögliches alternatives Modell für zuverlässiger zu halten.

2.7 Austauschbare Verteilungen*

Das pólyasche Urnenmodell soll nun noch etwas weiterverfolgt werden, da es uns zu einem nützlichen Begriff führen wird. Wir wollen uns zunächst überlegen, dass der Ansatz (2.15) im Spezialfall des pólyaschen Modells zu einer einigermaßen übersichtlichen Formel führt:

Satz 2.8 *Im pólyaschen Urnenmodell ist für jedes $\omega = (\omega_1, \dots, \omega_n)$ mit k schwarzen Kugeln in der Stichprobe*

$$P(\omega) = \prod_{\nu=0}^{n-1} (W + S + \nu c)^{-1} \cdot \prod_{i=0}^{n-k-1} (W + ic) \cdot \prod_{j=0}^{k-1} (S + jc). \quad (2.18)$$

Beweis: Betrachten wir z.B. $\omega = (w, s, s, w, w)$, so ist

$$P(\omega) = \frac{W}{S+W} \cdot \frac{S}{S+W+c} \cdot \frac{S+c}{S+W+2c} \cdot \frac{W+c}{S+W+3c} \cdot \frac{W+2c}{S+W+4c}.$$

Man sieht, dass die Folge der Nenner unabhängig von ω ist und zum ersten Faktor in (2.18) führt. Bei Ziehung der $(i+1)$ -ten weißen Kugel erscheint im Zähler des entsprechenden Bruchs der Faktor $W+ic$, ganz gleich, wie viele schwarze Kugeln vorher gezogen wurden. Entsprechend erscheint $S+jc$ im Zähler des Quotienten, der zur Ziehung der $(j+1)$ -ten schwarzen Kugel gehört. $\square$

Sei $\Omega_1 = \Omega_2 = \dots = \Omega_n$ und $\Omega = \Omega_1 \times \dots \times \Omega_n$. Ein Wahrscheinlichkeitsmaß P auf Ω heißt **austauschbar**, wenn für alle Permutationen τ von $\{1, \dots, n\}$ und für alle $\omega \in \Omega$ gilt

$$P(\omega) = P((\omega_{\tau(1)}, \omega_{\tau(2)}, \dots, \omega_{\tau(n)})).$$

Aus Satz 2.8 folgt, dass die Verteilung auf $\{w, s\}^n$, die zum pólyaschen Modell gehört, austauschbar ist.

Über austauschbare Verteilungen gibt es viele interessante Resultate. Wir betrachten hier nur eine einfache Eigenschaft, die es aber immerhin gestatten wird, etwas Neues über das Ziehen von Kugeln ohne Rücklegen zu lernen.

Zieht man aus einer Urne, die zu Beginn W weiße und S schwarze Kugeln enthält, sukzessive ohne Rücklegen $n \leq W+S$ Kugeln, so entspricht dies, wie erwähnt, dem Fall $c = -1$. Es ist

$$P(\{X_1 = w\}) = \frac{W}{W+S}$$

und

$$\begin{aligned}
P(\{X_2 = w\}) &= P(\{X_1 = w\}) P(\{X_2 = w\} | \{X_1 = w\}) \\
&\quad + P(\{X_1 = s\}) P(\{X_2 = w\} | \{X_1 = s\}) \\
&= \frac{W}{W+S} \cdot \frac{W-1}{W+S-1} + \frac{S}{W+S} \cdot \frac{W}{W+S-1} \\
&= \frac{W}{W+S} = P(\{X_1 = w\}).
\end{aligned}$$

Rechnet man weiter, so erhält man auch $P(\{X_3 = w\}) = W/(W+S)$; nur werden die Rechnungen jetzt schnell aufwändiger. Wollte man auf diese Weise noch Wahrscheinlichkeiten wie $P(\{X_{10} = w\})$ oder $P(\{X_7 = s\} | \{X_{10} = w\})$ ausrechnen, so brauchte man allerhand Geduld. Alles wird aber ganz einfach, wenn man sich überlegt, dass die Austauschbarkeit der Verteilung es gestattet, die Zeitpunkte nicht nur für Elementarereignisse sondern auch für andere Ereignisse zu vertauschen. Aus dem folgenden Satz folgt dann z.B.

$$\begin{aligned}
P(\{X_7 = s\} | \{X_{10} = w\} \cap \{X_{12} = s\}) &= P(\{X_3 = s\} | \{X_1 = w\} \cap \{X_2 = s\}) \\
&= \frac{S-1}{W+S-2}.
\end{aligned}$$

Satz 2.9 Sei $\Omega_1 = \Omega_2 = \dots = \Omega_n \neq \emptyset$ abzählbar und P auf $\Omega = \prod_{i=1}^n \Omega_i$ austauschbar. Bezeichne X_i die i -te Koordinaten-Abbildung, und $X = (X_1, \dots, X_n)$. Für eine beliebige Permutation τ von $\{1, \dots, n\}$ sei $X^\tau = (X_{\tau(1)}, \dots, X_{\tau(n)})$. Für jedes $A \subset \Omega$ gilt

$$P(X \in A) = P(X^\tau \in A).$$

Insbesondere gilt für beliebige $A_i \subset \Omega_i$

$$P\left(\bigcap_{i=1}^n \{X_i \in A_i\}\right) = P\left(\bigcap_{i=1}^n \{X_{\tau(i)} \in A_i\}\right).$$

Beweis: Die Austauschbarkeit bedeutet gerade, dass für beliebige $a \in \Omega$ $P(X = a) = P(X^\tau = a)$ für alle τ gilt. Man braucht dies nur über alle $a \in A$ zu summieren. $\square$

2.8 Genetische Modelle*

Im einfachsten Fall verläuft bei so genannten „diploiden“ Organismen, z.B. beim Menschen, die Vererbung von Eigenschaften nach dem folgenden Schema:

Jede Körperzelle des Organismus enthält in ihrem Zellkern die gleichen Chromosomen. Entlang der Chromosomen sind **Gene** wie in einer Kette nebeneinander aufgereiht. Dies sind die Erbfaktoren, die jeweils für ein Merkmal verantwortlich sind, entweder allein oder gemeinsam mit anderen Genen. Ein Gen kann in zwei oder mehr Zustandsformen auftreten, die man als **Allele** bezeichnet. Bei diploiden Organismen treten die Chromosomen paarweise auf. Zu jedem Chromosom gibt es ein zweites so genanntes homologes Chromosom mit den gleichen Genen, wobei das gleiche Gen auf den beiden homologen Chromosomen entweder als gleiches oder als verschiedenes Allel vorhanden sein kann. Gibt es für ein Gen zwei Allele A und a , so sind also die Kombinationen AA , Aa , aa möglich, die wir **Genotypen** nennen. aA lässt sich von Aa nicht unterscheiden.

Ein Beispiel eines durch ein einziges Genpaar bestimmten Merkmals beim Menschen ist die Schmeckfähigkeit für Phenylthioharnstoff (PTH). In diesem Fall ist die Schmeckfähigkeit „dominant“. Für Personen mit den Genotypen AA oder Aa schmeckt PTH leicht bitter, für Personen mit aa nicht. Es wird für uns aber keine Rolle spielen, ob wir die zu den drei Genotypen gehörigen Erscheinungsbilder (**Phänotypen**) unterscheiden können oder nicht.

Sei nun zu einem bestimmten Zeitpunkt in einer großen Bevölkerung der Genotyp AA mit der relativen Häufigkeit $u > 0$ vorhanden, der Genotyp Aa mit der relativen Häufigkeit $2v > 0$, und aa mit relativer Häufigkeit $w > 0$. Dann ist $u + 2v + w = 1$.

Ist das Gen wie im obigen Beispiel nicht geschlechtsgebunden, so überträgt bei jedem Fortpflanzungsvorgang jedes Elternteil ein Gen seines Genpaares, und zwar wird jedes der beiden Gene gerade mit Wahrscheinlichkeit $1/2$ ausgewählt, unabhängig vom anderen Elternteil. Hat z.B. der Vater den Genotyp Aa und die Mutter aa , so hat also der Nachkomme mit Wahrscheinlichkeit $1/2$ den Genotyp Aa und mit Wahrscheinlichkeit $1/2$ den Genotyp aa .

Wählt man Vater und Mutter unabhängig voneinander zufällig aus (und geht man davon aus, dass die obigen relativen Häufigkeiten $u, 2v, w$ nicht vom Geschlecht abhängen), so wird die Wahrscheinlichkeit, dass Vater und Mutter Genotyp AA haben, gerade u^2 sein. Tabelle 2.1 führt die möglichen Kombinationen der Genotypen mit deren Wahrscheinlichkeiten auf. Die letzte Spalte enthält die Wahrscheinlichkeit P_{AA} , dass diese Kombination von Genotypen zu einem Nachkommen vom Genotyp AA führt.

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

Tabelle 2.1

Mit der Formel von der totalen Wahrscheinlichkeit ergibt sich also für ein Mitglied der ersten Nachkommengeneration der Genotyp AA mit Wahrscheinlichkeit

$$P_1(AA) = u^2 + 2uv/2 + 2uv/2 + 4v^2/4 = (u + v)^2.$$

Aus Symmetriegründen gilt

$$P_1(aa) = (w + v)^2.$$

Dann ist $P_1(Aa) = 1 - (u + v)^2 - (w + v)^2$. Wegen $u + 2v + w = 1$ folgt

$$P_1(Aa) = ((u + v) + (v + w))^2 - (u + v)^2 - (w + v)^2 = 2(u + v)(v + w).$$

Gibt es Zeugungen von vielen jeweils zufällig ausgewählten Elternpaaren, so ergeben sich als neue relative Häufigkeiten von AA , Aa und aa also

$$u_1 = P_1(AA) = (u + v)^2, \quad 2v_1 = P_1(Aa) = 2(u + v)(v + w), \quad w_1 = (v + w)^2.$$

Für die nächste Generation folgen daraus die relativen Häufigkeiten

$$u_2 = (u_1 + v_1)^2, \quad 2v_2 = 2(u_1 + v_1)(v_1 + w_1), \quad w_2 = (v_1 + w_1)^2.$$

Setzt man ein, so ergibt sich

$$u_2 = ((u+v)^2 + (u+v)(v+w))^2 = (u+v)^2 ((u+v) + (v+w))^2 = (u+v)^2 = u_1.$$

Aus Symmetriegründen folgt $w_2 = w_1$, und dann muss auch $v_2 = v_1$ gelten. In der 2. Generation ist also die Häufigkeitsverteilung unverändert. Durch Induktion folgt jetzt für die k -te Generation

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

Die Häufigkeitsverteilung der Genotypen ist also in allen Nachkommengenerationen gleich. Dies ist das bekannte **Hardy-Weinberg-Theorem**, das 1908 von G. H. Hardy und W. Weinberg unabhängig voneinander entdeckt wurde. (Stillschweigend haben wir eine Reihe von Voraussetzungen gemacht: Es gibt weder Mutationen, noch Immigration, noch Selektion. Bei einer für Überleben und Partnerwahl so unwesentlichen Eigenschaft wie PTH-Schmeckfähigkeit dürfte die letztere Bedingung wohl gewährleistet sein.)

2.9 Bedingte Wahrscheinlichkeit und Scheinkorrelation*

Jeden Tag werden statistische Untersuchungen durchgeführt, die belegen sollen, ob irgendwelche Zusammenhänge zwischen gewissen Größen bestehen oder nicht. Z.B. könnte man versuchen, die Kriminalitätsrate von Ausländern und Inländern zu vergleichen, oder prüfen, ob die Ganzheitsmethode zum Lesen lernen besser geeignet ist als die klassische Methode mit dem Buchstabieren. Im einfachsten Fall würde man an n Versuchsobjekten nur beobachten, welche von zwei interessierenden Merkmalen an wie vielen Versuchsobjekten erfüllt sind. Man würde also z.B. zählen wie viele Schüler nach der Ganzheitsmethode unterrichtet worden sind, wie viele davon innerhalb von z.B. 8 Monaten ein bestimmtes Niveau der Lesefähigkeit erreicht haben, und wie hoch die entsprechenden Zahlen für die klassische Methode sind.

Wie man sorgfältig geplante Studien dieser Art auswerten kann, soll uns später beschäftigen. Hier soll uns ein Beispiel aus der Praxis zeigen, wie vorsichtig man mit solchen Zahlen sein muss. (Quelle: Bickel-O'Connell (1975) und Freedman-Pisani-Purves (1978).)

An der Universität von Berkeley wurde im Herbst 1973 gezählt, wie viele männliche und wie viele weibliche Bewerber um Zulassung zum „graduate“ Studium es gab, und wie viele jeweils zugelassen wurden. Es gab 8442 männliche Bewerber, von denen 3738 (44 %) zugelassen wurden, und 4321 weibliche Bewerber, von denen 1494 (35 %) zugelassen wurden. Da es keinerlei Anzeichen dafür gab, dass die Damen weniger qualifiziert waren als die Herren, lag der Schluss nahe, dies sei ein Fall von sexueller Diskriminierung.

Die Bewerbung erfolgte jeweils für ein bestimmtes Fach. Als man sich die einzelnen Fächer ansah, um festzustellen, wo die Hauptquelle der Diskriminierung lag, gab es eine Überraschung. In einzelnen Fächern gab es zwar eine leichte Bevorzugung der Männer, aber nicht von einer Größenordnung, die die obigen Daten erklären könnte. Und in anderen Fächern gab es eine Bevorzugung der Damen. Summa summarum sah es eher nach einer Bevorzugung der Damen aus.

Es gab ungefähr hundert Fächer, etwas zu viele, um sie hier aufzulisten. Aber das Bild aus den sechs größten Fächern, die zusammen mehr als ein Drittel aller Bewerber hatten, war relativ typisch. Auch bei Einschränkung auf diese sechs Fächer ergab sich schon das gleiche Paradoxon. Tabelle 2.2 enthält die Daten für diese sechs Fächer.

Fach	Männlich		Weiblich	
	Anzahl der Bewerber	Prozentsatz der Zulassungen	Anzahl der Bewerber	Prozentsatz der Zulassungen
1	825	62	108	82
2	560	63	25	68
3	325	37	593	34
4	417	33	375	35
5	191	28	393	24
6	373	6	341	7

Tabelle 2.2

Schränkt man sich auf diese sechs Fächer ein, so ist der Anteil der männlichen Bewerber, die zugelassen wurden, 44,5 % und der entsprechende Anteil der weiblichen Bewerber 32 %. Aber die einzigen Fächer, in denen der Prozentsatz der Zulassungen für die Männer höher ist als für die Frauen, sind die Fächer 3 und 5, und da sind die Unterschiede nur 3 % bzw. 4 %. Die Erklärung liegt woanders:

In den ersten beiden Fächern wurde ein ziemlich hoher Anteil der Bewerber zugelassen. Hier gab es eine große Zahl männlicher Bewerber und nur eine relativ kleine Zahl weiblicher Bewerber. In den anderen vier Fächern war die Zulassungsquote klein. Hier gab es überproportional viele weibliche Bewerber.

Also: *Es gab eine Tendenz, dass Frauen sich in Fächern bewarben, in denen die Zulassung schwer zu erlangen war.*

Oft studierten Frauen Fächer mit langen Studiendauern, schlechteren Arbeitsmarktchancen und geringeren Bestandteilen an Gebieten wie Mathematik.

Selbst wenn in jedem einzelnen Fach die Zulassungsquote für Frauen größer wäre als die für Männer, würde das noch nicht implizieren, dass dies auch für die Gesamtheit gilt. Man überlege sich dies mit nur zwei Fächern!

Formal lässt sich dieser zunächst verwirrende Sachverhalt folgendermaßen begründen: Sei A die Menge der zugelassenen, und B die Menge der männlichen Bewerber, sowie C_i die Menge der Bewerber im i -ten Fach. Für die zufällige Wahl einer Person aus der Gesamtheit ist nach der auf die bedingten Wahrscheinlichkeiten angewandten Formel von der totalen Wahrscheinlichkeit

$$P(A|B) = P(A|B \cap C_1)P(C_1|B) + P(A|B \cap C_2)P(C_2|B) + \dots$$

und

$$P(A|B^c) = P(A|B^c \cap C_1)P(C_1|B^c) + P(A|B^c \cap C_2)P(C_2|B^c) + \dots$$

Die Zulassungsquote $P(A|B)$ der Männer kann selbst dann wesentlich größer sein als die Zulassungsquote $P(A|B^c)$ der Frauen, wenn für alle i die Ungleichung

$$P(A|B \cap C_i) \leq P(A|B^c \cap C_i)$$

gilt, denn die „Gewichte“ $P(C_i|B)$ können gerade für die i groß sein, für die $P(A|B \cap C_i)$ relativ groß ist, und die Gewichte $P(C_i|B^c)$ können gerade für die i groß sein, für die $P(A|B^c \cap C_i)$ klein ist.

Der Schluss, dass statistische Vergleiche zur Klärung von Fragen wie der nach sexueller Diskriminierung ungeeignet seien, wäre allerdings voreilig. Man kann aus diesem Beispiel nur lernen, dass statistische Vergleiche so geplant werden sollten, dass der Einfluss anderer Größen — wie hier der Fächerwahl — möglichst eliminiert wird. Ist dies nicht möglich, so braucht man subtilere Auswertungsmethoden, die dem Rechnung tragen.

Anmerkungen*

Unabhängigkeit: Lange arbeitete man mit einem vagen Unabhängigkeitsbegriff. Man sagte „ A heißt von B unabhängig, wenn die Wahrscheinlichkeit von A nicht davon abhängt ob B eintritt“, ohne dass gesagt wurde, was dies heißen sollte. Die Produktformel galt als beweisbarer Satz (Multiplikationstheorem). Georg Bohlmann (1869–1928) gab 1908 die heute übliche präzise Definition der Unabhängigkeit an.

Zur Formel von Bayes: Die Formel von Bayes in Satz 2.2, die es erlaubt, aus den Wahrscheinlichkeiten $P(B_k)$ der Elemente B_k einer Zerlegung von Ω und aus den Wahrscheinlichkeiten $P(A|B_k)$ die „inversen“ bedingten Wahrscheinlichkeiten $P(B_k|A)$ zu berechnen, wurde 1713 von Thomas Bayes publiziert. Für wohldefinierte Zufallsmechanismen ist sie völlig unproblematisch.

Beginnend mit Laplace hat man die Formel später aber vielfach auch recht willkürlich angewandt. In den B_k wurden Ursachen gesehen, für die man **A-priori-Wahrscheinlichkeiten** $P(B_k)$ festlegte, z.B. mangels Vorwissen die Gleichverteilung. Die berechneten Größen $P(B_k|A)$ waren dann die **A-posteriori-Wahrscheinlichkeiten**, die sich aus der Information ergaben, dass A geschehen war.

Problematisch wird dies, sobald man nicht mehr von Wahrscheinlichkeiten wohldefinierter *Ereignisse* spricht, sondern *Hypothesen* Wahrscheinlichkeiten zuordnet. Es passt schlecht in den Rahmen der hier behandelten Zufallsexperimente, etwa die Erfolgswahrscheinlichkeit p für ein Heilmittel als zufällig zu betrachten, und von der A-priori- (oder A-posteriori-) Wahrscheinlichkeit zu sprechen, dass $p \leq 0,4$ ist.

Damit soll nicht gesagt sein, dass Entscheidungsverfahren, die aus Bayes-Ansätzen abgeleitet werden, nicht oft von großem Wert sind. Es lassen sich auch Beispiele angeben, bei denen man durchaus über empirisch stützbar Vorinformationen verfügt. Zum Beispiel können in der Qualitätskontrolle in langfristigen Beobachtungen ermittelte Kenntnisse über typische Schwankungen der Zahl der Ausschussexemplare vorliegen, die bei der Schätzung des Schlechtanteils einer bestimmten Lieferung mit herangezogen werden können, vgl. Stange (1977).

Zur Forderung der σ -Additivität: Die Forderung, dass Wahrscheinlichkeiten P additiv sein sollen, ist völlig unumstritten. Die Forderung, dass P σ -additiv sein soll, wird von der großen Mehrheit der Stochastiker nicht in Frage gestellt, aber sie ist zweifellos nicht so selbstverständlich wie die Additivität. Zumindest ist sie wenig intuitiv. Wer den Begriff der Wahrscheinlichkeit mit Hilfe von relativen Häufigkeiten motiviert, kann so zwar die Forderung der Additivität, aber nicht die der σ -Additivität begründen. Sind $E_1, E_2, \dots$ disjunkte Teilmengen von $\mathbb{N}$ und konvergieren die relativen Häufigkeiten $h_n(E_i) = n^{-1} \text{card}(E_i \cap \{1, \dots, n\})$, so folgt daraus nicht die Konvergenz von $h_n(\bigcup_{i=1}^{\infty} E_i)$.

Der eigentliche Grund dafür, dass σ -Additivität als Axiom ziemlich generell akzeptiert wird, dürfte praktischer Natur sein: Erstens erfüllen die für die meisten Anwendungen ausreichenden Verteilungen, nämlich die diskreten und die in Kapitel II diskutierten Verteilungen mit Dichten, dieses Axiom. Zweitens ist σ -Additivität eine ständig benutzte Eigenschaft in sehr vielen Beweisen, auf die man ungern verzichtet. Es gibt allerdings auch Problemstellungen in der Stochastik, für die sich der additive Rahmen als zweckmäßig erwiesen hat; siehe Dubins-Savage (1976).

Ein Bereich, in dem sich die Benutzung additiver aber nicht σ -additiver „Wahrscheinlichkeiten“ sehr natürlich ergibt, ist die „wahrscheinlichkeitstheoretische Zahlentheorie“. Ist $T_k = \{k, 2k, 3k, \dots\}$ die Menge der durch k teilbaren natürlichen Zahlen, so existiert offenbar der Limes $Q(T_k)$ der relativen Häufigkeiten $h_n(T_k)$, und er ist $= 1/k$. Da jede k -te Zahl durch k teilbar ist, ist in gewissem Sinn die „Wahrscheinlichkeit“ der Eigenschaft, durch k teilbar zu sein, gerade $1/k$. Sind p und q verschiedene Primzahlen, so ist $T_{pq} = T_p \cap T_q$, und also $Q(T_p \cap T_q) = Q(T_p)Q(T_q)$. Die Mengen T_p sind also „unabhängig“ unter Q . Dies mag zunächst nur wie eine oberflächliche Analogie aussehen, aber es hat zu vielen tiefliegenden interessanten Aussagen u.a. über die Zahl der Primteiler von natürlichen Zahlen geführt (siehe Kac (1959), Elliott (1979/80)).

$Q(F)$ lässt sich nicht für alle $F \subset \mathbb{N}$ wie oben definieren, sondern nur für solche F , die sich in endlich vielen Schritten durch Bildung von Komplementen, Durchschnitten und Vereinigungen aus Mengen der Form T_k konstruieren lassen. Q lässt sich jedoch mit Hilfe des Satzes von Hahn-Banach zu einer auf der Menge aller Teilmengen $F \subset \mathbb{N}$ definierten Funktion Q derart fortsetzen, dass $0 \leq Q(F) \leq 1$ für alle F

gilt, Q additiv ist, und $Q(F) = \lim h_n(F)$ für die F gilt, für die der Limes existiert. Dieses Q ist nicht σ -additiv.

Aufgaben

- Die Seiten von zwei Würfeln sind mit den folgenden Zahlen beschriftet:

Würfel 1: 6, 3, 3, 3, 3, 3 Würfel 2: 5, 5, 5, 2, 2, 2.

Man gebe eine Beschriftung für einen dritten Würfel so an, dass das folgende Spiel für den zweiten Spieler vorteilhaft ist: Spieler 1 darf einen der drei Würfel wählen; dann darf Spieler 2 einen der verbleibenden Würfel wählen. Jeder würfelt mit dem von ihm gewählten Würfel, und wer die höhere Augenzahl hat, hat gewonnen.

- Wie groß ist die Wahrscheinlichkeit
 - beim sechsmaligen Würfeln eines Würfels mindestens eine Sechs zu erzielen, bzw.
 - beim zwölfmaligen Würfeln eines Würfels mindestens zwei Sechsen zu erzielen?
- Seien die Wahrscheinlichkeiten für die Kinderzahlen $0, 1, \dots, 5$ einer Familie durch $0,3; 0,2; 0,2; 0,15; 0,1$ und $0,05$ gegeben. (Die Wahrscheinlichkeiten noch höherer Kinderzahlen werden hier vernachlässigt.) Wie groß ist, wenn man Jungengeburten und Mädchengeburten als gleich wahrscheinlich annimmt, die Wahrscheinlichkeit, dass ein zufällig ausgewählter Junge mindestens eine Schwester hat? (L)
- Ein Würfel wird dreimal geworfen. Wie groß ist die Wahrscheinlichkeit, dass mindestens eine Drei gewürfelt wird, gegeben mindestens einer der Würfe ist eine Sechs? (L)
- Die k -te von $n \geq 2$ Urnen enthält k schwarze und $n - k$ weiße Kugeln. Eine der Urnen wird zufällig gewählt und eine Kugel daraus gezogen. Wie groß ist die bedingte Wahrscheinlichkeit, dass nach diesem Ziehen die gewählte Urne noch mindestens so viel schwarze Kugeln enthält wie weiße, gegeben die gezogene Kugel ist schwarz.
- Peter und Paul werfen wiederholt eine Münze. Wenn Kopf kommt, erhält Peter einen Punkt, wenn Zahl kommt Paul. Wer zuerst N Punkte hat gewinnt. Wie groß ist Peters Gewinnwahrscheinlichkeit zu einem Zeitpunkt, an dem er $m < N$ Punkte hat und Paul $n < N$ Punkte.
- Für r Personen, die im Erdgeschoss eines Hauses in einen Aufzug einsteigen (und voneinander unabhängig sind), ist jeweils die Wahrscheinlichkeit $1/n$, dass sie in Stockwerk $k = 1, \dots, n$ aussteigen. Wie groß ist die Wahrscheinlichkeit, dass mindestens 2 Passagiere im gleichen Stockwerk aussteigen?
- Ein Würfel ist nicht ganz gleichmäßig. Es sei $P(k) = 1/6 + \varepsilon_k$ mit $|\varepsilon_k| < 1/12$. Man zeige, dass die maximale Abweichung $\max |\varepsilon_k|$ von der Gleichverteilung mindestens halbiert wird, wenn man *ein* Wurf Ergebnis dadurch ermittelt, dass man zweimal würfelt und die zur Augensumme mod 6 kongruente Zahl aus $\{1, \dots, 6\}$ als Ergebnis angibt.
- Sei $n \geq 2$, $\Omega = \{0, 1\}^n$, und für alle $\omega = (\omega_1, \dots, \omega_n) \in \Omega$ sei $P(\omega) = 2^{-n}$. Man betrachte die Ereignisse

$$A_j = \{\omega \in \Omega : \omega_j = 1\} \quad (j = 1, \dots, n) \quad \text{und} \quad B = \{\omega \in \Omega : \omega_1 + \dots + \omega_n = 1 \bmod 2\}.$$

Welche der folgenden drei Familien sind unabhängig:

$$\text{a) } \{A_1, \dots, A_n, B\}; \quad \text{b) } \{A_1, \dots, A_n\}; \quad \text{c) } \{A_2, \dots, A_n, B\}?$$

- Auf einer Prüfstation werden Produkte getestet. Man weiß, dass 2% aller erzeugten Produkte einen Fehler haben. Beim Prüfen wird bei 95% der defekten Teile der Fehler festgestellt, aber auch 1% der fehlerfreien Produkte wird aussortiert. Mit welcher Wahrscheinlichkeit ist ein nicht aussortiertes Produkt wirklich fehlerfrei?
- Ein die Blutgruppen des Menschen bestimmendes Gen kann 3 Zustände annehmen: $A, B, 0$. Damit gibt es 6 mögliche Genotypen $AA, BB, 00, A0, B0, AB$, die in der Elterngeneration die Wahrscheinlichkeiten $p_1, p_2, \dots, p_6$ haben mögen. Berechne die Verteilung der Genotypen unter den Nachkommen bei zufälliger Paarung und zeige, dass das Hardy-Weinberg-Gesetz auch hier gilt.

12. Auf einer Ausstellung sind von 12 Gemälden 10 Originale. Ein Besucher wählt zufällig ein Bild aus, befragt aber, bevor er es kauft, einen Experten nach dessen Meinung. Dieser gibt im Mittel bei 9 von 10 Werken eine richtige Beurteilung ab, unabhängig davon, ob das vorgelegte Bild ein Original oder eine Fälschung ist.
Wenn der Experte entscheidet, dass das Bild eine Fälschung sei, gibt der Besucher das Bild zurück und wählt ein anderes. Mit welcher Wahrscheinlichkeit ist dieses dann ein Original? (L)
13. Zwei Spieler A, B würfeln in der Reihenfolge $A B B A B A B A \dots$. Wer die erste Sechs würfelt gewinnt. Mit welcher Wahrscheinlichkeit w_A gewinnt Spieler A ? (L)

§ 3 Zufallsvariable, Erwartungswert, Varianz

3.1 Verteilungen von Zufallsvariablen

In vielen Zufallsexperimenten interessiert nicht so sehr das Ergebnis ω , sondern nur eine bestimmte Größe $X(\omega)$, die durch ω bestimmt ist. Bei der zufälligen Auswahl einer Person könnte z.B. ω der Name oder die Passnummer der ausgewählten Person sein und $X(\omega)$ ihr Einkommen. Andere Beispiele wären die Augensumme beim zweifachen Würfeln oder die Anzahl der aus einer Urne gezogenen weißen Kugeln. Gelegentlich interessieren auch Kennzeichen qualitativer Art wie Religion, Augenfarbe usw.

Definition 3.1 Ist (Ω, P) ein diskreter Wahrscheinlichkeitsraum und $\mathcal{X}$ eine beliebige Menge, so nennen wir eine Abbildung $X : \Omega \rightarrow \mathcal{X}$ eine $\mathcal{X}$ -wertige **Zufallsvariable**.

Eine Zufallsvariable ist also mathematisch nichts anderes als eine Funktion. Nur weil ω das Ergebnis eines Zufallsexperiments ist, ist auch der Wert $X(\omega)$ zufällig. Denkt man an die übliche Bedeutung des Terms „Variable“ in der Mathematik, so erscheint die Bezeichnung Zufallsvariable unpassend, aber sie ist weithin üblich, und sie entspricht auch den Bezeichnungen in anderen Sprachen (random variable, variable aléatoire). Man hat bei der Wahl des Namens wohl eher an die variierenden Werte bei mehrfachen Wiederholungen des Experiments gedacht. Immerhin sei erwähnt, dass Zufallsvariable auch **Zufallsgrößen** genannt werden.

Gewöhnlich wählt man große Buchstaben wie X, Y, N, S zur Bezeichnung von Zufallsvariablen, und kleine für ihre möglichen oder konkret beobachteten Werte, aber es ist nützlich, Ausnahmen zuzulassen.

Am wichtigsten sind reellwertige Zufallsvariable, so dass diese eigentlich immer gemeint sind, wenn nichts über den Wertebereich gesagt wird. $\mathbb{R}^d$ -wertige Zufallsvariable nennt man oft **Zufallsvektoren**.

An Zufallsvariablen interessiert vor allem ihre **Verteilung**. Die Verteilung von X gibt an, wie wahrscheinlich die einzelnen Werte von X sind. Da Ω abzählbar ist, ist der Wertebereich $\mathcal{X}_X := \{X(\omega) : \omega \in \Omega\}$ von X abzählbar. Unter der Verteilung von X versteht man das Wahrscheinlichkeitsmaß P_X auf $\mathcal{X}_X$ mit

$$P_X(x) = P(\{\omega \in \Omega : X(\omega) = x\}) \quad (x \in \mathcal{X}_X).$$

Es führt auch nicht zu Schwierigkeiten, wenn wir P_X als Wahrscheinlichkeitsmaß auf der evtl. überabzählbaren Menge $\mathcal{X}$ auffassen, also z.B. auf $\mathcal{X} = \mathbb{R}$. Man kann dann

$$P_X(A) = \sum_{x \in A} P_X(x) \quad (A \subset \mathcal{X})$$

setzen. Diejenigen x , die nicht als Werte von X vorkommen, haben die Wahrscheinlichkeit $P_X(x) = 0$, so dass in Wirklichkeit nur über abzählbar viele x summiert wird. Wir brauchen so bei der Betrachtung verschiedener Zufallsvariablen nicht immer deren Wertebereiche zu unterscheiden.

Stellt man sich P als eine Massenverteilung auf Ω vor, bei der jeder Punkt ω die Masse $P(\omega)$ trägt, so ist P_X die Massenverteilung auf $\mathcal{X}$, die man dadurch erhält, dass für jedes ω die auf ω liegende Masse nach $X(\omega)$ transportiert wird. Im Allgemeinen können mehrere ω auf das gleiche x abgebildet werden, so dass deren Masse sich dann dort aufsummiert.

Wir kennen bereits eine Reihe von Verteilungen, wenn wir auch in den Beispielen in Abschnitt 2.4 die Zufallsvariablen nicht immer formal definiert hatten. Ist etwa P auf $\{0, 1\}^n$ das Wahrscheinlichkeitsmaß zu n Bernoulli-Experimenten mit Erfolgswahrscheinlichkeit p , so ist die Zahl der Erfolge die Zufallsvariable $S(\omega) = \omega_1 + \omega_2 + \dots + \omega_n$. Die Wahrscheinlichkeit für k Erfolge war $\binom{n}{k} p^k (1-p)^{n-k}$. Dies ist also in der neuen Bezeichnungsweise $P_S(k)$. P_S ist die Binomialverteilung $b_{n,p}$.

Ein anschauliches Bild der Verteilung einer reellwertigen Zufallsvariablen X lässt sich gewinnen, wenn man das zugehörige **Stabdiagramm** zeichnet. Dazu legt man zunächst eine Liste der möglichen Werte x_k von X an, rechnet für jedes x_k die Wahrscheinlichkeit $P_X(x_k)$ aus, und zeichnet dann senkrecht über den Punkten x_k der x -Achse Striche der Länge $P_X(x_k)$. Ist z.B. $n = 3$ und $p = 0,6$, so erhalten wir für die Zahl S der Erfolge die Tabelle 3.1 und das daneben abgebildete Stabdiagramm.

$x_k = k$	$P_S(k)$
0	0,064
1	0,288
2	0,432
3	0,216

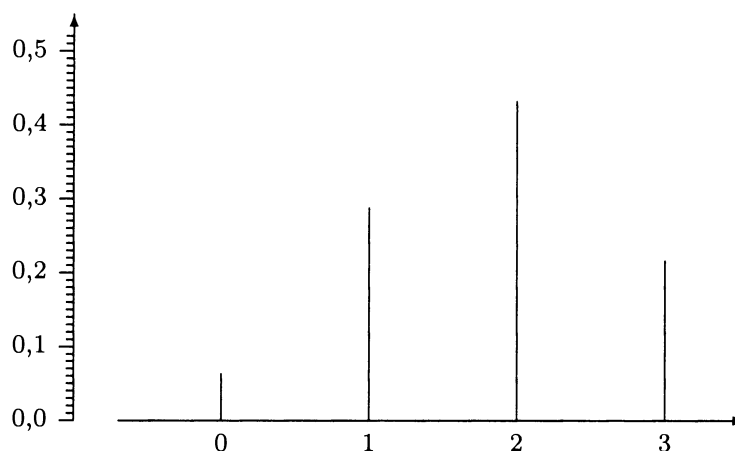


Tabelle 3.1

Bild 3.1 Stabdiagramm

Zur Notation: Neben der obigen kurzen Notation P_X für die Verteilung von X ist auch die ausführlichere Schreibweise $P \circ X^{-1}$ üblich und nützlich. Es ist nämlich $P_X(A) = P(X^{-1}(A))$, wenn X^{-1} die Mengenabbildung bezeichnet, die jedem $A \subset \mathcal{X}$ die Menge aller Urbilder von Punkten aus A zuordnet.

Obwohl schon $\{X = 3\}$ eine Abkürzung für $\{\omega \in \Omega : X(\omega) = 3\}$ war, werden wir jetzt öfter auch noch die geschweiften Klammern in Ausdrücken wie $P(\{X = 3\})$ weglassen und $P(X = 3)$ schreiben. $P(X_1 \in A_1, X_2 \in A_2)$ ist also Abkürzung für $P(\{\omega \in \Omega : X_1(\omega) \in A_1, X_2(\omega) \in A_2\})$.

Gemeinsame Verteilungen mehrerer Zufallsvariablen: Sind auf einem Wahrscheinlichkeitsraum (Ω, P) mehrere Zufallsvariable $X_1, X_2, \dots, X_n$ mit evtl. verschiedenen Wertebereichen $\mathcal{X}_1, \dots, \mathcal{X}_n$ definiert, so kann man sie zu einer Zufallsvariablen X mit Wertebereich $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \dots \times \mathcal{X}_n$ zusammenfassen, indem man $X(\omega) = (X_1(\omega), X_2(\omega), \dots, X_n(\omega))$ setzt. Die Verteilung von X in $\mathcal{X}$ nennt man dann die **gemeinsame Verteilung** von $X_1, \dots, X_n$. Sie ist also durch Angabe aller $P(X_1 = x_1, \dots, X_n = x_n)$ mit $(x_1, \dots, x_n) \in \mathcal{X}$ bestimmt.

Beispiel 3.2 Sei wieder P die Bernoulli-Verteilung zu $0 < p < 1$ auf $\Omega = \{0, 1\}^n$, und sei $S(\omega) = \omega_1 + \dots + \omega_n$ die Zahl der Erfolge. Für ω mit $S(\omega) \geq 1$ sei $N(\omega) = \inf\{j \geq 1 : \omega_j = 1\}$ die Wartezeit bis zum ersten Erfolg, und für ω mit $S(\omega) = 0$ sei $N(\omega) = n+1$. Die gemeinsame Verteilung von S und N ist also beschrieben durch Angabe aller Wahrscheinlichkeiten $p(k, h) := P(S = k, N = h)$ mit $0 \leq k \leq n, 1 \leq h \leq n+1$.

Offenbar ist $p(0, n+1) = (1-p)^n$ und $p(0, h) = 0$ für $1 \leq h \leq n$. Ist $S(\omega) = k \geq 1$ und $N(\omega) = h$, so muss $\omega_i = 0$ für $i < h$ und $\omega_h = 1$ gelten, und es müssen genau $k-1$ Einsen unter den $\omega_{h+1}, \dots, \omega_n$ vorkommen. Es gibt $\binom{n-h}{k-1}$ solche Elemente ω , und für jedes davon gilt $P(\omega) = p^k (1-p)^{n-k}$. Also ist für $k \geq 1$

$$p(k, h) = \binom{n-h}{k-1} p^k (1-p)^{n-k}. \quad (3.1)$$

Ist $k-1 > n-h$, so gibt es keine solchen ω , und es ist $\binom{n-h}{k-1} = 0$; die Formel (3.1) bleibt also richtig. Man beachte, dass $P(N = h | S = k)$ nicht von p abhängt!

Um Wahrscheinlichkeiten von Ereignissen zu berechnen, die nur von einer Zufallsvariablen X_1 abhängen, d.h., die die Form $\{X_1 \in A\}$ haben, genügt die Kenntnis der Verteilung von X_1 . Will man aber Wahrscheinlichkeiten von Ereignissen bestimmen, die von X_1 und X_2 abhängen, so genügt es nicht, die Verteilungen von X_1 und von X_2 zu kennen, sondern man braucht die gemeinsame Verteilung.

Ist $\{i_1, i_2, \dots, i_k\}$ eine k -elementige Teilmenge von $\{1, \dots, n\}$, so nennt man die Verteilung von $(X_{i_1}, X_{i_2}, \dots, X_{i_k})$ eine k -dimensionale **Randverteilung** oder **Marginalverteilung** von $(X_1, \dots, X_n)$. Es ist klar, dass sie aus der gemeinsamen Verteilung berechenbar ist. Z.B. ist die Verteilung von $(X_1, \dots, X_k)$ durch die Zahlen $P(X_1 = x_1, \dots, X_k = x_k)$ bestimmt, und diese kann man aus den Zahlen $P(X_1 = x_1, \dots, X_n = x_n)$ berechnen, indem man über alle $x_{k+1} \in \mathcal{X}_{k+1}, \dots, x_n \in \mathcal{X}_n$ summiert. Speziell sind die Verteilungen der X_i die 1-dimensionalen Marginalverteilungen von $(X_1, \dots, X_n)$.

Die gemeinsame Verteilung von zwei Zufallsvariablen lässt sich als Matrix auffassen. Die beiden 1-dimensionalen Marginalverteilungen ergeben sich dann als Zeilensummen und Spaltensummen. Tabelle 3.2 enthält die gemeinsame Verteilung der Zufallsvariablen S und N aus Beispiel 3.2 für $n = 3$, $p = 1/2$.

		N			
		1	2	3	4
S	0	0	0	0	1/8
	1	1/8	1/8	1/8	0
	2	2/8	1/8	0	0
	3	1/8	0	0	0
		1/2	1/4	1/8	1/8

Tabelle 3.2 Gemeinsame Verteilung von S und N für $n = 3$, $p = 1/2$

Als Zeilensumme zum Wert 1 von S liest man z.B. $P(S = 1) = 3/8$ ab, als Spaltensumme zum Wert 1 von N erhält man $P(N = 1) = 1/2$. Dies erklärt wohl die Bezeichnung „Randverteilung“. In diesem Beispiel ist es natürlich einfacher, die Verteilungen von S und N direkt zu berechnen, aber das muss nicht immer so sein. Z.B. haben wir in Abschnitt 1.7 die gemeinsame Verteilung der Zahl der D -Runs und der Zahl der H -Runs abgeleitet und könnten daraus nun ohne Mühe die uns noch unbekannte Verteilung der Zahl der D -Runs berechnen.

3.2 Unabhängigkeit

Wir kommen nun zu einem ganz zentralen Begriff.

Definition 3.3 Sei (Ω, P) ein diskreter Wahrscheinlichkeitsraum. Eine Familie $\{X_i, i \in I\}$ von Zufallsvariablen $X_i : \Omega \rightarrow \mathcal{X}_i$ heißt **unabhängig**, wenn für jede Wahl von $A_i \subset \mathcal{X}_i$ die Ereignisse $\{X_i \in A_i\}$, $i \in I$, unabhängig sind.

Es sei nun $\{X_i, i \in I\}$ solch eine unabhängige Familie. Indem wir $\mathcal{X}_i$ durch $X_i(\Omega)$ ersetzen, können wir annehmen, die $\mathcal{X}_i$ seien abzählbar. Wir betrachten o.B.d.A. nur den Fall $I = \{1, \dots, n\}$. Dann gilt also speziell für alle $(x_1, \dots, x_n) \in \mathcal{X} = \mathcal{X}_1 \times \dots \times \mathcal{X}_n$:

$$P(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^n P(X_i = x_i). \quad (3.2)$$

Gilt diese Identität, so ist offenbar die Verteilung von $X = (X_1, \dots, X_n)$ gerade das Produkt $P_{X_1} \times \dots \times P_{X_n}$ der Verteilungen der X_i . Wie in Abschnitt 2.3 folgt dann für alle $A_i \subset \mathcal{X}_i$

$$P\left(\bigcap_{i=1}^n \{X_i \in A_i\}\right) = \prod_{i=1}^n P(X_i \in A_i) \quad (3.3)$$

und dann die Unabhängigkeit der Ereignisse $\{X_i \in A_i\}$.

Mit anderen Worten: Es gilt

Satz 3.4 Sei (Ω, P) ein diskreter Wahrscheinlichkeitsraum. Dann sind die folgenden Aussagen äquivalent:

- (i) Die Zufallsvariablen $X_1, \dots, X_n$ sind unabhängig.
- (ii) Für beliebige $x_1, \dots, x_n$ gilt (3.2).
- (iii) Für beliebige $A_i \subset \mathcal{X}_i$ gilt (3.3).

Sei $f \circ X$, wie üblich, die Hintereinanderausführung von X und f , d.h. $f \circ X(\omega) = f(X(\omega))$.

Satz 3.5 Sind die Zufallsvariablen $X_1, \dots, X_n$ unabhängig, und sind f_i für $i = 1, \dots, n$ beliebige Funktionen auf $\mathcal{X}_i$ (mit beliebigen Wertebereichen $\mathcal{Y}_i$), so sind die Zufallsvariablen $Y_i = f_i \circ X_i$ ($i = 1, \dots, n$) unabhängig.

Beweis: Für beliebige $y_1, \dots, y_n$ sei $A_i = \{x_i \in \mathcal{X}_i : f_i(x_i) = y_i\}$. Dann ist $\{Y_i = y_i\} = \{X_i \in A_i\}$. Aus (3.3) folgt

$$P(Y_1 = y_1, \dots, Y_n = y_n) = P(Y_1 = y_1) \dots P(Y_n = y_n). \quad \square$$

Ohne Schwierigkeiten lässt sich dieser Satz noch im Sinne von Satz 2.6 verallgemeinern. Sind z.B. $X_1, X_2, \dots, X_5$ unabhängig und definiert man $Y_1(\omega) = f_1(X_1(\omega), X_2(\omega))$, $Y_2(\omega) = f_2(X_3(\omega), X_4(\omega), X_5(\omega))$, so sind Y_1 und Y_2 unabhängig.

Ist (Ω, P) das Produkt der Wahrscheinlichkeitsräume $(\Omega_1, P_1), \dots, (\Omega_n, P_n)$, und hängen Zufallsvariable Y_i auf Ω nur von der i -ten Koordinate ab, so garantiert Satz 3.5 die Unabhängigkeit von $Y_1, \dots, Y_n$. Wir können uns so immer unabhängige Zufallsvariable mit vorgegebenen Verteilungen P_{Y_i} verschaffen.

3.3 Erwartungswerte

Will man einen „mittleren Wert“ für eine reellwertige Zufallsvariable X angeben, so ist es sinnvoll, die Werte $X(\omega)$ mit den Wahrscheinlichkeiten $P(\omega)$ zu gewichten, also $\sum_{\omega \in \Omega} X(\omega) P(\omega)$ als Mittelwert zu wählen. Damit diese Summe wohldefiniert und reellwertig ist, sollte sie unabhängig von Umordnungen in der Summationsreihenfolge sein. Dies ist genau dann der Fall wenn $\sum_{\omega \in \Omega} |X(\omega)| P(\omega)$ konvergiert.

Definition 3.6 Ist (Ω, P) ein diskreter Wahrscheinlichkeitsraum, so sagen wir, dass die Erwartung einer reellwertigen Zufallsvariablen X existiert, wenn $\sum_{\omega \in \Omega} |X(\omega)| P(\omega)$ konvergiert. Wir definieren dann

$$EX = E(X) = \sum_{\omega \in \Omega} X(\omega) P(\omega)$$

als den **Erwartungswert** von X .

Ist $x_1, x_2, \dots$ eine Abzählung des Wertebereichs von X , so ist

$$EX = \sum_{i=1}^{\infty} \sum_{\{\omega: X(\omega)=x_i\}} X(\omega) P(\omega) = \sum_{i=1}^{\infty} x_i P(X = x_i). \quad (3.4)$$

Wir sehen also, dass EX nur von der Verteilung von X abhängt, und dass wir EX ebenso gut durch die Summe auf der rechten Seite von (3.4) hätten definieren können.

Satz 3.7 Sei (Ω, P) diskret, und seien X, Y reellwertige Zufallsvariable, deren Erwartung existiert. Dann gilt:

- (i) Für $\lambda \in \mathbb{R}$ existiert die Erwartung von λX , und es ist $E(\lambda X) = \lambda EX$.
- (ii) Die Erwartung von $X + Y$ existiert, und es ist $E(X + Y) = EX + EY$.
- (iii) Sind X, Y unabhängig, so existiert die Erwartung von $X \cdot Y$ und es ist $E(X \cdot Y) = (EX)(EY)$.

Beweis: (i) und (ii) folgen unmittelbar aus den Definitionen.

(iii): Sind (x_i) und (y_j) die Werte von X und Y , so gilt wegen der Nichtnegativität aller Terme

$$\begin{aligned} \sum_{\omega \in \Omega} |X(\omega) \cdot Y(\omega)| P(\omega) &= \sum_i \sum_j |x_i| |y_j| P(X = x_i, Y = y_j) \\ &= \sum_i \sum_j |x_i| |y_j| P(X = x_i) P(Y = y_j) \\ &= \left(\sum_i |x_i| P(X = x_i) \right) \left(\sum_j |y_j| P(Y = y_j) \right). \end{aligned}$$

Diese Rechnung ist sogar richtig, wenn die Erwartungen von X oder Y nicht existieren. Da sie aber existieren, ist das letzte Produkt endlich, und wegen der absoluten Konvergenz können wir die gleiche Rechnung ohne Absolutstriche wiederholen. $\square$

Stellen wir uns die Verteilung einer reellwertigen Zufallsvariablen X als Massenverteilung auf $\mathbb{R}$ vor, bei der die Werte x_k von X das Gewicht $P(X = x_k)$ tragen, so ist EX gerade der Schwerpunkt der Massenverteilung. Hat X z.B. die Werte 0,1; 2,4 und 6, und sind die zugehörigen Wahrscheinlichkeiten 0,1; 0,1; 0,4; 0,15 und 0,25, so ergibt sich $EX = 3$ und das Stabdiagramm in Bild 3.2.

Wir müssen die x -Achse an der durch den Pfeil markierten Stelle $EX = 3$ unterstützen, wenn Gleichgewicht herrschen soll.

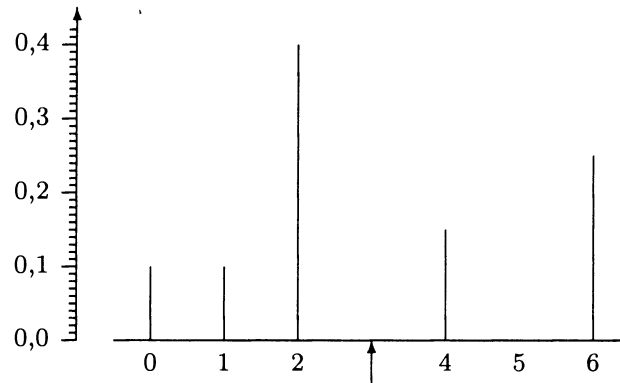


Bild 3.2

Beispiele: (1) Ist X *binomialverteilt*, so ist

$$\begin{aligned} EX &= \sum_{k=0}^n k \binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_{k=0}^n np \frac{(n-1)!}{(k-1)!((n-1)-(k-1))!} p^{k-1} (1-p)^{(n-1)-(k-1)} = np. \end{aligned}$$

Man erhält das gleiche Ergebnis auch leicht mit einem kleinen Trick, der auch in anderen Beispielen nützlich ist. Dazu schreiben wir X als Summe $X_1 + X_2 + \dots + X_n$, wobei $X_i(\omega) = 1$ sein soll, wenn das i -te Teilerperiment ein Erfolg ist, und $= 0$ sonst. Dann ist $EX_i = p$ und $EX = E(X_1 + \dots + X_n) = EX_1 + \dots + EX_n = np$.

(2) Ist X *hypergeometrisch* verteilt mit Parametern n, N, S , also $P(X = s) = \binom{S}{s} \binom{N-S}{n-s} / \binom{N}{n}$, so liefert der gleiche Trick $EX = nS/N$. ($X_i(\omega)$ ist 1, wenn die i -te ohne Rücklegen gezogene Kugel schwarz ist.)

(3) Ist X *geometrisch* verteilt, so ist $P(X = k) = p(1-p)^{k-1}$ ($k = 1, 2, \dots$). Also ist $EX = \sum_{k=1}^{\infty} kp(1-p)^{k-1}$. Um die unendliche Summe auszuwerten, differenzieren wir die für $|x| < 1$ gültige Identität

$$\frac{1}{1-x} = \sum_{k=0}^{\infty} x^k$$

auf beiden Seiten und erhalten so

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

Setzt man darin $x = 1 - p$, so folgt $EX = p/p^2 = 1/p$.

(4) (*Wartezeit beim Ziehen ohne Rücklegen*)* In einer Urne befinden sich wieder S schwarze und $N - S$ weiße Kugeln. T_1 sei die Nummer der Ziehung, bei der zum ersten Mal eine schwarze Kugel erscheint. Wird mit Rücklegen gezogen, so ist T_1 geometrisch verteilt mit $p = S/N$ und also $ET_1 = N/S$. Nun soll aber *ohne* Rücklegen gezogen werden.

Es ist nicht schwer, $P(T_1 = k)$ zu berechnen, aber dieser Weg zur Berechnung von ET_1 wird rasch unübersichtlich. Wir gehen anders vor. Bei $n = N$ Ziehungen werden alle Kugeln gezogen, also S schwarze. Sei $\omega_i = s$, wenn die i -te gezogene Kugel schwarz ist, und $= w$ sonst. Aus Symmetriegründen sind alle Elemente $\omega = (\omega_1, \dots, \omega_N)$ gleichwahrscheinlich.

Sei U_i der Zeitpunkt, an dem die i -te schwarze Kugel erscheint, $T_1 = U_1$, $T_i = U_i - U_{i-1}$ ($i = 2, \dots, S$), $T_{S+1} = (N + 1) - U_S$. Für das Ergebnis

$$\omega = (w, w, s, w, w, w, w, s, s, w)$$

wäre also wie in Bild 3.3

$$T_1(\omega) = 3, \quad T_2(\omega) = 5, \quad T_3(\omega) = 1, \quad T_4(\omega) = 2.$$

(Hier ist $S = 3$, $N = 10$.)

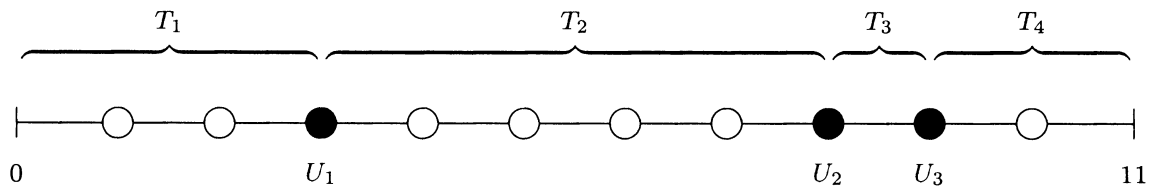


Bild 3.3 Wartezeiten

Wir wollen uns überlegen, dass die Verteilungen der Zufallsvariablen $T_1, \dots, T_{S+1}$ und also auch die zugehörigen Erwartungswerte gleich sind. Es ist klar, dass aus Symmetriegründen T_1 und T_{S+1} die gleiche Verteilung haben. Ist $\tilde{\omega}$ das Element, das aus $\omega = (\omega_1, \dots, \omega_N)$ hervorgeht, indem man die Teilfolgen

$$(\omega_1, \dots, \omega_{T_1(\omega)}) \quad \text{und} \quad (\omega_{T_1(\omega)+1}, \dots, \omega_{U_2(\omega)})$$

vertauscht, so ist $T_1(\omega) = T_2(\tilde{\omega})$ und $T_2(\omega) = T_1(\tilde{\omega})$. Da die Verteilung P austauschbar ist, ist $P(\omega) = P(\tilde{\omega})$. (In unserem Beispiel wäre $\tilde{\omega} = (w, w, w, w, s, w, w, s, s, w)$.) Die Zuordnung von ω und $\tilde{\omega}$ ist umkehrbar eindeutig. Also ist für jedes k

$$P(T_1 = k) = \sum_{\{\omega: T_1(\omega)=k\}} P(\omega) = \sum_{\{\tilde{\omega}: T_2(\tilde{\omega})=k\}} P(\tilde{\omega}) = P(T_2 = k).$$

Analog zeigt man, dass auch die übrigen T_i die gleiche Verteilung haben. Also folgt $ET_1 = ET_2 = \dots = ET_{S+1}$. Nun ist aber $T_1 + \dots + T_{S+1} = N + 1$, und daher $N + 1 = E(T_1 + \dots + T_{S+1}) = (S + 1) \cdot ET_1$. Damit gilt

$$ET_1 = \frac{N + 1}{S + 1}.$$

Im Mittel findet man die erste schwarze Kugel etwas schneller als beim Ziehen mit Rücklegen.

Funktionen von Zufallsvariablen: Ist $X : \Omega \rightarrow \mathcal{X}$ eine Zufallsvariable mit beliebigem Wertebereich $\mathcal{X}$, und $f : \mathcal{X} \rightarrow \mathbb{R}$ eine reellwertige Funktion auf $\mathcal{X}$, so ist $Y = f \circ X$ eine reellwertige Zufallsvariable. Bezeichne E_P die (oben E genannte) Erwartung bzgl. P , und E_{P_X} die bzgl. P_X . Es gilt

Satz 3.8 $E_P Y$ existiert genau dann, wenn $E_{P_X} f$ existiert, und es ist $E_P Y = E_{P_X} f$.

Beweis: Die Existenz von $E_P Y$ besagt $\sum_{\omega} |f(X(\omega))| P(\omega) < \infty$, und die von $E_{P_X} f$ besagt $\sum_x |f(x)| P(\{\omega : X(\omega) = x\}) < \infty$. Das ist offenbar dasselbe, und im Fall der absoluten Konvergenz sind die beiden Summen auch ohne Absolutstriche konvergent und einander gleich. $\square$

3.4 Das Rechnen mit Indikatorfunktionen

Für beliebiges $A \subset \Omega$ definieren wir die **Indikatorfunktion** 1_A von A durch

$$1_A(\omega) = \begin{cases} 1, & \text{falls } \omega \in A, \\ 0, & \text{falls } \omega \notin A. \end{cases}$$

Die Zuordnung von 1_A zu A ist offenbar umkehrbar eindeutig. Der Nutzen dieser einfachen Zuordnung besteht darin, dass man mit Funktionen rechnen kann. Sei $\mathbf{1} = 1_{\Omega}$ die konstante Funktion mit Wert 1 auf Ω . Die folgenden Identitäten und Aussagen sind sehr leicht nachzuprüfen.

$$1_{A^c} = \mathbf{1} - 1_A, \quad (3.5)$$

$$1_{A \cap B} = 1_A 1_B, \quad (3.6)$$

$$1_A^2 = 1_A, \quad (3.7)$$

$$A \subset B \iff 1_A \leq 1_B, \quad (3.8)$$

$$A \cap B = \emptyset \Rightarrow 1_{A \cup B} = 1_A + 1_B. \quad (3.9)$$

All dies gilt für beliebige Mengen. Ist auf Ω ein Wahrscheinlichkeitsmaß gegeben, so gilt außerdem

$$E 1_A = P(A), \quad (3.10)$$

denn es ist $A = \{\omega : 1_A(\omega) = 1\}$ und also $E 1_A = 0 \cdot P(A^c) + 1 \cdot P(A)$. Durch Induktion folgt aus (3.6) leicht

$$1_{A_1 \cap A_2 \cap \dots \cap A_n} = 1_{A_1} 1_{A_2} \dots 1_{A_n} \quad (3.11)$$

und mittels der de Morganschen Regel $(A_1 \cup \dots \cup A_n)^c = A_1^c \cap \dots \cap A_n^c$ ist dann

$$1_{A_1 \cup A_2 \cup \dots \cup A_n} = \mathbf{1} - (\mathbf{1} - 1_{A_1})(\mathbf{1} - 1_{A_2}) \dots (\mathbf{1} - 1_{A_n}). \quad (3.12)$$

Als erste Anwendung ergibt sich ein Beweis von Gl. (1.18), indem man das Produkt auf der rechten Seite von (3.12) ausmultipliziert, und dann links und rechts den Erwartungswert bildet.

Als zweite Anwendung geben wir ein Beispiel aus der Zuverlässigkeitstheorie.

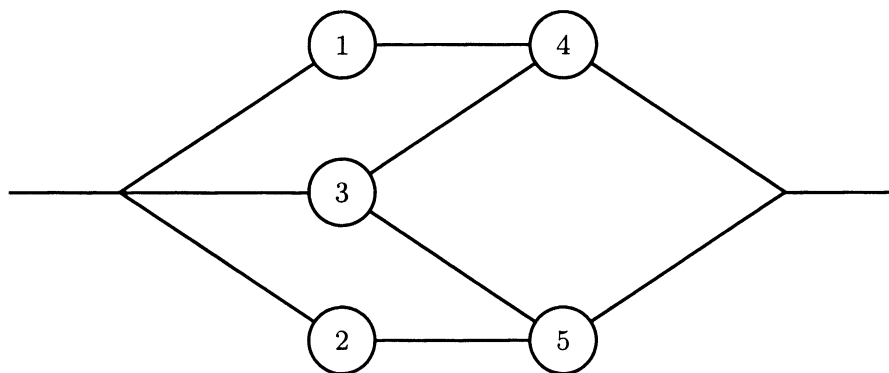


Bild 3.4

Beispiel 3.9* (*Zuverlässigkeit von Kühlsystemen*) In einem Kraftwerk sind zum Betrieb von Kühlanlagen zwei voneinander unabhängige Pumpen installiert. Zum Betrieb jeder der Pumpen dient ein eigener Generator. Falls einer davon ausfällt, kann er noch kurzfristig durch das vorhandene Notstromaggregat ersetzt werden. Wir können diese Funktionsweise durch das Diagramm in Bild 3.4 beschreiben.

An den Stellen 1, 2 sind die Generatoren, an Stelle 3 ist das Notstromaggregat, und an den Stellen 4, 5 sind die Pumpen. An den Stellen 1–5 können unabhängig voneinander Störungen eintreten, und zwar an Stelle i mit Wahrscheinlichkeit q_i . Der Weg von links nach rechts wird dann an Stelle i unterbrochen. Die Wahrscheinlichkeit p , dass das Gesamtsystem funktioniert, ist die Wahrscheinlichkeit, dass es einen ununterbrochenen Weg von links nach rechts gibt. Wie groß ist p ?

Es gibt die Wege W_1 über 1 und 4, W_2 über 3 und 4, W_3 über 3 und 5 und W_4 über 2 und 5. Sei A_i das Ereignis, dass Weg W_i offen ist, A die Vereinigung der A_i , und B_j das Ereignis, dass es an Stelle j keine Störung gibt. Dann ist $p = P(A)$, $A_1 = B_1 \cap B_4$, $A_2 = B_3 \cap B_4$, $A_3 = B_3 \cap B_5$ und $A_4 = B_2 \cap B_5$. Nun braucht man nur noch einzusetzen. Nach den Formeln (3.12) und (3.6) ist

$$1_A = 1 - (1 - 1_{B_1} 1_{B_4})(1 - 1_{B_3} 1_{B_4})(1 - 1_{B_3} 1_{B_5})(1 - 1_{B_2} 1_{B_5}).$$

Multipliziert man aus, und nutzt man $1_{B_i} = 1_{B_i}^2$ aus, so heben sich ein paar Terme weg, und es bleibt

$$\begin{aligned} 1_A = & 1_{B_3} 1_{B_5} + 1_{B_2} 1_{B_5} + 1_{B_1} 1_{B_4} + 1_{B_3} 1_{B_4} - 1_{B_2} 1_{B_3} 1_{B_5} - 1_{B_3} 1_{B_4} 1_{B_5} \\ & - 1_{B_1} 1_{B_3} 1_{B_4} - 1_{B_1} 1_{B_2} 1_{B_4} 1_{B_5} + 1_{B_1} 1_{B_2} 1_{B_3} 1_{B_4} 1_{B_5}. \end{aligned}$$

Mit $p_i = 1 - q_i = P(B_i)$ ergibt sich nach Satz 3.7

$$p = E 1_A = p_3 p_5 + p_2 p_5 + p_1 p_4 + p_3 p_4 - p_2 p_3 p_5 - \dots$$

Unzuverlässigkeiten in der Schätzung der p_i addieren sich in der Formel für p natürlich, und hinzu kommt, dass die Unabhängigkeitsannahmen in konkreten Systemen manchmal fragwürdig sind.

Wir geben noch eine weitere Anwendung:

Die Einschluss-Ausschluss-Formel*: Zu einer Feier bringt jeder Gast ein Geschenk mit. Die Geschenke werden zufällig auf die Anwesenden verteilt. Mit welcher Wahrscheinlichkeit bekommen genau n Gäste ihr Geschenk zurück?

Ist N die Zahl der Gäste, so können wir als Stichprobenraum Ω die Menge der Permutationen $\omega = (\omega_1, \dots, \omega_N)$ der Zahlen $1, \dots, N$ wählen. P ist die Gleichverteilung auf Ω . Der k -te Gast bekommt sein Geschenk zurück, wenn $\omega_k = k$ ist. Sei also $X_k(\omega) = 1$, wenn $\omega_k = k$ ist, und $= 0$ sonst. Sei $A_k = \{\omega : X_k(\omega) = 1\}$. Die Wahrscheinlichkeiten von endlichen Durchschnitten A von Mengen $A_{k_1}, \dots, A_{k_n}$ lassen sich leicht berechnen. Für ein $\omega \in A$ muss $\omega_{k_1} = k_1, \dots, \omega_{k_n} = k_n$ sein. Für die übrigen $N - n$ Koordinaten gibt es $(N - n)!$ Möglichkeiten, sie auf die übrigen $N - n$ Plätze zu verteilen. Also ist $P(A) = (N - n)!/N!$. Wir stehen also vor dem Problem, die Wahrscheinlichkeit des Ereignisses zu bestimmen, dass genau n Ereignisse A_k eintreten, wenn die Wahrscheinlichkeiten der endlichen Durchschnitte der A_k bekannt sind. Dieses Problem tritt relativ häufig auf, und so wollen wir es gleich allgemein lösen.

Sei also jetzt (Ω, P) ein Wahrscheinlichkeitsraum, $A_1, \dots, A_N$ seien Ereignisse, und für beliebige $\{k_1, \dots, k_n\} \subset \{1, \dots, N\}$ sei $P(A_{k_1} \cap \dots \cap A_{k_n})$ bekannt. Sei

$$B_n = \{\omega \in \Omega : \omega \in A_k \text{ für genau } n \text{ Werte von } k\},$$

$$C_n = \{\omega \in \Omega : \omega \in A_k \text{ für mindestens } n \text{ Werte von } k\}.$$

Speziell ist $C_1 = A_1 \cup \dots \cup A_N$. Wie in Abschnitt 1.5 setzen wir $S_0 := 1$ und

$$S_n := \sum_{\{k_1, \dots, k_n\}} P(A_{k_1} \cap \dots \cap A_{k_n}),$$

wobei die Summe über alle n -elementigen Teilmengen $\{k_1, \dots, k_n\} \subset \{1, \dots, N\}$ erstreckt wird.

Satz 3.10 (Einschluss-Ausschluss-Formel) *Es gilt:*

$$(i) \quad P(B_n) = \sum_{k=n}^N (-1)^{k-n} \binom{k}{n} S_k;$$

$$(ii) \quad P(C_n) = \sum_{k=n}^N (-1)^{k-n} \binom{k-1}{n-1} S_k.$$

Beweis: Ist $\{i_1, \dots, i_n\}$ eine Teilmenge von $\{1, \dots, N\}$, so sei $\{i_{n+1}, \dots, i_N\}$ das Komplement. Nun ist

$$\begin{aligned} 1_{B_n} &= \sum_{\{i_1, \dots, i_n\}} 1_{A_{i_1}} \cdot \dots \cdot 1_{A_{i_n}} (1 - 1_{A_{i_{n+1}}}) \cdot \dots \cdot (1 - 1_{A_{i_N}}) \\ &= \sum_{\{i_1, \dots, i_n\}} 1_{A_{i_1}} \cdot \dots \cdot 1_{A_{i_n}} \sum_{k=0}^{N-n} (-1)^k \sum_{\{j_1, \dots, j_k\}} 1_{A_{j_1}} \cdot \dots \cdot 1_{A_{j_k}}. \end{aligned}$$

Es wird über alle $\{j_1, \dots, j_k\} \subset \{i_{n+1}, \dots, i_N\}$ summiert; dabei ist für $k = 0$ $\{j_1, \dots, j_k\}$ die leere Menge und $1_{A_{j_1}} \cdot \dots \cdot 1_{A_{j_k}} = 1$. Es gibt $\binom{n+k}{n}$ Möglichkeiten, eine Menge $\{h_1, \dots, h_{n+k}\} \subset \{1, \dots, N\}$ in disjunkte Teilmengen $\{i_1, \dots, i_n\}$, $\{j_1, \dots, j_k\}$ zu zerlegen. Also ist

$$1_{B_n} = \sum_{k=0}^{N-n} (-1)^k \sum_{\{h_1, \dots, h_{n+k}\}} \binom{n+k}{n} 1_{A_{h_1}} \cdot \dots \cdot 1_{A_{h_{n+k}}}.$$

Aus $P(B_n) = E(1_{B_n})$ und der Additivität des Erwartungswerts folgt (i).

(ii) C_n ist die disjunkte Vereinigung der Mengen $B_n, B_{n+1}, \dots, B_N$. Für $n = N$ ist $C_N = B_N$, und beide Formeln in Satz 3.10 ergeben wegen $\binom{N}{N} = \binom{N-1}{N-1} = 1$ den Wert S_N . Also gilt (ii) für $n = N$. Ist (ii) für ein $n > 1$ bereits bewiesen, so folgt

$$\begin{aligned} P(C_{n-1}) &= P(B_{n-1}) + P(C_n) \\ &= \sum_{k=n-1}^N (-1)^{k-n+1} \binom{k}{n-1} S_k + \sum_{k=n}^N (-1)^{k-n} \binom{k-1}{n-1} S_k \\ &= S_{n-1} + \sum_{k=n}^N (-1)^{k-(n-1)} S_k \left[\binom{k}{n-1} - \binom{k-1}{n-1} \right]. \end{aligned}$$

Aus $\binom{k-1}{n-1} + \binom{k-1}{n-2} = \binom{k}{n-1}$ folgt (ii) für $n-1$. $\square$

Wir kehren nun zu dem Problem mit den Gästen zurück. $Y(\omega) = \sum_{k=1}^N X_k(\omega)$ ist die Zahl der Gäste, die ihr Geschenk zurückerhalten. (Allgemeiner spricht man von der Zahl der **Koinzidenzen**, engl. „match“.) Die Verteilung von Y heißt daher auch **Koinzidenzverteilung**. Sie ist nun leicht ableitbar:

Satz 3.11 Sei $Y(\omega)$ in einer zufälligen Permutation $\omega = (\omega_1, \dots, \omega_N)$ von $\{1, \dots, N\}$ die Zahl der k mit $\omega_k = k$. Dann ist

$$P(Y = n) = \frac{1}{n!} \sum_{i=0}^{N-n} \frac{(-1)^i}{i!}.$$

Beweis: Es ist $P(A_{k_1} \cap \dots \cap A_{k_n}) = (N-n)!/N!$, also $S_n = \binom{N}{n} (N-n)!/N! = 1/n!$. Daraus folgt

$$P(Y = n) = P(B_n) = \sum_{k=n}^N \frac{(-1)^{k-n}}{k!} \binom{k}{n} = \sum_{k=n}^N \frac{(-1)^{k-n}}{n! (k-n)!}. \quad \square$$

Die Wahrscheinlichkeit, dass mindestens ein Gast sein Geschenk zurückerhält, ist

$$P(Y \geq 1) = P(C_1) = \sum_{k=1}^N (-1)^{k-1} \binom{k-1}{0} S_k = \sum_{k=1}^N \frac{(-1)^{k-1}}{k!}.$$

Für großes N ist offenbar $P(Y \geq 1) \approx 1 - e^{-1} \approx 0,632$ und $P(Y = n) \approx e^{-1}/n!$.

3.5 Varianz und Kovarianz

Während der Erwartungswert eine Maßzahl für den Schwerpunkt einer Verteilung ist, ist die Varianz eine Maßzahl für die Streuung um diesen Schwerpunkt.

Definition 3.12 Ist X eine reellwertige Zufallsvariable auf dem diskreten Wahrscheinlichkeitsraum (Ω, P) , und existiert $E(X^2)$, so heißt $\text{Var}(X) = E((X - EX)^2)$ die **Varianz** von X , und $\sigma_X = \sqrt{\text{Var}(X)}$ die **Streuung** oder **Standardabweichung** von X . Ist Y eine weitere reellwertige Zufallsvariable, für die $E(Y^2)$ existiert, so heißt

$$\text{Cov}(X, Y) = E((X - EX)(Y - EY))$$

die **Kovarianz** und $\rho_{XY} = \text{Cov}(X, Y)/(\sigma_X \sigma_Y)$ der **Korrelationskoeffizient** von X und Y . Die Zufallsvariablen X und Y heißen **unkorreliert**, wenn $\text{Cov}(X, Y) = 0$ ist.

Bemerkungen: (1) Wegen $|X| \leq 1 + X^2$ existiert EX , und wegen $(X - EX)^2 \leq X^2 + 2|EX| \cdot |X| + (EX)^2$ auch $E((X - EX)^2)$. Die Existenz von $E((X - EX)(Y - EY))$ folgt aus $|X \cdot Y| \leq X^2 + Y^2$.

(2) Wir werden gleich sehen, dass $\text{Var}(\lambda X) = \lambda^2 \text{Var}(X)$ gilt. Von einem Streuungsmaß würde man sich aber wünschen, dass es sich proportional zu X ändert. Dies ist für σ_X der Fall. Warum benutzt man aber nicht $E(|X - EX|)$? Im Prinzip wäre dies möglich, aber σ_X hat viele Vorteile, nicht nur, weil es sich meist leichter berechnen lässt. σ_X tritt auch auf natürliche Weise in vielen theoretischen Resultaten auf. Ein anderer Vorteil mag sein, dass σ_X gerade die störenden großen Abweichungen stärker bewertet.

Ist X Zufallsvariable mit $EX = 0$, und haben die Werte $x_1, \dots, x_n$ die Wahrscheinlichkeit $P(X = x_i) = 1/n$, so ist $\sigma_X = \sqrt{\sum x_i^2/n}$ Vielfaches des euklidischen Abstandes des Punktes $(x_1, \dots, x_n)$ vom Nullpunkt. Die Verallgemeinerung dieser Idee ermöglicht es später, mit Zufallsvariablen geometrisch zu argumentieren.

(3) Für komplexwertige X, Y definiert man Erwartungswerte wie oben, und setzt $\text{Var}(X) = E(|X - EX|^2)$ und $\text{Cov}(X, Y) = E((X - EX)(\overline{Y - EY}))$. Dabei ist $\bar{c}$ die zu c komplex konjugierte Zahl. Offenbar ist $\text{Var}(X) = \text{Cov}(X, X)$.

(4) Für jede positive ganze Zahl k nennt man $E(X^k)$ das **k -te Moment** von X . Wenn dieser Erwartungswert existiert, sagt man, dass das k -te Moment existiere. Existiert EX , so sagt man, dass $X - EX$ aus X durch **Zentrieren** hervorgehe. Da die Varianz das zweite Moment von $X - EX$ ist, nennt man sie auch das **zweite zentrale Moment** von X .

(5) Wir hatten EX als Schwerpunkt der Verteilung von X gedeutet. Analog lässt sich $\text{Var}(X)$ als Trägheitsmoment bei Drehung um den Schwerpunkt deuten.

Satz 3.13 (Rechenregeln) Sind X, Y, X_i komplexwertige Zufallsvariable, für die die zweiten Momente existieren, und sind a, b, c, d reelle oder komplexe Zahlen, so gilt:

- (i) $\text{Var}(X) = E(|X|^2) - |EX|^2$.
- (ii) $\text{Var}(aX + b) = |a|^2 \text{Var}(X)$.
- (iii) $\text{Cov}(X, Y) = E(X\bar{Y}) - EX\bar{EY}$.
- (iv) $\text{Cov}(aX + b, cY + d) = a\bar{c} \text{Cov}(X, Y)$.
- (v) $\text{Cov}(X, Y) = \overline{\text{Cov}(Y, X)}$.
- (vi) $\text{Var}(X_1 + \dots + X_n) = \sum_{i=1}^n \text{Var}(X_i) + \sum_{i \neq j} \text{Cov}(X_i, X_j)$.
- (vii) Sind X, Y unabhängig, so sind sie auch unkorreliert.

(viii) (Bienaymé). Für unabhängige $X_1, \dots, X_n$ gilt $\text{Var}(X_1 + \dots + X_n) = \sum_{i=1}^n \text{Var}(X_i)$.

Beweis: Beginnen wir mit (iii). Benutzt man $E(\bar{Y}) = \overline{EY}$, so erhält man

$$\begin{aligned} E((X - EX)(\bar{Y} - \overline{EY})) &= E(X\bar{Y}) - E(X\overline{EY}) - E((EX)\bar{Y}) + E((EX)(\overline{EY})) \\ &= E(X\bar{Y}) - (EX)(\overline{EY}) - (EX)(\overline{EY}) + (EX)(\overline{EY}) \\ &= E(X\bar{Y}) - (EX)(\overline{EY}). \end{aligned}$$

Die Aussage (i) ist der Spezialfall $X = Y$, denn es gilt $|Z|^2 = Z\bar{Z}$. (iv) ergibt sich durch eine einfache direkte Rechnung, und (ii) ist wieder Spezialfall. Auch (v) ist klar.

(vi) Wegen (iv) können wir $EX_i = 0$ annehmen. Dann ist

$$\begin{aligned} \text{Var}(X_1 + \dots + X_n) &= E((X_1 + \dots + X_n)(\overline{X_1 + \dots + X_n})) \\ &= \sum_{i=1}^n E(X_i \bar{X}_i) + \sum_{i \neq j} E(X_i \bar{X}_j). \end{aligned}$$

Da $EX_i = 0$ ist, ist $E(X_i \bar{X}_i) = \text{Var}(X_i)$ und $E(X_i \bar{X}_j) = \text{Cov}(X_i, X_j)$.

(vii) Nach Satz 3.5 sind $X - EX$ und $\bar{Y} - \overline{EY}$ unabhängig, und daher ist nach Satz 3.7

$$\text{Cov}(X, Y) = E((X - EX)(Y - EY)) = E(X - EX) \overline{E(Y - EY)} = 0 \cdot 0.$$

Nun folgt (viii) aus (vi). □

Satz 3.14 (Cauchy-schwarzsche Ungleichung) *Existieren $E(|X|^2)$ und $E(|Y|^2)$, so gilt*

$$|E(X\bar{Y})|^2 \leq E(|X|^2) E(|Y|^2). \quad (3.13)$$

In dieser Ungleichung gilt Gleichheit genau dann, wenn X und Y auf der Menge Ω' der ω mit $P(\omega) > 0$ linear abhängig sind, d.h. Zahlen a, b existieren, die nicht beide $= 0$ sind, so dass $P(aX + bY \neq 0) = 0$ ist.

Beweis: Sei $\alpha = E(|Y|^2)$ ($= E(Y\bar{Y})$), und $\beta = -E(X\bar{Y})$. Wir können $\alpha \neq 0$ und also $\alpha > 0$ annehmen, denn sonst ist auf Ω' $Y(\omega) = 0$, und dann ist auch $E(X\bar{Y}) = 0$. Es ist $\alpha = \bar{\alpha}$. Nach den Rechenregeln erhalten wir

$$\begin{aligned} 0 &\leq E(|\alpha X + \beta Y|^2) = E((\alpha X + \beta Y)(\overline{\alpha X + \beta Y})) \\ &= \alpha \bar{\alpha} E(X\bar{X}) + \alpha \bar{\beta} E(X\bar{Y}) + \bar{\alpha} \beta E(\bar{X}Y) + \bar{\beta} \beta E(Y\bar{Y}) \\ &= \alpha (E(X\bar{X}) E(Y\bar{Y}) - E(\bar{X}Y) E(X\bar{Y}) - E(\bar{X}Y) E(X\bar{Y}) + E(X\bar{Y}) E(\bar{X}Y)) \\ &= \alpha (E(|X|^2) E(|Y|^2) - |E(X\bar{Y})|^2). \end{aligned}$$

Wegen $\alpha > 0$ gilt (3.13).

Gilt Gleichheit in (3.13), so ergibt die obige Rechnung $E(|\alpha X + \beta Y|^2) = 0$, d.h. auf Ω' , muss $\alpha X + \beta Y = 0$ gelten. Ist $\alpha > 0$, so können wir $a = \alpha$, $b = \beta$ wählen. Ist $\alpha = 0$, so können wir $a = 0$, $b = 1$ wählen. Sind umgekehrt X und Y auf Ω' linear abhängig, so ist auf Ω' $aX = -bY$, und daraus folgt leicht Gleichheit in (3.13). □

Bemerkungen: (1) Wendet man (3.13) auf die Zufallsvariablen $|X|$ und $|Y|$ an, so folgt sogar die formal stärkere Ungleichung

$$E(|XY|)^2 \leq E(|X|^2) E(|Y|^2). \quad (3.14)$$

(2) Durch Anwendung von (3.13) auf die zentrierten Zufallsvariablen $X - EX$ und $Y - EY$ ergibt sich

$$|\text{Cov}(X, Y)| \leq \sigma_X \sigma_Y. \quad (3.15)$$

Interpretation: Für reellwertige X, Y bedeutet positive Kovarianz, dass eine Tendenz besteht, nach der $X(\omega)$ für diejenigen ω die größeren Werte annimmt, für die auch $Y(\omega)$ die größeren Werte annimmt. Dann wird nämlich häufig $X(\omega) - EX$ das gleiche Vorzeichen haben wie $Y(\omega) - EY$ und damit

$$\text{Cov}(X, Y) = \sum_{\omega} (X(\omega) - EX) (Y(\omega) - EY) P(\omega)$$

positiv sein. Negative Kovarianz deutet auf die umgekehrte Tendenz hin.

Positive Kovarianz ist das Gleiche wie positive Korrelation. Nach (3.15) nimmt der Korrelationskoeffizient $\rho_{XY} = \text{Cov}(X, Y) / (\sigma_X \sigma_Y)$ nur Werte zwischen -1 und $+1$ an. In den Extremfällen $\rho_{XY} = +1$ und $\rho_{XY} = -1$ gilt für die zentrierten Variablen $X' = X - EX$ und $Y' = Y - EY$ die Gleichung $|E(X' Y')|^2 = \sigma_{X'}^2 \sigma_{Y'}^2$, und also müssen X und Y einer Gleichung $Y = cX + d$ genügen. Je näher ρ_{XY} bei $+1$ oder bei -1 liegt, umso besser liegen die Wertepaare $(X(\omega), Y(\omega))$ auf einer Geraden.

Besonders gut lässt sich dies veranschaulichen, wenn $\Omega = \{1, \dots, n\}$ und P die Gleichverteilung ist. Zeichnet man jedes Paar $(X(\omega), Y(\omega))$ als Punkt in der Ebene, so wird das Paar der Zufallsvariablen durch eine Punktwolke beschrieben. Die Gleichverteilung bewirkt, dass jeder Punkt gleichwertig ist. Betrachten wir einmal ein paar Punktwolken.

In Bild 3.5 a) ist die Punktwolke aufwärts gerichtet. Es ist $\text{Cov}(X, Y) > 0$. Allerdings liegen die Punkte nicht gut auf einer Geraden. ρ_{XY} nimmt einen kleinen positiven Wert an.

In Bild 3.5 b) ist keine aufwärts oder abwärts gerichtete Tendenz zu erkennen. $\text{Cov}(X, Y)$ und ρ_{XY} sind 0 oder nahe bei 0.

In Bild 3.5 c) gibt es eine abwärts gerichtete Tendenz, und die Punkte liegen gut auf einer Geraden. Es ist $\text{Cov}(X, Y)$ negativ und ρ_{XY} nahe bei -1 .

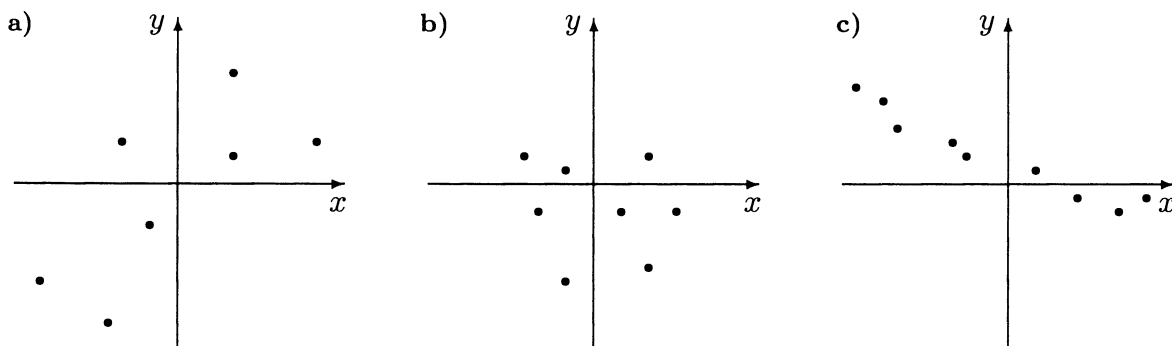


Bild 3.5

Bemerkung: Die Umkehrung der Aussage, dass unabhängige Zufallsvariable unkorreliert sind, gilt nicht. Sei etwa P das Wahrscheinlichkeitsmaß auf $\Omega = \{1, 2, 3, 4\}$, das durch $P(1) = P(2) = 2/5$ und $P(3) = P(4) = 1/10$ gegeben ist, und seien X, Y die Zufallsvariablen mit

$$\begin{aligned} X(1) = 1, \quad Y(1) = -1, & \quad X(2) = -1, \quad Y(2) = 1, \\ X(3) = Y(3) = 2, & \quad X(4) = Y(4) = -2. \end{aligned}$$

Dann ist $EX = EY = 0$ und

$$\text{Cov}(X, Y) = E(XY) = (-1) \cdot \frac{2}{5} + (-1) \cdot \frac{2}{5} + 4 \cdot \frac{1}{10} + 4 \cdot \frac{1}{10} = 0.$$

Aber Y ist extrem von X abhängig, denn $X(\omega)$ bestimmt ω eindeutig und damit auch $Y(\omega)$. Also ist Y sogar eine Funktion von X . Formal folgt die Abhängigkeit z.B. aus

$$P(X = 1, Y = -1) = \frac{2}{5} \neq \frac{4}{25} = P(X = 1) P(Y = -1).$$

Beispiele für Varianzen: (1) Ist X binomialverteilt mit Parametern n und p , so ist $\text{Var}(X) = np(1-p)$. Um dies nachzuprüfen, schreiben wir wie bei der Berechnung des Erwartungswerts $X = X_1 + \dots + X_n$ mit Bernoulli-verteilten X_i . Dann ist

$$\text{Var}(X) = \sum \text{Var}(X_i) = n \text{Var}(X_1)$$

und

$$\text{Var}(X_1) = E(X_1^2) - (EX_1)^2 = p - p^2 = p(1-p).$$

(2) Sei X die Zahl der schwarzen Kugeln in einer Stichprobe vom Umfang n ohne Rücklegen aus einer Urne mit S schwarzen und W weißen Kugeln. Sei $X_i = 1$, wenn die i -te Kugel schwarz ist, sonst $= 0$. Setze $N = S + W$, $p = S/N$, $q = W/N$. Es ist

$$\text{Var}(X_1) = EX_1^2 - (EX_1)^2 = p - p^2 = p \cdot q$$

und

$$\text{Cov}(X_1, X_2) = E(X_1 X_2) - EX_1 EX_2 = p \cdot \frac{S-1}{N-1} - p^2 = p \left(\frac{S-1}{N-1} - p \right).$$

Wegen der Austauschbarkeit der Verteilung ist

$$\text{Var}(X_i) = \text{Var}(X_1) \quad \text{und} \quad \text{Cov}(X_i, X_j) = \text{Cov}(X_1, X_2) \quad \text{für} \quad i \neq j.$$

Mit Satz 3.13 (vi) ergibt sich nach etwas Rechnung

$$\text{Var}(X) = npq \frac{N-n}{N-1}$$

als Varianz der hypergeometrischen Verteilung mit Parametern S , W , n .

3.6 Das schwache Gesetz der großen Zahlen

Mit Hilfe der obigen Begriffe ist es nun nicht schwer, das so genannte schwache Gesetz der großen Zahlen für Zufallsvariable mit endlicher Varianz zu beweisen. Es wird uns helfen, den Zusammenhang von Wahrscheinlichkeiten und relativen Häufigkeiten besser zu verstehen.

Das entscheidende Hilfsmittel ist die zwar elementare, aber für theoretische Überlegungen enorm wichtige **tschebyschewsche Ungleichung**:

Satz 3.15 Sei (Ω, P) ein diskreter Wahrscheinlichkeitsraum und X eine reell- oder komplexwertige Zufallsvariable mit endlicher Varianz. Dann gilt für jedes $\varepsilon > 0$:

$$P(|X - EX| \geq \varepsilon) \leq \text{Var}(X)/\varepsilon^2.$$

Beweis: Sei $Z = X - EX$. Setze $Y(\omega) = 0$ für ω mit $|Z(\omega)| < \varepsilon$, und $Y(\omega) = \varepsilon^2$ für ω mit $|Z(\omega)| \geq \varepsilon$. Dann ist $Y \leq |Z|^2$ und also

$$\text{Var}(X) = E(|Z|^2) \geq E(Y) = \varepsilon^2 P(Y = \varepsilon^2) = \varepsilon^2 P(|X - EX| \geq \varepsilon). \quad \square$$

Der Wert der tschebyschewschen Ungleichung liegt in ihrer Allgemeinheit. Die Abschätzung ist ziemlich grob.

[Die Tschebyschew-Ungleichung wird auch (berechtigt aber weniger verbreitet) Ungleichung von Bienaymé-Tschebyschew genannt. Bienaymé stellte sie 1853 in Zusammenhang mit der Methode der kleinsten Quadrate auf. Tschebyschew fand 1867, vermutlich unabhängig, einen anderen Zugang und verwandte sie erstmals zum Beweis des Gesetzes der großen Zahlen.]

Gelegentlich ist eine Variante der tschebyschewschen Ungleichung nützlich, die sog. **markowsche Ungleichung**:

Ist ϕ eine auf $[0, \infty)$ definierte nichtnegative monoton wachsende Funktion, und $\phi(\varepsilon) > 0$, so gilt für jede Zufallsvariable Z

$$P(|Z| \geq \varepsilon) \leq \frac{1}{\phi(\varepsilon)} E(\phi(|Z|)).$$

Der Beweis ist der gleiche wie oben. Für ω mit $|Z(\omega)| \geq \varepsilon$ setzt man $Y(\omega) = \phi(\varepsilon)$, und sonst $Y(\omega) = 0$.

Satz 3.16 (Schwachere Gesetz der großen Zahlen für unabhängige Zufallsvariable mit beschränkter Varianz) Seien $X_1, X_2, \dots, X_n$ unabhängige Zufallsvariable mit gleichem Erwartungswert und $\text{Var}(X_i) \leq M < \infty$. Dann gilt für alle $\varepsilon > 0$

$$P\left(\left|\frac{1}{n}(X_1 + X_2 + \dots + X_n) - EX_1\right| \geq \varepsilon\right) \leq \frac{M}{\varepsilon^2 n} \rightarrow 0.$$

Beweis: Sei $X = (X_1 + \dots + X_n)/n$. Dann ist $EX = EX_1$ und

$$\text{Var}(X) = n^{-2} \text{Var}(X_1 + \dots + X_n) = n^{-2} (\text{Var}(X_1) + \dots + \text{Var}(X_n)) \leq M/n.$$

Nun braucht man nur noch die tschebyschewsche Ungleichung anzuwenden. $\square$

Durch das schwache Gesetz der großen Zahlen können wir jetzt besser verstehen, weshalb die relative Häufigkeit der Sechsen bei vielen Würfeln mit einem fairen Würfel gewöhnlich nahe bei $1/6$ liegt, oder allgemeiner, warum bei n Bernoulli-Versuchen die relative Häufigkeit der Erfolge meistens nahe bei p liegt. Ist nämlich $X_i = 1$, wenn der i -te Versuch einen Erfolg ergibt, und $= 0$ sonst, so ist $EX_i = p$ und $\text{Var}(X_i) = p(1-p) \leq 1/4$. $h_n = (X_1 + \dots + X_n)/n$ ist die relative Häufigkeit der Erfolge. Also ist

$$P(|h_n - p| \geq \varepsilon) \leq \frac{1}{4\varepsilon^2 n}.$$

Für großes n ist also die Wahrscheinlichkeit dafür, dass die relative Häufigkeit der Erfolge sich um mehr als ε von der Erfolgswahrscheinlichkeit p unterscheidet, sehr klein. Dies ist das schon erwähnte Gesetz der großen Zahlen von J. Bernoulli. Um seine Leistung zu würdigen, muss man sich vor Augen halten, dass damals noch direkte Rechnungen erforderlich waren, und die hier benutzten Begriffe noch nicht verfügbar waren.

Wir haben so einen Zusammenhang von Wahrscheinlichkeiten und relativen Häufigkeiten gefunden, aber nicht in der Weise, dass Wahrscheinlichkeiten als relative Häufigkeiten oder als deren Grenzwerte definiert werden, sondern umgekehrt, von einem axiomatischen Wahrscheinlichkeitsbegriff ausgehend.

Konvention über Folgen von Zufallsvariablen: Bisher haben wir nur Modelle für eine endliche Folge von Teilerperimenten konstruiert. Modelle für unendliche Folgen von Teilerperimenten lassen sich nicht mehr im diskreten Rahmen konstruieren. Wir können uns aber hier mit einer Konvention behelfen: Unendliche Folgen von Teilerperimenten werden beschrieben, indem man für jedes n ein Modell für die ersten n Teilerperimente angibt. Wahrscheinlichkeiten, die nur von den ersten n Teilerperimenten abhängen, werden im Modell für die ersten n Teilerperimente bestimmt. Genau genommen müsste man dann $P^{(n)}$ statt P schreiben, um diese Abhängigkeit auszudrücken, aber dies ist zu schwerfällig. Wir hängen den Wahrscheinlichkeiten, Zufallsvariablen usw. deshalb keinen zusätzlichen Index, der das Modell angibt, an.

Als erste Anwendung des Gesetzes der großen Zahlen betrachten wir:

Beispiel 3.17 (*Ein vorteilhaftes Spiel, bei dem man auf Dauer verliert*) Ein Spiel, bei dem in jeder Runde der Erwartungswert des Gewinns gleich dem des Verlustes ist, heißt **fair**, eins, bei dem der erwartete Gewinn jeweils größer ist als der erwartete Verlust, heißt vorteilhaft. Das folgende Spiel ist vorteilhaft: Man beginnt mit dem Startkapital $X_0 = 1$. In jeder Runde wird eine Münze geworfen. Das Kapital X_n nach der n -ten Runde sei $X_{n-1}/2$, wenn der n -te Wurf Kopf ist, und $5X_{n-1}/3$, wenn im n -ten Wurf Zahl kommt. (Man verliert also $X_{n-1}/2$ mit Wahrscheinlichkeit $1/2$, und man gewinnt $2X_{n-1}/3$ mit Wahrscheinlichkeit $1/2$, so dass das Spiel vorteilhaft ist.)

Sei $Y_n = 1/2$, wenn im n -ten Wurf Kopf kommt, und $= 5/3$, wenn im n -ten Wurf Zahl kommt. Dann ist $X_n = Y_1 \cdot Y_2 \cdot \dots \cdot Y_n$. Die Y_i sind unabhängig. Aus

$$EY_i = \left(\frac{1}{2}\right)^2 + \frac{1}{2} \cdot \frac{5}{3} = \frac{13}{12} > 1 \quad \text{folgt} \quad EX_n = \prod_{k=1}^n EY_k \rightarrow \infty.$$

Sei $\mu = E(\log Y_i)$. Dann ist $\mu = (\log \frac{1}{2} + \log \frac{5}{3}) \frac{1}{2} < (\log \frac{1}{2} + \log 2) \frac{1}{2} = 0$. Setzt man $\varepsilon = |\mu/2| = -\mu/2$, so gilt nach dem Gesetz der großen Zahlen

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

Es folgt $P(n^{-1} \log X_n - \mu \leq -\mu/2) \rightarrow 1$. Für großes n ist also mit Wahrscheinlichkeit nahe bei 1 $X_n \leq e^{\mu n/2}$. Wegen $\mu < 0$ strebt $e^{\mu n/2}$ (exponentiell schnell) gegen 0. Auf lange Sicht strebt der Kapitalstand ziemlich schnell gegen 0. (EX_n ist trotz der kleinen Gewinnwahrscheinlichkeit groß, weil man mit kleiner Wahrscheinlichkeit enorm große Gewinne machen kann.)

Aufgaben

1. Seien X_1, X_2 zwei unabhängige Zufallsvariable mit

$$P(X_i = k) = (1-p)^k p \quad (k = 0, 1, 2, \dots; 0 \leq p \leq 1).$$

Sei $Z = \max(X_1, X_2)$. Man bestimme die gemeinsame Verteilung von Z und X_1 und die Verteilung von Z . (L)

2. Man zeige: Ist X eine Zufallsvariable mit Werten in $\mathbb{Z}^+$, so gilt

$$\text{a) } EX = \sum_{n=1}^{\infty} P(X \geq n), \quad \text{b) } EX^2 = \sum_{n=1}^{\infty} (2n-1) P(X \geq n).$$

3. Sei X_n gleichverteilt auf $\{-n, \dots, 0, 1, \dots, n\}$. Man vergleiche für große n $P(|X_n| \geq n/2)$ und $P(|X_n| \geq n/10)$ mit den Abschätzungen, die man aus der Tschebyschew-Ungleichung erhält.
4. Eine Urne enthält eine weiße und zwei schwarze Kugeln. Es werden drei Kugeln mit Rücklegen gezogen. Sei $X_k = 0$, wenn die k -te gezogene Kugel weiß ist, und sonst $X_k = 1$.
- Man beschreibe die gemeinsame Verteilung von X_1, X_2, X_3 .
 - Man beschreibe die zu (X_2, X_3) gehörige Marginalverteilung.
 - Man bestimme die Verteilung von $X_1 + X_2 + X_3$.
5. $X_1, \dots, X_n$ seien unabhängige Zufallsvariable mit positiven Werten, die identisch verteilt sind. Man zeige, dass $E(X_1/(X_1 + \dots + X_n)) = 1/n$ ist.
6. X und Y seien zwei Zufallsvariable mit endlicher Varianz σ_x^2 bzw. σ_y^2 . Für welche reellen Zahlen ist $E((Y - aX - b)^2)$ minimal und wie groß ist dieses Minimum?
7. Es wird so lange gewürfelt, bis jede der Zahlen $1, \dots, 6$ mindestens einmal gekommen ist. Wie groß ist der Erwartungswert der Zahl der benötigten Würfe? Wie groß ist die Varianz der Zahl der Würfe, die man braucht, wenn gerade das zweite verschiedene Wurf Ergebnis beobachtet wurde, bis das dritte kommt? (L)
8. Es gibt Leitungen zwischen den Orten 2 und 3 und von jedem dieser Orte zu den Orten 1 und 4. Jede dieser Leitungen wird unabhängig von den anderen mit Wahrscheinlichkeit p gestört. Mit welcher Wahrscheinlichkeit kann man noch eine Nachricht von 1 nach 4 übermitteln?
9. Man würfelt wiederholt mit zwei Würfeln. Berechnen Sie die Wahrscheinlichkeit, dass das Ereignis „Augensumme ist 7“ vor dem Ereignis „Augensumme ist 5“ erscheint. (L)
10. Die Herstellung eines Artikels verursacht Kosten $c > 0$. Kann er verkauft werden, so ist dafür ein Erlös $e > c$ erzielbar. Die Nachfrage N nimmt mit Wahrscheinlichkeit p_k den Wert $k \geq 0$ an. Sei G_a der Gewinn, wenn a Exemplare des Artikels hergestellt werden. Bestimmen Sie G_a als Funktion von N . Für welches a ist EG_a maximal? (L)
11. Die folgende schöne Aufgabe über Wartezeiten beim iterierten Koinzidenzproblem verdanke ich Herrn Kollegen M. R. Schroeder:
Bei einem Tanzfest von N Ehepaaren werden beim ersten Tanz die Partner einander zugelost. Wer mit seiner Gattin getanzt hat, verlässt mit dieser das Parkett, und beim nächsten Tanz werden die verbleibenden Damen und Herren neu einander zugelost, usw. Wie groß ist die erwartete Anzahl der Tänze, die die Kapelle spielen muss? (L)

§ 4 Grundbegriffe der Schätztheorie

Wir wollen nun auch ein paar wichtige Begriffsbildungen der Statistik kennen lernen. Ein klassisches Beispiel soll uns dabei helfen.

Beispiel 4.1 (*Schätzung eines Fischbestandes*) Ein Teich enthält eine unbekannte Zahl N von Fischen, die geschätzt werden soll. Dazu werden W Fische gefangen, mit einem weißen Fleck markiert und wieder ausgesetzt. Man wartet eine Weile, dann werden in einem zweiten Fischzug n Fische gefangen, und die Zahl x der markierten Fische in diesem zweiten Fang wird ermittelt.

Eine plausible Schätzung $\hat{N}(x)$ von N ergibt sich aus folgender Überlegung: Wenn x nicht zu klein ist, müsste der Anteil x/n der markierten Fische am zweiten Fang ungefähr gleich dem Anteil W/N der markierten Fische am Gesamtbestand sein. Wenn $\hat{N}(x)$ eine gute Schätzung von N ist, müsste also $x/n \approx W/\hat{N}(x)$ sein. Es liegt daher nahe, als Schätzwert $\hat{N}(x)$ die zu Wn/x nächstgelegene ganze Zahl zu wählen. (Dass diese Schätzung für kleine x nicht sehr zuverlässig ist, sieht man schon an der Form von $\hat{N}(x)$. Ist x klein, so bewirken kleine Änderungen von x große Änderungen von $\hat{N}(x)$. Man sollte dann n erhöhen.)

Die eben heuristisch abgeleitete Schätzung ergibt sich unwesentlich modifiziert auch aus einem ganz anderen Ansatz, der den Vorteil hat, stark verallgemeinerungsfähig zu sein. Wir beschreiben dazu den zweiten Fang durch das Modell des Ziehens von n Kugeln aus einer Urne mit W weißen und $S = N - W$ schwarzen Kugeln ohne Rücklegen. Die Wahrscheinlichkeit, genau x markierte Fische im zweiten Fang zu haben, ist dann

$$P_N(x) = \binom{W}{x} \binom{N-W}{n-x} / \binom{N}{n} \quad (0 \leq x \leq n).$$

Die Abhängigkeit der Wahrscheinlichkeit von N ist von prinzipiell anderer Art als bei bedingten Wahrscheinlichkeiten. N ist nicht Ereignis, sondern ein unbekannter Wert eines Parameters der Verteilung.

Der **Maximum-Likelihood-Ansatz** zur Schätzung von N besagt, man solle den Wert von N als Schätzwert nehmen, der $P_N(x)$ für das beobachtete x maximiert. Um ihn zu finden, bilden wir

$$\frac{P_N(x)}{P_{N-1}(x)} = \frac{\binom{W}{x} \binom{N-W}{n-x} \binom{N-1}{n}}{\binom{N}{n} \binom{W}{x} \binom{N-W-1}{n-x}} = \frac{(N-W)(N-n)}{N(N-W-n+x)}.$$

Es ergibt sich, dass $P_N(x) > P_{N-1}(x)$ äquivalent zu $(N-W)(N-n) > N(N-W-n+x)$ ist, und also zu $nW > Nx$. Die gleiche Äquivalenz gilt für die Relationen $<$ und $=$.

Daraus lässt sich folgern: $P_N(x)$ ist maximal für $\hat{N}(x) = \lceil nW/x \rceil$, wobei $\lceil t \rceil$ die größte ganze Zahl ist, die $\leq t$ ist. Ist nW/x keine ganze Zahl, so ist dieses $\hat{N}(x)$ eindeutig. Andernfalls sind nW/x und $(nW/x) - 1$ Werte von N , für die $P_N(x)$ maximal ist, und es gibt keine weiteren. Das neue $\hat{N}(x)$ unterscheidet sich nicht wesentlich von dem der obigen heuristischen Überlegung. Wir nennen $\hat{N}(x) = \lceil nW/x \rceil$ den **Maximum-Likelihood-Schätzer** von N . Die Idee hinter diesem Schätzer ist, dass Werte des Parameters N , für die $P_N(x)$ groß ist, plausibler scheinen als solche, für die $P_N(x)$ klein ist. Sie bieten gewissermaßen die bessere Erklärung für die Beobachtung x . Aber diese Idee hat nichts Zwingendes. Wir werden versuchen müssen, auch andere Kriterien mit ins Spiel zu bringen.

Gewarnt werden muss vor der Versuchung, $\hat{N}(x)$ als den „wahrscheinlichsten“ Wert von N zu betrachten. Da auf der Menge der Werte, die N annehmen kann, kein Wahrscheinlichkeitsmaß definiert ist, hätte eine solche Aussage keinerlei mathematischen Inhalt.

Es soll auch gleich auf eine fundamentale Problematik hingewiesen werden, die analog auch in anderen statistischen Problemen auftritt: Selbst wenn gezeigt wäre, dass $\hat{N}(x)$ im obigen Modell mit großer Wahrscheinlichkeit nahe am wahren Parameter liegt, wäre damit noch nicht sicher, dass das Verfahren zur Schätzung des Fischbestandes sehr zuverlässig ist. Ist es wirklich klar, dass die hypergeometrische Verteilung ein gutes Modell ist? Fische schwimmen oft in Schwärmen, und gewisse Fische könnten sich leichter fangen lassen als andere. Wir können daher nicht ausschließen, dass das Urnenmodell nicht passt. Ob es passt, kann letztlich nur durch Erfahrung entschieden werden. (Zumindest sollte man wohl beide Fänge und das Aussetzen der Fische möglichst gleichmäßig über die Fläche des Sees verteilen.)

4.1 Der allgemeine Rahmen von Schätzproblemen

Der entscheidende Unterschied zu den bisher untersuchten Modellen ist, dass wir nicht mehr nur *ein* Wahrscheinlichkeitsmaß betrachten, sondern eine ganze Familie von Wahrscheinlichkeitsmaßen. Zur formalen Beschreibung eines Schätzproblems mit diskretem Stichprobenraum benötigen wir

- eine nichtleere höchstens abzählbare Menge $\mathcal{X}$, den Stichprobenraum,
- eine Familie $\{P_\vartheta : \vartheta \in \Theta\}$ von Wahrscheinlichkeitsmaßen auf $\mathcal{X}$,
- eine zu schätzende Funktion $g(\vartheta)$.

$\mathcal{X}$ ist die Menge der möglichen Beobachtungsergebnisse. Während in der Wahrscheinlichkeitstheorie Ω eine allgemein akzeptierte Notation für den Stichprobenraum ist, wählt man in der Statistik oft andere Buchstaben, wie z.B. $\mathcal{X}$. Durch die unterschiedliche Bezeichnung wollen wir darauf hinweisen, dass jedes $x \in \mathcal{X}$ beobachtbar sein muss. In Wahrscheinlichkeitstheoretischen Modellen werden wir zunehmend auch Stichprobenräume Ω benutzen, deren Elemente ω , nicht beobachtbar sind.

Θ kann eine beliebige abstrakte Menge sein, mit deren Elementen die für möglich gehaltenen Verteilungen parametrisiert sind, aber in der Praxis ist Θ meist ein Intervall in $\mathbb{R}$, $\mathbb{R}^d$ oder $\mathbb{N}$.

In Beispiel 4.1 war $\mathcal{X} = \{0, 1, 2, \dots, n\}$, wenn n die Zahl der Fische im zweiten Fang war. Ferner war $\vartheta = N$ die unbekannte Zahl der Fische im See, $P_\vartheta = P_N$ die hypergeometrische Verteilung $h(\cdot; n, N, W)$ und $g(\vartheta) = \vartheta = N$.

Dass der Parameter ϑ selbst zu schätzen ist, also $g(\vartheta) = \vartheta$ gilt, ist natürlich der häufigste Fall. Aber andere Beispiele werden auch vorkommen. Z.B. könnte die Varianz $np(1-p)$ einer Binomialverteilung zu schätzen sein. Dann wäre $\vartheta = p$ und $g(p) = np(1-p)$. Später werden wir auch die durch zweidimensionale Parameter $\vartheta = (\mu, \sigma^2)$ bestimmte „Normalverteilung“ kennen lernen und z.B. $g(\vartheta) = \mu$ schätzen.

Ist $\mathcal{Y}$ der Wertebereich von g , so heißt jede Abbildung $T : \mathcal{X} \rightarrow \mathcal{Y}$ ein **Schätzer** von $g(\vartheta)$. Diese Sprechweise lässt es natürlich zu, dass auch solche Funktionen T Schätzer genannt werden, die statistisch völlig unsinnig sind.

Häufig deutet man in der Notation bereits an, was geschätzt wird, und setzt ein „Dach“ über die zu schätzende Größe. $\hat{N}$ wäre ein Schätzer von N , $\hat{p}$ ein Schätzer für p und $\hat{g}$ ein Schätzer für $g(\vartheta)$. Man unterscheidet dann nicht immer ganz sorgfältig den Schätzer $\hat{g}$ von dem **Schätzwert** $\hat{g}(x)$ aufgrund der Beobachtung x . Z.B. schreibt man meist kurz $\hat{p} = 0,3$, wenn sich aus dem beobachteten x der Schätzwert $\hat{p}(x) = 0,3$ ergibt.

4.2 Maximum-Likelihood-Schätzer

Wir haben in Beispiel 4.1 bereits ein sehr wichtiges Verfahren zur Auffindung von Schätzern kennen gelernt. Allgemein können wir es folgendermaßen beschreiben:

Die Funktion L_x , die jedem ϑ den Wert $L_x(\vartheta) := P_\vartheta(x)$ zuordnet, nennen wir **Likelihood-Funktion**. Wenn L_x einen Maximalwert in $\hat{\vartheta}(x)$ annimmt, also

$$L_x(\hat{\vartheta}) := \sup\{L_x(\vartheta) : \vartheta \in \Theta\} \quad (4.1)$$

gilt, nennen wir $\hat{\vartheta}(x)$ eine **Maximum-Likelihood-Schätzung** von ϑ und $g(\hat{\vartheta}(x))$ eine Maximum-Likelihood-Schätzung von $g(\vartheta)$.

In den meisten Fällen in der statistischen Praxis gibt es einen eindeutig bestimmten Maximum-Likelihood-Schätzer $\hat{\vartheta}$, und er ist gewöhnlich ein guter Schätzer von ϑ .

Besonders gilt dies, wenn sich die Schätzung auf viele unabhängige Einzelbeobachtungen stützt. Im allgemeinen Fall braucht jedoch ein Maximum nicht zu existieren, oder es kann mehrere Maximum-Likelihood-Schätzer geben. Selbst wenn ein eindeutiger Maximum-Likelihood-Schätzer existiert, braucht er nicht immer „gut“ zu sein.

In vielen Fällen ist Θ ein Intervall in $\mathbb{R}$, und eine Maximum-Likelihood-Schätzung kann durch Differenziation gefunden werden. Dabei ist es häufig zweckmäßig, statt L_x die Funktion $\mathcal{L}_x = \log L_x$ zu betrachten, die wegen der Monotonie der Funktion $t \mapsto \log t$ das Maximum an der gleichen Stelle hat. ($\mathcal{L}_x$ heißt **log-Likelihood-Funktion**; $\log$ bezeichnet wie üblich den natürlichen Logarithmus).

Beispiel 4.2 (*Schätzung einer Erfolgswahrscheinlichkeit*) In n Bernoulli-Experimenten soll die Erfolgswahrscheinlichkeit p aus der Zahl x der Erfolge geschätzt werden. Es ist

$$L_x(p) = \binom{n}{x} p^x (1-p)^{n-x}.$$

Daher ist $\mathcal{L}_x(p) = \log L_x(p) = \log \binom{n}{x} + x \log p + (n-x) \log(1-p)$, und

$$\frac{d}{dp} \mathcal{L}_x(p) = \frac{x}{p} - \frac{n-x}{1-p}. \quad (4.2)$$

Als Nullstelle in (4.2) findet man $\hat{p}(x) = x/n$, und man prüft leicht, dass es sich um ein Maximum von $\mathcal{L}_x(p)$ handelt. Also ist x/n Maximum-Likelihood-Schätzung von p .

In vielen Fällen besitzt die **Maximum-Likelihood-Gleichung** $\frac{d}{d\vartheta} \mathcal{L}_x(\vartheta) = 0$ keine explizit angebbare Lösung $\hat{\vartheta}(x)$. Für diesen Fall sind gute Näherungsverfahren entwickelt worden. (Vgl. van der Waerden (1971), Kapitel VIII.)

Zum Abschluss dieser Diskussion soll uns ein einfaches Beispiel zeigen, dass man mit Maximum-Likelihood-Schätzern systematische Fehler machen kann.

Beispiel 4.3* (*Taxiproblem*) In einer großen Stadt gibt es N Taxis, die — von außen gut lesbar — die Nummern $1, \dots, N$ tragen. Ein Passant steht an einer viel befahrenen Straße und beobachtet die Nummern der n Taxis $x_1, \dots, x_n$. Wiederholungen werden ignoriert. Nehmen wir an, dass noch alle Taxis in Betrieb sind, so ist $\mathcal{X}$ die Menge der Teilmengen $x = \{x_1, \dots, x_n\}$ vom Umfang n aus $\{1, 2, 3, \dots\}$, die wir uns so hingeschrieben denken können, dass $x_1 < x_2 < \dots < x_n$ gilt. Nun soll N geschätzt werden.

Natürlich muss $N \geq x_n$ sein. Dann ist $P_N(x) = \binom{N}{n}^{-1}$, denn wenn es N Taxis gibt, sind die Teilmengen von $\{1, \dots, N\}$ der Mächtigkeit n gleichwahrscheinlich. $\binom{N}{n}^{-1}$ ist umso größer, je kleiner N ist. Also ist $\hat{N}(x) = x_n = \max_i x_i$ der Maximum-Likelihood-Schätzer. Man schätzt die Zahl der Taxis durch die höchste beobachtete Nummer. Wegen $\hat{N}(x) \leq N$ gibt man nie eine zu hohe Schätzung ab. Aber besonders realistisch scheint die Schätzung nicht zu sein.

Versuchen wir mit heuristischen Argumenten Schätzer zu finden, die nicht systematisch N unterschätzen. Aus Symmetriegründen müsste „im Durchschnitt“ von vielen Stichproben die Zahl $x_1 - 1$ der nicht beobachteten Taxis mit Nummern $< x_1$ etwa gleich der Zahl $N - x_n$ der nicht beobachteten Taxis mit Nummern $> x_n$ sein. Also bietet sich $N_1(x) = x_n + x_1 - 1$ als Schätzer an.

Noch besser ist es, die Länge der nicht beobachteten Lücke $\{x_n + 1, \dots, N\}$ oberhalb von x_n durch die mittlere Länge der vorangegangenen Lücken zu schätzen, also durch

$$\frac{1}{n} ((x_1 - 1) + (x_2 - x_1 - 1) + \dots + (x_n - x_{n-1} - 1)) = \frac{x_n - n}{n}.$$

Dies führt zu dem Schätzer

$$N_2(x) = x_n + (x_n - n)/n.$$

(In der Praxis würde man die nächstgelegene ganze Zahl als Schätzer wählen.)

Bisher können wir noch nicht sagen, ob und in welchem Sinn N_2 ein besserer Schätzer ist als der Maximum-Likelihood-Schätzer $\hat{N}$. Darauf kommen wir später zurück.

Auch in diesem Problem wird wieder sichtbar, dass alles davon abhängt, ob das Modell stimmt. Wenn z.B. gewisse Taxis schon verschrottet worden sind und die Nummern nicht neu vergeben wurden, oder man sich in einem Stadtteil aufhält, in dem nur die kleinen Nummern vergeben worden sind, wird die Schätzung entsprechend unzuverlässig sein.

Das Taxiproblem lässt sich leicht so variieren, dass $\{1, \dots, N\}$ durch ein unbekanntes Intervall $\{M, M + 1, \dots, N\}$ ersetzt wird. Es heißt, dass Varianten davon im zweiten Weltkrieg eine Rolle gespielt haben. Aus den Seriennummern der abgeschossenen Panzer wurde die Zahl der Panzer der Wehrmacht von den Alliierten zuverlässiger geschätzt als durch Geheimdienstmethoden.

4.3 Erwartungstreue

Ist T eine Abbildung von $\mathcal{X}$ in $\mathbb{R}$, so bezeichnen wir den Erwartungswert bzgl. P_ϑ mit E_ϑ . Es ist also

$$E_\vartheta T = \sum_{x \in \mathcal{X}} T(x) P_\vartheta(x).$$

Schreiben wir solche Erwartungswerte hin, so soll damit zugleich deren Existenz vorausgesetzt sein.

Die Beobachtung des Ergebnisses des Zufallsexperiments, das dem Schätzproblem zugrundeliegt, lässt sich natürlich als Beobachtung der Zufallsvariablen X mit $X(x) = x$ auffassen. Wenn wir augenfällig machen wollen, dass T Funktion des Beobachtungsergebnisses ist, schreiben wir statt T auch $T(X)$.

In einem Schätzproblem heißt ein Schätzer $\hat{g}$ von $g(\vartheta)$ **erwartungstreu**, wenn für alle $\vartheta \in \Theta$ die Gleichung

$$E_{\vartheta}(\hat{g}(X)) = g(\vartheta) \quad (4.3)$$

gilt. Speziell heißt $\hat{\vartheta}$ erwartungstreuer Schätzer von ϑ , wenn für alle $\vartheta \in \Theta$ die Gleichung

$$\vartheta = E_{\vartheta}(\hat{\vartheta}(X))$$

gilt.

$b(\vartheta, \hat{g}) = E_{\vartheta}(\hat{g}(X)) - g(\vartheta)$ heißt **Bias** der Schätzung $\hat{g}$. Ein Schätzer ist also genau dann erwartungstreu (engl. „unbiased“), wenn sein Bias = 0 ist.

Ist X binomialverteilt mit Parametern n und p , so ist $E(X/n) = p$. Also ist in Beispiel 4.2 die Schätzung $\hat{p}(X) = X/n$ erwartungstreu.

Wir erwähnen noch zwei weitere einfache und doch sehr wichtige Beispiele: Unabhängige Messungen einer Größe μ beschreibt man oft durch unabhängige Zufallsvariable $X_1, \dots, X_n$. Es ist im Moment nicht nötig, die unbekannte Verteilung P_{ϑ} von $X = (X_1, \dots, X_n)$ genau festzulegen. Wir nehmen nur an, dass $E_{\vartheta} X_i = \mu$ für $i = 1, \dots, n$ sei. (Wenn μ der wahre zu messende Wert ist, soll die Verteilung so beschaffen sein, dass die X_i um μ herum streuen.) Sei $g_1(\vartheta) = E_{\vartheta}(X_i) = \mu$ und

$$\bar{X} = \frac{1}{n} \sum_{k=1}^n X_k$$

der **Mittelwert**. Dann ist $E_{\vartheta}(\bar{X}) = n^{-1}(E_{\vartheta} X_1 + \dots + E_{\vartheta} X_n) = \mu$. Also ist der Mittelwert eine erwartungstreue Schätzung von μ .

Haben die X_i eine unbekannte Varianz σ^2 , so bestimmt P_{ϑ} auch $\sigma^2 = g_2(\vartheta) = \text{Var}_{\vartheta}(X_i)$. (Im einfachsten Fall hängt die Verteilung nur von μ und σ^2 ab, und man kann $\vartheta = (\mu, \sigma^2)$ wählen. Var_{ϑ} bezeichnet die Varianz bei Benutzung von P_{ϑ} .)

Man wählt gewöhnlich

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad (4.4)$$

als Schätzung von σ^2 . Hierbei fällt auf, dass durch $n-1$ und nicht durch n geteilt wird. Der Grund ist, dass man σ^2 erwartungstreu schätzen möchte. Um zu zeigen, dass s^2 dieser Forderung genügt, berechnen wir $E_{\vartheta} s^2$. Da die X_i unkorreliert sind, ist für $i \neq j$ $E_{\vartheta}((X_i - \mu)(X_j - \mu)) = 0$. Ferner ist

$$E_{\vartheta}((\bar{X} - \mu)^2) = \text{Var}_{\vartheta}(\bar{X}) = \frac{1}{n^2} \text{Var}_{\vartheta}(X_1 + \dots + X_n) = \frac{\sigma^2}{n}.$$

Damit ergibt sich

$$\begin{aligned} E_{\vartheta}((X_i - \bar{X})^2) &= E_{\vartheta}(((X_i - \mu) - (\bar{X} - \mu))^2) \\ &= \sigma^2 - \frac{2}{n} \sum_{j=1}^n E_{\vartheta}(((X_i - \mu)(X_j - \mu))) + \frac{\sigma^2}{n} = \sigma^2 \left(1 - \frac{1}{n}\right). \end{aligned}$$

Es folgt

$$E_{\vartheta} s^2 = \frac{1}{n-1} \cdot n \cdot \sigma^2 \cdot \frac{n-1}{n} = \sigma^2. \quad (4.5)$$

Zugleich ergibt sich, dass s^2/n eine erwartungstreue Schätzung der Varianz $\text{Var}_{\vartheta}(\bar{X}) = \sigma^2/n$ des Mittelwertes $\bar{X}$ ist.

Erwartungstreue ist sicher eine wünschenswerte Eigenschaft eines Schätzers, aber man sollte nicht nur erwartungstreue Schätzer in Betracht ziehen. Es gibt Schätzprobleme, in denen gar keine erwartungstreuen Schätzer existieren. Ist z.B. X binomialverteilt, so ist die Standardabweichung $\sqrt{np(1-p)}$ nicht erwartungstreu schätzbar, denn für jeden Schätzer T ist

$$E_p T = \sum_{x=0}^n T(x) \binom{n}{x} p^x (1-p)^{n-x}$$

ein Polynom in p . (Damit ist die Ableitung von $E_p T$ nach p in $p = 0$ endlich, aber die von $\sqrt{np(1-p)}$ ist unendlich.)

Wenn erwartungstreue Schätzer existieren, brauchen sie von anderen Kriterien aus betrachtet nicht unbedingt gute Schätzer zu sein. Natürlich wird man in jedem Fall versuchen, den Bias klein zu halten.

Ergänzung zum Taxiproblem*: Im obigen Taxiproblem (Beispiel 4.3) ist die Maximum-Likelihood-Schätzung $\hat{N}$ nicht erwartungstreu. Aber die Schätzer N_1 und N_2 sind erwartungstreu. Dies lässt sich mit Hilfe der Überlegungen im letzten Beispiel von Abschnitt 3.3 über die Wartezeit beim Ziehen ohne Rücklegen zeigen. Setzt man dort $S = n$, so ist der Stichprobenraum Ω die Menge der Folgen $\omega = (\omega_1, \dots, \omega_N)$, in denen genau n Koordinaten ω_i den Wert s und $N - n$ Koordinaten den Wert w haben. Wir können ω die (wiederum aufsteigend geordnete) Menge $x = \{x_1, \dots, x_n\}$ der Indizes i mit $\omega_i = s$ zuordnen. Die Stichproben x im Taxiproblem und ω im Urnenproblem sind zwar auf ganz verschiedene Art zustandegekommen, aber es ist $P_N(x) = \binom{N}{n}^{-1} = P(\omega)$.

Es ist $U_i(\omega) = x_i$, und wir hatten

$$E(U_n) = E(T_1 + \dots + T_n) = n E T_1 = n \frac{N+1}{S+1}$$

gezeigt. Wegen $n = S$ gilt also

$$E_N(N_2) = E\left(U_n + \frac{U_n - n}{n}\right) = \frac{n+1}{n} E(U_n) - 1 = N.$$

Analog zeigt man die Erwartungstreue von N_1 .

4.4 Der mittlere quadratische Fehler

Die wohl wichtigste Forderung, die man an einen Schätzer T von $g(\vartheta)$ stellen wird, ist, dass mit großer Wahrscheinlichkeit die Schätzwerte $T(x)$ nahe bei der zu schätzenden Größe $g(\vartheta)$ liegen. Ein gutes Maß für die Abweichung ist der **mittlere quadratische Fehler**

$$R(\vartheta, T) = E_{\vartheta}((T(X) - g(\vartheta))^2).$$

Für die Wahl dieser Größe als Maßzahl für die Ungenauigkeit der Schätzung sprechen ähnliche Gründe wie wir sie nach Definition 3.12 für die Benutzung der Varianz angeführt hatten.

Der mittlere quadratische Fehler lässt sich aus der Varianz und dem Bias berechnen, denn es ist

$$R(\vartheta, T) = E_{\vartheta}(((T - E_{\vartheta}T) - (g(\vartheta) - E_{\vartheta}T))^2) = \text{Var}_{\vartheta}(T) + b(\vartheta, T)^2.$$

Natürlich möchte man Schätzer T benutzen, für die $R(\vartheta, T)$ klein ist. Aber dieses Kriterium erlaubt oft keine eindeutige Wahl. Sind T_1, T_2 zwei Schätzer, so ist oft $R(\vartheta, T_1)$ für einige Werte von ϑ kleiner als $R(\vartheta, T_2)$ und für andere ist es gerade umgekehrt. Es ist nicht schwer, Schätzer anzugeben, die für spezielle Parameterwerte gut sind.

Man kann z.B. ein $\vartheta_0 \in \Theta$ auswählen und $T_0(x) = g(\vartheta_0)$ setzen. Dann ist $R(\vartheta_0, T_0) = 0$, aber $R(\vartheta, T_0) = (g(\vartheta) - g(\vartheta_0))^2$ kann für andere ϑ groß sein. Solche unsinnigen Schätzer sollte man natürlich von vorneherein ausschließen.

Immerhin gelingt es in vielen Fällen, wenn man nur Schätzer betrachtet, die gewisse Nebenbedingungen (wie z.B. Erwartungstreue) erfüllen, in dieser Klasse von Schätzern einen zu finden, der — verglichen mit allen anderen in dieser Klasse — für *alle* ϑ einen kleineren mittleren quadratischen Fehler aufweist.

Hier müssen wir uns mit einem Beispiel begnügen. In dem gleich folgenden Abschnitt über die Informationsungleichung wollen wir zeigen, dass für binomialverteiltes X der Schätzer $\hat{p}(X) = X/n$ der Erfolgswahrscheinlichkeit p in diesem Sinn optimal unter allen erwartungstreuen Schätzern ist.

(Dies schließt nicht aus, dass es auch Argumente für andere (nicht erwartungstreue) Schätzer gibt: Wenn man keine Erfolge beobachtet, sollte man vielleicht nicht gleich auf $p = 0$ schließen — wobei man sich nur „nach unten“ irren könnte — sondern bloß darauf, dass p klein ist. Ein Beispiel eines Schätzers, der dies leistet, wäre $\hat{p}_1(X) = (X+1)/(n+2)$. Man kann nachrechnen, dass $R(p, \hat{p}) > R(p, \hat{p}_1)$ für alle p aus einem offenen Intervall $]p_1, p_2[$ mit $0 < p_1 < p_2 < 1$ gilt.)

Natürlich kann auch ohne Optimalitätsbeweise die Betrachtung des mittleren quadratischen Fehlers eine Hilfe sein. Z.B. waren im Taxiproblem N_1 und N_2 erwartungstreu. Es lässt sich zeigen, dass der mittlere quadratische Fehler für N_2 kleiner ist als für N_1 . Also wird man N_2 bevorzugen.

4.5 Die Informationsungleichung*

Wir wollen nun eine untere Abschätzung für die Varianz $\text{Var}_{\vartheta}(T)$ eines Schätzers $T = T(X)$ von $g(\vartheta)$ herleiten.

Wir nehmen an, dass $P_{\vartheta}(x) > 0$ für alle ϑ und x gilt, und dass die Parametermenge Θ ein offenes Intervall in $\mathbb{R}$ ist. Es war $L_x(\vartheta) = P_{\vartheta}(x)$ und $\mathcal{L}_x(\vartheta) = \log L_x(\vartheta)$. Nehmen wir an, dass die Ableitung nach ϑ existiert, und bezeichnen wir sie mit einem Strich, so ist

$$\mathcal{L}'_x(\vartheta) = \frac{d}{d\vartheta} \log L_x(\vartheta) = \frac{L'_x(\vartheta)}{L_x(\vartheta)}.$$

Wir verwenden die Gleichung

$$1 = \sum_x L_x(\vartheta). \quad (4.6)$$

Aus der Definitionsgleichung des Bias ergibt sich

$$g(\vartheta) + b_T(\vartheta) = E_{\vartheta}T = \sum_x T(x) L_x(\vartheta). \quad (4.7)$$

Nimmt man an, dass in (4.6) und (4.7) Differenziation und Summation vertauschbar sind, was bei endlichem $\mathcal{X}$ immer erfüllt ist, so ist

$$\begin{aligned} g'(\vartheta) + b'_T(\vartheta) &= \sum_x T(x) L'_x(\vartheta) = \sum_x T(x) \mathcal{L}'_x(\vartheta) L_x(\vartheta) \\ &= E_{\vartheta}(T(X) \mathcal{L}'_X(\vartheta)) \end{aligned} \quad (4.8)$$

und

$$0 = \sum_x L'_x(\vartheta) = \sum_x \mathcal{L}'_x(\vartheta) L_x(\vartheta) = E_{\vartheta}(\mathcal{L}'_X(\vartheta)). \quad (4.9)$$

Multipliziert man (4.9) mit $E_{\vartheta}T$, und subtrahiert man die Gleichung dann von (4.8), so ergibt sich

$$g'(\vartheta) + b'_T(\vartheta) = E_{\vartheta}((T(X) - E_{\vartheta}T) \mathcal{L}'_X(\vartheta)).$$

Durch Anwendung der cauchy-schwarzschen Ungleichung erhalten wir

$$\begin{aligned} (g'(\vartheta) + b'_T(\vartheta))^2 &\leq E_{\vartheta}((T(X) - E_{\vartheta}T)^2) \cdot E_{\vartheta}((\mathcal{L}'_X(\vartheta))^2) \\ &= \text{Var}_{\vartheta}(T) \cdot E_{\vartheta}((\mathcal{L}'_X(\vartheta))^2). \end{aligned}$$

Die Größe $I(\vartheta) = E_{\vartheta}((\mathcal{L}'_X(\vartheta))^2)$ nennt man **Fisher-Information**. Ist sie positiv — was in der Regel der Fall ist —, so kann man sie auf die andere Seite bringen und erhält die (von Fréchet, Cramér und Rao entdeckte) **Informationsungleichung**

$$\text{Var}_{\vartheta}(T) \geq \frac{(g'(\vartheta) + b'_T(\vartheta))^2}{I(\vartheta)}. \quad (4.10)$$

Für erwartungstreue Schätzer von $g(\vartheta) = \vartheta$ ist unter den oben gemachten Voraussetzungen $g'(\vartheta) = 1$ und $b'_T(\vartheta) = 0$; also

$$\text{Var}_{\vartheta}(T) \geq \frac{1}{I(\vartheta)}. \quad (4.11)$$

Oft ist X ein Vektor von unabhängigen Zufallsvariablen $X_1, \dots, X_n$ und $P_{\vartheta}(x)$ für $x = (x_1, \dots, x_n)$ von der Form $Q_{\vartheta}(x_1) Q_{\vartheta}(x_2) \dots Q_{\vartheta}(x_n)$. Setzt man wieder Vertauschbarkeit von Summation und Differenziation nach ϑ voraus, so ist analog zu (4.9)

$$E_{\vartheta}\left(\frac{d}{d\vartheta} \log Q_{\vartheta}(X_i)\right) = 0.$$

Für $i \neq j$ folgt damit

$$E_{\vartheta}\left(\left(\frac{d}{d\vartheta} \log Q_{\vartheta}(X_i)\right) \left(\frac{d}{d\vartheta} \log Q_{\vartheta}(X_j)\right)\right) = 0.$$

Wir erhalten

$$\begin{aligned} I(\vartheta) &= E_{\vartheta}\left(\left(\frac{d}{d\vartheta} \log(Q_{\vartheta}(X_1) \cdot Q_{\vartheta}(X_2) \cdot \dots \cdot Q_{\vartheta}(X_n))\right)^2\right) \\ &= E_{\vartheta}\left(\left(\sum_{i=1}^n \frac{d}{d\vartheta} \log Q_{\vartheta}(X_i)\right)^2\right) = \sum_{i=1}^n E_{\vartheta}\left(\left(\frac{d}{d\vartheta} \log Q_{\vartheta}(X_i)\right)^2\right). \end{aligned}$$

Ist also $I^Q(\vartheta) = E_{\vartheta} \left(\left(\frac{d}{d\vartheta} \log Q_{\vartheta}(X_i) \right)^2 \right)$ die Fisher-Information zur Familie $\{Q_{\vartheta}, \vartheta \in \Theta\}$, so ist $I(\vartheta) = n I^Q(\vartheta)$.

Sei nun speziell $(X_1, \dots, X_n)$ Bernoulli-verteilt mit Parameter $\vartheta = p$. Dann ist $Q_p(1) = p$, $Q_p(0) = 1 - p$ und

$$\frac{d}{dp} \log Q_p(1) = \frac{d}{dp} \log p = \frac{1}{p}, \quad \frac{d}{dp} \log Q_p(0) = -\frac{1}{1-p}.$$

Es folgt

$$I^Q(p) = E_p \left(\left(\frac{d}{dp} \log Q_p(X_1) \right)^2 \right) = p \cdot \frac{1}{p^2} + (1-p) \cdot \frac{1}{(1-p)^2} = \frac{1}{p(1-p)}.$$

Damit ist $I(p) = n/(p(1-p))$. Für jeden erwartungstreuen Schätzer T von p muss also $\text{Var}_p(T) \geq p(1-p)/n$ gelten. Andererseits ist $\hat{p}(X) = (X_1 + \dots + X_n)/n$ erwartungstreuer Schätzer mit $\text{Var}_p(\hat{p}) = p(1-p)/n$. $\hat{p}$ ist also unter den erwartungstreuen Schätzern optimal.

4.6 Konsistenz*

Zur Schätzung einer Erfolgswahrscheinlichkeit p haben wir eigentlich nicht nur einen sondern — genauer — für jedes n einen Schätzer angegeben, nämlich bei Bernoulli-verteilten $X_1, \dots, X_n$ den Schätzer $(X_1 + \dots + X_n)/n$. Nach dem schwachen Gesetz der großen Zahl konvergieren diese Schätzer in dem Sinne gegen p , dass die Wahrscheinlichkeit eines Schätzfehlers, der $\varepsilon > 0$ übertrifft, für jedes $\varepsilon > 0$ gegen 0 strebt. Diese Eigenschaft wollen wir Konsistenz nennen.

Sei allgemeiner $X_1, X_2, \dots$ eine Folge von Zufallsvariablen, die nacheinander ermittelte Teilergebnisse beschreibt. Sei $X^{(n)} = (X_1, \dots, X_n)$ und $\hat{g}^{(n)} = \hat{g}^{(n)}(X^{(n)})$ ein Schätzer von $g(\vartheta)$, der sich auf $X^{(n)}$ stützt. Die Folge $(\hat{g}^{(n)})$ von Schätzern heißt **konsistent** (für $g(\vartheta)$), wenn für alle ϑ und alle $\varepsilon > 0$

$$P_{\vartheta}(|\hat{g}^{(n)}(X^{(n)}) - g(\vartheta)| > \varepsilon) \rightarrow 0 \quad (4.12)$$

gilt. Dies ist eine Art Minimalforderung an Folgen von Schätzern. Wenn man nicht einmal aus einer ganzen Folge von Beobachtungsergebnissen $g(\vartheta)$ immer genauer schätzen kann, kann man kaum zufrieden sein.

Betrachten wir noch einmal das in Abschnitt 4.3 gegebene Beispiel der unabhängigen Messungen. Wir hatten den Mittelwert $\bar{X} = (X_1 + \dots + X_n)/n$ von n Messungen einer Größe μ als Schätzer für μ gewählt. Schreiben wir dafür ausführlicher $\bar{X}(n)$, so ist also $\hat{g}^{(n)}(X^{(n)}) = \bar{X}(n)$ Schätzer für $g(\vartheta) = \mu$. Nehmen wir an, dass die Varianzen der X_i beschränkt seien, so folgt (4.12) aus dem Gesetz der großen Zahlen. Der Mittelwert ist daher konsistenter Schätzer von μ .

Als Schätzung von σ^2 hatten wir die durch (4.4) definierte Größe s^2 gewählt. Im Fall der Schätzung von σ^2 wäre also $g(\vartheta) = \sigma^2$

$$\hat{g}^{(n)}(X^{(n)}) = s^2 = s^2(n) = \frac{1}{n-1} \sum_{i=1}^n ((X_i - \mu) - (\bar{X}(n) - \mu))^2.$$

Setzen wir voraus, dass die vierten Momente von $X_i - \mu$ beschränkt sind, so sind die Varianzen der Zufallsvariablen $(X_i - \mu)^2$ beschränkt. Also gilt nach dem Gesetz der großen Zahlen wegen $E_{\vartheta}((X_i - \mu)^2) = \sigma^2$ für alle $\varepsilon' > 0$

$$P_{\vartheta} \left(\left| \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 - \sigma^2 \right| > \varepsilon' \right) \rightarrow 0.$$

Ferner gilt $P_{\vartheta}(|\bar{X}(n) - \mu| > \varepsilon') \rightarrow 0$ für alle $\varepsilon' > 0$. Da $\varepsilon' > 0$ beliebig war, lässt sich nun ohne große Mühe auch in diesem Fall (4.12) zeigen. $s^2(n)$ ist also konsistenter Schätzer für σ^2 .

(Nach Satz 11.9 braucht man im ersten dieser Beispiele nur die Existenz des Erwartungswerts und im zweiten nur die der Varianz vorauszusetzen. Die X_i werden als identisch verteilt angenommen.)

Auf das Fischfangbeispiel und das Taxibeispiel können wir den Begriff der Konsistenz nicht unmodifiziert anwenden, da dort bei festem Parameterwert die Zahl n der Beobachtungen beschränkt war.

4.7 Konfidenzintervalle

Schätzt man eine Erfolgswahrscheinlichkeit p aus einer Bernoulli-verteilten Folge $X = (X_1, \dots, X_n)$ durch $\hat{p}(X) = (X_1 + \dots + X_n)/n$, und ist n hinreichend groß, so ist mit großer Wahrscheinlichkeit $|\hat{p}(X) - p|$ klein. Dies legt es nahe zu versuchen, bei Beobachtung von x ein kleines Intervall $C(x)$ um den Schätzwert $\hat{p}(x)$ herum festzulegen, in dem man p vermuten darf. Man könnte etwa fordern, dass — wenn p der unbekannte wahre Parameter ist — das Ereignis, dass $C(x)$ p enthält, mindestens eine Wahrscheinlichkeit von z.B. 0,95 haben soll.

Im Prinzip könnten wir zur Lösung dieser Aufgabe die tschebyschewsche Ungleichung heranziehen. Es ist

$$E_p(\hat{p}(X)) = p \quad \text{und} \quad \text{Var}(\hat{p}(X)) = \frac{p(1-p)}{n} \leq \frac{1}{4n}.$$

Für jedes $\delta > 0$ ist also $P_p(|\hat{p}(X) - p| \geq \delta) \leq 1/(4n\delta^2)$. Wir erreichen, dass

$$P_p(|\hat{p}(X) - p| \geq \delta) \leq 0,05$$

gilt, indem wir $1/(4n\delta^2) = 0,05$, also $\delta = \sqrt{5/n}$ setzen. Setzt man also $C(x) = [\hat{p}(x) - \delta, \hat{p}(x) + \delta]$, so gilt — wie gewünscht —

$$P_p(p \in C(X)) \geq 0,95.$$

Beobachtet man etwa eine Folge x der Länge $n = 50$ mit $k = 20$ Erfolgen, so ergibt sich $\hat{p}(x) = 0,4$ als Schätzung der Erfolgswahrscheinlichkeit, $\delta = 1/\sqrt{10} \approx 0,32$ und das „Konfidenzintervall“ $C(x) = [0,08, 0,72]$.

Das so konstruierte Intervall ist unnötig lang, da die Tschebyschew-Ungleichung überhaupt nicht die genaue Form der vorliegenden Verteilung ausnutzt. Wir wollen daher jetzt ein besseres allgemeines Verfahren zur Konstruktion von Konfidenzintervallen beschreiben. Im eben besprochenen Beispiel liefert es das Intervall $[0,26, 0,54]$. Um noch kürzere Intervalle zu erhalten, muss man n erhöhen.

Sei also ein Schätzproblem mit Stichprobenraum $\mathcal{X}$, Parameterbereich Θ und der Familie $\{P_\vartheta : \vartheta \in \Theta\}$ gegeben, und sei $g(\vartheta)$ zu schätzen. Ist $\{C(x) : x \in \mathcal{X}\}$ eine Familie von Teilmengen $C(x)$ des Wertebereichs $g(\Theta)$ von g und gilt für ein gegebenes $\alpha > 0$ (etwa $\alpha = 0,05$)

$$P_\vartheta(\{x \in \mathcal{X} : g(\vartheta) \in C(x)\}) \geq 1 - \alpha \quad \text{für alle } \vartheta \in \Theta, \quad (4.13)$$

so heißt $C(x)$ **Konfidenzbereich** für $g(\vartheta)$ zum **Konfidenzniveau** $1 - \alpha$ (zur Beobachtung x). Oft ist $g(\Theta)$ ein Intervall und $C(x)$ ein Teilintervall. Dann nennt man $C(x)$ auch **Konfidenzintervall** zum Niveau $1 - \alpha$. Ist $C(x)$ ein Intervall der Form $(-\infty, c(x)]$, so heißt $c(x)$ eine **obere Konfidenzschranke** zum Niveau α .

Es ist wichtig, sich diese Definition genau anzusehen, damit die Angabe von $C(x)$ nicht falsch interpretiert wird:

Nicht ϑ ist zufällig, sondern x und damit $C(x)$. Wird also etwa ein Konfidenzintervall wie $[0,26, 0,54]$ für eine Erfolgswahrscheinlichkeit p angegeben, so ist nicht etwa die Wahrscheinlichkeit $\geq 0,95$, dass es p enthält. Für ein festes Intervall I gilt entweder $p \in I$ oder nicht, aber $\{p \in [0, 1] : p \in I\}$ ist kein „Ereignis“, dem wir eine Wahrscheinlichkeit zugeordnet haben. Die Aussage über das Niveau $1 - \alpha$ ist vielmehr eine Aussage über die gesamte Familie $\{C(x) : x \in \mathcal{X}\}$, d.h. über die Vorschrift, nach der das Intervall aus der Beobachtung bestimmt wird. Wenn wir für jedes x das Intervall $C(x)$ als Konfidenzintervall angeben, wird — was auch immer $\vartheta \in \Theta$ ist — das zufällige Intervall $C(x)$ in ca. 95 % der Fälle $g(\vartheta)$ enthalten. Das bedeutet aber nicht, dass in ca. 95 % der Fälle, in denen die Beobachtung zum Intervall $[0,26, 0,54]$ führt, $g(\vartheta) \in [0,26, 0,54]$ gilt.

Wie kann man nun eine Familie $\{C(x) : x \in \mathcal{X}\}$ mit den gewünschten Eigenschaften konstruieren?

Manchmal kann man ganz naiv vorgehen. Man bestimmt eine Schätzung $\hat{g}$ von $g(\vartheta)$ und versucht mit Hilfe der Verteilungen P_ϑ Zahlen $\delta(x)$ so zu bestimmen, dass für alle ϑ die Ungleichungen

$$P_\vartheta(\{x : |g(\vartheta) - \hat{g}(x)| \leq \delta(x)\}) \geq 1 - \alpha$$

gelten. Dann ist $C(x) = [\hat{g}(x) - \delta(x), \hat{g}(x) + \delta(x)]$. Bei erwartungstreuem $\hat{g}$ führt gelegent-

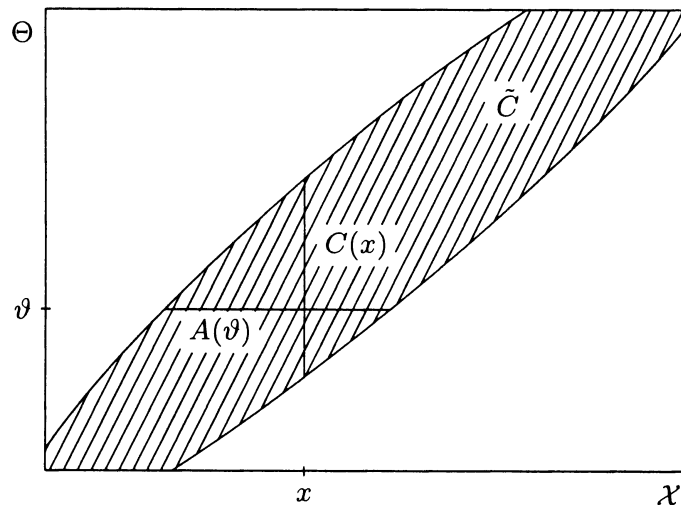


Bild 4.1 Konfidenzbereiche $C(x)$

lich ein Ansatz $\delta(x) = \beta \hat{s}(x)$ zum Erfolg, wenn $\hat{s}(x)$ ein Schätzer der Standardabweichung von $\hat{g}$ ist. Beispiele für solche relativ einfachen Konstruktionen werden wir in § 13 kennen lernen.

Im allgemeinen Fall müssen wir leider etwas mehr arbeiten. Wir nehmen jetzt an, dass $g(\vartheta) = \vartheta$ zu schätzen sei. Die Angabe der Familie $\{C(x) : x \in \mathcal{X}\}$ ist gleich bedeutend mit der Angabe von

$$\tilde{C} := \{(x, \vartheta) \in \mathcal{X} \times \Theta : \vartheta \in C(x)\}$$

und daher auch mit der Angabe aller Mengen

$$A(\vartheta) := \{x \in \mathcal{X} : (x, \vartheta) \in \tilde{C}\} = \{x : \vartheta \in C(x)\}.$$

$A(\vartheta)$ enthält die Stichproben x , in deren Konfidenzbereich ϑ enthalten ist. Zeichnen wir zur Veranschaulichung Θ und $\mathcal{X}$ als Intervalle, so kann sich Bild 4.1 ergeben. Die Mengen $C(x)$ sind die Schnitte durch $\tilde{C}$, die wir durch Festhalten der x -Koordinate erhalten, und die $A(\vartheta)$ sind die Schnitte durch $\tilde{C}$, die wir durch Festhalten der ϑ -Koordinate erhalten.

Offenbar sind die drei Aussagen $\vartheta \in C(x)$, $(x, \vartheta) \in \tilde{C}$ und $x \in A(\vartheta)$ äquivalent. Die Bedingung (4.13) lässt sich also auch in der Form

$$P_{\vartheta}(A(\vartheta)) \geq 1 - \alpha \quad \text{für alle } \vartheta \in \Theta \quad (4.14)$$

schreiben. Wir brauchen also nur für jedes ϑ eine Menge $A(\vartheta) \subset \mathcal{X}$ mit (4.14) anzugeben.

Nun wollen wir natürlich, dass für festes x der Konfidenzbereich $C(x)$ möglichst klein ist. Ist z.B. $C(x)$ ein Intervall, so ist die Aussage $\vartheta \in C(x)$ umso präziser, je kleiner $C(x)$ ist. (Im Extremfall $C(x) = \Theta$ besagt $\vartheta \in C(x)$ gar nichts mehr, was wir nicht ohnehin wissen.)

Insgesamt wird $\tilde{C}$ klein sein, wenn die Mengen $A(\vartheta)$ möglichst wenige Punkte enthalten. Damit wir trotzdem (4.14) erfüllen können, ist es nahe liegend, die Mengen $A(\vartheta)$ so zu wählen, dass sie die Punkte x enthalten, für welche $P_{\vartheta}(x)$ besonders groß ist.

Ist nicht nur $\mathcal{X}$ sondern auch Θ diskret, so führt dieser Ansatz zu dem folgenden Verfahren: Für jedes ϑ betrachtet man alle $P_{\vartheta}(x)$. Ist $x_1, x_2, \dots$ eine (von ϑ abhängige) Nummerierung der Punkte von $\mathcal{X}$ mit

$$P_{\vartheta}(x_1) \geq P_{\vartheta}(x_2) \geq P_{\vartheta}(x_3) \geq \dots \quad (4.15)$$

so wählt man $A(\vartheta) = \{x_1, \dots, x_k\}$ mit dem minimalen k , für das das so bestimmte $A(\vartheta)$ die Ungleichung $P_{\vartheta}(A(\vartheta)) \geq 1 - \alpha$ erfüllt. (Eine Nummerierung mit der Eigenschaft (4.15) existiert bei festem ϑ immer, weil für jedes n nur endlich viele x (höchstens n) die Eigenschaft $P_{\vartheta}(x) \geq 1/n$ haben.)

Wir wollen das Verfahren an einem Beispiel illustrieren.

Beispiel 4.4 (*Bestimmung von Konfidenzintervallen für die hypergeometrische Verteilung*) In einer Urne liegen W weiße und $S = 5$ schwarze Kugeln. Es werden $n = 5$ Kugeln ohne Rücklegen gezogen. x sei die Zahl der gezogenen weißen Kugeln. Der unbekannte Parameter ist $\vartheta = W$. Es sei bekannt, dass $W \leq 7$ ist. (Wir wollen uns mit dieser Bedingung zusätzliche Arbeit sparen.)

Die Wahrscheinlichkeiten $P_{\vartheta}(x) = \binom{\vartheta}{x} \binom{5}{5-x} / \binom{\vartheta+5}{5}$ sind in Tabelle 4.1 dargestellt.

Will man eine Familie $\{C(x), x \in \mathcal{X}\}$ von Konfidenzbereichen zum Niveau $1 - \alpha = 0,9$ bestimmen, so schreibt das oben dargestellte Verfahren Folgendes vor: Man betrachtet zuerst bei festem ϑ die zugehörige Zeile. Neben die größte Wahrscheinlichkeit in dieser

$\vartheta \backslash x$	0	1	2	3	4	5	$A(\vartheta)$
0	1.	0	0	0	0	0	$\{0\}$
1	0,167.	0,833.	0	0	0	0	$\{0, 1\}$
2	0,047	0,476.	0,476.	0	0	0	$\{1, 2\}$
3	0,018	0,268.	0,536.	0,179.	0	0	$\{1, 2, 3\}$
4	0,008	0,159.	0,476.	0,317.	0,040	0	$\{1, 2, 3\}$
5	0,004	0,099.	0,397.	0,397.	0,099.	0,004	$\{1, 2, 3, 4\}$
6	0,002	0,065	0,325.	0,433.	0,162.	0,013	$\{2, 3, 4\}$
7	0,001	0,044	0,265.	0,442.	0,221.	0,026	$\{2, 3, 4\}$
$C(x)$	$\{0, 1\}$	$\{1, 2, \dots, 5\}$	$\{2, 3, \dots, 7\}$	$\{3, 4, \dots, 7\}$	$\{5, 6, 7\}$	$\emptyset$	

Tabelle 4.1

Zeile macht man einen Punkt, dann neben die zweitgrößte, falls die größte noch nicht $\geq 0,9$ ist, usw. So macht man der Größe der Wahrscheinlichkeiten nach in dieser Zeile weiter bis die „punktierten“ Wahrscheinlichkeiten eine Summe $\geq 0,9$ haben. $A(\vartheta)$ ist die Menge der x , für die neben $P_\vartheta(x)$ ein Punkt steht. $C(x)$ ist die Menge der ϑ , für die in Spalte x ein Punkt neben $P_\vartheta(x)$ steht.

Es mag überraschen, dass sich in diesem Beispiel $C(5) = \emptyset$ ergibt. Das obige Verfahren kann dazu führen, dass für gewisse x der Konfidenzbereich $C(x)$ leer ist. In der Definition von Konfidenzbereichen war auch nicht gefordert, dass die Mengen $C(x)$ nicht leer sind. Dennoch stört dies bei der Interpretation. Man sollte deshalb die leeren Mengen $C(x)$ besser durch Mengen $\{\hat{\vartheta}(x)\}$ ersetzen, in denen $\hat{\vartheta}(x)$ ein geeignet erscheinender Schätzwert ist. Das Niveau wird dabei höchstens erhöht.

Allgemeine Anmerkungen über Konfidenzbereiche: Bei der Wahl des Konfidenzniveaus $1 - \alpha$ ist zu bedenken, dass die Mengen $C(x)$ umso größer werden, je kleiner man α wählt. Zwischen den Zielen, ein hohes Niveau einzuhalten und kleine $C(x)$ zu finden, muss man einen Kompromiss finden.

Die oben beschriebene Konstruktion liefert Konfidenzbereiche $\{C(x), x \in \mathcal{X}\}$ für ϑ . Setzt man nun $C_g(x) = g(C(x))$, so ist $\{C_g(x), x \in \mathcal{X}\}$ ein System von Konfidenzbereichen zu $g(\vartheta)$, das ebenfalls das Niveau $1 - \alpha$ einhält.

Noch ein anderes Verfahren, aus bekannten Konfidenzbereichen andere zu erzeugen, ist erwähnenswert. Ist für $i = 1, \dots, k$ $\{C_i(x), x \in \mathcal{X}\}$ ein System von Konfidenzbereichen zum Niveau $1 - \alpha_i$ für $g_i(\vartheta)$ und $C(x) = C_1(x) \times \dots \times C_k(x)$, so ist $\{C(x), x \in \mathcal{X}\}$ ein System von Konfidenzbereichen für den Vektor $g(\vartheta) = (g_1(\vartheta), \dots, g_k(\vartheta))$ zum Niveau $1 - (\alpha_1 + \dots + \alpha_k)$, denn es gilt

$$\begin{aligned}
 P_\vartheta(\{x : g(\vartheta) \notin C(x)\}) &= P_\vartheta\left(\bigcup_{i=1}^k \{x : g_i(\vartheta) \notin C_i(x)\}\right) \\
 &\leq \sum_{i=1}^k P_\vartheta(\{x : g_i(\vartheta) \notin C_i(x)\}) \leq \sum_{i=1}^k \alpha_i.
 \end{aligned}$$

Die geschilderten Methoden zur Konstruktion von Konfidenzbereichen erfüllen zwar die geforderte Niveau-Bedingung, aber wir können noch nicht sagen, ob nicht in geeignetem Sinne „bessere“ Methoden existieren. Die Frage nach der Optimalität würde jedoch an dieser Stelle zu weit führen.

Konfidenzintervalle für die Erfolgswahrscheinlichkeit*: Für das Problem, Konfidenzintervalle für eine Erfolgswahrscheinlichkeit p zu bestimmen, ist es zweckmäßig, die oben geschilderte Methode der Wahl der $A(\vartheta)$ zu modifizieren, da die praktische Durchführung sonst sehr schwierig ist. Als Vorbereitung dient

Lemma 4.5 *Sei X (auf (Ω, P_p)) $b_{n,p}$ -verteilt und $x < n$. Dann ist $P_p(X \leq x)$ als Funktion von p stetig und strikt monoton fallend, und es gilt $P_0(X \leq x) = 1$, $P_1(X \leq x) = 0$.*

Beweis: Nicht trivial ist nur die Monotonieaussage. Um für $p_1 < p_2$ die Ungleichung

$$P_{p_1}(X \leq x) > P_{p_2}(X \leq x)$$

zu beweisen, wählen wir p_3 mit $p_1 = p_2 p_3$ und auf einem beliebigen geeigneten Wahrscheinlichkeitsraum (Ω, P) zwei voneinander unabhängige Bernoulli-Folgen $X_{i,1}, \dots, X_{i,n}$ ($i = 2, 3$) mit $P(X_{i,j} = 1) = p_i$.

Sei $X_{1,j} = X_{2,j} X_{3,j}$. Durch die Wahl von p_3 ist $X_{1,1}, \dots, X_{1,n}$ eine Bernoulli-Folge mit Erfolgswahrscheinlichkeit p_1 . Aus

$$\{X_{2,1} + \dots + X_{2,n} \leq x\} \subset \{X_{1,1} + \dots + X_{1,n} \leq x\}$$

folgt

$$P_{p_1}(X \leq x) = P(X_{1,1} + \dots + X_{1,n} \leq x) \geq P(X_{2,1} + \dots + X_{2,n} \leq x) = P_{p_2}(X \leq x).$$

Die Ungleichung ist sogar strikt, da für $x < n$ das Ereignis

$$\{X_{2,1} + \dots + X_{2,n} = n, X_{3,1} + \dots + X_{3,n} \leq x\}$$

positive Wahrscheinlichkeit hat. □

Sei nun $0 < \alpha < 1$, und sei $a = a(p)$ maximal mit $P_p(X < a) < \alpha/2$. Sei ferner $b = b(p)$ minimal mit $P_p(X > b) < \alpha/2$. Setzt man

$$A(p) = \{a(p), a(p) + 1, \dots, b(p)\},$$

so gilt $P_p(A(p)) > 1 - \alpha$, so dass die Bedingung (4.14) erfüllt ist. Wir wollen uns nun überlegen, wie die zugehörigen Konfidenzintervalle aussehen.

Nach Lemma 4.5 ist $a(p)$ monoton wachsend. Sei $p_o(x)$ für $x < n$ der nach Lemma 4.5 eindeutig bestimmte Wert von p , welcher $P_p(X \leq x) = \alpha/2$ erfüllt. Sei ferner $p_o(n) = 1$. Dann ist $p \leq p_o(x)$ äquivalent zu $P_p(X \leq x) \geq \alpha/2$.

Ist $x \geq a(p)$, so ist $P_p(X \leq x) \geq P_p(X < a(p) + 1) \geq \alpha/2$. Ist umgekehrt $P_p(X \leq x) \geq \alpha/2$, so ist $x \geq a(p)$. Daher ist $x \geq a(p)$ äquivalent zu $p \leq p_o(x)$.

Sei $p_u(x)$ für $x > 0$ der Wert von p mit $P_p(X \geq x) = \alpha/2$, und sei $p_u(0) = 0$. Die analoge Überlegung zeigt, dass $x \leq b(p)$ äquivalent zu $p \geq p_u(x)$ ist. Damit ist $x \in A(p)$ äquivalent zu $p \in [p_u(x), p_o(x)]$. Das bedeutet, dass die durch die $A(p)$ bestimmten Mengen $C(x)$ gerade die Form

$$C(x) = [p_u(x), p_o(x)]$$

haben. Diese Mengen bilden also eine Familie von Konfidenzintervallen zum Niveau $1 - \alpha$ für die Erfolgswahrscheinlichkeit p .

Hinweise zur Berechnung der Konfidenzintervalle*: Bei der Bestimmung von p_o und p_u kann man eine Beziehung zur (erst in § 14 diskutierten) F -Verteilung ausnutzen, für die es am Ende des Buches Tabellen gibt. Ist $F_{\mu, \nu}(c)$ der Wert der Verteilungsfunktion der F -Verteilung mit μ und ν Freiheitsgraden in c , und setzt man $\mu = 2(x + 1)$, $\nu = 2(n - x)$ und $c = (n - x)p / ((1 - p)(x + 1))$, so gilt für eine $b_{n, p}$ -verteilte Zufallsvariable X

$$P_p(X \leq x) = 1 - F_{\mu, \nu}(c). \quad (4.16)$$

(Vgl. z.B. K. Stange (1970), S. 433 oder A. Hald (1952).) Die Bestimmungsgleichung für p_o ist $P_{p_o}(X \leq x) = \alpha/2$. Man sucht also den Wert c_0 mit $F_{\mu, \nu}(c_0) = 1 - \alpha/2$ in der Tabelle auf und löst die Gleichung $c_0 = (n - x)p_o / ((1 - p_o)(x + 1))$ nach p_o auf.

Die Bestimmungsgleichung für p_u ist $P_{p_u}(X \geq x) = \alpha/2$. Setze $X' = n - X$, $x' = n - x$, $\mu' = 2(x' + 1) = 2(n - x + 1)$, $\nu' = 2(n - x') = 2x$. Ist X $b_{n, p}$ -verteilt so ist X' $b_{n, 1-p}$ -verteilt. Man sucht daher den Wert c_u mit $F_{\mu', \nu'}(c_u) = 1 - \alpha/2$ in der Tabelle auf und bestimmt p_u aus der Gleichung $c_u = x(1 - p_u) / (p_u(n - x + 1))$.

Blyth und Still (1983) haben Konfidenzintervalle tabelliert, die sich von den hier angegebenen unterscheiden und die nach verschiedenen dort diskutierten Kriterien Optimalitätseigenschaften aufweisen. Diese gelten für $n \leq 30$ und $\alpha = 0,05$ bzw. $\alpha = 0,01$.

Für große n stützt man sich auf Approximationen der Binomialverteilung. Diesen wenden wir uns im nächsten Paragraphen zu.

Aufgaben

1. Bei einer Qualitätskontrolle wurden 100 Blitzlichtlämpchen geprüft. In dieser Stichprobe waren 4 Exemplare defekt. Man bestimme die Maximum-Likelihood-Schätzung für die Anzahl der defekten Lämpchen unter den produzierten 10 000 Lämpchen.
2. In einem Bernoulli-Experiment treten vor dem r -ten Erfolg genau x Misserfolge auf. Man berechne hieraus den Maximum-Likelihood-Schätzer für die Erfolgswahrscheinlichkeit. Ist er erwartungstreu? (L)
3. Man vergleiche die Varianz der Schätzung aus Aufgabe 2 mit der unteren Schranke, die sich aus der Informationsungleichung ergibt.
4. In einer Urne sind $N = 10$ Kugeln, darunter eine unbekannte Zahl S schwarzer Kugeln. $n = 6$ Kugeln werden ohne Rücklegen gezogen. Sei x die Zahl der schwarzen Kugeln unter diesen sechs. Man gebe Konfidenzintervalle $C(x)$ zum Niveau 0,9 an.
5. Man betrachte das Taxiproblem in Beispiel 4.3, wenn n Taxis beobachtet werden und diese nicht notwendig verschieden sein müssen (Ziehen mit Rücklegen).
6. $X_1, X_2, \dots$ seien unabhängig und $b_{m, p}$ -verteilt mit unbekanntem m und unbekanntem p (mit $0 < p < 1$). Man gebe zu jedem n einen Schätzer $\hat{g}(X_1, \dots, X_n)$ von m so an, dass diese Folge von Schätzern konsistent ist. (L)
7. Seien $a < b$ natürliche Zahlen. $X_1, \dots, X_n$ seien unabhängige, auf $\{a, a + 1, \dots, b\}$ gleichverteilte Zufallsvariable. Bei unbekanntem a und b bieten sich

$$\bar{X} = \frac{1}{n} (X_1 + \dots + X_n) \quad \text{und} \quad Y = \frac{1}{2} (\max X_i + \min X_i)$$

als Schätzer für $c = (a + b)/2$ an. Man untersuche, ob diese Schätzer erwartungstreu sind und vergleiche $\text{Var}(\bar{X})$ und $\text{Var}(Y)$.

8. Ein Gerät enthält die stör anfälligen Teile A_1 , A_2 und B . Es funktioniert, wenn B und mindestens eins der Teile A_1 , A_2 funktioniert. Aus Untersuchungen einzelner Bauteile sind Konfidenzintervalle für die Wahrscheinlichkeiten p_A und p_B bekannt, dass ein Teil vom Typ A bzw. B nach einem Jahr noch funktioniert. Unter der Annahme, dass Störungen in den Bauteilen unabhängig voneinander auftreten, leite man — notfalls zu einem kleineren Niveau — ein möglichst kleines Konfidenzintervall für die Wahrscheinlichkeit her, dass das Gerät nach einem Jahr noch funktioniert.

9. Sei X die Anzahl der Unfälle in einer bestimmten Stadt in einer Woche. Wir betrachten X als Poisson-verteilt mit Parameter $\lambda > 0$. Wir wollen aus der Beobachtung von X die Wahrscheinlichkeit schätzen, dass in den folgenden drei Wochen kein Unfall geschieht, also $g(\lambda) = P(X = 0)^3$. Zeigen Sie: Ist T erwartungstreuer Schätzer, so liefert T unsinnige Schätzwerte. (L)
10. Seien $X_1, X_2, \dots, X_n$ unabhängig gleichverteilt in $\{1, 2, \dots, b\}$ und $X = \max(X_1, \dots, X_n)$. Geben Sie ein Konfidenzintervall für b zum Niveau $1 - \alpha$ auf Grund der Beobachtung von X an. (L)

§ 5 Approximationen der Binomialverteilung

Für großes n ist die exakte Berechnung der Wahrscheinlichkeit

$$b_{n,p}(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad (5.1)$$

in n unabhängigen Versuchen mit Erfolgswahrscheinlichkeit p genau k Erfolge zu haben, mühsam. Wie wahrscheinlich ist es, bei $n = 80$ Würfeln einer Münze $k = 40$ mal Kopf zu erhalten? Am Ergebnis $\binom{80}{40} 2^{-80}$ lässt sich nicht einmal die Größenordnung so ohne weiteres erkennen. Noch unübersichtlicher ist die Berechnung von Summen solcher Wahrscheinlichkeiten, also etwa der Wahrscheinlichkeit zwischen 40 und 50 mal Kopf zu erhalten. Wir wollen uns daher nun mit Approximationen für solche Wahrscheinlichkeiten beschäftigen.

5.1 Approximation von $n!$ und $b_{n,p}(k)$

Wir brauchen vor allem Approximationen für die in $\binom{n}{k} = n!/(k!(n-k)!)$ mehrfach auftretenden Fakultäten. Wir nennen zwei Folgen (a_n) und (b_n) **asymptotisch gleich** (oder **asymptotisch äquivalent**) für $n \rightarrow \infty$ und schreiben dafür

$$a_n \sim b_n, \quad \text{wenn} \quad \lim_{n \rightarrow \infty} \frac{a_n}{b_n} = 1$$

ist. Unser wichtigstes Hilfsmittel ist:

Satz 5.1 (Stirlingsche Formel) *Ist*

$$\eta_n := \sqrt{2\pi n} (n/e)^n,$$

so gilt

$$n! \sim \eta_n.$$

Der Ausdruck für η_n sieht selbst etwas kompliziert aus, aber $n!$ besteht aus n *verschiedenen* Faktoren und ist daher schwerer zu berechnen, wenn n groß ist. In der Approximation $\eta_n/(\eta_k \eta_{n-k})$ von $\binom{n}{k}$ ergibt sich noch eine Vereinfachung, weil e^{-n} im Zähler gegen $e^{-k} \cdot e^{-(n-k)}$ im Nenner gekürzt werden kann.

Betrachten wir ein Beispiel. Die Wahrscheinlichkeit bei $2n$ Würfeln einer Münze genau n -mal Kopf zu erhalten, ist $\binom{2n}{n} 2^{-2n}$. Als Approximation ergibt sich

$$\frac{\eta_{2n}}{\eta_n^2} 2^{-2n} = \frac{(2n)^{2n+1/2}}{\sqrt{2\pi} (n^{n+1/2})^2 2^{2n}} = \frac{1}{\sqrt{\pi n}}. \quad (5.2)$$

Die Approximation von $n!$ durch η_n ist schon für kleine Werte von n recht genau, aber es gibt noch schärfere Abschätzungen, die wir zwar nicht brauchen, aber doch erwähnen wollen. Definiert man $\rho(n)$ durch

$$n! = \eta_n \cdot e^{\rho(n)},$$

so besagt Satz 5.1, dass $\rho(n) \rightarrow 0$ für $n \rightarrow \infty$ gilt. Wir zeigen im Anhang zu § 5

$$\frac{1}{12n+1} < \rho(n) < \frac{1}{12n}. \quad (5.3)$$

Also ist $\eta_n \cdot \exp(1/(12n))$ eine noch bessere Abschätzung von $n!$. Die folgenden Werte zeigen, wie gut die Approximationen schon für kleine n sind:

n	$n!$	η_n	$\eta_n \cdot \exp(1/(12n))$
2	2	1,919	2,0007
5	120	118,019	120,0026

Der relative Fehler $(n! - \eta_n)/n!$ strebt sehr schnell gegen 0. Satz 5.1 steht in einigen Lehrbüchern der Differential- und Integralrechnung (z.B. Courant (1955), S. 317), könnte also vorausgesetzt werden. Will man das nicht, so ist es zweckmäßig, in dem im Anhang zu diesem Paragraphen gegebenen Beweis der stirlingschen Formel zuerst nur Schritt 1 zu lesen, in dem die folgende Aussage bewiesen wird:

$$\text{Es existiert ein } \tau > 0 \text{ mit } n! \sim \tau e^{-n} n^{n+1/2}. \quad (5.4)$$

Zum Beweis von Satz 5.1 fehlt dann noch der Nachweis, dass $\tau = \sqrt{2\pi}$ ist. Dies wird im Anhang in Schritt 2 gezeigt, der allerdings die folgenden Überlegungen bis zum Satz 5.4 voraussetzt. Wir benutzen jetzt nur (5.4).

Sei $0 < p < 1$ und $q = 1 - p$. Es liegt nahe, dass vor allem solche Werte k von Interesse sind, für die k/n ungefähr p ist. Wir betrachten daher Folgen (k_n) mit $k_n/n \rightarrow p$, schreiben aber zur Abkürzung k statt k_n . Offenbar gilt $k \rightarrow \infty$, $n - k \rightarrow \infty$. Aus (5.1) und (5.4) folgt daher

$$b_{n,k}(k) \sim \frac{1}{\tau} \sqrt{\frac{n}{k(n-k)}} \left(\frac{np}{k}\right)^k \left(\frac{nq}{n-k}\right)^{n-k}. \quad (5.5)$$

Aus $k \sim np$ und $n - k \sim nq$ ergibt sich

$$\sqrt{\frac{n}{k(n-k)}} \sim \frac{1}{\sigma_n},$$

wobei $\sigma_n = \sqrt{npq}$ die Standardabweichung der $b_{n,p}$ -Verteilung ist. Es genügt also nun, das Grenzverhalten von

$$\chi(n, k) = \left(\frac{np}{k}\right)^k \left(\frac{nq}{n-k}\right)^{n-k}$$

zu studieren. Dazu gehen wir zum (natürlichen) Logarithmus über. Sei $t = k/n$. t ist Abkürzung für $t_n = k_n/n$. Es gilt $t \rightarrow p$. Wir erhalten

$$-\log \chi(n, k) = n \left(t \log \frac{t}{p} + (1-t) \log \frac{1-t}{q} \right).$$

Die Funktion $g(t) = (\dots)$ in der Klammer hat an der Stelle $t = p$ den Wert $g(p) = 0$ und die Ableitungen $g'(p) = 0$, $g''(p) = 1/p + 1/q = 1/(pq)$ (Übung). Nach der Taylorformel ist daher

$$g(t) = \frac{1}{2pq} (t-p)^2 + \psi(t-p),$$

wobei in einer Umgebung von $t = p$ die Abschätzung $|\psi(t - p)| \leq c|t - p|^3$ mit einer geeigneten Konstanten $c > 0$ gilt.

Nehmen wir nun nicht nur $t \rightarrow p$ an, sondern sogar $n(t - p)^3 \rightarrow 0$, so folgt $n\psi(t - p) \rightarrow 0$ und also

$$\left| -\log \chi(n, k) - \frac{n(t - p)^2}{2pq} \right| \rightarrow 0.$$

Setzt man

$$x(n, k) = \frac{k - np}{\sigma_n}, \quad (5.6)$$

so ist $n(t - p)^2/(2pq) = x(n, k)^2/2$. Wir erhalten dann also $\chi(n, k)/\exp(-x(n, k)^2/2) \rightarrow 1$. Die Bedingung $n(t - p)^3 \rightarrow 0$ ist äquivalent zu der Bedingung

$$\frac{x(n, k)^3}{\sqrt{n}} \rightarrow 0. \quad (5.7)$$

Fassen wir alles zusammen, so haben wir gezeigt, dass aus (5.7)

$$b_{n,p}(k) \sim \frac{1}{\tau\sigma_n} \exp\left(-\frac{x(n, k)^2}{2}\right) \quad (5.8)$$

folgt. Sind (α_n) und (β_n) zwei Folgen mit

$$\frac{x(n, \alpha_n)^3}{\sqrt{n}} \rightarrow 0 \quad \text{and} \quad \frac{x(n, \beta_n)^3}{\sqrt{n}} \rightarrow 0, \quad (5.9)$$

so gilt, wie man an den benutzten Abschätzungen sieht, die Konvergenz sogar gleichmäßig für alle Folgen (k_n) mit $\alpha_n \leq k_n \leq \beta_n$. Wir führen nun noch als Abkürzung die Funktion

$$\varphi(x) := \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) \quad (5.10)$$

ein. φ heißt **Dichte der Standard-Normalverteilung**. Mit der Einschränkung, dass der Beweis von $\tau = \sqrt{2\pi}$ noch nachgetragen werden muss, wenn die stirlingsche Formel nicht als bekannt vorausgesetzt werden soll, haben wir den folgenden Satz bewiesen:

Satz 5.2 (Lokaler Grenzwertsatz für die Binomialverteilung) *Ist $0 < p < 1$ und (k_n) eine Folge mit $x(n, k_n)^3/\sqrt{n} \rightarrow 0$, so gilt*

$$b_{n,p}(k_n) \sim \frac{1}{\sigma_n} \varphi(x(n, k_n)). \quad (5.11)$$

Sind (α_n) und (β_n) Folgen, die der Bedingung (5.9) genügen, so konvergiert der Quotient der linken und der rechten Seite in (5.11) sogar gleichmäßig für alle Folgen (k_n) mit $\alpha_n \leq k_n \leq \beta_n$ gegen 1.

Tabelle 5.1 soll einen Eindruck von der Qualität der Approximation von $b_{n,p}(k)$ durch $\varphi(x(n, k))/\sigma_n$ vermitteln. Die Approximation ist gut, wenn σ_n nicht zu klein ist und $|k - np|/\sigma_n$ nicht zu groß. Bei $p = 0,2$ oder $p = 0,8$ braucht man also größere n als bei $p = 0,5$, und für Werte von k in der Nähe von np (die am wahrscheinlichsten sind) ist die Approximation am besten.

k	$n = 8, p = 0,2$		$n = 8, p = 0,5$		$n = 25, p = 0,2$	
	Approximation	Exakt	Approximation	Exakt	Approximation	Exakt
0	0,130	0,168	0,005	0,004	0,009	0,004
1	0,306	0,336	0,030	0,031	0,027	0,024
2	0,331	0,294	0,104	0,109	0,065	0,071
3	0,164	0,147	0,220	0,219	0,121	0,136
4	0,037	0,046	0,282	0,273	0,176	0,187
5	0,004	0,009	0,220	0,219	0,199	0,196
6	0+	0,001	0,104	0,109	0,176	0,163
7	0+	0+	0,030	0,031	0,121	0,111
8	0+	0+	0,005	0,004	0,065	0,062
9					0,027	0,029
10					0,009	0,012
11					0,002	0,004

Tabelle 5.1 Vergleich der Binomialverteilung mit ihrer Approximation
(nach Meyer (1971), S. 249)

Stellt man die $b_{n,p}$ -Verteilung anschaulich dar, indem über jedem Intervall $[k - 1/2, k + 1/2]$ auf der x -Achse ein Rechteck mit dem Flächeninhalt $b_{n,p}(k)$ gezeichnet wird, so werden für wachsendes n diese Schaubilder (**Histogramme**) immer flacher und die Schwerpunkte np wandern nach ∞ ab.

Man ändert daher zweckmäßig die Skalen und betrachtet $x(n, k)$ statt k . Man zeichnet also über den Intervallen $[x(n, k - 1/2), x(n, k + 1/2)]$ Rechtecke vom Flächeninhalt $b_{n,p}(k)$. Da deren Breite $1/\sigma_n$ ist, muss die Höhe $\sigma_n b_{n,p}(k)$ sein. Für großes n ist nach Satz 5.2

$$\sigma_n b_{n,p}(k) \approx \varphi(x(n, k)) .$$

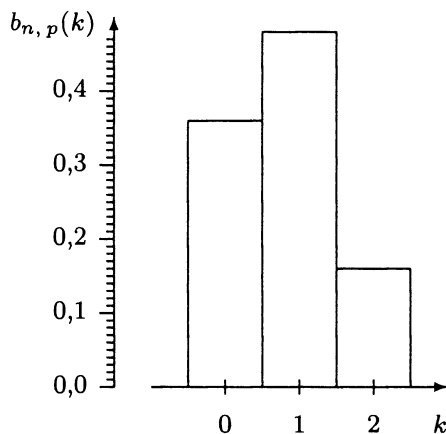


Bild 5.1 Binomialverteilung für $p = 0,4, n = 2$

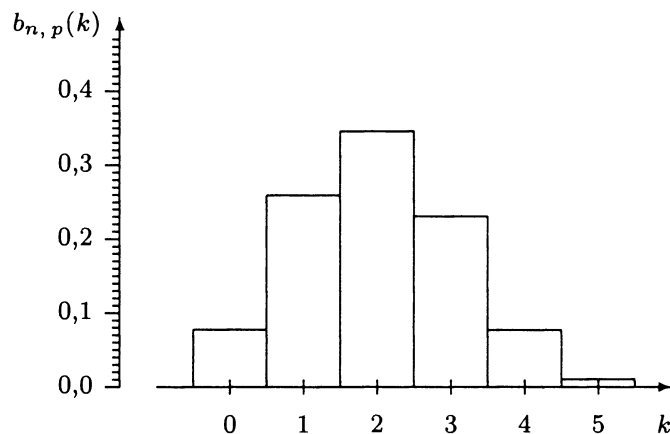


Bild 5.2 Binomialverteilung für $p = 0,4, n = 5$

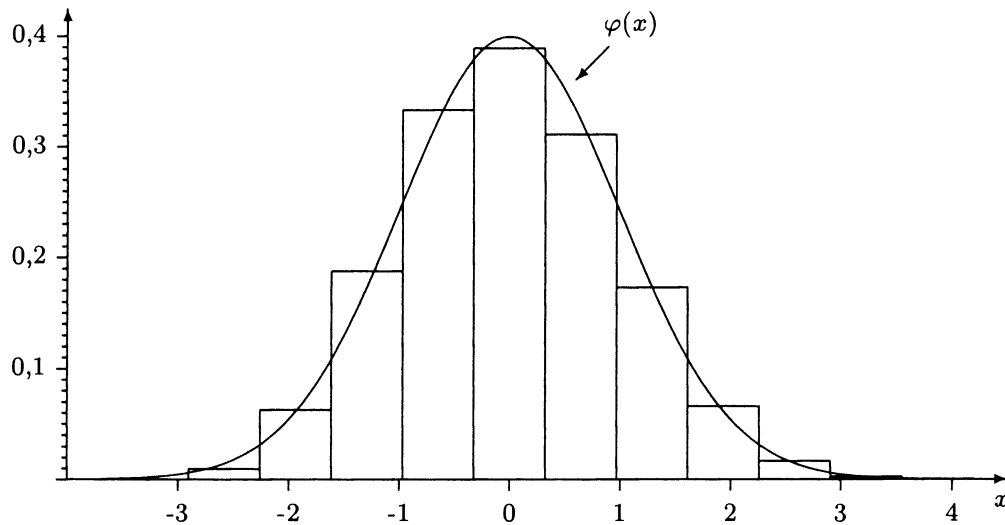


Bild 5.3 Binomialverteilung mit transformierter Skala für $p = 0,4$, $n = 10$ und die Approximation durch $\varphi(x)$

Die Histogramme für die $x(n, k)$ werden daher für $n \rightarrow \infty$ der **gaußschen Glockenkurve** $\varphi(x)$ immer ähnlicher. Bild 5.3 zeigt das Histogramm für die $x(n, k)$ für $n = 10$, $p = 0,4$ und die Approximation durch $\varphi(x)$.

Die Funktion φ wird im Folgenden noch eine große Rolle spielen. Insbesondere brauchen wir

Lemma 5.3 Es ist

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

Beweis: Es ist

$$\left(\int_{-\infty}^{\infty} e^{-t^2/2} dt \right)^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-x^2/2} e^{-y^2/2} dx dy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(x^2+y^2)/2} dx dy.$$

Wir gehen zu Polarkoordinaten über und erhalten wegen $dx dy = r d\vartheta dr$ für das Doppelintegral den Wert

$$\int_0^{2\pi} \int_0^{\infty} r e^{-r^2/2} dr d\vartheta = 2\pi \left[-e^{-r^2/2} \right]_0^{\infty} = 2\pi.$$

□

5.2 Der Satz von de Moivre-Laplace

Wir wollen nun Summen von Wahrscheinlichkeiten $b_{n,p}(k)$ für großes n approximieren. Wir benötigen dazu die durch

$$\Phi(x) = \int_{-\infty}^x \varphi(t) dt$$

definierte **Verteilungsfunktion Φ der Standard-Normalverteilung**. Offenbar ist

$$\int_a^b \varphi(t) dt = \Phi(b) - \Phi(a).$$

Bild 5.4 gibt ein Bild vom Verlauf von Φ , und daneben sind ein paar Werte angegeben, die relativ häufig gebraucht werden.

Ausführlich ist Φ am Ende des Buches tabelliert. Es genügt eine Tabelle, die $\Phi(x)$ für $x \geq 0$ angibt. Wegen der Symmetrie $\varphi(t) = \varphi(-t)$ ist nämlich

$$\Phi(-x) = \int_{-\infty}^{-x} \varphi(t) dt = \int_x^{\infty} \varphi(t) dt = 1 - \Phi(x). \quad (5.12)$$

Für eine $b_{n,p}$ -verteilte Zufallsvariable S_n sei

$$S_n^* = \frac{S_n - np}{\sigma_n} = \frac{S_n - ES_n}{\sqrt{\text{Var}(S_n)}}.$$

Nimmt S_n den Wert k an, so hat S_n^* den Wert $x(n, k)$. (Man überlegt sich leicht, dass S_n^* gerade die Zufallsvariable der Form $(S_n - c)/d$ mit $d > 0$ ist, für die $ES_n^* = 0$ und $\text{Var}(S_n^*) = 1$ ist. S_n^* heißt daher die **standardisierte** oder **normierte** Form von S_n .)

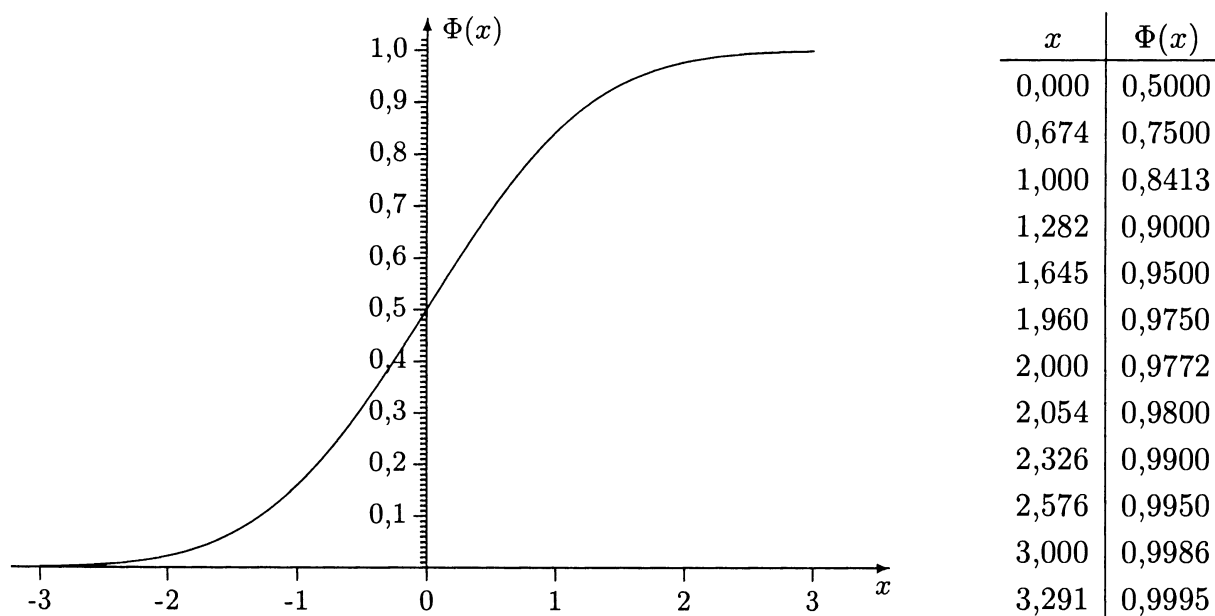


Bild 5.4 Verteilungsfunktion der Standard-Normalverteilung

Satz 5.4 (Satz von de Moivre-Laplace) Sei $0 < p < 1$, und S_n $b_{n,p}$ -verteilt. Dann gilt für alle $a < b$

$$\lim_{n \rightarrow \infty} P(a \leq S_n^* \leq b) = \Phi(b) - \Phi(a). \quad (5.13)$$

Beweis: Offenbar ist $a \leq S_n^* \leq b$ äquivalent zu $a\sigma_n + np \leq S_n \leq b\sigma_n + np$. Sei α_n die kleinste ganze Zahl $\geq a\sigma_n + np$ und β_n die größte ganze Zahl $\leq b\sigma_n + np$. Dann ist

$$\{a \leq S_n^* \leq b\} = \{\alpha_n \leq S_n^* \leq \beta_n\}.$$

Wegen $|x(n, \alpha_n) - a| \leq 1/\sigma_n$ und $|x(n, \beta_n) - b| \leq 1/\sigma_n$ sind die Folgen $(x(n, \alpha_n))$ und $(x(n, \beta_n))$ beschränkt, so dass (5.9) gilt. Aus (5.8) folgt daher die Existenz einer Folge $\varepsilon_n \rightarrow 0$ mit

$$1 - \varepsilon_n < \frac{b_{n,p}(k)}{\sqrt{2\pi} \varphi(x(n, k)) / (\tau \sigma_n)} < 1 + \varepsilon_n$$

für $\alpha_n \leq k \leq \beta_n$. Setzt man

$$R_n = \sum_{k=\alpha_n}^{\beta_n} \frac{1}{\sigma_n} \varphi(x(n, k)),$$

so gilt also

$$(1 - \varepsilon_n) R_n \sqrt{2\pi} / \tau \leq P(a \leq S_n^* \leq b) \leq (1 + \varepsilon_n) R_n \sqrt{2\pi} / \tau. \quad (5.14)$$

Die $x(n, k)$ sind die Mittelpunkte von Intervallen der Länge $1/\sigma_n$, in die das Intervall $[x(n, \alpha_n - 1/2), x(n, \beta_n + 1/2)]$ unterteilt ist. Also ist R_n eine Riemann-Summe, die das Integral

$$\int_{x(n, \alpha_n - 1/2)}^{x(n, \beta_n + 1/2)} \varphi(x) dx = \Phi(x(n, \beta_n + 1/2)) - \Phi(x(n, \alpha_n - 1/2)) \quad (5.15)$$

approximiert. Für $n \rightarrow \infty$ gilt $x(n, \alpha_n - 1/2) \rightarrow a$, $x(n, \beta_n + 1/2) \rightarrow b$ und also $R_n \rightarrow \Phi(b) - \Phi(a)$. Aus (5.14) folgt daher (5.13), wenn wir wieder $\tau = \sqrt{2\pi}$ als bekannt voraussetzen. Andernfalls muss man nun noch im Anhang zu diesem Paragraphen den zweiten Beweisschritt zu Satz 5.1 lesen. $\square$

Bemerkung: Der Ausdruck in (5.15) strebt zwar gegen $\Phi(b) - \Phi(a)$, aber selbst für große n ist er noch eine bessere Approximation für $P(\alpha_n \leq S_n \leq \beta_n)$ als $\Phi(b) - \Phi(a)$.

Praktisch werden die obigen Ergebnisse z.B. folgendermaßen angewandt: Will man für bestimmte $\alpha < \beta$ und nicht zu kleines n die Wahrscheinlichkeit $P(\alpha \leq S_n \leq \beta)$ abschätzen, so rechnet man um:

$$P(\alpha \leq S_n \leq \beta) = P\left(\frac{\alpha - np}{\sigma_n} \leq S_n^* \leq \frac{\beta - np}{\sigma_n}\right),$$

setzt dann $a = (\alpha - np)/\sigma_n$, $b = (\beta - np)/\sigma_n$ und gibt $\Phi(b) - \Phi(a)$ als approximativen Wert der gesuchten Wahrscheinlichkeit an. Das sieht ziemlich verboten aus, weil a und b als Funktionen von n definiert sind. Aber aus Satz 5.4 folgt sogar die folgende formal stärkere Aussage:

Satz 5.5* *Unter den Bedingungen von Satz 5.4 gilt*

$$\lim_{n \rightarrow \infty} \sup_{a < b} |P(a \leq S_n^* \leq b) - (\Phi(b) - \Phi(a))| = 0. \quad (5.16)$$

Beweis: Wir setzen $\Phi(-\infty) = 0$ und $\Phi(+\infty) = 1$. Sei $\varepsilon > 0$ fest vorgegeben. Es existieren endlich viele Punkte

$$-\infty = t_0 < t_1 < \dots < t_{K-1} < t_K = +\infty$$

mit $\Phi(t_{i+1}) - \Phi(t_i) < \varepsilon$. Für hinreichend großes $n \geq N(\varepsilon)$ gilt

$$|P(t_i \leq S_n^* \leq t_j) - (\Phi(t_j) - \Phi(t_i))| < \varepsilon \quad \text{für } 1 \leq i \leq j \leq K-1.$$

Sei $a < b$ beliebig gewählt. Ist i der Index mit $t_i \leq a < t_{i+1}$ und j der Index mit $t_j \leq b < t_{j+1}$, so ist für großes $n \geq N(\varepsilon)$

$$P(a \leq S_n^* \leq b) \geq P(t_{i+1} \leq S_n^* \leq t_j) \geq \Phi(t_j) - \Phi(t_{i+1}) - \varepsilon \geq \Phi(b) - \Phi(a) - 3\varepsilon.$$

Sei nun $a' = a$ falls $a \geq t_1$ ist, und $= t_1$ sonst. Sei $b' = b$ falls $b \leq t_{K-2}$ ist, und $= t_{K-2}$ sonst. Dann gilt für $n \geq N(\varepsilon)$, wenn i und j nun mit a' und b' definiert werden,

$$\begin{aligned} P(a \leq S_n^* \leq b) &\leq P(a' \leq S_n^* \leq b') + 1 - P(t_1 \leq S_n^* \leq t_{K-2}) \\ &\leq P(t_i \leq S_n^* \leq t_{j+1}) + 1 - P(t_1 \leq S_n^* \leq t_{K-2}) \\ &\leq \Phi(t_{j+1}) - \Phi(t_i) + \varepsilon + 1 - \Phi(t_{K-2}) + \Phi(t_1) + \varepsilon \\ &\leq \Phi(b) + \varepsilon - \Phi(a) + \varepsilon + 1 - (1 - 2\varepsilon) + 2\varepsilon. \quad \square \end{aligned}$$

5.3 Anwendungen

(1) **Würfelexperiment, Bedeutung der Korrekturterme $\pm 1/2$:** Wie groß ist näherungsweise die Wahrscheinlichkeit bei 600 Würfeln mit einem korrekten Würfel mindestens 90 und höchstens 100 Sechsen zu erhalten?

Es ist $n = 600$ und $p = 1/6$, also $np = 100$ und $\sigma_n = \sqrt{600 \cdot (1/6) \cdot (5/6)} \approx 9,13$. Daraus folgt

$$\begin{aligned} P(90 \leq S_n \leq 100) &= P\left(\frac{90 - 100}{\sigma_n} \leq S_n^* \leq \frac{100 - 100}{\sigma_n}\right) \\ &\approx \Phi(0) - \Phi(-10/9,13) \\ &\approx 0,5 - \Phi(-1,095) = 0,5 - (1 - \Phi(1,095)) \approx 0,36, \end{aligned}$$

wobei der Wert $\Phi(1,095) \approx 0,8632$ der Tafel am Ende dieses Buches entnommen wurde. (Es wurden nicht mehr Stellen im Ergebnis angegeben, da es sich ohnehin um eine Approximation handelt.)

Dieses Ergebnis ist noch unnötig ungenau. Wir haben für $\beta = 100$ und $\alpha = 90$ die Wahrscheinlichkeit $P(\alpha \leq S_n \leq \beta)$ durch $\Phi(x(n, \beta)) - \Phi(x(n, \alpha))$ approximiert. Wie wir gleich sehen werden, liefert die auch in der Herleitung von Satz 5.4 verwendete Approximation durch $\Phi(x(n, \beta + 1/2)) - \Phi(x(n, \alpha - 1/2))$ ein genaueres Ergebnis. Mit dieser ist

$$\begin{aligned} P(90 \leq S_n \leq 100) &\approx \Phi(+0,5/9,13) - \Phi(-10,5/9,13) \\ &\approx \Phi(0,055) - \Phi(-1,15) \\ &= \Phi(0,055) - (1 - \Phi(1,15)) \approx 0,397. \end{aligned}$$

Nach Tafeln ist der exakte Wert der Wahrscheinlichkeit 0,4025. Wir sehen, dass die Berücksichtigung der „Korrekturterme“ $\pm 1/2$ selbst bei $n = 600$ noch einen merkbaren Unterschied bewirkt, obwohl natürlich für $n \rightarrow \infty$ der Einfluss verschwindet.

Berücksichtigt man die Korrekturterme, so lassen numerische Vergleichsrechnungen mit den exakten Werten von $b_{n,p}(k)$ erkennen, dass die Approximation durch die Normalverteilung schon für relativ kleine Werte n , etwa $n = 10$, ziemlich gut ist, vorausgesetzt, dass p nicht nahe an 0 oder 1 liegt. Ist etwa $n = 10$, $p = 0,4$, so kann man die $b_{n,p}(k)$ exakt berechnen oder Tafeln entnehmen und kommt z.B. auf

$$P(S_n \leq 3) = 0,3822.$$

Mit Berücksichtigung der Korrektur kommt man auf die Approximation $P(S_n \leq 3) \approx \Phi(x(n, 3 + 1/2))$. Dabei ist

$$x(n, 3 + 1/2) = (3 - 10 \cdot 0,4 + 1/2) / \sqrt{10 \cdot 0,4 \cdot 0,6} \approx -0,3227.$$

Aus Tafeln ergibt sich $\Phi(-0,3227) \approx 0,3734$. Der Fehler ist also nur etwa 0,01, also 3 %.

Ohne Korrektur ergäbe sich $P(S_n \leq 3) \approx 0,26$, was sich um etwa 30 % vom wahren Wert unterscheidet.

Wir haben in dieser Vergleichsrechnung eine „einseitige“ Approximation verwendet, also $P(-\infty \leq S_n \leq 3)$ statt $P(0 \leq S_n \leq 3)$ approximiert. Für die exakten Wahrscheinlichkeiten macht dies keinen Unterschied, da S_n keine negativen Werte annehmen kann. Für die Approximation mit der Normalverteilung ist der Unterschied so klein, dass es rechnerisch sinnvoller ist, in der Approximation die untere Grenze $-\infty$ einzusetzen.

(2) **Wahlvorhersage, Bestimmung eines Stichprobenumfangs:** Wir wollen den Prozentsatz der Wähler einer Partei A schätzen. Werden n Wähler befragt und sind darunter S_n Wähler der Partei A , so sei S_n/n der Schätzer für die Wahrscheinlichkeit p , dass ein zufällig ausgewählter Wähler für die Partei A stimmt. Wie groß muss n sein, damit die Wahrscheinlichkeit eines Irrtums von mehr als 1 % nicht größer ist als 0,05?

Es soll also gelten $P(-0,01 \leq S_n/n - p \leq 0,01) \approx 0,95$. Mit $\sigma_n = \sqrt{npq}$ ergibt sich nach Satz 5.5

$$0,95 \approx P\left(-\frac{0,01 n}{\sigma_n} \leq S_n^* \leq \frac{0,01 n}{\sigma_n}\right) \approx \Phi\left(\frac{0,01 n}{\sigma_n}\right) - \Phi\left(-\frac{0,01 n}{\sigma_n}\right) = 2 \Phi\left(\frac{0,01 n}{\sigma_n}\right) - 1,$$

also $\Phi(0,01 n / \sigma_n) \approx 0,975$.

Zu jedem $z \in (0, 1)$ existiert eine eindeutig bestimmte Zahl $x = \Phi^{-1}(z)$ mit $\Phi(x) = z$. (Φ^{-1} sei also die Umkehrfunktion zu Φ .) Aus der Tabelle entnehmen wir $\Phi^{-1}(0,975) = 1,96$. Also muss $0,01 \sqrt{n} / \sqrt{pq} \approx 1,96$ sein, d.h. $n \approx p \cdot q \cdot 10\,000 \cdot 1,96^2$. Nun ist p leider unbekannt. Aber der Maximalwert von $pq = p(1-p)$ ist $1/4$. Man kommt also in jedem Fall mit

$$n = 2\,500 \cdot 1,96^2 \approx 9\,600$$

Befragungen aus. Hat man schon von vorneherein die Information, dass jedenfalls $p \leq 0,1$ ist, so kommt man mit $n = 0,1 \cdot 0,9 \cdot 10\,000 \cdot 1,96^2 \approx 3\,450$ Befragungen aus.

(3) **Konfidenzintervalle für die Erfolgswahrscheinlichkeit:** In § 4 hatten wir das folgende Verfahren zur Konstruktion von Konfidenzintervallen zum Niveau $1 - \alpha$ für eine Erfolgswahrscheinlichkeit p angegeben:

Eine $b_{n,p}$ -verteilte Zufallsvariable X wird beobachtet. Sei für $x < n$ $p_o(x)$ der Wert von p , für den $P_p(X \leq x) = \alpha/2$ ist, und sei $p_o(n) = 1$. Symmetrisch dazu sei $p_u(x)$ für $x > 0$ der Wert von p , der $P_p(X \geq x) = \alpha/2$ erfüllt, und $p_u(0) = 0$. Dann ist

$$C(x) = [p_u(x), p_o(x)]$$

ein Konfidenzintervall zum Niveau $1 - \alpha$ für p .

Das in § 4 angegebene Verfahren zur Berechnung von $p_u(x)$ und $p_o(x)$ ist allerdings ohne Computerhilfe für große n nicht mehr praktikabel, da die F -Verteilung nicht für beliebig hohe Zahlen von Freiheitsgraden vertafelt ist. Für große n benutzt man die Approximation mit der Normalverteilung.

Die $b_{n,p}$ -verteilte Größe heißt jetzt nicht S_n , sondern X . Die Approximation für $P_p(X \leq x)$ ist also $\Phi((x + 1/2 - np)/\sigma_n)$. Wir suchen daher nach einem Wert von p , für den

$$\Phi((x + 1/2 - np)/\sigma_n) = \alpha/2$$

ist. Setzt man $c = \Phi^{-1}(1 - \alpha/2)$, so ist diese Gleichung gleich bedeutend mit der Gleichung

$$c = (np - x - 1/2)/\sqrt{np(1-p)}.$$

Die Lösung dieser Gleichung dient gewöhnlich als Approximation von $p_o(x)$. Analog bestimmt man die Approximation von $p_u(x)$. Nach etwas Rechnung findet man so die approximativen Konfidenzintervalle mit den Grenzen

$$(n + c^2)^{-1} \left((x \pm \tfrac{1}{2}) + \tfrac{1}{2} c^2 \pm c \sqrt{(x \pm \tfrac{1}{2}) - n^{-1} (x \pm \tfrac{1}{2})^2 + \tfrac{1}{4} c^2} \right). \quad (5.17)$$

(Das Vorzeichen $+$ gilt für das obere Ende des Konfidenzintervalls, das Vorzeichen $-$ für das untere Ende. Ist der kleinere dieser Werte kleiner als 0, so ist er durch 0 zu ersetzen, und der größere durch 1, wenn er größer als 1 ist.)

Blyth und Still (1983) haben allerdings darauf hingewiesen, dass diese Konfidenzintervalle selbst für $n \rightarrow \infty$ die geforderte Niveaubedingung nicht erfüllen. Das liegt daran, dass die Normalapproximation nicht gleichmäßig in p gilt. Für sehr kleine p und für p nahe bei 1 ist sie auch für sehr große n ungeeignet. Zum Beispiel ist für $\alpha = 0,01$ das Niveau der durch (5.17) bestimmten Intervalle asymptotisch statt 0,99 nur 0,9677.

Häufig verwendet man statt der Intervalle (5.17) Konfidenzintervalle mit den Grenzen

$$\frac{x}{n} \pm \frac{c}{\sqrt{n}} \sqrt{\frac{x}{n} \left(1 - \frac{x}{n}\right)}.$$

Sie ergeben sich aus (5.17), indem man für $n \rightarrow \infty$ kleine Terme vernachlässigt. Gegen diese Intervalle lassen sich zusätzliche Einwände geltend machen. Sie brauchen selbst für Werte von p nahe bei 1/2 nicht mit Wahrscheinlichkeit $1 - \alpha$ den wahren Parameter zu überdecken. Die Verwendung der Intervalle (5.17) ist daher vorzuziehen, aber sie sollten nur angewandt werden, wenn x/n nicht nahe an 0 oder 1 liegt.

Will man sichere quantitative Fehlerabschätzungen durchführen, so reichen die hier bewiesenen Resultate nicht aus. Eine bekannte Fehlerabschätzung (Satz von Berry-Esseen) ist im Anhang von § 12 angegeben. Sie bestätigt unsere Beobachtung aus Tabelle 5.1, dass für kleine Werte von pq die Approximation erst für größere n gut ist. Eine Faustregel besagt, dass man die Approximation für $npq \geq 6$ verwenden kann, aber dies hängt natürlich davon ab, welche Fehler man noch tolerieren will.

5.4 Die Poisson-Approximation

Für kleine Erfolgswahrscheinlichkeiten ist eine andere Approximation der Binomialverteilung vorteilhaft. Wir wollen sogar den etwas allgemeineren Fall betrachten, in dem die Erfolgswahrscheinlichkeiten in den einzelnen Teilversuchen verschieden sein dürfen. Wir beginnen mit einem kleinen Satz, der auch für andere Anwendungen von Interesse ist.

Satz 5.6 Sind X und Y unabhängige Zufallsvariable mit ganzzahligen Werten, so ist

$$P(X + Y = k) = \sum_i P(X = i) P(Y = k - i).$$

Sind $u = (u_i)$ und $v = (v_j)$ zwei Wahrscheinlichkeitsverteilungen auf $\mathbb{Z}$, so nennt man die Verteilung $w = (w_k)$ mit

$$w_k = \sum_i u_i v_{k-i} \quad (5.18)$$

die **Faltung** $u * v$ von u und v . Die Verteilung von $X + Y$ ist also die Faltung der Verteilung von X mit der von Y , wenn X und Y unabhängig sind.

Beweis: Es ist

$$\begin{aligned} P(X + Y = k) &= \sum_i P(X = i, X + Y = k) \\ &= \sum_i P(X = i, Y = k - i) \\ &= \sum_i P(X = i) P(Y = k - i). \end{aligned} \quad \square$$

Definition 5.7 Eine Zufallsvariable X heißt **Poisson-verteilt** mit Parameter $\lambda \geq 0$ (kurz: $P(\lambda)$ -verteilt), wenn

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!} \quad (k = 0, 1, \dots)$$

gilt. (Abkürzend schreiben wir auch $p(k | \lambda) = e^{-\lambda} \lambda^k / k!$.)

Lemma 5.8 Sind X_1, X_2 unabhängig und ist X_i $P(\lambda_i)$ -verteilt, so ist die Zufallsvariable $X_1 + X_2$ $P(\lambda_1 + \lambda_2)$ -verteilt.

Beweis: In $\sum_i P(X_1 = i) P(X_2 = k - i)$ sind nur die Terme mit $i \geq 0$ von Null verschieden, da X_1 und X_2 nur nichtnegative Werte annehmen. Also ist

$$\begin{aligned} P(X_1 + X_2 = k) &= \sum_{i=0}^k e^{-\lambda_1} \frac{\lambda_1^i}{i!} \cdot e^{-\lambda_2} \frac{\lambda_2^{k-i}}{(k-i)!} \cdot \frac{k!}{k!} \\ &= e^{-(\lambda_1 + \lambda_2)} \cdot \frac{1}{k!} \sum_{i=0}^k \binom{k}{i} \lambda_1^i \lambda_2^{k-i} \\ &= e^{-(\lambda_1 + \lambda_2)} \frac{(\lambda_1 + \lambda_2)^k}{k!}. \end{aligned} \quad \square$$

Wir zeigen nun, dass die Poisson-Verteilung eine gute Approximation für die Verteilung der Zahl der Erfolge ist, wenn die Erfolgswahrscheinlichkeiten klein sind:

Satz 5.9 $X_1, \dots, X_n$ seien unabhängige Zufallsvariable mit $P(X_i = 1) = p_i$ und $P(X_i = 0) = 1 - p_i$. Sei $S = X_1 + \dots + X_n$ und $\lambda = p_1 + \dots + p_n$. Dann gilt

$$\sum_{k=0}^{\infty} \left| P(S = k) - e^{-\lambda} \frac{\lambda^k}{k!} \right| \leq 2 \sum_{i=1}^n p_i^2. \quad (5.19)$$

Beweis: Es ist für die Berechnung der Verteilung von S egal auf welchem Wahrscheinlichkeitsraum die Zufallsvariablen definiert sind. Also können wir uns einen aussuchen, der für den Beweis vorteilhaft ist.

Wir setzen $\Omega_i = \{-1, 0, 1, 2, \dots\}$, $P_i(0) = 1 - p_i$, $P_i(-1) = e^{-p_i} - (1 - p_i)$ und $P_i(k) = e^{-p_i} p_i^k / k!$ für $k \in \mathbb{N}$. Sei $\Omega = \Omega_1 \times \dots \times \Omega_n$ und $P = P_1 \times \dots \times P_n$, d.h. für $\omega = (\omega_1, \omega_2, \dots, \omega_n) \in \Omega$ sei

$$P(\omega) = P_1(\omega_1) P_2(\omega_2) \dots P_n(\omega_n).$$

Wir setzen

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

Dann haben die X_i die geforderte Verteilung. Die Y_i sind unabhängig und $P(\lambda_i)$ -verteilt. Die Wahl von (Ω_i, P_i) wurde so getroffen, dass unter der Nebenbedingung, dass diese Verteilungen vorliegen, $P(X_i = Y_i)$ möglichst groß ist. Es ist

$$P(X_i = Y_i) = P_i(0) + P_i(1) = 1 - p_i + e^{-p_i} p_i.$$

Daher ist $P(X_i \neq Y_i) = p_i - e^{-p_i} p_i = p_i(1 - e^{-p_i}) \leq p_i^2$. Nach Lemma 5.8 ist $T = Y_1 + \dots + Y_n$ $P(\lambda)$ -verteilt. Die abzuschätzende Summe in (5.19) lässt sich nun schreiben als

$$\begin{aligned} & \sum_{k=0}^{\infty} |P(S = k) - P(T = k)| \\ &= \sum_{k=0}^{\infty} |P(S = k, T = k) + P(S = k, T \neq k) - P(T = k, S = k) - P(T = k, S \neq k)| \\ &\leq \sum_{k=0}^{\infty} (P(S = k, T \neq k) + P(T = k, S \neq k)) = 2P(S \neq T) \\ &\leq 2 \sum_{i=1}^n P(X_i \neq Y_i) \leq 2 \sum_{i=1}^n p_i^2. \end{aligned} \quad \square$$

Folgerung 5.10 Ist $p(n)$ eine Folge mit $0 \leq p(n) \leq 1$ und $np(n) \rightarrow \lambda$, so gilt

$$b_{n,p(n)}(k) = \binom{n}{k} p(n)^k (1 - p(n))^{n-k} \rightarrow e^{-\lambda} \frac{\lambda^k}{k!}.$$

Beweis: Man setzt $p_i = p(n)$ ($i = 1, \dots, n$). Dann ist $P(S = k) = b_{n,p(n)}(k)$, und es gilt

$$2 \sum_{i=1}^n p(n)^2 = 2p(n) \cdot np(n) \rightarrow 0. \quad \square$$

Die Folgerung lässt sich auch leicht direkt beweisen. Setzt man $\lambda_n = np(n)$, so ist

$$b_{n,p(n)}(k) = \frac{1}{k!} \cdot \frac{n}{n} \cdot \frac{n-1}{n} \cdot \dots \cdot \frac{n-k-1}{n} \cdot \lambda_n^k \cdot \left(1 - \frac{\lambda_n}{n}\right)^n \cdot \left(1 - \frac{\lambda_n}{n}\right)^{-k}.$$

Die Quotienten $n/n, (n-1)/n, \dots$ und der letzte Faktor streben gegen 1 und λ_n^k gegen λ^k . Die Behauptung folgt daher aus $(1 - \lambda_n/n)^n \rightarrow e^{-\lambda}$. Dieses bekannte Resultat ergibt sich aus der folgenden Überlegung: Da $\log x$ in $x = 1$ die Ableitung 1 hat, gilt $\log(1-h) = -h + h \Delta(h)$ mit $\Delta(h) \rightarrow 0$ für $h \rightarrow 0$. Für $h = \lambda_n/n$ folgt

$$\left(1 - \frac{\lambda_n}{n}\right)^n = \exp\left\{n\left(-\frac{\lambda_n}{n} + \frac{\lambda_n}{n} \Delta\left(\frac{\lambda_n}{n}\right)\right)\right\} \rightarrow e^{-\lambda}.$$

Aus der Tabelle 5.2 ergibt sich ein Bild von der Güte der Approximation, wenn die p_i alle gleich p sind, und $np = \lambda = 1$ gilt.

k	$p(k 1)$	$b_{100, 1/100}(k)$	$b_{10, 1/10}(k)$
0	0,367	0,366	0,349
1	0,367	0,369	0,387
2	0,184	0,184	0,194
3	0,061	0,061	0,057

Tabelle 5.2 Vergleich Poisson-Verteilung/Binomialverteilung

In der praktischen Anwendung verwendet man die Poisson-Verteilung als Modell überall dort, wo gezählt wird wie viele von vielen möglichen, aber einzeln relativ unwahrscheinlichen unabhängigen Ereignissen eintreten.

Beispiele: (1) In einem Hörsaal seien $n = 91$ Studenten (bzw. Studentinnen). Die Wahrscheinlichkeit p , heute Geburtstag zu haben, ist $p = 1/365$. Die Zahl X derer, die heute Geburtstag haben, ist praktisch $P(\lambda)$ -verteilt mit $\lambda = 91/365 \approx 0,25$.

(2) In einem Land gibt es ca. 30 Selbstmorde pro 100 000 Einwohner pro Jahr. In einer typischen Stadt mit 120 000 Einwohnern wäre dann damit zu rechnen, dass die Zahl der Selbstmorde im kommenden Jahr $P(\lambda)$ -verteilt ist mit $\lambda = 36$.

(3) Von einer Ware, z.B. Blitzlichtlampen, ist ein kleiner Anteil $p = 0,015$ schon bei der Produktion defekt. Wie groß muss man n wählen, damit ein Karton mit n (unabhängigen) Exemplaren mit Wahrscheinlichkeit $\geq 0,8$ mindestens 100 intakte Exemplare enthält? n ist minimal zu wählen mit der Eigenschaft

$$0,8 \leq \sum_{k=0}^{n-100} b_{n,p}(k).$$

Setzt man $\lambda_n = np$, so ist nach der Poisson-Approximation die rechte Seite ungefähr gleich

$$r_n := e^{-\lambda_n} \sum_{k=0}^{n-100} \frac{\lambda_n^k}{k!}.$$

Mit Hilfe des Taschenrechners stellt man fest, dass $n = 102$ die kleinste ganze Zahl ist, für die $r_n \geq 0,8$ gilt. Man muss daher 102 Blitzlichtlampen in den Karton geben.

Oft ist es klar, dass gezählt wird, wie viele von einzeln unwahrscheinlichen Ereignissen eintreten, obwohl man über n und p keine genauen Angaben machen kann. Dann verwendet man Schätzwerte von λ . Z.B. ist die Zahl der Meteoriteneinschläge in einem Gebiet in einem bestimmten Zeitraum Poisson-verteilt, weil viele Meteoriten die Erde treffen und jeder mit kleiner Wahrscheinlichkeit in das Gebiet trifft. Den Parameter λ kann man schätzen, wenn man die Zahl der Einschläge in einem großen Gebiet in einem langen

Zeitraum kennt. Ein ähnliches Beispiel wäre die Zahl der Chromosomenbeschädigungen bei Bestrahlung.

Der Parameter λ der Poisson-Verteilung hat eine doppelte Interpretation. Ist X $P(\lambda)$ -verteilt, so ist

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

Fast die gleiche Rechnung ergibt $E(X(X-1)) = \lambda^2$. Daher ist

$$\text{Var}(X) = E(X^2) - (EX)^2 = E(X(X-1)) + EX - (EX)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

λ ist also zugleich Erwartungswert und Varianz der $P(\lambda)$ -Verteilung.

Anhang*

Beweis der stirlingschen Formel (Satz 5.1)

1. Schritt (Beweis von (5.4)): Grundidee ist der Vergleich von

$$\log n! = \log 1 + \log 2 + \dots + \log n$$

mit einem Integral. Da $\log x$ konkav und monoton wachsend ist, gilt für $k = 1, \dots, n$

$$\int_{k-1/2}^{k+1/2} \log x \, dx \leq \log k \leq \int_k^{k+1} \log x \, dx.$$

Summation ergibt

$$\int_{1/2}^{n+1/2} \log x \, dx \leq \log n! \leq \int_1^{n+1} \log x \, dx.$$

Nun ist $x \log x - x$ eine Stammfunktion von $\log x$. Also ist

$$\left(n + \frac{1}{2}\right) \log\left(n + \frac{1}{2}\right) - n - \frac{1}{2} \log \frac{1}{2} \leq \log n! \leq (n+1) \log(n+1) - n.$$

Wir setzen

$$d_n = \log n! - (n + 1/2) \log n + n \quad (n \geq 1).$$

Die Ungleichungen für $\log n!$ zeigen sofort, dass d_n durch die positive Konstante $-(1/2) \log(1/2)$ nach unten beschränkt ist. Es gilt

$$d_n - d_{n+1} = \left(n + \frac{1}{2}\right) \log \frac{n+1}{n} - 1 = \int_n^{n+1} \log x \, dx - \frac{1}{2} (\log(n+1) + \log n) \geq 0. \quad (5.20)$$

Also ist (d_n) monoton fallend und konvergiert gegen eine Konstante $C > 0$. Sei $\tau = e^C$. Dann folgt (5.4) aus

$$\tau = \lim \exp(\log n! - (n + 1/2) \log n + n) = \lim \frac{n!}{n^{n+1/2} e^{-n}}.$$

2. Schritt (Bestimmung von τ): Nun benötigen wir Lemma 5.3 und die Ungleichungen (5.14). Zu beliebig kleinem $\eta > 0$ existieren beliebig große b mit $1/b^2 < \eta$. Nach der tschebyschewschen Ungleichung ist $P(|S_n^*| > b) \leq 1/b^2 < \eta$. Sei $a = -b$ und seien die Folgen $(\alpha_n), (\beta_n)$ von ganzen Zahlen wie im Beweis von Satz 5.4 gewählt, also so dass

$$\{a \leq S_n^* \leq b\} = \{\alpha_n \leq S_n \leq \beta_n\}$$

ist. Dann gilt $x(n, \alpha_n) \rightarrow -b$, $x(n, \beta_n) \rightarrow b$ und $R_n \rightarrow \Phi(b) - \Phi(-b)$. Aus $P(a \leq S_n^* \leq b) \geq 1 - \eta$ und (5.14) folgt nun für $n \rightarrow \infty$

$$(\Phi(b) - \Phi(-b)) \sqrt{2\pi}/\tau \leq 1 \quad \text{und} \quad 1 - \eta \leq (\Phi(b) - \Phi(-b)) \sqrt{2\pi}/\tau.$$

Nach Lemma 5.3 liegt $\Phi(b) - \Phi(-b)$ für großes b nahe bei 1. Da $\eta > 0$ beliebig klein sein konnte, folgt $\tau = \sqrt{2\pi}$. $\square$

Zusatz: Der obige Beweis gestattet mit wenig Mehraufwand auch den Beweis der quantitativen Abschätzung (5.3). Es ist

$$\frac{n+1}{n} = \frac{1 + 1/(2n+1)}{1 - 1/(2n+1)}.$$

Setzt man in der Entwicklung

$$\frac{1}{2} \log \frac{1+t}{1-t} = t + \frac{t^3}{3} + \frac{t^5}{5} + \dots \quad (|t| < 1)$$

$t = 1/(2n+1)$, so folgt also aus (5.20)

$$\begin{aligned} d_n - d_{n+1} &= \frac{1}{3(2n+1)^2} + \frac{1}{5(2n+1)^4} + \dots \\ &< \frac{1}{3} \cdot \frac{1}{(2n+1)^2} \cdot \left(1 - \frac{1}{(2n+1)^2}\right)^{-1} = \frac{1}{12n} - \frac{1}{12(n+1)}. \end{aligned}$$

Die Folge $(d_n - 1/(12n))$ ist also wachsend. Andererseits ist wegen

$$d_n - d_{n+1} > \frac{1}{3(2n+1)^2} > \frac{1}{12n+1} - \frac{1}{12(n+1)+1}$$

die Folge $(d_n - 1/(12n+1))$ monoton fallend. Zusammen mit der monotonen Konvergenz von (d_n) gegen C ergibt dies $C + 1/(12n+1) < d_n < C + 1/(12n)$ und daraus folgt (5.3).

P. Buchner (1951) hat mit einem sehr ähnlichen Argument sogar bewiesen, dass man für $n \geq 2$ in (5.3) $12n+1$ durch $12n+1/4$ ersetzen kann.

Historische Anmerkung: Der lokale Grenzwertsatz und der Satz von de Moivre-Laplace waren die ersten Grenzwertsätze für Verteilungen. *Abraham de Moivre* publizierte sie in der 1733 erschienenen dritten Auflage seines Buches „Doctrin de chances“. *Laplace* (1812) erweiterte diese Ergebnisse (u.a. Satz 5.5) und demonstrierte ihre Bedeutung. De Moivre kannte die stirlingsche Formel bis auf die Aussage $\tau = \sqrt{2\pi}$ schon vor Stirling. (Er konnte τ als unendliche Reihe bestimmen.) *Siméon Denis Poisson* (1781–1840) entdeckte 1832 die Approximation der Binomialverteilung durch die Poisson-Verteilung. Sie geriet jedoch völlig in Vergessenheit. Erst 1898 demonstrierte *L. v. Bortkiewicz* in seinem Büchlein „Das Gesetz der kleinen Zahlen“ ihre Bedeutung für die Praxis. Die Verschärfung in Satz 5.9 haben *Hodges* und *Le Cam* (1960) angegeben; siehe auch *Serfling* (1975). Der im Beweis von Satz 5.9 verwendete, von *W. Doeblin* 1938 in anderem Kontext entwickelte Gedanke, die X_i mit leichter handhabbaren Zufallsvariablen zu vergleichen, spielt heute unter dem Namen „coupling“ eine wichtige Rolle beim Studium zahlreicher stochastischer Prozesse.

Aufgaben

1. Zeigen Sie: Sind (α_n) und (β_n) zwei Folgen von ganzen Zahlen mit $\alpha_n \leq \beta_n$ und (5.9), so ist $P(\alpha_n \leq S_n \leq \beta_n) \sim \Phi(x(n, \beta_n + 1/2)) - \Phi(x(n, \alpha_n - 1/2))$.
2. Eine Klasse von $2N$ Jungen und $2N$ Mädchen wird zufällig in zwei gleich große Gruppen unterteilt. Bestimmen Sie die Wahrscheinlichkeit, dass jede Gruppe gleich viele Jungen und Mädchen enthält, und schätzen Sie diese mit der stirlingschen Formel ab.
3. Ein Hotel hat 200 Betten. Wie viele Reservierungen darf der Hotelmanager akzeptieren, wenn erfahrungsgemäß eine Reservierung mit Wahrscheinlichkeit 0,2 annulliert wird, und die Wahrscheinlichkeit einer Überbuchung höchstens 0,025 sein soll. (Benutzen Sie die Normalapproximation.) (L)
4. Beweisen Sie mittels der stirlingschen Formel:

$$\lim_{\lambda \rightarrow \infty} \sum_{k=a_\lambda}^{b_\lambda} \frac{\lambda^k}{k!} e^{-\lambda} = \Phi(b) - \Phi(a) \quad \text{für } a < b.$$

Dabei sei $a_\lambda = [\lambda + a\sqrt{\lambda}]$ und $b_\lambda = [\lambda + b\sqrt{\lambda}]$.

5. Unter 100 Personen, die 1984 einen Führerschein erworben haben, haben 30 im Jahr nach Erwerb des Führerscheins einen Schaden verursacht. Man gebe ein Konfidenzintervall zum Niveau 0,8 für die Wahrscheinlichkeit p an, im Jahr nach Erwerb des Führerscheins einen Schaden zu verursachen.
6. Man zeige, dass sich das schwache Gesetz der großen Zahl für Bernoulli-Folgen aus dem Satz von de Moivre-Laplace folgern lässt. (L)
7. Unter n Neugeborenen sind S_n Knaben. Die Wahrscheinlichkeit einer Knabengeburt wird durch $\hat{p} = S_n/n$ geschätzt. Wie groß muss n sein, damit mit Wahrscheinlichkeit 0,95 die Ungleichung $|p - \hat{p}| \leq 0,002$ gilt?
8. Die Zufallsvariable X sei Poisson-verteilt mit Parameter $\lambda > 0$. Man zeige:

$$P(X \leq n) = \frac{1}{n!} \int_{\lambda}^{\infty} x^n e^{-x} dx \quad (n \in \mathbb{N}). \quad (\text{L})$$

9. Ein Geiger-Müller-Zählrohr Z und eine radioaktive Quelle Q seien so postiert, dass ein Teilchen, das von Q emittiert wird, von Z mit Wahrscheinlichkeit 10^{-4} registriert wird. Während der Beobachtungszeit emittiert Q 30 000 Teilchen. Man berechne approximativ die Wahrscheinlichkeit dafür, dass
 - a) Z kein Teilchen registriert;
 - b) Z mehr als 2 Teilchen registriert.Ferner schätze man den Approximationsfehler ab. (L)
10. Ein Buch mit 400 Seiten enthält 400 Druckfehler, die zufällig verteilt sind. Wie groß ist (approximativ) die Wahrscheinlichkeit, dass auf den ersten beiden Seiten je genau ein Druckfehler vorkommt.

§ 6 Tests

Es ist ein Grundgedanke der empirischen Wissenschaften, dass die Entscheidung zwischen konkurrierenden Modellen der Realität auf Beobachtungen eines Experiments gestützt werden soll, das unter den alternativen Modellannahmen verschiedene Versuchsergebnisse erwarten lässt. Im Idealfall ist nach der Idee von Francis Bacon ein „experimentum crucis“ möglich, das zu einer definitiven Entscheidung führt. Ein berühmtes Beispiel ist der michelsonsche Interferenzversuch.

Solche Experimente sind jedoch die Ausnahme. In vielen Fällen sind nur Beobachtungen möglich, die zahlreichen zufälligen Einflüssen unterliegen und keine mit absoluter Sicherheit gültigen Schlüsse erlauben. Soll etwa ein neues Medikament mit einem bisher gebräuchlichen Medikament verglichen werden, um zu entscheiden, welches bessere Erfolgchancen bringt, so ist es selbst dann, wenn dieses Medikament bei 10 Testpersonen Heilung bewirkt, möglich, dass das Zufall ist. Wenn es keiner der 10 Personen Heilung bringt, kann auch dies Zufall sein. Man kann sich so oder so irren. Die statistische Testtheorie kann diese Irrtümer nicht vermeiden. Sie kann aber Kriterien liefern, nach denen sich die Irrtumswahrscheinlichkeiten kontrollieren lassen.

Die Grundidee statistischer Tests soll an einem Beispiel diskutiert werden, das keinerlei praktisches, wohl aber historisches Interesse beansprucht. R. A. Fisher und J. Neyman haben daran ihre unterschiedlichen Vorstellungen vom Testbegriff diskutiert.

6.1 Beispiel der „tea tasting Lady“

Eine Lady behauptet, dass sie — wenn sie Tee probiert, der einen Zusatz Milch enthält — unterscheiden könne, ob zuerst die Milch oder zuerst der Tee eingegossen worden ist.

Fisher schlägt in seinem Buch „Design of Experiments“ (1935) den folgenden Test vor, um zu entscheiden, ob die Lady wirklich die behauptete Fähigkeit hat: Es wird in vier Tassen zuerst Tee und dann ein Zusatz Milch gegeben. Diese nennen wir vom Typ 1. In vier weitere Tassen wird zuerst etwas Milch und dann der Tee eingegossen. Diese nennen wir vom Typ 2. Es wird gut umgerührt. Der Lady werden die acht Tassen in zufälliger Reihenfolge (mit Gleichverteilung auf der Menge der $8!$ Permutationen) gereicht. Sie wird aufgefordert, genau vier Tassen zu benennen, die sie für die Tassen vom Typ 1 hält. Sei nun X die Zahl der von der Lady richtig klassifizierten Tassen vom Typ 1. Rät sie alle 4 Tassen vom Typ 1 korrekt, so soll damit der Beweis der Behauptung als erbracht gelten. Die Wahrscheinlichkeit, dass eine so große Trefferzahl zufällig durch reines Raten zustandekommt, ist nach der hypergeometrischen Verteilung $1/70$, so dass man nur mit dieser Wahrscheinlichkeit der Lady Recht gibt, wenn sie mit ihrer Behauptung nicht Recht hat.

Schwieriger wird das Problem, wenn die Lady behauptet, mit ihren Klassifikationen zwar nicht unfehlbar zu sein, aber doch gewisse, wenn auch undeutliche, Unterschiede gelegentlich zu bemerken, die sie in die Lage versetzen, öfter die richtige Klassifikation zu treffen als dies dem Zufall entspricht.

Würde man nun der Lady schon Recht geben, wenn sie drei der vier Tassen vom Typ 1 herausfindet, so wäre immerhin die Wahrscheinlichkeit, dass dies zufällig zustandekäme $= \left(\binom{4}{4} \binom{4}{0} + \binom{4}{3} \binom{4}{1} \right) / \binom{8}{4} = \frac{17}{70}$. Daher wären 3 richtig klassifizierte Tassen vom Typ 1 noch kein sonderlich überzeugender Nachweis der Fähigkeit der Lady. Man kann in

dieser Situation Irrtümer zwar nicht völlig ausschließen, aber man möchte sich doch eine Grenze für die Irrtumswahrscheinlichkeit setzen. Z.B. sollte sie nicht größer als $\alpha = 0,05$ sein.

Fisher argumentiert nun, dass sich diese Grenze einhalten lässt, indem man das Experiment durch Erhöhung der Zahl der Tassen sensitiver macht. Bei doppelter Anzahl der Tassen könne man immerhin schon die Behauptung der Lady für $X \geq 7$ akzeptieren, also *einen* Fehler zulassen, ohne die Irrtumswahrscheinlichkeit 0,05 zu überschreiten.

Neyman setzt sich in seinem Buch „First Course in Probability and Statistics“ (1950) ausführlich mit dem Verfahren von Fisher auseinander. Er kritisiert, dass die Lady, wenn sie nicht gerade unfehlbar ist, im Nachteil sei. Bei der Versuchsanordnung von Fisher ist nicht klar, wie viele Tassen genommen werden müssen, damit die Lady eine reelle Chance hat, ihre Behauptung zu beweisen.

Dieses Manko liegt nach Neymans Ansicht an einer entscheidenden Lücke im Testbegriff von Fisher. Während bei Fisher nur die Verteilung von X unter der „Nullhypothese“ ins Spiel kommt, d.h. unter der Annahme, dass die Lady keine besondere Fähigkeit hat, besteht nach Neyman das statistische Problem gerade darin, zwischen zwei verschiedenen stochastischen Modellen zu entscheiden.

Um auch Verteilungen unter der Annahme, die Lady habe öfter Erfolg als es dem Zufall entspricht, ableiten zu können, schlägt Neyman eine Abänderung der Ausführung des Tests vor: Der Lady soll n -mal die Aufgabe gestellt werden, zwei Tassen, von denen eine vom Typ 1 und eine vom Typ 2 ist, korrekt zu klassifizieren. Die beiden Tassen werden ihr jeweils in einer zufälligen, durch Münzwurf bestimmten Reihenfolge gegeben. Damit die Lady unabhängig von früheren Entscheidungen urteilen kann, wird jedes Teilexperiment an einem anderen Tag ausgeführt. X sei die Zahl der Tage, an denen sie die beiden Tassen richtig klassifiziert.

Als Modell für diese Versuchsanordnung bietet es sich an, X als binomialverteilt mit Parametern n und p anzunehmen. Die „Nullhypothese“ entspricht dem Fall $p = 1/2$ und die Alternative, dass die Lady tatsächlich bessere Erfolgschancen hat, als es dem reinen Zufall entspricht, kann man durch $p > 1/2$ beschreiben. Man nimmt also an, dass die Lady, wenn sie Recht hat, an jedem Tag unabhängig von den anderen Tagen mit Wahrscheinlichkeit $p > 1/2$ einen Erfolg erzielt.

Die Auswertung verläuft wieder wie bei Fisher. Es wird eine Zahl t festgelegt mit der Eigenschaft, dass unter der Nullhypothese

$$P(X \geq t) \leq \alpha$$

ist. Dabei ist α eine vorgegebene Schranke, etwa $\alpha = 0,05$. Im Fall $X \geq t$ wird die Behauptung der Lady akzeptiert. Wir sagen dann, dass die Nullhypothese verworfen wird. Im Fall $X < t$ gilt die Behauptung der Lady als nicht erwiesen. In der Sprechweise der Statistik sagt man, dass die Nullhypothese akzeptiert wird. Fisher und Neyman sind sich darin einig, dass dies nicht so gedeutet werden kann, dass die Nullhypothese bewiesen ist.

Der entscheidende Unterschied zu dem Ansatz von Fisher liegt darin, dass man nun auch mathematisch präzise Aussagen über die Wahrscheinlichkeit, die Nullhypothese zu akzeptieren, wenn die Alternative zutrifft, machen kann. Bevor wir darauf näher eingehen, soll der von Neyman benutzte Testbegriff allgemein dargestellt werden. Er ist von Jerzy Neyman und Egon Pearson in den 30er Jahren entwickelt worden.

6.2 Grundbegriffe der Testtheorie

Von einem Testproblem spricht man, wenn eine zufällige Größe X mit einer unbekannten Verteilung P_ϑ beobachtet wird, und man aufgrund des beobachteten Wertes x der Zufallsvariablen entscheiden soll, ob P_ϑ einer bestimmten Menge von Verteilungen angehört oder nicht.

Im Folgenden sei $\mathcal{X}$ die Menge der möglichen Werte x einer Zufallsvariablen X und $\{P_\vartheta : \vartheta \in \Theta\}$ die Menge der in Betracht gezogenen Verteilungen von X . Wie schon in § 4.3 nehmen wir an, dass X durch die identische Abbildung auf $\mathcal{X}$ gegeben sei. Unter den möglichen Verteilungen von X sei eine nichtleere echte Teilmenge $\{P_\vartheta : \vartheta \in H\}$ durch zusätzliche Bedingungen ausgezeichnet. Ein **Test** ist eine Entscheidungsregel, die für jeden möglichen Wert x von X festlegt, ob man sich für die **Hypothese** „ $\vartheta \in H$ “ oder für die **Alternative** „ $\vartheta \in \Theta \setminus H$ “ entscheiden soll. (Man nennt auch kurz H die Hypothese und $K = \Theta \setminus H$ die Alternative.) Die Entscheidung für die Hypothese nennt man **Annahme der Hypothese**, und die Entscheidung für die Alternative nennt man **Verwerfen der Hypothese**. Ein Test ist also (bis auf weiteres) beschrieben durch Angabe der Menge R derjenigen x , für die die Hypothese verworfen werden soll. R wird **Verwerfungsbereich** oder **kritischer Bereich** des Tests genannt.

Innerhalb des gewählten Modells sind also zwei Arten von Fehlern möglich: Ist $\vartheta \in H$ und wird die Hypothese verworfen, so spricht man von einem **Fehler erster Art**. Ist $\vartheta \in K$ und wird die Hypothese angenommen, so spricht man von einem **Fehler zweiter Art**.

Praktisch gibt man R meist mit Hilfe einer Funktion $T(x)$ an, der sog. Teststatistik, die so gewählt wird, dass besonders große Werte von $T(x)$ gegen die Hypothese sprechen. Man wählt dann einen **kritischen Wert** t und verwirft die Hypothese im Fall $T(x) \geq t$. Dann ist also $R = \{x : T(x) \geq t\}$.

Bisher haben wir das Testproblem so formuliert, dass H und K völlig symmetrische Rollen spielen. In konkreten Fragestellungen gibt es aber gewöhnlich eine Asymmetrie. Ist man z.B. daran interessiert, ob sich irgendwelche Daten innerhalb einer etablierten Theorie erklären lassen oder auf neue Effekte hindeuten, so sollte man auf neue Effekte erst schließen, wenn wirklich deutliche Hinweise darauf vorliegen. Soll ein gebräuchliches Medikament durch ein neues ersetzt werden, so wird man bei unklaren Vergleichswerten vorerst bei dem alten Medikament bleiben. In beiden Fällen erscheint ein *vorschneller* Wechsel nicht ratsam. Im Zweifel kann man sich ja gewöhnlich weitere Daten verschaffen.

In der Formulierung des Testproblems trägt man dem so Rechnung, dass man als Hypothese die Verteilung (oder die Verteilungen) wählt, die der etablierten Theorie bzw. reiner Zufälligkeit entsprechen. Man verwendet dann auch oft statt des Wortes Hypothese das Wort **Nullhypothese**, um anzudeuten, dass die Hypothese keine systematischen Abweichungen vom reinen Zufall annimmt.

Man zieht nur Verwerfungsbereiche R in Betracht, für die die Wahrscheinlichkeit eines Fehlers erster Art durch eine vorgegebene kleine Zahl $\alpha > 0$ begrenzt ist. Dadurch erreicht man, dass man neue Effekte oder wesentliche Vorteile des neuen Medikaments nur dann behauptet, wenn wirklich die Daten deutlich dafür sprechen. Wir werden sehen, dass sich die Wahrscheinlichkeit eines Fehlers zweiter Art gewöhnlich nicht simultan in gleicher Weise begrenzen lässt. Praktisch bedeutet dies, wie gesagt, dass die *Annahme der Hypothese nicht so gedeutet werden darf, dass sie als erwiesen oder einigermaßen sicher gelten darf. Die Annahme der Hypothese bedeutet nur, dass die beobachteten Daten keinen hinreichenden Anlass zur Verwerfung der Hypothese bieten.* (Die Formulierung

„die Hypothese wird nicht verworfen“ gibt dies besser wieder als die Formulierung „die Hypothese wird angenommen.“)

In der Regel ist aber auch das Ziel nicht der Beweis der Gültigkeit der Hypothese. Man will ja gerade nachweisen, dass z.B. hohe Trefferquoten der Lady mit kleinen Irrtumswahrscheinlichkeiten *gerade nicht* durch Zufall erklärbar sind, dass ein neues Medikament *echt besser* ist als ein herkömmliches. Macht man die Aussage zur Hypothese, dass das, was man nachweisen will *nicht* gilt, so ergibt die Verwerfung der Hypothese mit kleinen Fehlerwahrscheinlichkeiten gerade den gewünschten Nachweis.

Quantitative Aussagen erhält man durch Betrachtung der **Gütefunktion**

$$\beta(\vartheta) = P_{\vartheta}(X \in R)$$

zu R , die jedem ϑ die Verwerfungswahrscheinlichkeit unter P_{ϑ} zuordnet. Wir sagen, dass der Test **Niveau** α habe, wenn für alle $\vartheta \in H$ die Ungleichung $\beta(\vartheta) \leq \alpha$ gilt. Die Wahrscheinlichkeit eines Fehlers erster Art ist dann maximal α . Für $\vartheta \in K$ heißt $\beta(\vartheta)$ die **Macht** des Tests in ϑ . Ist die Macht $\beta(\vartheta)$ nahe bei 1, so ist die Wahrscheinlichkeit $1 - \beta(\vartheta)$ eines Fehlers zweiter Art klein.

Wählt man $R = \emptyset$, so heißt dies, dass für kein x die Hypothese verworfen wird. Die Wahrscheinlichkeit eines Fehlers erster Art ist dann 0, aber die eines Fehlers zweiter Art ist 1. Dies deutet an, dass in Wahrheit ein Kompromiss gefunden werden muss. α sehr klein anzusetzen, kann nicht generell gut sein, denn man wird berücksichtigen müssen, wie gravierend Fehler zweiter Art sind. Wenn sehr viele Daten vorliegen, wird man vielleicht auch bei kleinem α die Wahrscheinlichkeit eines Fehlers zweiter Art einigermaßen klein halten können. Üblich sind Werte wie $\alpha = 0,05$, $\alpha = 0,02$ oder $\alpha = 0,01$.

Diese abstrakten Überlegungen werden vielleicht klarer, wenn wir sie an dem Beispiel mit der „tea tasting Lady“ illustrieren.

6.3 Mehr zur „tea tasting Lady“

Das für die Versuchsanordnung von Neyman gewählte Modell ist durch

$$\mathcal{X} = \{0, 1, \dots, n\}, \quad \Theta = [1/2, 1], \quad \vartheta = p$$

und

$$P_p(X = x) = b_{n,p}(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

beschrieben. Die Hypothese ist $H = \{1/2\}$, und die Alternative ist $K =]1/2, 1]$. (Man sagt auch, dass die Hypothese $p = 1/2$ gegen die Alternative $p > 1/2$ getestet werden soll.) Wir hatten Verwerfungsbereiche der Form $R = \{x : x \geq t\}$ in Betracht gezogen. Die Lady muss also n Paare von Tassen klassifizieren, und die Hypothese wird verworfen, wenn die Lady mindestens t Paare richtig klassifiziert.

Sei $\beta(p|t, n) = P_p(X \geq t)$ die Gütefunktion dieses Tests. Setzt man als Niveau $\alpha = 0,05$ fest, und ist $n = 5$, so kommt nur $t = 5$ in Frage, denn für $t = 4$ ist bereits

$$\beta(1/2|4, 5) = \left(\frac{1}{2}\right)^5 + \binom{5}{1} \left(\frac{1}{2}\right)^5 \approx 0,187 > \alpha.$$

Für $t = 5$ erhält man die Gütefunktion $\beta(p|5, 5) = p^5$; siehe Bild 6.1.

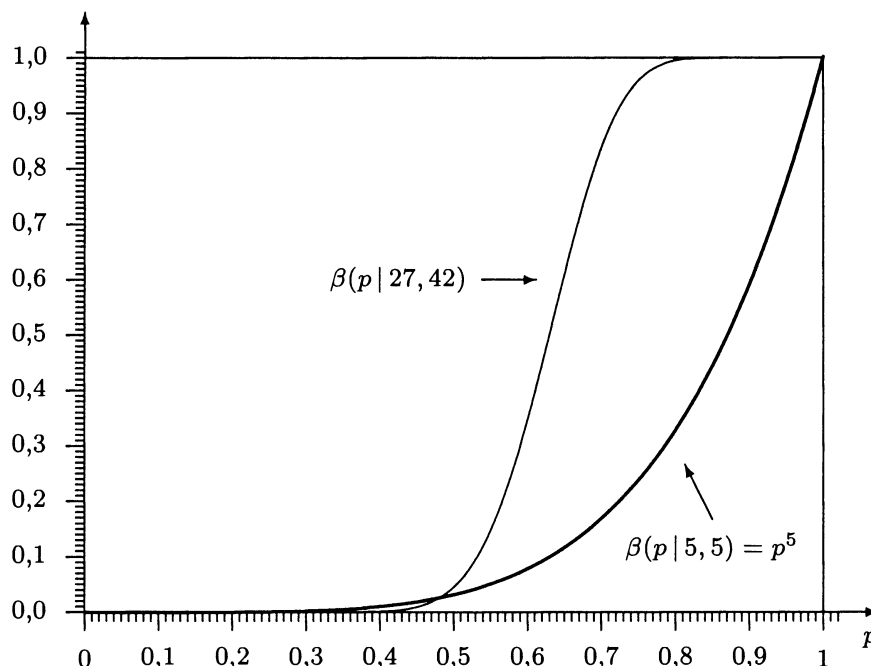


Bild 6.1 Gütefunktionen von zwei Tests der „tea tasting Lady“

Es ist $\beta(0,6 | 5, 5) \approx 0,08$ und $\beta(0,9 | 5, 5) \approx 0,59$. Wenn also die Lady immerhin eine Erfolgswahrscheinlichkeit von 0,6 pro Klassifikation hätte, würde sie doch nur mit Wahrscheinlichkeit 0,08 ihre Fähigkeit nachweisen können. Selbst bei $p = 0,9$ würde ihre Behauptung nur mit Wahrscheinlichkeit 0,59 akzeptiert. Man sieht so durch Betrachtung der Gütefunktion, ob der Test überhaupt in der Lage ist, eine möglicherweise vorhandene Abweichung von der Nullhypothese aufzuspüren. Ist etwa $p = 0,6$, so ist erst bei $n = 42$ zu klassifizierenden Tassenpaaren die Wahrscheinlichkeit wenigstens $1/3$, dass die Behauptung der Lady akzeptiert wird. (Das kleinste t mit $\beta(0,5 | t, 42) \leq 0,05$ ist $t = 27$.) Bei dem Ansatz von Fisher wäre es unklar, nach welchen Kriterien man entscheiden soll, wie viele Tassen Tee die Lady kosten muss.

Die Wahrscheinlichkeit eines Fehlers zweiter Art, also $1 - \beta(p | t, n)$, hängt offenbar von $p \in K$ ab. Hat der Test zu t und n das Niveau α , so gilt aus Stetigkeitsgründen für $p \rightarrow 1/2$

$$\lim(1 - \beta(p | t, n)) = 1 - \beta(0,5 | t, n) \geq 1 - \alpha.$$

Man kann daher auf keinen Fall — egal wie man n und t wählt — erreichen, dass für *alle* $p \in K$ die Wahrscheinlichkeit $1 - \beta(p | t, n)$ eine vorgegebene Schranke wie z.B. $\alpha' = 0,1$ unterschreitet. Man muss sich Gedanken darüber machen, wie große Abweichungen von der Hypothese man noch aufspüren will. Legt man z.B. Wert darauf, dass für alle $p \geq 0,7$ die Wahrscheinlichkeit eines Fehlers zweiter Art $\leq 0,1$ ist, so muss n einen bestimmten Mindestwert haben. Will man das Gleiche sogar für alle $p \geq 0,6$ erreichen, so ist dieser Mindestwert erheblich größer, und man muss abwägen, ob die höhere Zuverlässigkeit den höheren Aufwand rechtfertigt.

In diesem Beispiel bleibt ein wichtiger Vorteil des Vorgehens von Neyman und Pearson noch verborgen. Unser Bestreben, die Grundideen an einfachen Beispielen zu erklären, hat nämlich hier den Nachteil, dass die Wahl des Verwerfungsbereichs auf der Hand liegt.

Wer würde schon auf die Idee kommen, man würde sich seltener irren, wenn man als Teststatistik z.B. die Länge der längsten ununterbrochenen Trefferfolge verwenden würde? In komplizierteren Testproblemen ist es aber durchaus nicht immer offensichtlich, welche Teststatistik gewählt werden sollte. Wir werden sehen, dass der Ansatz von Neyman und Pearson es auch ermöglicht, verschiedene Teststatistiken zu vergleichen.

6.4 Ein verfeinertes Modell für den Tee-Test*

Zum Abschluss soll noch auf einen möglichen Einwand gegen Neymans Ansatz eingegangen werden, der ihm offenbar selbst bewusst war. Wer sagt uns, dass die verschiedenen von der Lady vorgenommenen Klassifizierungen wirklich unabhängig voneinander sind? Was wäre die Folge, wenn sie es nicht sind? Dies könnte bedeuten, dass die wahre Verteilung von X weder zur Hypothese $\{P_\vartheta : \vartheta \in H\}$ noch zur Alternative $\{P_\vartheta : \vartheta \in K\}$ gehört, und dass man eine größere Familie von Verteilungen hätte berücksichtigen müssen.

Man könnte die Benutzung eines nicht hinreichend realistischen Modells vielleicht einprägsam einen **Fehler dritter Art** nennen, damit man sich dieser Gefahr stets bewusst bleibt.

Würde durch Abhängigkeiten innerhalb der Folge der Entscheidungen der Lady die Binomialverteilung der Zahl X der richtigen Klassifizierungen auch in dem Fall gestört, dass die Lady die Unterschiede nicht schmecken kann, so hätten hohe Erfolgsquoten keine im Sinne von Irrtumswahrscheinlichkeiten quantifizierbare Beweiskraft. Man könnte argumentieren, dass die Verteilung der Hypothese nicht zutrifft *obwohl* die Lady die behauptete Fähigkeit nicht hat. Die Verwerfung der Hypothese wäre also nicht unbedingt so zu deuten, dass die Behauptung der Lady anerkannt werden sollte.

Offenbar wollte Neyman durch die zeitlichen Abstände zwischen den Teilversuchen diesem Einwand vorbeugen. Aber das wäre gar nicht nötig gewesen. Wir wollen uns nun überlegen, dass die Entscheidungen der Lady auch untereinander abhängig sein dürfen. Wir brauchen dazu ein Modell, in dem die Folge der Aussagen der Lady vorkommt:

Sei $\mathcal{X}$ nun die Menge der Folgen

$$x = (y_1, z_1, y_2, z_2, \dots, y_n, z_n)$$

mit $y_i, z_i \in \{1, 2\}$. Für $i = 1, \dots, n$ soll $y_i = k$ bedeuten, dass die im i -ten Teilexperiment zuerst gereichte Tasse vom Typ k ist. $z_i = h$ soll bedeuten, dass die Lady die im i -ten Teilexperiment zuerst gereichte Tasse als vom Typ h klassifiziert. Ein Treffer im i -ten Teilexperiment liegt also genau dann vor, wenn $y_i = z_i$ ist.

Sei $Y_i(x) = y_i$, $Z_i(x) = z_i$, $Y = (Y_1, \dots, Y_n)$ und $Z = (Z_1, \dots, Z_n)$. Durch die Versuchsanordnung ist sichergestellt, dass die Y_i untereinander unabhängig sind mit $P(Y_i = 1) = P(Y_i = 2) = 1/2$. Will man für die Folge der Entscheidungen der Lady beliebige Abhängigkeiten und Präferenzen zulassen, so kann Z eine *beliebige* Verteilung auf $\{1, 2\}^n$ haben. Die Hypothese wird zweckmäßig durch die Bedingung beschrieben, dass Z unabhängig von Y ist. Gibt es nämlich keinerlei feststellbare Geschmacksunterschiede zwischen den beiden Arten der Zubereitung, so dürfte es auch keinen Zusammenhang stochastischer Art zwischen der Folge Y und der Folge Z geben.

Wir wählen daher als Familie $\{P_\vartheta : \vartheta \in \Theta\}$ die Menge aller Verteilungen von X , für die die Zufallsvariablen $Y_1, \dots, Y_n$ unabhängig sind mit $P_\vartheta(Y_i = 1) = P_\vartheta(Y_i = 2) = 1/2$. $\{P_\vartheta : \vartheta \in H\}$ sei darin die Teilmenge der Wahrscheinlichkeitsmaße, für die Y und Z unabhängig sind. Der Parameter ϑ hat jetzt natürlich nicht mehr so eine anschauliche

Interpretation wie die Erfolgswahrscheinlichkeit. Aber es hindert uns nichts, als Θ die Menge der eben beschriebenen Verteilungen zu wählen und für eine Verteilung ϑ auch P_ϑ zu schreiben.

(Wir haben der Einfachheit halber angenommen, dass die Lady bis zum Abschluss des Experiments nichts über die Y_i erfährt. Dies steckt in der Formulierung der Hypothese. Die Unabhängigkeit von Y und Z besagt unter anderem, dass Z_{i+1} von $Y_1, \dots, Y_i$ unabhängig ist.)

Es fehlt noch die Angabe des Verwerfungsbereichs. Sei $T_i(x) = 1$ falls im i -ten Telexperiment ein Treffer erzielt wird (also im Fall $y_i = z_i$), und sonst sei $T_i(x) = 0$. Dann ist $T_i(x) = y_i - z_i + 1 \pmod{2}$. $T(x) = T_1(x) + \dots + T_n(x)$ ist die Zahl der Treffer. Unter der Hypothese gilt für beliebiges $(u_1, \dots, u_n) \in \{0, 1\}^n$

$$\begin{aligned} P_\vartheta(T_1 = u_1, \dots, T_n = u_n) &= \sum_{z \in \{1, 2\}^n} P_\vartheta(Z = z, Y_i = z_i - u_i + 1 \pmod{2}, i = 1, \dots, n) \\ &= \sum_{z \in \{1, 2\}^n} P_\vartheta(Z = z) P_\vartheta(Y_i = z_i - u_i + 1 \pmod{2}, i = 1, \dots, n) \\ &= \sum_{z \in \{1, 2\}^n} P_\vartheta(Z = z) 2^{-n} = 2^{-n}. \end{aligned}$$

Unter der Hypothese ist also $(T_1, \dots, T_n)$ Bernoulli-verteilt mit $p = 1/2$. Also ist unter der Hypothese die Zahl der Treffer binomialverteilt mit $p = 1/2$, genau wie bei Neyman, nur dass jetzt T die Bezeichnung für die Anzahl der Treffer ist. Als Verwerfungsbereich können wir also wieder $R = \{x \in \mathcal{X} : T(x) \geq t\}$ wählen, wobei wieder t minimal unter der Nebenbedingung gewählt wird, dass das Niveau $\leq \alpha$ sein soll. In der Ausführung unterscheidet sich der Test also überhaupt nicht von Neymans Vorschlag.

Als Nachteil unseres Vorgehens mag es erscheinen, dass es nun praktisch unmöglich ist, für alle $\vartheta \in K$ die Macht $\beta(\vartheta)$ anzugeben. Für gewisse $\vartheta \in K$ wird $\beta(\vartheta)$ sogar kleiner als α sein. Aber das ist nicht schlimm. Es war bequem, eine so große Menge von Alternativen zuzulassen, aber wir sind hauptsächlich daran interessiert, dass $\beta(\vartheta)$ für solche $\vartheta \in K$ groß ist, bei denen hohe Trefferzahlen wahrscheinlich werden. Man könnte sich zur Überprüfung, ob die Lady eine echte Chance hat, durchaus auf solche ϑ beschränken, unter denen T binomialverteilt mit $p \geq 0,6$ ist.

Wenn das Modell die Verteilungen unter der *Alternative* nicht vollständig enthält, ist dies in der Regel weniger problematisch. Die eigentliche Gefahr eines Fehlers dritter Art liegt darin, dass die Modell-Hypothese H nicht alle Verteilungen enthält, die zu einer Modellierung der „realen“ Hypothese passen könnten. (Oben war dies die Hypothese, dass die Lady die Unterschiede nicht schmecken könne.)

Durch die Betrachtung des verfeinerten Modells brauchen wir uns keine Sorgen über einen Fehler dritter Art mehr zu machen. In Neymans Modell hätte ein Skeptiker einwenden können, es sei wegen der potenziellen Abhängigkeiten gar kein Test zum Niveau α ausgeführt worden. Solche Skeptiker existieren in großer Zahl in einem sehr ähnlichen Fall:

6.5 Beispiel des Testens der Existenz von außersinnlicher Wahrnehmung*

Es soll getestet werden, ob eine Versuchsperson über außersinnliche Wahrnehmung (ASW) verfügt, speziell über „Präkognition“, d.h., ob sie zukünftige Ereignisse im Voraus wahrnehmen kann. In jedem Teilversuch nennt die Versuchsperson eins der Symbole 1, 2, und zwar das, von dem sie annimmt, dass es als Nächstes kommt. Nachdem das im i -ten Versuch genannte Symbol Z_i notiert ist, wird eine Münze geworfen. Kommt Kopf, so registrieren wir $Y_i = 1$, sonst $Y_i = 2$. Die Folge der Y_i nennt man Zielfolge und die der Z_i Ratefolge. Die Hypothese, dass keine ASW existiert, kann durch die Annahme der Unabhängigkeit von $Y = (Y_1, \dots, Y_n)$ und $Z = (Z_1, \dots, Z_n)$ beschrieben werden. Das Modell aus Abschnitt 6.4 passt also auch hier. Das ASW-Experiment ist daher mathematisch äquivalent zu dem Experiment mit der „tea tasting Lady“. Kritiker, die wegen der Abhängigkeiten innerhalb der Ratefolge Zweifel äußern, ob solch ein Test mathematisch korrekt ist, haben offenbar Unrecht.

Auch für andersartige Versuchsanordnungen für ASW-Versuche lassen sich Tests ableiten. Eine erste kleine Komplikation liegt beispielsweise darin, dass die Versuche meist mit *feedback* ausgeführt werden, d.h., der Versuchsperson wird vor Nennung von Z_{i+1} gesagt, was Y_i war. Dies erfordert aber nur eine modifizierte Hypothese. Um sie zu formulieren, benötigen wir den Begriff der bedingten Unabhängigkeit. Man nennt ein Ereignis A **bedingt unabhängig** von B bei gegebenem C , wenn A von B unter der bedingten Wahrscheinlichkeit $P(\cdot | C)$ unabhängig ist, also wenn $P(A \cap B | C) = P(A | C) P(B | C)$ ist.

Die passende neue Hypothese ist: Für jedes $i \leq n$ ist Z_i bedingt unabhängig von $(Y_i, \dots, Y_n)$ gegeben $(Y_1, Z_1, \dots, Y_{i-1}, Z_{i-1})$. Die neue Hypothese ist also die Menge der P_ϑ mit der folgenden Eigenschaft: Für alle i mit $1 \leq i \leq n$ und für alle $z = (z_1, \dots, z_n)$ und $y = (y_1, \dots, y_n)$ gilt

$$P_\vartheta(Z_i = z_i, Y_i = y_i, \dots, Y_n = y_n | C) = P_\vartheta(Z_i = z_i | C) P_\vartheta(Y_i = y_i, \dots, Y_n = y_n | C),$$

wobei

$$C = \begin{cases} \{Y_1 = y_1, Z_1 = z_1, \dots, Y_{i-1} = y_{i-1}, Z_{i-1} = z_{i-1}\} & \text{für } i \geq 2, \\ \mathcal{X} & \text{für } i = 1. \end{cases}$$

Es lässt sich zeigen, dass auch unter dieser Hypothese die Trefferzahl $T(X)$ binomialverteilt mit $p = 1/2$ ist.

Es werden auch Versuchsanordnungen benutzt, deren mathematische Behandlung aufwändiger ist, nämlich solche mit abhängigen *Zielfolgen*. Ein klassisches Experiment verwendet gemischte Kartendecks mit 25 Karten, auf denen 5 verschiedene Symbole je fünfmal vorkommen. Sei $Y_1, \dots, Y_{25}$ die Folge der nacheinander aufgedeckten Symbole vom ersten Deck, $Y_{26}, \dots, Y_{50}$ die vom zweiten Deck usw. Dies ist die Zielfolge. Es stellt sich heraus, dass die Verteilung der Trefferzahl auch unter der Hypothese (selbst ohne feedback) von der unbekannten Verteilung der Ratefolge Z abhängt und damit prinzipiell nicht ermittelbar ist.

Dennoch lassen sich Tests zu vorgegebenem Niveau angeben. Man kann nämlich zu jedem $z = (z_1, \dots, z_n)$ unter der Hypothese der Unabhängigkeit von Y und Z die *bedingte* Verteilung der Trefferzahl $T(X)$ bei gegebenem $\{Z = z\}$ bestimmen, und diese wird im Fall $\{Z = z\}$ benutzt. Praktisch arbeitet man mit Approximationen durch die Normalverteilung (vgl. Greenwood-Stuart (1940), Krengel-Liese (1978)).

Es liegen äußerst eindrucksvolle Testergebnisse für die Existenz von ASW vor. Das Hauptproblem ist (obwohl auch einzelne mathematische Fehler gemacht worden sind)

nicht die mathematische Korrektheit der Tests, sondern die mangelnde Wiederholbarkeit und die Beeinträchtigung durch Kontrollen.

In einem Artikel in *Science* (1978) hat der angesehene Statistiker und Zauberkünstler P. Diaconis sich sehr kritisch zur Praxis der ASW-Experimente geäußert. Andererseits hat das Argument Gewicht, man dürfe nicht pauschal in einem Forschungsbereich alle positiven Resultate für wertlos erklären, indem man durchweg Manipulationen und mangelnde Kontrollen unterstellt. Ein Konsens scheint nicht in Sicht.

6.6 Eine Erweiterung des Testbegriffs: Randomisierte Tests

Um ein vorgegebenes Signifikanzniveau voll ausnutzen zu können, haben sich die Mathematiker einen kleinen Trick ausgedacht, mit dem man gelegentlich die Wahrscheinlichkeit eines Fehlers zweiter Art noch etwas verkleinern kann.

Nehmen wir einmal an, die beobachtete Größe X sei binomialverteilt mit $n = 5$, die Hypothese sei $p = 1/2$ und die Alternative sei $p = 3/4$. Die Hypothese soll zum Niveau $\alpha = 0,05$ getestet werden. Große Werte von X sprechen gegen die Hypothese. Wählt man $R = \{5\}$, so wird die Hypothese nur im Fall $X = 5$ verworfen. Die Wahrscheinlichkeit eines Fehlers erster Art ist $P_{1/2}(R) = 2^{-5} = 1/32$, also kleiner als nötig. Wählt man aber $R = \{4, 5\}$, so ist $P_{1/2}(R) = (1 + \binom{5}{4}) 2^{-5} = \frac{6}{32} > 0,05$. Man kann also den Wert $x = 4$ nicht zum Verwerfungsbereich dazunehmen.

Der Trick besteht nun darin, dass man bei Beobachtung von $x = 4$ manchmal die Hypothese akzeptiert und sie manchmal verwirft. Genauer: Wird $x = 5$ beobachtet, so wird die Hypothese verworfen, wird $x \in \{0, 1, 2, 3\}$ beobachtet, so wird die Hypothese akzeptiert, und wird $x = 4$ beobachtet, so stellt man ein zusätzliches Zufallsexperiment an, das mit der Beobachtung von X und also mit der Erfolgswahrscheinlichkeit p absolut nichts zu tun hat. Dieses Zufallsexperiment habe 0 und 1 als mögliche Ausgänge. Ist das Ergebnis 1, so wird H verworfen und sonst akzeptiert. Bezeichnet man in diesem Zufallsexperiment die Wahrscheinlichkeit einer 1 mit $\varphi(4)$, so ist im Fall $p = 1/2$ die Verwerfungswahrscheinlichkeit insgesamt

$$P_{1/2}(5) + P_{1/2}(4) \cdot \varphi(4) = \frac{1}{32} (1 + 5 \varphi(4)) .$$

Damit das erlaubte Niveau voll ausgeschöpft wird, setzt man $(1 + 5 \varphi(4))/32 = \alpha = 1/20$. Dann ist $\varphi(4) = 3/25$.

Verglichen mit dem Test mit $R = \{5\}$ ist die Verwerfungswahrscheinlichkeit sowohl für $p = 1/2$ wie für $p = 3/4$ offenbar nun etwas vergrößert, und damit ist die Wahrscheinlichkeit eines Fehlers zweiter Art verkleinert.

Natürlich wäre es besser gewesen im Fall $x = 4$ weitere Daten zu beschaffen, die etwas mit p zu tun haben, aber man geht davon aus, dass solche nicht verfügbar sind.

Im Folgenden werde ein **Test** durch eine Funktion $\varphi : \mathcal{X} \rightarrow [0, 1]$ beschrieben. Wird x beobachtet, so soll die Hypothese mit Wahrscheinlichkeit $\varphi(x)$ verworfen werden. Unsere alten durch einen Verwerfungsbereich R gegebenen Tests ergeben sich, wenn wir für $x \in R$ $\varphi(x) = 1$, und für $x \in R^c$ $\varphi(x) = 0$ setzen. Die **Gütefunktion** hat jetzt die Form $\beta(\vartheta) = E_{\vartheta}(\varphi) = \sum_x \varphi(x) P_{\vartheta}(x)$, denn mit Wahrscheinlichkeit $P_{\vartheta}(x)$ wird x beobachtet und dann wird H mit Wahrscheinlichkeit $\varphi(x)$ verworfen.

Für die Praxis hat diese Erweiterung des Testbegriffs keine Bedeutung. Ehe man sich die Mühe macht, noch zusätzliche Zufallsexperimente durchzuführen, die der Entscheidung immer etwas Willkürliches geben würden, schöpft man lieber das erlaubte Signifikanzniveau nicht aus. In der Regel ist es ja ohnehin etwas willkürlich festgelegt worden.

Aber für die mathematische Theorie hat die Betrachtung randomisierter Tests einen gewissen Nutzen.

6.7 Tests einfacher Hypothesen gegen einfache Alternativen

Eine Hypothese oder Alternative heißt **einfach**, wenn sie aus einer einzigen Verteilung besteht, andernfalls nennt man sie **zusammengesetzt**. Wir betrachten jetzt das Problem, eine einfache Hypothese P_H gegen eine einfache Alternative P_K zu testen. Die Frage ist, ob es unter allen Tests φ mit Niveau α (also mit $E_H(\varphi) = \sum \varphi(x) P_H(x) \leq \alpha$) einen gibt, der die Macht $E_K(\varphi)$ maximiert, und wie er aussieht.

Wenn wir für ein x entscheiden, dass $\varphi(x) = 1$ gelten soll, ergibt dieses x einen Beitrag $P_H(x)$ zum Niveau und einen Beitrag $P_K(x)$ zur Macht. Fassen wir $P_H(x)$ als Kosten auf, mit denen wir die Macht $P_K(x)$ einkaufen können, indem wir $\varphi(x) = 1$ setzen, so haben wir den Betrag α zur Verfügung und werden möglichst solche Punkte x einkaufen, für die die Macht pro Kosteneinheit, also $q(x) = P_K(x)/P_H(x)$ groß ist. Dies legt es nahe, die folgende Klasse von Tests zu betrachten:

Definition 6.1 Ein Test φ^* heißt **Neyman-Pearson-Test**, wenn eine Konstante c^* mit $0 \leq c^* \leq \infty$ existiert mit

$$\varphi^*(x) = \begin{cases} 1, & \text{falls } P_K(x) > c^* P_H(x), \\ 0, & \text{falls } P_K(x) < c^* P_H(x). \end{cases} \quad (6.1)$$

Auf der Menge der x mit $P_K(x) = c^* P_H(x)$ soll φ^* beliebige Werte $\gamma(x)$ mit $0 \leq \gamma(x) \leq 1$ annehmen dürfen.

Ein Test φ_1 heißt **schärfer** als φ_2 , wenn $E_K(\varphi_1) > E_K(\varphi_2)$ ist.

Wir wollen annehmen, dass für jedes x $P_H(x) + P_K(x) > 0$ ist. Die übrigen Ergebnisse x haben offenbar keinerlei Einfluss auf die Irrtumswahrscheinlichkeiten. (Damit ist $q(x)$ wohldefiniert.)

Satz 6.2 (Neyman-Pearson-Lemma) *Für das Testen einer einfachen Hypothese gegen eine einfache Alternative gilt:*

- (i) *Ist φ^* Neyman-Pearson-Test, so ist φ^* mindestens so scharf wie alle anderen Tests φ mit $E_H(\varphi) \leq E_H(\varphi^*)$.*
- (ii) *Zu $0 \leq \alpha \leq 1$ existiert ein Neyman-Pearson-Test φ^* mit $E_H(\varphi^*) = \alpha$. (Dabei kann man φ^* auf der Menge $\{x : P_K(x) = c^* P_H(x)\}$ konstant wählen.)*

Beweis: (i) Auf $A := \{x : \varphi^*(x) > \varphi(x)\}$ ist $\varphi^*(x) > 0$ und damit $P_K(x) \geq c^* P_H(x)$. Auf $B := \{x : \varphi^*(x) < \varphi(x)\}$ ist $\varphi^*(x) < 1$ und damit $P_K(x) \leq c^* P_H(x)$. Es folgt:

$$\begin{aligned} E_K(\varphi^*) - E_K(\varphi) &= \sum_{x \in \mathcal{X}} (\varphi^*(x) - \varphi(x)) P_K(x) \\ &= \sum_{x \in A} (\varphi^*(x) - \varphi(x)) P_K(x) + \sum_{x \in B} (\varphi^*(x) - \varphi(x)) P_K(x) \\ &\geq \sum_{x \in A} (\varphi^*(x) - \varphi(x)) c^* P_H(x) + \sum_{x \in B} (\varphi^*(x) - \varphi(x)) c^* P_H(x) \\ &= c^* \sum_{x \in \mathcal{X}} (\varphi^*(x) - \varphi(x)) P_H(x) = c^* (E_H(\varphi^*) - E_H(\varphi)) \geq 0. \end{aligned}$$

(ii) Für $\alpha = 0$ setze $c^* = \infty$. Ist dann $P_H(x) > 0$, so ist $P_K(x) < c^* P_H(x)$ und damit $\varphi^*(x) = 0$. Es folgt $E_H(\varphi^*) = 0$.

Sei also $\alpha > 0$. Wir setzen für $c \geq 0$

$$\alpha(c) := P_H(q(X) > c) \quad \text{und} \quad \alpha(c-0) := P_H(q(X) \geq c).$$

Offenbar ist $\alpha(0-0) = 1$, und $\alpha(\cdot)$ ist fallend. Wir verwenden nun die Stetigkeitsaussagen aus § 2.5 für P_H . Für strikt wachsende Folgen (c_n) sei

$$C_n = \{x : P_H(x) > 0 \text{ und } q(x) > c_n\}.$$

Die Folge (C_n) ist fallend. Strebt (c_n) gegen ∞ , so ist der Durchschnitt C der C_n leer. Es folgt $\alpha(c_n) \rightarrow P_H(C) = 0$ für $c_n \rightarrow \infty$. Strebt (c_n) gegen ein $c > 0$, so ist

$$C = \{x : P_H(x) > 0 \text{ und } q(x) \geq c\}.$$

Also gilt $\alpha(c_n) \rightarrow \alpha(c-0)$.

Ist (b_n) eine strikt gegen b fallende Folge, so ist die durch

$$B_n = \{x : P_H(x) > 0 \text{ und } q(x) > b_n\}$$

definierte Folge (B_n) wachsend. Ihre Vereinigung ist $B = \{x : P_H(x) > 0, q(x) > b\}$. Daher ist $\alpha(\cdot)$ rechtsstetig.

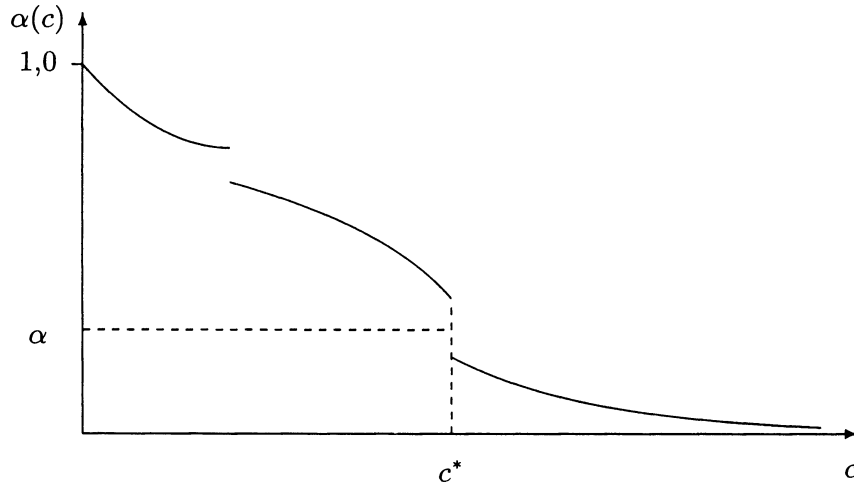


Bild 6.2 Bestimmung des kritischen Wertes für einen Test zu gegebenem Niveau

Sei nun $c^* = \inf\{c : \alpha(c) \leq \alpha\}$. Dann ist $\alpha(c^*-0) \geq \alpha \geq \alpha(c^*)$. Im Fall $\alpha(c^*-0) = \alpha(c^*)$ sei $\gamma^* = 0$. Im Fall $\alpha(c^*-0) > \alpha(c^*)$ sei

$$\gamma^* = \frac{\alpha - \alpha(c^*)}{\alpha(c^*-0) - \alpha(c^*)}.$$

Auf $\{x : P_K(x) = c^* P_H(x)\}$ sei $\varphi^*(x) = \gamma^*$ und auf dem Komplement dieser Menge sei φ^* durch (6.1) definiert. Dann ist

$$\begin{aligned} E_H(\varphi^*) &= P_H(q(X) > c^*) + \gamma^* P_H(q(X) = c^*) \\ &= \alpha(c^*) + \gamma^* (\alpha(c^*-0) - \alpha(c^*)) = \alpha. \end{aligned}$$

□

6.8 Anwendung auf zusammengesetzte Alternativen

Das gerade behandelte Problem, eine einfache Hypothese gegen eine einfache Alternative zu testen, taucht in der Praxis so gut wie nie auf. Aber das Neyman-Pearson-Lemma lässt sich oft mit Erfolg auch auf zusammengesetzte Hypothesen oder Alternativen anwenden. Wir wollen das an dem Beispiel der „tea tasting Lady“ illustrieren.

Sei also wieder X binomialverteilt mit Parametern n, p . Die Hypothese sei $p = 1/2$ und die Alternative sei $p > 1/2$.

Sei zuerst $p_H = 1/2$ und $1/2 < p_K < 1$ fest gewählt. Als Funktion von x ist

$$q(x) = \frac{P_K(x)}{P_H(x)} = \frac{\binom{n}{x} p_K^x (1 - p_K)^{n-x}}{\binom{n}{x} p_H^x (1 - p_H)^{n-x}}$$

strikt wachsend. Für jedes c ist also die Menge $\{x : q(x) > c\}$ ein Intervall $\{t, t+1, \dots, n\}$. (Ist die Menge $\{x : q(x) > c\}$ leer, so setzen wir $t = n + 1$.) Für einen optimalen Test φ existiert also ein t mit $\varphi(x) = 1$ für $x \geq t$ und $\varphi(x) = 0$ für $x < t - 1$. Im Fall $q(t-1) < c$ muss auch $\varphi(t-1) = 0$ sein. Im Fall $q(t-1) = c$ ist $\varphi(t-1)$ eine beliebige Zahl $\gamma \in [0, 1]$. Das Niveau dieses Tests ist

$$E_H(\varphi) = P_H(X \geq t) + \gamma P_H(X = t - 1).$$

Ist nun $\alpha > 0$ vorgegeben und soll ein *schärfster* Test zum Niveau α gefunden werden, so bestimmt man t und γ aus der Gleichung $E_H(\varphi) = \alpha$. (So sind wir auch schon bei der Einführung des Begriffs des randomisierten Tests vorgegangen, aber jetzt wissen wir, dass der Test *schärfster* Test zum Niveau α ist.) In der Gleichung $E_H(\varphi) = \alpha$ kommt aber p_K gar nicht vor. Der Test ist also für *alle* $p_K > 1/2$ *schärfster* Test zum Niveau α von $p = p_H$ gegen $p = p_K$. Das heißt, dass er auch *schärfster* Test gegen die *zusammengesetzte* Alternative $K = \{p : p > 1/2\}$ ist.

Wir brauchen also in der Tat für das Experiment mit der „tea tasting Lady“ nur Tests in Betracht zu ziehen, die die Hypothese verwerfen, wenn die Trefferzahl eine Schranke t übertrifft.

Auch für viele andere Verteilungen, etwa die Poisson-Verteilung, lässt sich die Monotonie von $q(x)$ zeigen, so dass man auch für sie optimale Tests der zusammengesetzten Alternativen finden kann. Auch zusammengesetzte *Hypothesen* geeigneter Form kann man zulassen, oben z.B. die Hypothese $\{p : p \leq 1/2\}$, denn der abgeleitete Test hält sogar für diese größere Hypothese das Niveau α ein.

6.9 Allgemeine Hinweise zur Testtheorie

All unsere Überlegungen setzten voraus, dass die Beobachtung X erst angestellt wird, wenn alles andere festliegt. Die Versuchung liegt manchmal nahe, aus bestimmten Daten Vermutungen abzuleiten und die gleichen Daten zum „Beweis“ dieser Vermutungen zu nutzen. Jemand der in einer bestimmten Schule große Erfolge mit der Ganzheitsmethode erzielt hat, und dann *diese* Daten nutzt, um die Hypothese zu verwerfen, dass die Buchstabiermethode mindestens so gut sei wie die Ganzheitsmethode, arbeitet statistisch fehlerhaft. Wenn die „tea tasting Lady“ über einen längeren Zeitraum Geschmacksproben durchführt und ihre Resultate im August eine besonders hohe Trefferquote aufweisen, kann man auch nicht diese ausgewählten Daten nutzen, um ihre Fähigkeit zu beweisen.

Auch bei der Auswertung von eigens für ein bestimmtes Testproblem erhobenen Daten kann man noch ähnliche Fehler machen: In komplexeren Fragestellungen gibt es oft mehrere Teststatistiken $T_1(X), T_2(X), \dots, T_k(X)$, die in Betracht kommen und von denen jede gewisse Vorzüge gegenüber den anderen aufweist. Nicht immer gibt es schließlich so nahe liegende Kriterien wie die Trefferzahl. Wählt man die Teststatistik T_i als Kriterium, so könnte der Verwerfungsbereich die Form $R_i = \{x : T_i(x) \geq t_i\}$ für geeignetes t_i haben. Wissenschaftler weisen gerne „signifikante“ Ergebnisse vor, d.h. solche, bei denen die Nullhypothese verworfen wird. So kommt es vor, dass nach Beobachtung von x der Verwerfungsbereich R_i gewählt wird, für den x signifikant ist. Ist jeder der Verwerfungsbereiche R_i so gewählt, dass für ihn die Fehlerwahrscheinlichkeit erster Art maximal α ist, so könnte bei diesem Vorgehen die Fehlerwahrscheinlichkeit evtl. bis zu $k\alpha$ sein.

Bei sauberer Arbeitsweise sind die folgenden *Phasen in der Bearbeitung eines Testproblems* zu trennen:

Phase 1: Vorüberlegungen

Man fragt sich: Was soll getestet werden? Welches Experiment und welche aus den Daten berechenbare Größe könnte Aufschluss über die Gültigkeit der realen Hypothese geben? Welches Modell ist geeignet? Welche Teststatistik sollte man wählen? Hier, wie überall in der Statistik, ist der gesunde Menschenverstand und die Rückkopplung mit den Fachleuten aus dem Gebiet, aus dem die Fragestellung stammt, von entscheidender Bedeutung.

Zu den Vorüberlegungen gehört auch noch die Wahl des Signifikanzniveaus. Üblich sind Werte wie $\alpha = 0,05$, $\alpha = 0,02$ oder $\alpha = 0,01$, gelegentlich auch $\alpha = 0,001$. Bei der Wahl von α sollte überlegt werden, wie gravierend ein Fehler erster Art im Vergleich zu einem Fehler zweiter Art ist und wie sich die Festlegung von α auf den Test und damit auf die Gütefunktion auswirkt.

Will man die Überlegenheit einer Therapie A gegenüber einer Therapie B zeigen, so wird man ein kleines α wählen, um sich im Fall der Ablehnung der Hypothese „ B ist mindestens so gut wie A “ mit der Schlussfolgerung „ A ist besser als B “ möglichst sicher zu fühlen. Will man bei einer Untersuchung auf Nebenwirkung eines Medikaments die Hypothese „keine Nebenwirkung“ testen, so wird in der Praxis sogar $\alpha = 0,2$ benutzt, da ein Fehler zweiter Art schwer wiegende Konsequenzen haben könnte.

Wenn die Macht des Tests nicht ausreicht, sollte man bereit sein, die vorgesehene Zahl n der Messungen zu erhöhen oder auch sonst das Experiment abzuändern. Erst nachdem der Verwerfungsbereich festgelegt ist, folgen

Phase 2: Erhebung der Daten

Phase 3: Auswertung der Daten und Interpretation

In der Praxis ähneln sich natürlich viele Testprobleme, so dass man, statt jedes Mal Phase 1 zu durchlaufen, lieber auf Bewährtes zurückgreift.

6.10 p -Werte*

Es kann sein, dass sich in ein und demselben Testproblem ein Statistiker für das Niveau 0,05 entscheidet und ein anderer für das Niveau 0,01. Die gleiche Beobachtung x kann dann bei dem ersten Statistiker zur Verwerfung von H führen und bei dem zweiten zur Annahme von H , selbst wenn beide die gleiche Teststatistik T heranziehen. Viele Praktiker geben daher gerne bei Beobachtung von x die kleinste Zahl α an, für die die Wahl von α als Niveau gerade noch zur Verwerfung führt. Mathematische Statistiker betrachten

dieses Vorgehen oft kritisch, oder sie ignorieren diese Praxis, da ja α vorgegeben werden muss, wenn der Test und damit die Irrtumswahrscheinlichkeiten wohldefiniert sein sollen. Wir wollen aber trotzdem noch etwas darauf eingehen:

Wählt man t als kritischen Wert und also $R_t = \{x : T(x) \geq t\}$ als Verwerfungsbereich, so ist

$$\alpha_t = \sup\{P_{\vartheta}(R_t) : \vartheta \in H\}$$

die kleinste Zahl α , für die dieser Test noch das Niveau α hat. (Man nennt sie auch **effektives Niveau**). $\alpha_{T(x)}$ heißt das **beobachtete Signifikanzniveau** oder der **p-Wert** zur Beobachtung x . Gilt für das vorgegebene α die Ungleichung $\alpha \geq \alpha_{T(x)}$, und hat man zum Niveau α den kritischen Wert t so bestimmt, dass der Verwerfungsbereich möglichst groß und damit t möglichst klein ist, so ist $t \leq T(x)$. Das heißt im Fall $\alpha \geq \alpha_{T(x)}$ kommt es zur Verwerfung von H . Im Fall $\alpha < \alpha_{T(x)}$ überlegt man sich analog, dass die Beobachtung x zur Annahme von H führt. (Randomisierung schließen wir hier aus.)

Man kann also die jetzt bei statistischen Auswertungen von Computern berechneten p -Werte als Entscheidungsanweisungen für den Statistiker auffassen, der α fest gewählt hat. Je nach Wahl von α wird die Anweisung zu verschiedenen Entscheidungen führen.

Kritisch an der Verwendung von p -Werten ist vor allem, dass sie leicht fehlinterpretiert werden. Völliger Unsinn wäre es, $\alpha_{T(x)}$ als die Wahrscheinlichkeit zu betrachten, dass die Hypothese richtig ist. Nicht ganz so offensichtlich ist der Fehler, im Falle $\alpha_{T(x)} = 0,023$ zu schließen, nun sei H zum Niveau 0,023 abzulehnen. Das Niveau soll ja nicht vom Ergebnis x abhängen.

Ist die Hypothese einfach, so ist die folgende Interpretation richtig: $\alpha_{T(x)}$ ist die Wahrscheinlichkeit (unter H) dafür, dass $T(X) \geq T(x)$ ist, also die Wahrscheinlichkeit, dass die Teststatistik einen mindestens so großen Wert annimmt wie den aktuell beobachteten.

Ein Vorteil der p -Werte ist es, dass man mit ihrer Hilfe verschiedene unabhängig voneinander durchgeführte Tests der gleichen Hypothese zusammenfassen kann. (Vgl. van Zwet und Osterhoff (1967).)

Aufgaben

1. Neymans Test der „tea tasting Lady“ werde mit $n = 6$ und $R = \{5, 6\}$ durchgeführt. Wie groß ist das (effektive) Niveau des Tests? Wie groß ist die Macht des Tests in $p = 0,6$?
2. Für welchen randomisierten Neyman-Pearson-Test der „tea tasting Lady“ mit $n = 6$ ist das Niveau 0,02? (L)
3. Die Hypothese, dass eine Zufallsvariable X zum Parameter $p_H = 2/3$ geometrisch verteilt ist, soll gegen die Alternative getestet werden, dass sie geometrisch mit $p < 2/3$ verteilt ist. Geben Sie möglichst mächtige randomisierte und nichtrandomisierte Tests zum Niveau $\alpha = 0,1$ an, und vergleichen Sie die Macht in $p = 1/3$.
4. Um zu testen, ob in einem Paket, das 100 Glühbirnen enthält, weniger als 10 defekte Glühbirnen enthalten sind, prüft ein Händler beim Kauf 10 der Birnen und nimmt das Paket nur an, wenn alle 10 in Ordnung sind. Beschreiben Sie das Verhalten des Händlers testtheoretisch. Welches ist das Niveau des Tests? (L)
5. Ein möglicherweise inhomogener Würfel wird zweimal geworfen. Es soll die Hypothese der Gleichverteilung gegen die Alternative getestet werden, dass der Würfel verfälscht ist. Ist φ ein beliebiger Test, so gilt für den Test $\bar{\varphi}$ mit $\bar{\varphi}(x_1, x_2) = (\varphi(x_1, x_2) + \varphi(x_2, x_1))/2$ die Gleichung $E_{\vartheta}(\bar{\varphi}) = E_{\vartheta}(\varphi)$ für alle ϑ . Wir beschränken uns daher auf Tests mit $\varphi(x_1, x_2) = \varphi(x_2, x_1)$ für alle $x = (x_1, x_2)$. Man zeige: Ist φ ein randomisierter Test, der die Hypothese verwirft, wenn die beiden Wurfsergebnisse gleich sind, so gibt es keinen Test $\varphi' \neq \varphi$ mit $E_H(\varphi') \leq E_H(\varphi)$ und $E_{\vartheta}(\varphi') \geq E_{\vartheta}(\varphi)$ für alle $\vartheta \in K$. (L)

6. Man beweise die Behauptung aus Abschnitt 6.5, dass die Trefferzahl binomialverteilt mit $p = 1/2$ ist, wenn für jedes i Z_i und $\{Y_i, \dots, Y_n\}$ bedingt unabhängig bei gegebenem $(Y_1, Z_1, Y_2, \dots, Z_{i-1})$ ist.
7. Für das verfeinerte Modell des Tee-Tests in Abschnitt 6.4 zeige man: Notwendig dafür, dass unter der Hypothese der Unabhängigkeit der Zielfolge von der Ratefolge, die Verteilung der Trefferzahl unabhängig von der Verteilung der Ratefolge ist, ist, dass die Zielfolge auf $\{1, 2\}^n$ gleichverteilt ist. (Wir hatten uns überlegt, dass die Bedingung hinreichend ist.)
8. Bestimmen Sie unter Verwendung der Normalapproximation den Verwerfungsbereich eines Tests zum Niveau 0,05, der bei 1000 Würfeln einer Münze überprüfen soll, ob die Münze fair ist.
9. $X = (X_1, \dots, X_n)$ sei Bernoulli-verteilt mit Erfolgswahrscheinlichkeit p . Wie groß muss n sein, damit ein Test der Hypothese $p = 0,2$ gegen die Alternative $p = 0,8$ existiert, der $E_H(\varphi) \leq 0,05$ und $E_K(\varphi) \geq 0,95$ erfüllt. Zeigen Sie mit Hilfe des schwachen Gesetzes der großen Zahl, dass für jedes $p_H \neq p_K$ ein Test φ der Hypothese $p = p_H$ gegen $p = p_K$ mit $E_H(\varphi) \leq 0,05$ und $E_K(\varphi) \geq 0,95$ existiert, wenn nur n hinreichend groß ist.
10. Sei X eine zum Parameter $\lambda > 0$ poissonverteilte Zufallsvariable.
 - (a) Geben Sie den schärfsten nichtrandomisierten Neyman-Pearson-Test für $H = \{\lambda = 2\}$ gegen $K = \{\lambda = 1/2\}$ zum Niveau $\alpha = 0,2$ an.
 - (b) Gibt es einen schärferen nichtrandomisierten Test, der ebenfalls das Niveau $\alpha = 0,2$ einhält?
(L)

§ 7 Erzeugende Funktionen*

Wir wollen nun ein einfaches und doch erstaunlich schlagkräftiges Hilfsmittel für das Studium von Verteilungen auf $\mathbb{Z}^+ = \{0, 1, 2, \dots\}$ kennen lernen.

Definition 7.1 Als **erzeugende Funktion** (der Verteilung) einer Zufallsvariablen X mit ganzzahligen nichtnegativen Werten bezeichnen wir die Potenzreihe

$$g(t) = \sum_{n=0}^{\infty} P(X = n) t^n.$$

Zur Unterscheidung von anderen erzeugenden Funktionen schreiben wir auch manchmal ausführlicher $g_X(t)$ statt $g(t)$. Offenbar ist $g(t) = E(t^X)$. Da die Koeffizienten nichtnegativ sind und ihre Summe 1 ist, konvergiert die Reihe mindestens für alle t mit $|t| \leq 1$. Ist $g^{(n)}(t)$ die n -te Ableitung von g an der Stelle t und $g^{(0)}(t) = g(t)$, so ist aufgrund der Eindeutigkeit der Potenzreihenentwicklung

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

Die Beziehung zwischen der erzeugenden Funktion von X und der Verteilung von X ist also umkehrbar eindeutig, so dass man die erzeugende Funktion als kompakte Beschreibung der Verteilung betrachten kann.

Beispiele: (a) Ist X Poisson-verteilt mit Parameter $\lambda > 0$, so ist $P(X = n) = e^{-\lambda} \lambda^n / n!$, also

$$g(t) = e^{-\lambda} \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} t^n = e^{-\lambda} e^{\lambda t} = e^{\lambda(t-1)}.$$

(b) Ist X geometrisch verteilt, so ist $P(X = 0) = 0$ und $P(X = n) = p(1-p)^{n-1}$ für $n \geq 1$, also

$$g(t) = \sum_{n=1}^{\infty} p(1-p)^{n-1} t^n = \frac{pt}{1 - (1-p)t}.$$

Satz 7.2 (i) Für $0 \leq t \leq 1$ ist g stetig, monoton wachsend und konvex, und es ist $g(0) = P(X = 0)$ und $g(1) = 1$.

(ii) Sei $k \geq 1$, und sei $g^{(k)}(1-)$ der linksseitige Grenzwert

$$g^{(k)}(1-) = \lim_{t \rightarrow 1-0} g^{(k)}(t).$$

Dann gilt

$$E(X(X-1)\dots(X-k+1)) = g^{(k)}(1-). \quad (7.1)$$

(Beide Seiten sind genau dann endlich, wenn $E(X^k)$ endlich ist.)

Beweis: (i) ist trivial. Zum Beweis von (ii) betrachten wir für $0 \leq t < 1$

$$g^{(k)}(t) = \sum_{n=0}^{\infty} n(n-1)\dots(n-k+1)P(X=n)t^{n-k}. \quad (7.2)$$

Die Glieder der Reihe sind durch $n^k P(X = n)$ beschränkt, und die Summe dieser Terme ist EX^k . Ist $EX^k < \infty$, so erhält man — wegen der gleichmäßigen Konvergenz der Reihe — die Gleichung (7.1) aus (7.2) durch Grenzübergang $t \rightarrow 1$. Ist $EX^k = \infty$, so ist auch $\sum_{n=2k}^{\infty} n^k P(X = n) = \infty$. Für $t \rightarrow 1$ folgt dann $g^{(k)}(t) \rightarrow \infty$, da für $n \geq 2k$

$$n(n-1) \dots (n-k+1) \geq n^k/2^k$$

ist. □

Bemerkung: Man überlegt sich leicht, dass $g^{(1)}(1-)$ die linksseitige Ableitung $g'(1)$ in $t = 1$ ist. Diese ist also genau dann endlich, wenn EX endlich ist. Entsprechendes gilt für die höheren linksseitigen Ableitungen

$$g^{(k)}(1-) = \lim_{t \rightarrow 1-0} \frac{g^{(k-1)}(1-) - g^{(k-1)}(t)}{1-t}.$$

Anwendung: Mittels (7.1) lassen sich Momente der Verteilung von X oft leichter berechnen als direkt aus der Verteilung. Man geht rekursiv vor:

$$\begin{aligned} EX &= g'(1-), \\ EX^2 &= E(X(X-1)) + EX = g''(1-) + g'(1-), \end{aligned}$$

usw. Man erhält so z.B. für die obigen Beispiele der geometrischen Verteilung und der Poisson-Verteilung alternative Herleitungen des Erwartungswerts und der Varianz, aber es ist interessanter ein Beispiel zu betrachten, bei dem es die Methode der erzeugenden Funktionen gestattet, EX zu berechnen, ohne dass man die Verteilung von X so ohne weiteres explizit angeben kann.

Beispiel 7.3* (*Wartezeit auf den ersten Erfolgs-Run vorgegebener Mindestlänge*) Seien $X_1, X_2, \dots$ Bernoulli-verteilt, also unabhängig mit $P(X_i = 1) = p = 1 - P(X_i = 0)$, und sei für eine vorgegebene natürliche Zahl r

$$T(\omega) = \inf\{k \geq r : X_k(\omega) = X_{k-1}(\omega) = \dots = X_{k-r+1}(\omega) = 1\}$$

der erste Zeitpunkt k , an dem die Folge $X_1(\omega), \dots, X_k(\omega)$ r aufeinander folgende Einsen enthält. (Wenn man die Folge zum Zeitpunkt T abbricht, kann man ein diskretes Modell angeben, aber diese Frage lassen wir hier einmal beiseite.)

Für $n < r$ ist offenbar $P(T = n) = 0$. Ferner ist $P(T = r) = p^r$. Sei nun $n > r$. Ist $T(\omega) = n$, so enthält die Folge $X_1(\omega), \dots, X_r(\omega)$ mindestens eine 0. Zerlegt man $\{T = n\}$ in die Mengen, für die die erste 0 zu den Zeitpunkten $j = 1, \dots, r$ auftritt, so ist

$$P(T = n) = \sum_{j=1}^r P(X_1 = 1, \dots, X_{j-1} = 1, X_j = 0, T = n).$$

Ist für $m \geq r$ A_m die Menge der Folgen $a = (a_1, \dots, a_m) \in \{0, 1\}^m$ mit

$$m = \inf\{k : a_k = a_{k-1} = \dots = a_{k-r+1} = 1\},$$

so ist

$$\begin{aligned}
P(X_1 = 1, \dots, X_{j-1} = 1, X_j = 0, T = n) \\
&= P(X_1 = 1, \dots, X_{j-1} = 1, X_j = 0, (X_{j+1}, \dots, X_n) \in A_{n-j}) \\
&= p^{j-1} (1-p) P((X_{j+1}, \dots, X_n) \in A_{n-j}) \\
&= p^{j-1} (1-p) P((X_1, \dots, X_{n-j}) \in A_{n-j}) \\
&= p^{j-1} (1-p) P(T = n-j).
\end{aligned}$$

Für $n > r$ gilt also die Rekursionsformel

$$P(T = n) = \sum_{j=1}^r p^{j-1} (1-p) P(T = n-j).$$

Sei nun $g(t) = \sum_{n=1}^{\infty} t^n P(T = n)$. Dann ist

$$\begin{aligned}
g(t) &= t^r p^r + t(1-p) \sum_{j=1}^r t^{j-1} p^{j-1} \sum_{n>j} t^{n-j} P(T = n-j) \\
&= t^r p^r + t(1-p) \frac{1 - t^r p^r}{1 - tp} g(t),
\end{aligned}$$

denn die letzte Summe ist gerade wieder $g(t)$. Nach kurzer Rechnung ergibt sich

$$g(t) = \frac{t^r p^r (1 - tp)}{1 - t + t^{r+1} p^r (1 - p)}.$$

Durch Partialbruchzerlegung und Taylorentwicklung sind nun die $P(T = n)$ prinzipiell berechenbar, aber es ist einfacher, $ET = g'(1-)$ direkt zu bestimmen. Es ist

$$g'(1-) = \frac{1 - p^r}{p^r (1 - p)}.$$

Zur Illustration geben wir in Tabelle 7.1 ein paar Werte von ET an, die man für den Münzwurf bzw. den Würfelwurf bei einem Wurf pro Sekunde erhält.

	ET	
r	$p = 1/2$	$p = 1/6$
5	1 Min.	2,6 Std.
10	34 Min.	28 Monate
15	18 Std.	18 098 Jahre
20	24 Tage	141 Mill. Jahre

Tabelle 7.1 (nach Feller (1950), I, S. 300)

Nun leiten wir ein paar einfache Rechenregeln für erzeugende Funktionen ab, die beim Studium von Summen von unabhängigen Zufallsvariablen nützlich sind.

Satz 7.4 Sind X und Y unabhängige Zufallsvariable mit Werten in $\mathbb{Z}^+$, so gilt für die erzeugenden Funktionen zu X , Y und $X + Y$

$$g_{X+Y}(t) = g_X(t) g_Y(t).$$

Beweis: Da t^X und t^Y unabhängig sind, ist $E(t^{X+Y}) = E(t^X t^Y) = E(t^X) E(t^Y)$. $\square$

(Ein zweiter Beweis ergibt sich leicht mittels der Faltungsformel (Satz 5.6).)

Beispiel: Ist $P(X_i = 1) = p = 1 - P(X_i = 0)$, so ist offenbar $g_{X_i}(t) = (1 - p) + pt$. Die erzeugende Funktion der Summe S_n von n Bernoulli-verteilten Variablen $X_1, \dots, X_n$, d.h. die der $b_{n,p}$ -Verteilung, ist also $((1 - p) + pt)^n$.

Als nächstes betrachten wir Summen von unabhängigen Zufallsvariablen, für die die Anzahl der Summanden selbst zufällig sein darf. Solche Summen treten in der Praxis nicht selten auf. Z.B. ist die bei einer Versicherung gemeldete Gesamt-Schadenshöhe die Summe der Schadenshöhen aus den einzelnen Schadensfällen, und die Zahl der Schadensfälle ist zufällig.

Seien $X_1, X_2, \dots, T$ unabhängige Zufallsvariable mit Werten in $\mathbb{Z}^+$. Die X_i sollen alle die gleiche Verteilung und damit auch die gleiche erzeugende Funktion $g_{X_i}(t) = g(t)$ haben. Sei $S_0 = 0$ und $S_n = X_1 + \dots + X_n$. Sei ferner $S_T(\omega) = S_{T(\omega)}(\omega)$.

Satz 7.5 Unter den genannten Bedingungen gilt für die erzeugende Funktion g_{S_T} von S_T

$$g_{S_T}(t) = g_T(g(t)). \quad (7.3)$$

Beweis: Wegen der Unabhängigkeit von T und S_n ist

$$P(S_T = k) = \sum_{n=0}^{\infty} P(\{T = n\} \cap \{S_n = k\}) = \sum_{n=0}^{\infty} P(T = n) P(S_n = k).$$

Nach dem vorigen Satz ist $g_{S_n}(t) = g(t)^n$. Also ist

$$\begin{aligned} g_{S_T}(t) &= \sum_{k=0}^{\infty} P(S_T = k) t^k \\ &= \sum_{n=0}^{\infty} P(T = n) \left(\sum_{k=0}^{\infty} P(S_n = k) t^k \right) \\ &= \sum_{n=0}^{\infty} P(T = n) (g(t))^n = g_T(g(t)). \end{aligned} \quad \square$$

Beispiele: (a) Ein radioaktives Präparat sendet in einer Minute T Teilchen aus. Für jedes davon ist die Wahrscheinlichkeit, dass es von einem Messgerät wahrgenommen wird, gleich p . T ist Poisson-verteilt mit Parameter $\lambda > 0$. Sei $X_i = 1$, wenn das i -te Teilchen wahrgenommen wird, und gleich 0 sonst. $X_1, X_2, \dots$ ist Bernoulli-verteilt, und S_T die Zahl der registrierten Teilchen. Also ist $g(t) = 1 - p + pt$, $g_T(t) = e^{\lambda(t-1)}$ und $g_{S_T}(t) = \exp(\lambda(1 - p + pt - 1)) = \exp(\lambda p(t - 1))$, d.h. S_T ist Poisson-verteilt mit Parameter λp .

(b) Für die Zufallsvariablen in Satz 7.5 gilt die **waldsche Identität**

$$E(S_T) = (ET)(EX_1). \quad (7.4)$$

Nach der Kettenregel ist nämlich $g'_{S_T}(t) = g'_T(g(t))g'(t)$, und beim Grenzübergang $t \rightarrow 1 - 0$ strebt $g(t)$ gegen 1. Setzt man $0 \cdot \infty = \infty \cdot 0 = 0$, so ist dabei sogar gestattet, dass ET oder EX_1 oder beide unendlich sind.

Verzweigungsprozesse

Eine schöne Anwendung erzeugender Funktionen ist die Bestimmung der Aussterbewahrscheinlichkeit bei Verzweigungsprozessen.

Francis Galton studierte 1873 das Phänomen des Aussterbens berühmter Familiennamen. Er stellte die Frage nach der Wahrscheinlichkeit des Aussterbens der männlichen Linie der Nachkommenschaft eines Mannes, wenn dieser und jeder seiner Söhne, Enkel usw. unabhängig voneinander mit Wahrscheinlichkeit p_k genau k Söhne hat.

Er fand Unterstützung bei Reverend Henry William Watson, der 1874 mit Hilfe von erzeugenden Funktionen eine (nicht ganz fehlerfreie und erst viel später korrigierte) Lösung des Problems gab. Erst 1972 wurde entdeckt, dass Irénée Jules Bienaymé das gleiche Problem schon 1845 beantwortet hatte.

Sei $Z_0 = 1$. Ist Z_n die Anzahl der männlichen Nachkommen (in männlicher Linie) in der n -ten Nachkommengeneration, und hat der j -te dieser Nachkommen $X_{n+1}^{(j)}$ Söhne, so ist

$$Z_{n+1} = \sum_{j=1}^{Z_n} X_{n+1}^{(j)}. \quad (7.5)$$

Es ist bequem anzunehmen, die $X_{n+1}^{(j)}$ seien auf ganz Ω definiert, obwohl die gegebene Interpretation natürlich nur auf $\{Z_n \geq j\}$ gültig ist. (Sonst müsste man mit bedingter Unabhängigkeit arbeiten.)

Seien also $\{X_n^{(j)}, n \geq 1\}$ unabhängige Zufallsvariable mit $P(X_n^{(j)} = k) = p_k$ ($k \geq 0$), und sei Z_n rekursiv durch $Z_0 = 1$ und (7.5) definiert. (Z_n) heißt dann **Bienaymé-Galton-Watson-Prozess**, oft traditionell auch Galton-Watson-Prozess. $q_n = P(Z_n = 0)$ ist die Wahrscheinlichkeit, dass es in der n -ten Generation in der männlichen Linie keine männlichen Nachkommen mehr gibt. Da aus $Z_n(\omega) = 0$ offenbar $Z_{n+1}(\omega) = 0$ folgt, ist $q_{n+1} \geq q_n$. $q = \lim q_n$ ist die gesuchte Aussterbewahrscheinlichkeit.

[**Bemerkung:** Bisher können wir noch keinen Wahrscheinlichkeitsraum angeben, auf dem unendlich viele unabhängige $X_n^{(j)}$ definierbar sind. Diese Problematik lassen wir hier beiseite. Gibt es ein $k_0 \in \mathbb{N}$ derart, dass $p_k = 0$ für alle $k > k_0$ ist, so hat niemand mehr als k_0 Söhne und zur Definition von Z_n für $n \leq N < \infty$ benötigen wir dann nur endlich viele unabhängige Zufallsvariable. In diesem Fall lässt sich also schon mit unseren bisher verfügbaren Mitteln ein Modell angeben.]

Heute interessiert man sich für Verzweigungsprozesse, von denen die obigen Prozesse den einfachsten Fall darstellen, natürlich nicht wegen der Familiennamen, sondern weil ähnliche Verzweigungen auch in anderen Situationen auftreten. Z.B. macht ein Neutron bei der Kernspaltung eine zufällige Zahl weiterer Neutronen frei.

Bestimmung der erzeugenden Funktion von Z_n : Da die $X_n^{(j)}$ alle die gleiche Verteilung haben, haben sie auch alle die gleiche erzeugende Funktion

$$g(t) = \sum_{k=0}^{\infty} p_k t^k.$$

Bezeichnet h_n die erzeugende Funktion von Z_n , so ist wegen $P(Z_0 = 1) = 1$ natürlich $h_0(t) = t$. Aus der rekursiven Definition (7.5) folgt nach Satz 7.5

$$h_{n+1}(t) = h_n(g(t)).$$

Also ist $h_1(t) = g(t)$, $h_2(t) = (g \circ g)(t)$ und allgemein $h_n(t) = (g \circ g \circ \dots \circ g)(t)$ die Funktion, die man durch n -fache iterierte Anwendung der Abbildung g erhält. Offenbar ist $h_{n+1}(t) = g(h_n(t))$. Nach Satz 7.2 ist $P(Z_n = 0) = h_n(0)$ und daher

$$q = \lim h_n(0).$$

Damit haben wir bei gegebenem g nur noch ein rein analytisches Problem zu lösen.

Bestimmung der Aussterbewahrscheinlichkeit: Ist $p_1 = 1$, so wird immer mit Wahrscheinlichkeit 1 genau ein Sohn geboren, und es gilt $q = 0$. Dies ist ein Sonderfall.

Satz 7.6 Die Aussterbewahrscheinlichkeit q ist die kleinste nichtnegative Lösung der Gleichung $g(t) = t$. Ist $g'(1) \leq 1$ und $p_1 < 1$, so ist $q = 1$; ist $g'(1) > 1$, so ist $q < 1$.

($g'(1)$ ist die erwartete Zahl der Söhne jedes Mitglieds der Nachkommenschaft. Der Prozess stirbt also — abgesehen vom Fall $p_1 = 1$ — mit Wahrscheinlichkeit 1 aus, wenn im Mittel höchstens ein Sohn geboren wird, und sonst nur mit Wahrscheinlichkeit < 1 .)

Beweis: Aufgrund der Stetigkeit von g ist

$$g(q) = g(\lim h_n(0)) = \lim g(h_n(0)) = \lim h_{n+1}(0) = q.$$

q ist demnach Lösung der Gleichung $g(t) = t$. Ist $u \geq 0$ eine weitere Lösung, so ist $u = g(u) \geq g(0) = h_1(0)$, und durch Induktion folgt aus $u \geq h_n(0)$ dann $u = g(u) \geq g(h_n(0)) = h_{n+1}(0)$. Durch den Grenzübergang $n \rightarrow \infty$ ergibt sich $u \geq q$. Damit ist die erste Teilaussage bewiesen.

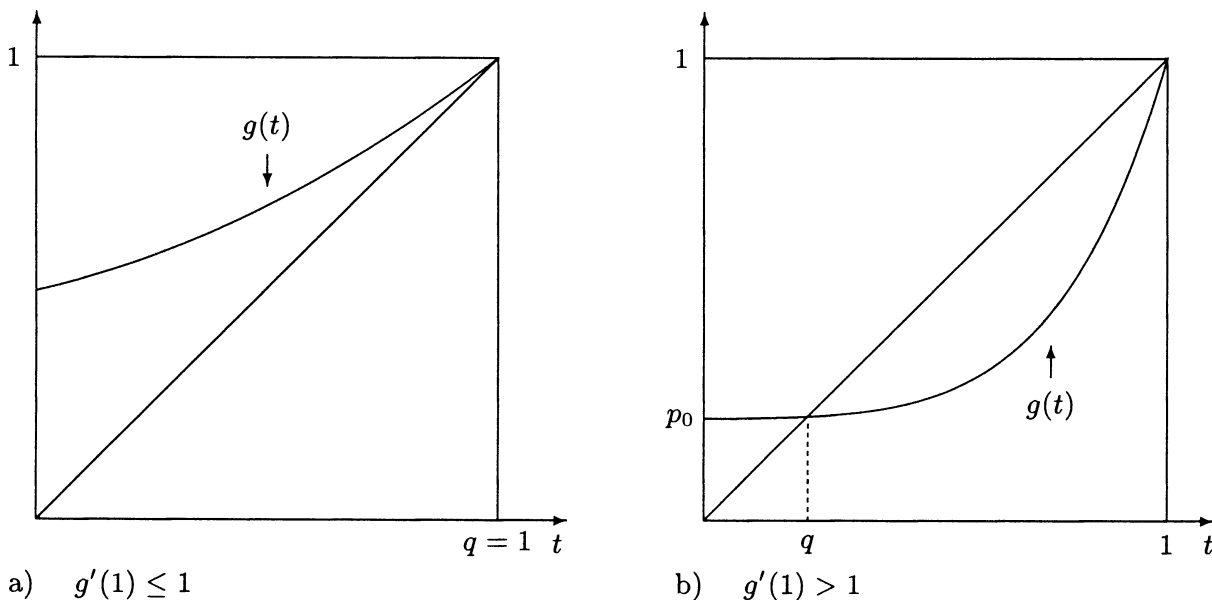


Bild 7.1 Bestimmung der Aussterbewahrscheinlichkeit

Ist $p_0 + p_1 = 1$, so kann in jeder Generation maximal ein männlicher Nachfahre existieren. Aus $P(Z_{n+1} = 1) = P(Z_n = 1)P(X_{n+1}^{(1)} = 1) = p_1 P(Z_n = 1)$ folgt induktiv $P(Z_n = 1) = p_1^n$. Damit gilt $q = \lim(1 - p_1^n)$. In diesem Fall ist $g'(1) = p_1 \leq 1$. Ist $p_1 < 1$, so ist $q = 1$.

Sei also nun $p_0 + p_1 < 1$. Dann ist mindestens eins der p_k mit $k \geq 2$ positiv. $g'(t) = \sum_{k=1}^{\infty} k p_k t^{k-1}$ ist dann auf $[0, 1)$ strikt monoton und $g(t)$ dort strikt konvex. Die

Gleichung $g(t) = t$ kann dann maximal zwei Lösungen haben. Wir betrachten wieder zwei Fälle; siehe Bild 7.1 a) und b).

Zu a): Ist $g'(1) \leq 1$, so ist $g'(t) < 1$ für $0 \leq t < 1$. Nach dem Mittelwertsatz muss dann $g(t)$ in $(0, 1)$ strikt oberhalb der Diagonalen verlaufen. Also ist 1 die einzige Lösung von $g(t) = t$ und damit $q = 1$.

Zu b): Ist $g'(1) > 1$, so ist $g'(t) > 1$ für hinreichend nahe bei 1 liegende $t < 1$. In diesem Bereich muss die Funktion $g(t)$ dann unterhalb der Diagonalen verlaufen. Ist $p_0 > 0$, so muss nach dem Zwischenwertsatz ein $t < 1$ mit $g(t) = t$ existieren. Da q die kleinste nichtnegative Lösung von $g(t) = t$ war, ist $0 < q < 1$. (Watson hatte diese Lösung nicht bemerkt und geschlossen, dass alle Familiennamen irgendwann aussterben.) Ist $p_0 = 0$, so ist $g(0) = 0$ und also $q = 0$. $\square$

Aufgaben

1. Sei T eine Zufallsvariable mit Wertebereich $\{2, 3, \dots\}$ und $P(T = k) = c/(k^2(k-1))$. Bestimmen Sie ET mit Hilfe der erzeugenden Funktion.
2. $T, X_1, X_2, \dots$ seien unabhängige Zufallsvariable mit Werten in $\mathbb{Z}^+$ und erzeugenden Funktionen g_T und $g = g_{X_i}$, für die $E(T^2)$ und $E(X_i^2)$ existieren. Zeigen Sie:

$$\text{Var}(S_T) = \text{Var}(T) (EX_1)^2 + (ET) \text{Var}(X_1). \quad (L)$$

3. $Z_0, Z_1, \dots$ sei ein Verzweigungsprozess mit $Z_0 = 1$, $EZ_1 = \mu$ und $\text{Var}(Z_1) = \sigma^2 < \infty$. Man zeige (mit Hilfe von Aufgabe 2):
 - a) $EZ_n = \mu^n$,
 - b) $\text{Var}(Z_n) = \sigma^2 \sum_{k=0}^{n-1} \mu^{n-1+k} \quad (n \in \mathbb{N})$.
4. Für die Nachkommenverteilung eines Bienaymé-Galton-Watson-Prozesses mit $Z_0 = 1$ gelte $p_k = p(1-p)^k = P(Z_1 = k)$. Wie groß ist die Aussterbewahrscheinlichkeit?
5. Bestimmen Sie die erzeugende Funktion einer negativ binomialverteilten Zufallsvariablen, indem Sie diese als Summe von unabhängigen Zufallsvariablen darstellen.
6. Seien $X_1, X_2, \dots$ unabhängig mit $P(X_i = 1) = p$, $P(X_i = -1) = 1 - p$ ($0 < p < 1$). Sei $S_i = X_1 + \dots + X_i$ und $w_n^{(k)} = P(S_1 < k, S_2 < k, \dots, S_{n-1} < k, S_n = k)$ ($k \geq 1$). Zeigen Sie für die erzeugende Funktion $g^{(k)}$ von $(w_n^{(k)})$, dass $g^{(k)} = g^k$ mit $g = g^{(1)}$ ist. Zeigen Sie ferner $g(t) = pt + (1-p)t g^2(t)$. (L).

§ 8 Entropie und Codierung*

Wir wollen wenigstens kurz auf einen Begriff der Stochastik eingehen, der weniger anschaulich ist als z.B. die Begriffe Wahrscheinlichkeit und Unabhängigkeit, der aber ebenfalls fundamentale Bedeutung hat: den Begriff der Entropie. Er ist eng mit dem der Information verknüpft.

Unser Ziel ist es, die Unbestimmtheit des Ausgangs eines Zufallsexperiments auf sinnvolle Weise quantitativ zu fassen. Wir fragen zunächst, wie viel Aufwand nötig ist, um das Ergebnis eines Zufallsexperiments mitzuteilen. Das Ergebnis eines Münzwurfexperiments lässt sich mitteilen, indem man eine Null oder eine Eins übermittelt. Das Ergebnis eines Experiments mit vier gleichwahrscheinlichen Ausgängen lässt sich mitteilen, indem man eine der Folgen 00, 01, 10 oder 11 übermittelt, also eine 01-Folge der Länge 2. Ein solches Experiment könnte z.B. in der Ausführung von zwei Würfeln einer Münze bestehen. Dies legt es nahe, als Maß für die Unbestimmtheit eines Zufallsexperiments die Länge der 01-Folge zu wählen, die man braucht, um das Ergebnis anzugeben. Diese Überlegung weist in die richtige Richtung, aber es ist zweckmäßig, sie noch zu modifizieren.

Fassen wir die Buchstaben eines Textes als Ergebnisse von Zufallsexperimenten auf, die z.B. ein Funker zu übermitteln hat, so wird er die Buchstaben so „codieren“, dass die häufiger auftretenden Buchstaben *kurzen* Morsefolgen entsprechen. Ähnlich arbeiten Datenkomprimierungsprogramme. Man ist daran interessiert so zu codieren, dass die erwartete Länge der codierten Nachricht minimal ist. Wir wollen nun untersuchen, was sich in dieser Hinsicht bei bekannter Wahrscheinlichkeitsverteilung auf der Menge der Nachrichten erreichen lässt.

8.1 Der Quellen-Codierungssatz

Sei $\mathcal{X}$ eine nichtleere endliche Menge und $\mathbb{A} = \{0, 1, \dots, D-1\}$ eine Menge mit $D \geq 2$ Elementen. Wir nennen $\mathbb{A}$ **Alphabet**.

Ein **Code** ist eine injektive (d.h. eindeutige) Abbildung von $\mathcal{X}$ in die Menge der nichtleeren endlichen Folgen von Elementen aus $\mathbb{A}$.

Ist z.B. $\mathcal{X} = \{1, 2, 3, 4\}$ und $\mathbb{A} = \{0, 1\}$, so ist die Abbildung

$$\begin{aligned} 1 &\mapsto c(1) = 0 \\ 2 &\mapsto c(2) = 10 \\ 3 &\mapsto c(3) = 110 \\ 4 &\mapsto c(4) = 111 \end{aligned} \tag{8.1}$$

ein Code. Wir lassen manchmal die Kommas und die Klammern in der Folge $c(x) = (c_1(x), c_2(x), \dots, c_n(x))$ weg, wenn keine Missverständnisse entstehen können.

Besteht eine Nachricht aus einer Folge $x_1, x_2, \dots, x_k$ von Elementen von $\mathcal{X}$, so kann man sie codieren, indem einfach die Codewörter $c(x_1), c(x_2), \dots, c(x_k)$ hintereinander geschrieben werden. Zur Folge $(2, 1, 4)$ gehört dann die Folge 100111.

Der Code (8.1) hat eine angenehme Eigenschaft: Kein Codewort ist Anfangsstück eines anderen Codewortes. Solche Codes nennen wir **Präfixcodes**. Hat ein Code diese Eigenschaft, so lässt sich aus der codierten Nachricht die Nachricht eindeutig zurückgewinnen: Ist $(a_1, a_2, \dots, a_k)$ die codierte Nachricht, so ist genau eine der Teilfolgen $(a_1, \dots, a_i)$ ein Codewort $c(x_1)$. Man notiert x_1 und sucht in der verkürzten Folge $(a_{i+1}, \dots, a_k)$ analog $c(x_2)$, auf, usw. Man überlegt sich leicht Beispiele von Codes, die nicht Präfixcodes sind, und für die dennoch jede Nachricht eindeutig (z.B. von hinten) decodiert werden kann.

Betrachtet man den klassischen Morse-Code als Code, der mit den Symbolen „·“ und „–“ codiert, so sind Nachrichten, die aus mehreren unmittelbar hintereinander geschriebenen Codewörtern bestehen, nicht decodierbar. In Wirklichkeit arbeitet der Morse-Code mit dem Alphabet $\{·, –, ○\}$, wobei „○“ für „Pause“ steht. Der Buchstabe *a* wird z.B. durch „· – ○“ codiert. Wird eins der Symbole dafür reserviert, das Ende eines Codewortes zu signalisieren, so ist der Code natürlich Präfixcode.

[Wir verlieren nichts, wenn wir uns auf Präfixcodes einschränken, denn zu jedem Code, für den alle Nachrichten eindeutig decodiert werden können, gibt es einen Präfixcode mit genau den gleichen Wortlängen. Ein Beweis dieser Aussage findet sich z.B. bei Jacobs (1983).]

Einen Präfixcode, dessen maximale Wortlänge N ist, kann man konstruieren, indem man die Folgen $(a_1, \dots, a_n)$ mit $n \leq N$ und $a_i \in \mathbb{A}$ in Form eines Baumes hinschreibt; siehe Bild 8.1.

Der Code (8.1) besteht aus den unterstrichenen 01-Folgen. Die Präfix-Eigenschaft bedeutet, dass, wenn ein Verzweigungspunkt als Codewort gewählt wurde, an Ästen, die von diesem Punkt ausgehen, keine weiteren Codeworte mehr vorkommen.

Satz 8.1 (Ungleichung von Kraft) *Einen Präfixcode mit Wortlängen $n_1, \dots, n_M$ zur Alphabetlänge D gibt es genau dann, wenn*

$$\sum_{i=1}^M D^{-n_i} \leq 1 \quad (8.2)$$

ist.

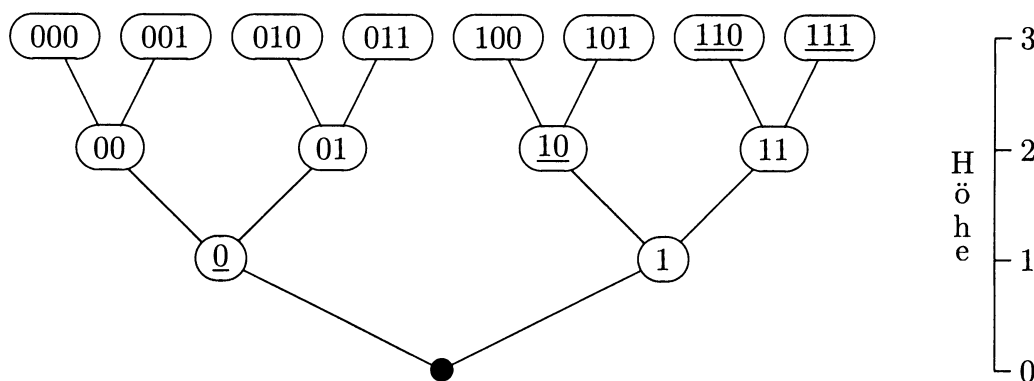


Bild 8.1 Codewortbaum

Beweis: Wir zeigen zuerst, dass (8.2) hinreichend ist. Dabei können wir $n_1 \leq n_2 \leq \dots \leq n_M$ annehmen. Sei $c(x_1)$ die Folge, die aus n_1 Nullen besteht. Im Baum zur Alphabetlänge D und zur maximalen Wortlänge n_M liegt der zu dieser Folge gehörige Punkt auf Höhe n_1 ganz links. Durch Wahl dieses Codewortes ist für jedes $h \geq n_1$ der Anteil D^{-n_1} der D^h auf Höhe h liegenden Punkte von der Wahl als Codewort ausgeschlossen, und zwar diejenigen D^{h-n_1} Punkte, die am weitesten links liegen. Als $c(x_2)$ wählt man den am weitesten links liegenden noch verfügbaren Punkt der Höhe n_2 . Nun sind für $h \geq n_2$ die $D^h (D^{-n_1} + D^{-n_2})$ am weitesten links liegenden Punkte der Höhe h für die weitere Wahl von Codeworten ausgeschlossen. So macht man weiter. Ist $m < M$ und sind m Codeworte auf die angegebene Weise bestimmt, so sind für $h \geq n_m$ die am weitesten

links liegenden $D^h (D^{-n_1} + \dots + D^{-n_m})$ Punkte der Höhe h „verboten“. Wegen (8.2) muss mindestens einer noch verfügbar sein. Man wählt auf Höhe n_{m+1} den am weitesten links liegenden noch verfügbaren Punkt. Die zugehörige Folge $(a_1, \dots, a_{n_{m+1}})$ wird als Codewort $c(x_{m+1})$ gewählt.

Die andere Richtung des Beweises ist noch einfacher. Ein Codewort der Länge n_m gehört zu einem Punkt der Höhe n_m . An den von diesem Punkt ausgehenden Ästen gibt es $D^{n_M - n_m}$ Punkte zur maximalen Höhe n_M . Da es sich um einen Präfixcode handelt, müssen die auf diese Weise verschiedenen Codeworten zugeordneten Punkte maximaler Höhe verschieden sein. Da es nur D^{n_M} Punkte maximaler Höhe gibt, ist

$$\sum_{m=1}^M D^{n_M - n_m} \leq D^{n_M}.$$

Daraus folgt (8.2). □

Um die erwartete Wortlänge optimaler Präfixcodes abzuschätzen, benötigen wir noch

Lemma 8.2 *Sind P und Q zwei Wahrscheinlichkeitsmaße auf $\mathcal{X}$, so gilt*

$$-\sum P(x) \log P(x) \leq -\sum P(x) \log Q(x),$$

und Gleichheit gilt genau im Fall $P = Q$. (Dabei sei $0 \log 0 = 0$.)

Beweis: Wegen $\log_b t = \log_b e \log_e t$ ist es egal, zu welcher Basis wir den Logarithmus wählen. Sei im Moment e die Basis. Es gilt $\log(1+t) \leq t$, und Gleichheit gilt nur für $t = 0$, denn die zweite Ableitung von $\log y$ ist strikt negativ, die erste Ableitung hat in $y = 1$ den Wert 1, und es ist $\log 1 = 0$.

Bezeichne $\sum'$ die Summe über alle x mit $P(x) > 0$. Es gilt

$$\sum' P(x) \log \frac{Q(x)}{P(x)} = \sum' P(x) \log \left(1 + \frac{Q(x) - P(x)}{P(x)} \right) \leq \sum' (Q(x) - P(x)). \quad (8.3)$$

Ist für ein x mit $P(x) = 0$ $Q(x) > 0$, so ist $\sum' Q(x) < 1 = \sum' P(x)$. Dann ist die rechte Summe in (8.3) negativ. Ist für ein x mit $P(x) > 0$ $Q(x) \neq P(x)$, so ist die Ungleichung in (8.3) strikt. Wir sehen also, dass die linke Seite in (8.3) ≤ 0 ist, und im Fall $P \neq Q$ sogar < 0 . Dies ist die Behauptung. □

Definition 8.3 Ist P Wahrscheinlichkeitsmaß auf $\mathcal{X}$, so heißt

$$H(P) = - \sum_{x \in \mathcal{X}} P(x) \log_2 P(x)$$

die **Entropie** von P .

In der Informationstheorie wird gewöhnlich der Logarithmus zur Basis 2 zugrundegelegt, und dem schließen wir uns nun hier an. Es ist nur eine Normierungsfrage. Schreibt man $H_b(P)$, wenn die Basis b gewählt wird, so ist offenbar $H_b(P) = (\log_b 2) H(P)$.

Sind P_X und $P_{X,Y}$ die Verteilung von X und die gemeinsame Verteilung von (X, Y) , so schreibt man auch $H(X)$ statt $H(P_X)$ und $H(X, Y)$ statt $H(P_{X,Y})$. Für einen Code c sei $n(x)$ die Länge des Codeworts $c(x)$ und $En = \sum P(x) n(x)$ die erwartete Codewortlänge. Der folgende Satz sagt uns ziemlich genau, was erreichbar ist:

Satz 8.4 (Quellen-Codierungssatz) *Ist c ein beliebiger Präfixcode zu einem Alphabet mit Alphabetlänge D , so gilt $H_D(P) \leq En$. Es gibt einen Präfixcode c mit $En < H_D(P) + 1$.*

Beweis: Ist c ein Präfixcode, so folgt aus der Ungleichung von Kraft für $\sigma = \sum D^{-n(x)}$ die Ungleichung $\sigma \leq 1$. Durch

$$Q(x) = D^{-n(x)}/\sigma$$

ist ein Wahrscheinlichkeitsmaß Q auf $\mathcal{X}$ definiert. Nach Lemma 8.2 gilt

$$H_D(P) \leq - \sum P(x) \log_D Q(x) = \sum P(x) n(x) + \sum P(x) \log_D \sigma \leq En.$$

Dabei gilt Gleichheit genau dann, wenn $\sigma = 1$ und $P = Q$ ist, also im Fall $P(x) = D^{-n(x)}$.

Ist P gegeben, und hat P diese Form, so existiert nach der Ungleichung von Kraft ein Präfixcode mit den Wortlängen $n(x)$. Zu gegebenem P existiert ein Präfixcode mit $H_D(P) = En$ also genau dann, wenn $-\log_D P(x)$ für jedes x eine ganze Zahl ist. Im allgemeinen Fall sei, wenn P gegeben ist, $n(x)$ die kleinste ganze Zahl $\geq -\log_D P(x)$. Dann gilt $D^{-n(x)} \leq P(x)$ für alle x und also (8.2). Zu diesen Wortlängen existiert also ein Präfixcode c . Die Abschätzung

$$\sum n(x) P(x) < \sum P(x) (-\log_D P(x) + 1) = H_D(P) + 1$$

zeigt, dass c die geforderte Eigenschaft hat. $\square$

Der so konstruierte Code c braucht nicht optimal zu sein. Er wird aber ausreichen, um im Folgenden zu zeigen, dass $H(P)$ ein natürliches Maß für die mit der Verteilung P verbundene Ungewissheit ist.

8.2 Anwendung auf mehrstufige Zufallsexperimente

Wir betrachten nun zunächst ein paar einfache Eigenschaften der Entropie. Da die Funktion $-x \log x$ im Einheitsintervall nichtnegativ und stetig ist, ist auch $H(P)$ nichtnegativ, und bei stetiger Änderung der $P(x)$ ändert sich $H(P)$ stetig. Hat $\mathcal{X}$ M Elemente, und setzt man in Lemma 8.2 $Q(x) = 1/M$, so folgt

$$H(P) \leq - \sum P(x) \log_2 \frac{1}{M} = \log_2 M.$$

Gleichheit gilt genau dann, wenn alle $P(x)$ gleich $1/M$ sind.

Satz 8.5 *Für Zufallsvariable X, Y mit endlich vielen Werten gilt*

$$H(X, Y) \leq H(X) + H(Y),$$

und Gleichheit gilt genau dann, wenn X und Y unabhängig sind.

Beweis: Hat X die Werte $x_1, \dots, x_m$ und Y die Werte $y_1, \dots, y_n$, so sei

$$p(i, j) = P(X = x_i, Y = y_j), \quad p_i = P(X = x_i) \quad \text{und} \quad q_j = P(Y = y_j).$$

Nach Lemma 8.2 ist

$$\begin{aligned}
H(X, Y) &= - \sum_i \sum_j p(i, j) \log p(i, j) \\
&\leq - \sum_i \sum_j p(i, j) \log(p_i q_j) \\
&= - \sum_i \sum_j p(i, j) \log p_i - \sum_i \sum_j p(i, j) \log q_j \\
&= - \sum_i p_i \log p_i - \sum_j q_j \log q_j = H(X) + H(Y)
\end{aligned}$$

Gleichheit gilt genau im Fall $p(i, j) = p_i q_j$. $\square$

Man nennt $I(X \parallel Y) = H(X) + H(Y) - H(X, Y)$ die **Information** von Y über X . Sie ist ein Maß der Abhängigkeit von X und Y . Offenbar ist $I(Y \parallel X) = I(X \parallel Y)$.

Durch Induktion folgt aus Satz 8.5

$$H(X_1, \dots, X_K) \leq \sum_{k=1}^K H(X_k).$$

Im Fall der Unabhängigkeit gilt Gleichheit. Beschreiben die Zufallsvariablen X_k die Ergebnisse von K identischen unabhängigen Teilerperimenten, so ist also $H(X_1, \dots, X_K) = K H(X_1)$.

Es gibt nach Satz 8.4 einen Präfixcode für das Gesamtexperiment, dessen erwartete Codewortlänge $\leq K H(X_1) + 1$ ist. Die pro Teilerperiment benötigte erwartete Codewortlänge ist also $\leq H(X_1) + 1/K$. Andererseits ist sie nach der unteren Abschätzung für jeden Präfixcode $\geq H(X_1)$. Für $K \rightarrow \infty$ ist also $H(X_1)$ genau die pro Teilerperiment benötigte erwartete Zahl der Symbole, die man zur Übermittlung der Ergebnisse braucht. Die Abschätzung $H(X_1) + 1/K$ für die pro Teilerperiment benötigte erwartete Zahl der Symbole bleibt sogar richtig, wenn die Teilerperimente abhängig sind. Die X_i müssen nur alle die gleiche Verteilung haben.

Verwandte Fragen: Dem Entropiebegriff kann man sich noch auf verschiedene andere Arten nähern. Der hier gewählte Zugang lässt sich auch so fassen: Wie viele nur mit Ja oder Nein zu beantwortende Fragen muss man stellen, um das Ergebnis eines Zufallsexperiments zu erfragen? Hat man einen Code c zum Alphabet $\{0, 1\}$, so könnte die erste Frage lauten: „Ist $c_1(x) = 1$?“ Hat man ein erfolgreiches Fragesystem, so kann man $c_1(x) = 1$ für die x setzen, für die die erste Frage mit Ja zu beantworten ist. Man überlegt sich leicht, dass durch Fortsetzung dieses Verfahrens jedem erfolgreichen Fragesystem ein Code zugeordnet werden kann und umgekehrt. Die erwartete Zahl der benötigten Fragen entspricht der erwarteten Codewortlänge. Eine weitere Interpretation ergibt sich in der Informatik, wenn man fragt: „Wie viel Bits benötigt man im Mittel, um das Ergebnis eines Zufallsexperiments abzuspeichern?“

In der von C. E. Shannon 1948 in einer bahnbrechenden Arbeit begründeten Informationstheorie geht es dann vor allem darum, Nachrichten so zu codieren, dass durch die Nachrichtenübertragung verursachte Fehler erkannt und korrigiert werden können.

Vielleicht reicht dies, um die Neugier zu wecken. Elementare Einführungen haben Jaglom-Jaglom (1960) und Topsøe (1974) gegeben. Ausführlicher ist Ash (1965). Ein Standardwerk ist das Buch von Gallager (1968). Ein geeignetes deutschsprachiges Lehrbuch ist Mathar (1996).

Aufgaben

Alle Zufallsvariablen $X, Y, \dots$ sollen nur endlich viele Werte haben.

1. Mit der Notation aus dem Beweis von Satz 8.5 sei $p(j|i) = P(Y = y_j | X = x_i)$, $H(p(\cdot|i)) = - \sum_j p(j|i) \log_2 p(j|i)$ und $H(Y|X) = \sum_i p_i H(p(\cdot|i))$. (Dann ist $H(Y|X) = H(X, Y) - H(X)$.) Man zeige:

- a) $H(Y, Z | X) = H(Y | X) + H(Z | X, Y)$;
- b) $H(Z | X, Y) \leq H(Z | X)$.
2. Sei $Y = g(X)$ eine beliebige Funktion von X . Man zeige: $H(Y) \leq H(X)$. (L)
3. Der aus den folgenden Codeworten bestehende Code ist kein Präfixcode: ab, bc, aaa, ca, b . Ist dennoch jede Nachricht eindeutig entzifferbar?
4. Seien $X_1, X_2, \dots$ unabhängig identisch verteilt, $H := H(X_1)$. Man zeige, dass es zu jedem $\varepsilon > 0$ und für hinreichend großes n eine Menge T_n von Folgen $(x_1, \dots, x_n)$ gibt mit

$$P\left((X_1, \dots, X_n) \in T_n\right) \geq 1 - \varepsilon$$

und

$$\text{card}(T_n) \leq 2^{(H+\varepsilon)n}.$$

(Wenn man eine Menge von Nachrichten mit Gesamtwahrscheinlichkeit ε weglässt, reichen $2^{(H+\varepsilon)n}$ Codewörter.)

Anleitung: Sei $k_n(i, x)$ die Häufigkeit von i in $x = (x_1, \dots, x_n)$. Für hinreichend kleines $\delta > 0$ kann man die Folgen x mit $|k_n(i, x) - nP(X_1 = i)| < n\delta$ (für alle Werte i) nehmen.

5. Sei $\lambda > 1$ und $\mathcal{P}_\lambda$ die Menge aller Wahrscheinlichkeitsverteilungen $P = (p_k)_{k \in \mathbb{N}}$ auf $\mathbb{N}$ mit vorgegebenem Erwartungswert λ . Sei $Q = (q_k)_{k \in \mathbb{N}}$ definiert durch $q_k = (1 - 1/\lambda)^{k-1}/\lambda$. Zeigen Sie $H(Q) = \sup\{H(P) : P \in \mathcal{P}_\lambda\}$. (Hinweis: Lemma 8.2 und $Q \in \mathcal{P}_\lambda$.) (L)
6. Sei $\mathcal{X} := \{1, 2, 3, 4, 5\}$ und $\mathbb{A} := \{\alpha, \beta, \gamma\}$. Sei c der Code mit $c(1) = \alpha\beta$, $c(2) = \beta\gamma$, $c(3) = \alpha\alpha\alpha$, $c(4) = \gamma\alpha$, $c(5) = \beta$.
 - a) Warum ist c kein Präfixcode?
 - b) Ist trotzdem jede Nachricht eindeutig entzifferbar?
 - c) Gibt es einen Präfixcode mit den gleichen Wortlängen?

§ 9 Laufzeitanalysen von rekursiven Algorithmen*

In diesem Abschnitt wollen wir eine erste Einführung in ein Thema geben, das angesichts des Vordringens der Computer von großer Bedeutung ist. Uns interessieren Aussagen über die Laufzeit von rekursiven Algorithmen. Als Beispiel dienen Sortieralgorithmen. Die Resultate dieses Abschnitts werden in den verbleibenden Teilen des Buches nicht benötigt.

Häufig stellt sich das Problem, eine größere Zahl von Daten in eine wohldefinierte Reihenfolge zu bringen, z.B. eine Namensliste alphabetisch zu ordnen. Sind $x_1, \dots, x_n$ die Daten in der Reihenfolge ihrer Beobachtung, so gelte für eine Permutation $\pi = (\pi_1, \dots, \pi_n)$ von $(1, \dots, n)$

$$x_{\pi_1} < x_{\pi_2} < \dots < x_{\pi_n},$$

und es ist unsere „Sortieraufgabe“, die Permutation, die dies leistet, herauszufinden, indem wir paarweise Vergleiche durchführen. Man könnte etwa x_1 mit x_2 vergleichen, dann den größeren der beiden Werte mit x_3 usw. Wir lassen zu, dass der Algorithmus selbst Gebrauch vom Zufall macht.

Bei der Frage nach der alphabetischen Ordnung wäre x_{π_1} der Name, der alphabetisch an erster Stelle stünde. Da der Algorithmus nur Vergleiche anstellt und sonst keine Information verwerten kann, können wir hier annehmen, $x_1, \dots, x_n$ sei eine Permutation von $1, \dots, n$. (Ordnet man Wörter alphabetisch, so wird man das Wort „Zyste“ gleich „weit hinten“ einordnen, aber solche Schlüsse aus der Beobachtung einzelner x_i sind hier nicht erlaubt.)

Die Laufzeit L_n eines Sortieralgorithmus sei daher definiert als die Zahl der paarweisen Vergleiche, die bis zur Lösung der Sortieraufgabe durchzuführen sind. L_n hängt von der (möglicherweise zufälligen) Reihenfolge der zu ordnenden Daten und gegebenenfalls von den Zufallsschritten im Algorithmus ab.

Wir wollen hier die Algorithmen **Mergesort** und **Quicksort** betrachten. Beides sind Algorithmen vom „divide and conquer“-Typ. Die Datenmenge wird geteilt, der Algorithmus wird auf die Teilmengen angewandt, und dann benötigt man zusätzliche Schritte, um die Lösungen der Teilprobleme zu einer Lösung des Gesamtproblems zusammenzufügen.

(a) **Mergesort:** Der Grundgedanke von Mergesort besteht darin das Sortieren einer langen Liste von Daten induktiv auf das Sortieren kürzerer Listen zurückzuführen. Um zwei bereits geordnete Teillisten zu einer geordneten Gesamtliste zusammenzufügen („to merge“), geht man dabei wie folgt vor:

1. Vergleiche die jeweils kleinsten Zahlen beider sortierten Teillisten.
2. Entferne die kleinere der beiden Zahlen aus der Liste und speichere sie als kleinstes Element einer Zielliste ab.
3. Wende dieses Verfahren rekursiv auf die nach und nach kürzer werdenden Teillisten an, bis eine davon leer ist. Die verbleibende Teilliste wird der Zielliste „oben“ angefügt.

Sei hier der Einfachheit halber $n = 2^m$ eine Zweierpotenz. Dann kann man in jedem Schritt die Listenlängen halbieren. Mergesort führt rekursiv die folgenden Schritte aus:

- (M1) Aufteilung der Liste der Daten in zwei gleichgroße Teillisten.
- (M2) Anwendung von Mergesort auf jede der beiden Teillisten.

(M3) Erstellen der geordneten Gesamtliste nach dem oben beschriebenen Schema.

Bezeichnet M_n die Laufzeit von Mergesort, $M_{n/2}$ die Laufzeit, die zur Sortierung der ersten Teilliste benötigt wird, $\bar{M}_{n/2}$ die Laufzeit, die zur Sortierung der zweiten Teilliste benötigt wird, und Z_n die zum Ineinandersortieren fertig sortierter Teillisten benötigte Zeit, so gilt

$$M_n = M_{n/2} + \bar{M}_{n/2} + Z_n. \quad (9.1)$$

Offenbar ist $M_1 = 0$, denn Listen der Länge 1 sind stets geordnet. Bevor wir irgendwelche wahrscheinlichkeitstheoretischen Annahmen machen, können wir die Schlüsselgleichung (9.1) benutzen, um Abschätzungen von M_n nach oben und nach unten anzugeben.

Zum Ineinandersortieren von bereits geordneten Listen mit s und t Elementen benötigt man mindestens $s \wedge t = \min(s, t)$ Vergleiche, denn der günstigste Fall ist der, dass bei jedem Vergleich das Element aus der kürzeren Liste in die Zielliste gelangt. Die Anzahl der Vergleiche ist höchstens $s+t-1$. Dieser Wert tritt dann auf, wenn so lange wie möglich bei jedem Vergleich ein Element aus einer Liste mit noch mindestens 2 Elementen in die Zielliste gelangt.

Wendet man dies mit $s = t = n/2$ an, so ergibt sich $n/2 \leq Z_n \leq n-1$. Hieraus folgt induktiv

$$m 2^{m-1} \leq M_n \leq (m-1) 2^m + 1, \quad (9.2)$$

indem man die Induktionsvoraussetzung auf $M_{n/2}$ und $\bar{M}_{n/2}$ anwendet. Man nennt dies die Abschätzungen für den günstigsten und ungünstigsten Fall.

Diese Fälle sind oft nicht typisch. Von größerem Interesse ist daher der „durchschnittliche“ Fall. Wir betrachten daher nun die Daten als zufällig. Jede Permutation sei gleichwahrscheinlich. (Ist diese Voraussetzung nicht realistisch, so könnte man sie erfüllen, indem man die Daten vor dem Sortieren einer zufälligen Permutation unterwirft.)

$x = (x_1, x_2, \dots, x_n)$ sei also das Ergebnis einer zufälligen Permutation von $(1, 2, \dots, n)$. Eine solche zufällige Permutation können wir wie folgt erzeugen: Sei $l = n/2$. Zuerst wird eine Menge $A = \{a_1, a_2, \dots, a_l\}$ mit $a_1 < a_2 < \dots < a_l$ aus $\{1, \dots, n\}$ ausgewählt, so dass jede Teilmenge A mit $\text{card}(A) = l$ gleichwahrscheinlich ist, also die Wahrscheinlichkeit $\binom{n}{l}^{-1}$ hat. Dann wählt man unabhängig von der Wahl von A und unabhängig voneinander Permutationen $(\sigma_1, \sigma_2, \dots, \sigma_l)$ von $(1, 2, \dots, l)$ und $(\tau_{l+1}, \dots, \tau_n)$ von $(l+1, l+2, \dots, n)$, beides mit Gleichverteilung. Seien $a_{l+1} < a_{l+2} < \dots < a_n$ die Elemente von A^c in aufsteigender Ordnung. Setze $x_i = a_{\sigma_i}$ für $1 \leq i \leq l$ und $x_j = a_{\tau_j}$ für $l+1 \leq j \leq n$. Wir können nun also annehmen, die zufällige Permutation x sei auf diese Weise zustande gekommen.

$M_l(x)$ hängt nun nur von σ ab, denn bei einem Vergleich von x_i und x_j mit $i, j \leq l$ gilt $x_i < x_j$ genau dann, wenn $\sigma_i < \sigma_j$ gilt. Analog hängt $\bar{M}_l(x)$ nur von τ ab. Ferner hängt $Z_n(x)$ nur von der Wahl von A ab, denn vor dem Ineinandersortieren hat man ja die Teillisten schon einzeln sortiert. Die Zufallsvariablen M_l , $\bar{M}_l$ und Z_n sind also stochastisch unabhängig, und die Verteilung von M_l und $\bar{M}_l$ ist genau die Verteilung der Laufzeit von Mergesort angewandt auf eine zufällige Permutation von l („aufsteigenden“) Elementen. Insbesondere haben M_l und $\bar{M}_l$ die gleiche Verteilung.

Die Verteilung von M_n ergibt sich nun rekursiv, wenn wir zuerst die Verteilung von Z_n bestimmen. (Sie ist die Faltung der Verteilungen von M_l , $\bar{M}_l$ und Z_n .)

Zum Ineinandersortieren der Teillisten A und A^c brauchen wir genau dann maximal $n-r$ Vergleiche, wenn die größten r Elemente von $\{1, \dots, n\}$ entweder alle zu A oder alle zu A^c gehören ($1 \leq r \leq l$). Die Wahrscheinlichkeit, dass sie alle in A^c vorkommen, ist $\binom{n-r}{l} / \binom{n}{l}$, denn dann müssen die l Elemente von A alle aus $\{1, \dots, n-r\}$ stammen.

Die Wahrscheinlichkeit, dass sie alle zu A gehören ist genauso groß. Setzen wir $z = n - r$, so erhalten wir

$$P(Z_n \leq z) = 2 \binom{z}{l} / \binom{n}{l}.$$

Wir betrachten statt Z_n nun $Z'_n = n - Z_n$. Es ist

$$P(Z'_n \geq r) = 2 \binom{n-r}{l} / \binom{n}{l}.$$

Wir setzen $q_r = P(Z'_n \geq r)$ für $1 \leq r \leq l$. Unser nächstes Ziel ist die Berechnung des Erwartungswerts und der Varianz von M_n . Wir benötigen dazu das folgende kombinatorische

Lemma 9.1 Für ganze Zahlen $u, v \geq 0$ und $w \geq 1$ gilt:

$$(a) \quad \sum_{r=1}^v \binom{u+v-r}{u} = \binom{u+v}{u+1}; \quad (b) \quad \sum_{r=1}^v \binom{r}{w} \binom{u+v-r}{u} = \binom{u+v+1}{u+w+1}.$$

Beweis: (a): Für $v = 0$ ist die linke Summe leer und die rechte Seite 0. Also gilt (a). Für $v = 1$ sind beide Seiten = 1. Die Behauptung folgt dann induktiv mittels (1.21).

(b): Die Fälle $u = 0$ (v, w beliebig, $v = 0$ (u, w beliebig), $w \geq v$ und $v = 1$ sind leicht direkt nachzurechnen. Dann verwendet man beim Schluss von v auf $v + 1$ bei gegebenem u die Aussage für u und v und die Aussage für $u - 1$ und $v + 1$. $\square$

Nach Übungsaufgabe 2 aus § 3 ist

$$E(Z'_n) = \sum_{r=1}^l q_r \quad E((Z'_n)^2) = \sum_{r=1}^l (2r-1) q_r. \quad (9.3)$$

Wenden wir Lemma 9.1 (a) mit $u = v = l = n/2$ an, so erhalten wir für $n = 2^m$ mit $m \geq 1$:

$$E(Z'_n) = \frac{2}{\binom{n}{l}} \sum_{r=1}^l \binom{n-r}{l} = \frac{2}{\binom{n}{l}} \binom{n}{l+1} = \frac{2l}{l+1} = \frac{2n}{n+2}.$$

Aus der zweiten Gleichung in (9.3) ergibt sich unter Verwendung von Lemma 9.1

$$\begin{aligned} E((Z'_n)^2) &= \frac{2}{\binom{n}{l}} \left[\sum_{r=1}^l 2 \binom{r}{1} \binom{n-r}{l} - \sum_{r=1}^l \binom{n-r}{l} \right] \\ &= \frac{2}{\binom{n}{l}} \left[2 \binom{n+1}{l+2} - \binom{n}{l+1} \right] \\ &= 2 \left[\frac{2(n+1)l}{(l+1)(l+2)} - \frac{l}{l+1} \right] = \frac{6l^2}{(l+1)(l+2)} = \frac{6n^2}{(n+2)(n+4)}. \end{aligned}$$

Offenbar gilt $E(Z_n) = n - E(Z'_n)$ und $\text{Var}(Z_n) = \text{Var}(Z'_n)$. Die rekursive Gleichung (9.1) liefert nun

Satz 9.2 Für $n = 2^m$ mit $m \geq 1$ gilt:

$$(a) \quad E(M_n) = n \sum_{i=1}^m \frac{2^{i-1}}{2^{i-1} + 1}, \quad (b) \quad \text{Var}(M_n) = n \sum_{i=1}^m \frac{2^{i-1} (2^{i-1} - 1)}{(2^{i-1} + 1)^2 (2^{i-1} + 2)}.$$

Beweis: (a): Für $n = 2$ wird stets genau ein Vergleich durchgeführt. Dann ist $m = 1$ und auch die rechte Seite in (a) ist gleich 1. Nun erhalten wir induktiv mittels (9.1)

$$E(M_{2n}) = 2 E(M_n) + 2n - E(Z'_{2n}).$$

Der Term $2 E(M_n)$ liefert die ersten m Summanden in der Summe, die zu $E(M_{2n})$ gehört, und es ist

$$2n - E(Z'_{2n}) = 2n \left(1 - \frac{2}{2n + 2} \right) = 2n \frac{n}{n + 1}.$$

Dies steuert den neuen Summanden bei.

(b) ergibt sich induktiv aus $\text{Var}(M_{2n}) = \text{Var}(M_n) + \text{Var}(\bar{M}_n) + \text{Var}(Z_{2n}) = 2 \text{Var}(M_n) + \text{Var}(Z_{2n})$, wie der Leser leicht nachrechnet. $\square$

Der Faktor von n in der Formel (b) für die Varianz konvergiert für $m \rightarrow \infty$ gegen eine Konstante $\rho \approx 0,307$. Mittels Tschebyschew folgt $P(|M_n - EM_n| > a_n) \rightarrow 0$, wenn (a_n) eine Folge mit $a_n/\sqrt{n} \rightarrow \infty$ ist.

Wir erwähnen nur, dass sich mit Hilfe des Zentralen Grenzwertsatzes mit Lindebergbedingung (siehe Bauer (1991) oder Gänssler-Stute (1977)) beweisen lässt: Ist

$$M_n^* = \frac{M_n - EM_n}{\sqrt{\text{Var}(M_n)}},$$

so gilt für alle $x \in \mathbb{R}$ $P(M_n^* < x) \rightarrow \Phi(x)$.

(b) **Quicksort:** Für das oben beschriebene Sortierproblem wird häufig auch ein alternativer Algorithmus mit Namen „Quicksort“ verwendet, der rekursiv die folgenden Schritte ausführt:

- (Q1) Eine der Zahlen x_i ($1 \leq i \leq n$), die zu ordnen sind, wird zufällig (mit Gleichverteilung) ausgewählt.
- (Q2) Es werden durch paarweisen Vergleich mit dieser Zahl x_i die Menge $M^<$ der x_j die kleiner sind als x_i und die Menge $M^>$ der x_j , die größer sind als x_i bestimmt. Man kennt dann die Stelle Z_n , an der das zufällig gewählte x_i im geordneten Vektor steht. Z_n ist gleichverteilt auf $\{1, \dots, n\}$ und es ist $\text{card}(M^<) = Z_n - 1$ und $\text{card}(M^>) = n - Z_n$.
- (Q3) Der Algorithmus wird auf $M^<$ und $M^>$ angewandt, sofern diese noch zu ordnen sind, also sofern sie aus mindestens zwei Elementen bestehen.

Ist $Z_n = k$, und sind $y_1, \dots, y_{k-1}$ die Elemente von $M^<$ in der Reihenfolge, in der sie in der Folge $x = (x_1, \dots, x_n)$ stehen, und $y_{k+1}, \dots, y_n$ die Elemente von $M^>$, ebenfalls in der Reihenfolge aus x , so ist jede Reihenfolge von $y_1, \dots, y_{k-1}$ und jede Reihenfolge von $y_{k+1}, \dots, y_n$ gleichwahrscheinlich, und die beiden Reihenfolgen sind bedingt unabhängig, gegeben $\{Z_n = k\}$.

Als Laufzeit Q_n des Algorithmus Quicksort bezeichnen wir die (zufällige) Zahl der benötigten paarweisen Vergleiche. In Schritt 2 werden $n - 1$ Vergleiche benötigt.

Sei $Q_{Z_n-1}^<$ die Zahl der Vergleiche, die bei der rekursiven Anwendung für das Sortieren von $M^<$ benötigt wird, und sei $Q_{n-Z_n}^<$ die Zahl der für das Sortieren von $M^>$ benötigten Vergleiche. Dann ist

$$Q_n = n - 1 + Q_{Z_n-1}^< + Q_{n-Z_n}^>. \quad (9.4)$$

Der ungünstigste Fall tritt auf, wenn die zufällig gewählte Zahl jedes Mal gerade die kleinste oder größte ist. Dann ist eine der Mengen $M^<$, $M^>$ leer, die andere enthält $n - 1$ Elemente, und entsprechend in den anderen rekursiven Schritten. Man benötigt dann $(n - 1) + (n - 2) + \dots + 1 = n(n - 1)/2$ Vergleiche, d.h. am Schluss sind *alle* möglichen Vergleiche durchgeführt worden.

Der günstigste Fall ist der, in dem die Listen in jedem Schritt möglichst genau halbiert werden, also $M^<$ und $M^>$ jeweils $(n - 1)/2$ Elemente enthalten, wenn n ungerade ist, und $n/2$ bzw. $(n - 2)/2$ Elemente, wenn n gerade ist, und entsprechend in den späteren Schritten. Maximal $\lceil \log_2 n \rceil$ „Halbierungen“ sind möglich. Es ergeben sich rund $n \lceil \log_2 n \rceil$ Abfragen.

Für uns ist wieder der Fall zufälliger Permutationen am interessantesten. Sei $\mathbb{N}_n = \{1, 2, \dots, n\}$ und $\mathbb{N}_{n,k} = \mathbb{N}_n \setminus \{k\}$. Der Zufallsvektor $X = (X_1, \dots, X_n)$ beschreibe eine zufällige Permutation von $\mathbb{N}_n$. T sei gleichverteilt auf $\mathbb{N}_n$ und unabhängig von X . Ist $T = i$, so wird in Schritt 2 der Wert von X_i mit den übrigen X_j verglichen.

Sei Π_n die Menge der Permutationen von $\mathbb{N}_n$ und $\Pi_{n,k}$ die Menge der Permutationen von $\mathbb{N}_{n,k}$. Für jedes $(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) \in \Pi_{n,k}$ gilt:

$$\begin{aligned} P(X_1 = x_1, \dots, X_{i-1} = x_{i-1}, X_{i+1} = x_{i+1}, \dots, X_n = x_n, T = i, Z_n = k) \\ = P(X_1 = x_1, \dots, X_{i-1} = x_{i-1}, X_i = k, X_{i+1} = x_{i+1}, \dots, X_n = x_n, T = i) \\ = n^{-1} (n!)^{-1}. \end{aligned}$$

Summation über alle Elemente von $\Pi_{n,k}$ liefert $P(Z_n = k, T = i) = n^{-2}$. Daher ist die bedingte Wahrscheinlichkeit

$$P(X_1 = x_1, \dots, X_{i-1} = x_{i-1}, X_{i+1} = x_{i+1}, \dots, X_n = x_n \mid T = i, Z_n = k) \quad (9.5)$$

eine Gleichverteilung auf $\Pi_{n,k}$.

Seien $Y_{k,1}, \dots, Y_{k,k-1}$ die Zahlen $\{1, 2, \dots, k - 1\}$ in der Reihenfolge, in der sie in X vorkommen und $Y_{k,k+1}, \dots, Y_{k,n}$ die Zahlen $\{k + 1, \dots, n\}$ ebenfalls in der Reihenfolge, in der sie in X vorkommen. Wie in unseren Überlegungen über Mergesort sieht man nun, dass die Zufallsvektoren

$$Y_k^< = (Y_{k,1}, \dots, Y_{k,k-1}) \quad \text{und} \quad Y_{n-k}^> = (Y_{k,k+1}, \dots, Y_{k,n})$$

unabhängig und gleichverteilt auf der Menge der Permutationen von $\{1, \dots, k\}$ und $\{k + 1, \dots, n\}$ sind. Sei $Q_k^<$ die Zahl der Vergleiche, die Quicksort zum Sortieren von $Y_k^<$ braucht, und $Q_{n-k}^>$ die Zahl der Vergleiche, die Quicksort zum Sortieren von $Y_{n-k}^>$ braucht. Ist $Z_n = k$, so gibt $Y_k^<$ gerade die Elemente von $M^<$ und $Y_{n-k}^>$ die von $M^>$ an. Also gilt (9.4), und zwar nicht nur im Sinne von Verteilungsgleichheit, wie in der Literatur behauptet wird.

Damit wir bei gegebenem $\{Z_n = k\}$ die Verteilungen von $Q_k^<$ und $Q_{n-k}^>$ benutzen können, müssen wir natürlich noch prüfen, dass die bedingte Verteilung nicht verschieden ist. Das ergibt sich aber aus (9.5). Aus der Gleichverteilung folgt nämlich

$$\begin{aligned} P(Y_{k,1} = y_1, \dots, Y_{k,k-1} = y_{k-1}, Y_{k,k+1} = y_{k+1}, \dots, Y_{k,n} = y_n \mid T = i, Z_n = k) \\ = \frac{1}{(k-1)!(n-k)!} \end{aligned}$$

für jeden möglichen Wert von $Y_k^<$ und $Y_{n-k}^>$. Nach Satz 2.2 (iv) können wir in dieser Gleichung noch die Bedingung $\{T = i\}$ weglassen.

Setzen wir $w_{n,t} = P(Q_n = t)$, so ergibt sich nun für die Verteilungen die Rekursion

$$\begin{aligned} w_{n,n-1+q} &= \sum_{k=1}^n P(Z_n = k) P(Q_{k-1}^< + Q_{k-1}^> = q) \\ &= \frac{1}{n} \sum_{k=1}^n \sum_{r=1}^q w_{k-1,r} \cdot w_{n-k,q-r}. \end{aligned}$$

Wir können nun ziemlich leicht den Erwartungswert $a_n = E(Q_n)$ bestimmen. Es ist

$$\begin{aligned} a_n &= n - 1 + \sum_{k=1}^n \frac{1}{n} E(Q_{k-1}^< + Q_{k-1}^>) \\ &= n - 1 + \sum_{k=1}^n \frac{1}{n} (a_{k-1} + a_{n-k}) \\ &= n - 1 + \frac{2}{n} \sum_{i=1}^{n-1} a_i. \end{aligned}$$

Zieht man $(n-1)a_{n-1}/n$ ab, so ergibt sich nach kurzer Umformung

$$\frac{a_n}{n+1} = \frac{a_{n-1}}{n} + \frac{2(n-1)}{n(n+1)}.$$

Sei $c_n = \sum_{i=1}^n i^{-1}$. Aus der letzten Gleichung folgt nun induktiv

$$\frac{a_n}{n+1} = 2c_{n+1} + \frac{2}{n+1} - 4.$$

$c_n - \log n$ konvergiert bekanntlich gegen die eulersche Konstante $C = 0,5772 \dots$. Damit ergibt sich für $E(Q_n)$ eine Approximation der Form

$$E(Q_n) = 2n \log n + n(2C - 4) + 2 \log n + 2C + 1 + o(n^{-1} \log n)$$

Siehe Knuth (1973).

Dass die Laufzeit von Quicksort sich in der Regel nicht stark von $2n \log n$ unterscheidet, ergibt sich auch aus Resultaten von U. Rösler (1991): Sei

$$R_n = n^{-1} (Q_n - E(Q_n)).$$

Rösler zeigte mit einem Fixpunktargument in einem geeignet metrisierten Raum von Wahrscheinlichkeitsmaßen, dass $P(R_n \leq x)$ gegen eine von ihm charakterisierte Verteilungsfunktion konvergiert. Da dies ein wenig Maßtheorie voraussetzt, verweisen wir den interessierten Leser auf die Originalarbeit oder die Überblicksarbeit von Rösler und Rüschemdorf (1999).

Eine untere Abschätzung: Wir wollen nun noch die Ergebnisse aus § 8 nutzen, um eine untere Abschätzung für die erwartete Laufzeit von Sortieralgorithmen anzugeben. Wir beschränken uns auf Algorithmen ohne Zufallsschritte, (obwohl diese Einschränkung sich eliminieren lässt). Wird ein solcher Algorithmus auf eine zufällige Permutation

$x = (x_1, \dots, x_n)$ von $(1, 2, \dots, n)$ angewandt, so sind zuerst zwei Elemente x_i, x_j mit $i \neq j$ zu vergleichen. Sei $c_1 = 1$, falls $x_i < x_j$ gilt, und sonst $c_1 = 0$. Sind t Vergleiche durchgeführt und sind die Ergebnisse durch eine 0-1-Folge $c_1, \dots, c_t$ beschrieben, und ist die Folge x noch nicht fertig sortiert, so schreibt der Algorithmus den nächsten Vergleich vor. Sei $c_{t+1} = 0$ oder $= 1$ je nach Ausgang des Vergleichs. Ist $l = l(x)$ die Zahl der bis zur vollständigen Sortierung benötigten Vergleiche, so ist die Zuordnung $x \mapsto (c_1, \dots, c_l)$ ein Präfixcode. Die erwartete Zahl der Vergleiche ist also gleich der erwarteten Codewortlänge, und diese ist $\geq H(X)$, wobei X die Gleichverteilung auf der Menge der Permutationen von $(1, \dots, n)$ besitzt. Es ist also $H(X) = \log_2 n!$. Nach der Abschätzung (5.3) erhalten wir

$$\log_2 n! \geq n \log_2 n - n \log_2 e + \log_2 \sqrt{2\pi n} + (12n + 1)^{-1} \log_2 e.$$

Wir sehen, dass der größte Term dieser Abschätzung gleich dem der erwarteten Laufzeit von Mergesort ist. Quicksort ist im Mittel um den Faktor $2 \log n / \log_2 n = 2 \log 2 \approx 1,39$ langsamer, hat aber eine vorteilhafte einfache rekursive Struktur. Das Betriebssystem UNIX verwendet eine Variante von Quicksort, bei der nicht direkt mit einem zufällig gewählten x_i verglichen wird, sondern mit dem mittleren von 3 zufällig gewählten Elementen aus x .

Bei Simulationen erweist sich Mergesort trotz des kleineren Erwartungswerts der Laufzeit nicht als überlegen gegenüber Quicksort. Dies dürfte daran liegen, dass die *tatsächliche* Laufdauer der Sortierung nicht nur von der Zahl der angestellten Vergleiche abhängt.

Anmerkungen: Mergesort wurde 1945 von Johann von Neumann und Quicksort wurde von Hoare (1961) vorgeschlagen. Die Einschränkung $n = 2^m$ ist für die Analyse bequem, aber nicht nötig. Die erwähnte Beweismethode für Verteilungskonvergenz mittels Banachs Fixpunktsatz und geeigneten Wahrscheinlichkeitsmetriken, die unabhängig von Rösler (1991) und Rachev und Rüschendorf (1995) für Probleme dieses Typs entwickelt wurde, hat sich inzwischen als schlagkräftiges Instrument auch für andere Problemstellungen erwiesen. (Siehe Rösler-Rüschendorf (1999).)

Aufgaben

1. Jede Permutation $\pi = (\pi_1, \pi_2, \dots, \pi_N)$ von $(1, \dots, N)$ sei gleichwahrscheinlich. Die Elemente von π sollen geordnet werden. Der Algorithmus „Insertion Sort“ sortiert zuerst π_1 und π_2 . Sind $\pi_1, \dots, \pi_{i-1}$ sortiert, und stehen sie (aufsteigend) der Größe nach geordnet in $A[1], \dots, A[i-1]$, so vergleicht der Algorithmus $v = \pi_i$ mit $A[i-1]$. Ist $v > A[i-1]$, so kommt v in $A[i]$, und die ersten i Zahlen sind sortiert. Andernfalls wird $A[i] := A[i-1]$ gesetzt und v mit $A[i-2]$ verglichen, usw. Sei $Y_i(\pi) = \text{card}\{j < i : \pi_j > \pi_i\}$. Der Sortieraufwand wird durch $X = \sum_{i=1}^N Y_i$ beschrieben. (Genauer durch $X + N - 1$.)
 - a) Zeigen Sie, dass die Y_i unabhängig sind mit $P(Y_i = j) = 1/i$ für $0 \leq j < i$.
 - b) Bestimmen Sie die erzeugende Funktion von X .
 - c) Geben Sie ein Rekursionsverfahren zur Berechnung der Verteilung von X an.
2. (Algorithmus von Steinhaus) Sei $x_1, \dots, x_n$ zu sortieren. Vergleiche zuerst x_1 und x_2 . Ist $b_1 < b_2 < \dots < b_k$ das Ergebnis des Sortierens von $x_1, \dots, x_k$, so sortiere x_{k+1} wie folgt ein: Vergleiche x_{k+1} zuerst mit $b_{\lceil k/2 \rceil}$. (Dabei ist $\lceil \alpha \rceil$ die kleinste ganze Zahl $\geq \alpha$.) Ist $x_{k+1} > b_{\lceil k/2 \rceil}$, so vergleiche x_{k+1} mit $b_{\lceil 3k/4 \rceil}$, andernfalls mit $b_{\lceil k/4 \rceil}$, usw. Zeigen Sie, dass die vollständige Ordnung nach maximal $\log_2(n!) + n - 2$ Vergleichen gefunden ist. Bestimmen Sie eine Rekursionsformel für den Erwartungswert der Zahl der benötigten Vergleiche für das Einordnen von x_{k+1} in $b_1 < \dots < b_k$ unter Annahme der Gleichverteilung auf der Menge aller Permutationen.
3. Seien n Personen (z.B. eine Kompanie) auf Syphilis zu testen. Jede sei mit Wahrscheinlichkeit p infiziert, unabhängig von den übrigen. Folgendes Vorgehen ist möglich: Ein Teil des Bluts aus

den Blutproben von je k Personen wird gemischt und gemeinsam getestet. Ist eine Probe negativ, so sind alle Beteiligten frei von Syphilis. Ist eine Probe positiv, so wird der Rest des Blutes aus den Blutproben der beteiligten Personen separat getestet. Zeigen Sie: Für $n \geq 2$ ist es genau dann optimal nur Einzeltests durchzuführen, wenn $p \geq 3/2 - (1/2)\sqrt{5}$ ist. Zeigen Sie: Für jedes Testverfahren ist der Erwartungswert der Zahl der benötigten Tests

$$\geq n[-p \log_2 p - (1-p) \log_2(1-p)].$$

Kapitel II Allgemeine Modelle

Die bisher eingehaltene Einschränkung auf diskrete Wahrscheinlichkeitsräume ist in mancher Hinsicht unnatürlich. Sie war nur zweckmäßig, weil im diskreten Fall eine Reihe technischer Schwierigkeiten noch nicht auftreten. Aber auch nichtdiskrete Modelle spielen eine große Rolle. Bei ihrer Diskussion lassen wir einige Resultate über maßtheoretische Grundlagen unbewiesen, da deren Beweis für das Verständnis der wahrscheinlichkeitstheoretischen und statistischen Ideen nicht unbedingt erforderlich ist. Wer hinreichend motiviert ist, kann diese Beweise leicht später in der angegebenen Literatur nachlesen.

§ 10 Wahrscheinlichkeitsmaße mit Dichten

Neben den diskreten Wahrscheinlichkeitsmaßen werden uns vor allem solche mit Dichten interessieren. Es ist aber ökonomisch, die Grundbegriffe gleich allgemein zu formulieren.

10.1 σ -Algebren und allgemeine Wahrscheinlichkeitsmaße

Um ein mathematisches Modell für ein Zufallsexperiment mit beliebig vielen möglichen Ausgängen zu bilden, muss man wieder festlegen, welche nichtleere Menge Ω als Beschreibung der Menge der möglichen Versuchsergebnisse dienen soll. Ω braucht jetzt nicht mehr endlich oder abzählbar zu sein. Wir bestehen nicht darauf, dass die Ergebnisse $\omega \in \Omega$ tatsächlich beobachtbar sind. Z.B. kann man als Menge der möglichen Ergebnisse einer Längenmessung die Menge der nichtnegativen reellen Zahlen vorsehen, obwohl natürlich nicht unendlich genau gemessen werden kann. Als mathematische Beschreibung der Ergebnisse einer Folge von Münzwürfen, für deren Länge keine Grenze festgelegt ist, werden wir unendliche Folgen $\omega = (K, Z, Z, Z, K, Z, K, K, \dots)$ von Symbolen K (für Kopf) und Z (für Zahl) verwenden. Beobachtet man die Zeitpunkte $s_1 < s_2 < \dots$, an denen ein Geigerzähler eine radioaktive Emission anzeigt, so kann man als Ergebnismenge Ω die Menge all dieser Folgen von reellen Zahlen betrachten. Solche Beschreibungen sind mathematische Idealisierungen der Ergebnisse eines nur gedachten Gesamtexperiments, das konkret nie beendet werden könnte. Nichtsdestoweniger ist es in der Wahrscheinlichkeitstheorie sinnvoll, mit solchen Idealisierungen zu arbeiten, genauso wie die Geometer mit unendlichen Geraden arbeiten, obwohl noch nie jemand eine unendliche Gerade gesehen hat.

Will man nun — wie im diskreten Fall — *allen* Teilmengen A von Ω eine Wahrscheinlichkeit $P(A)$ zuordnen und dabei die in Kapitel I benutzten Eigenschaften von Wahrscheinlichkeitsmaßen bewahren, so trifft man auf prinzipielle mathematische Hindernisse, die im Anhang zu diesem Paragraphen kurz geschildert werden.

Der übliche Ausweg aus dem Dilemma besteht darin, dass man darauf verzichtet, $P(A)$ für *alle* $A \subset \Omega$ zu definieren. Man ordnet nicht allen Teilmengen eine Wahrscheinlichkeit zu, sondern nur allen Mengen eines Mengensystems $\mathcal{A}$, das hinreichend reichhaltig ist.

Hinreichend reichhaltig soll dabei zweierlei bedeuten: Erstens soll $\mathcal{A}$ gewisse Mengen A einfacher Bauart enthalten, z.B. alle Intervalle, wenn $\Omega = \mathbb{R}^+$ ist, oder die Menge der

Ergebnisse einer unendlichen Folge von Münzwürfen, für die die ersten vier Würfe die Folge (K, Z, Z, Z) ergeben. Man kann damit rechnen, dass sich $P(A)$ für solche einfachen Mengen A leichter festlegen lässt als für kompliziertere Mengen, z.B. beliebige offene Mengen.

Zweitens soll „hinreichend reichhaltig“ heißen, dass wir wie bisher mit Mengen operieren wollen. Ist $A \in \mathcal{A}$, so soll z.B. auch $A^c \in \mathcal{A}$ liegen. Wenn man in $\mathcal{A}$ alle abzählbaren Mengenoperationen durchführen kann, ohne $\mathcal{A}$ zu verlassen, nennen wir $\mathcal{A}$ σ -Algebra. Man braucht aber nur etwas weniger zu fordern, weil der Rest dann folgt:

Definition 10.1 Sei $\Omega \neq \emptyset$ beliebig. Eine Familie $\mathcal{A}$ von Teilmengen von Ω heißt **σ -Algebra** (gelegentlich auch σ -Körper), wenn gilt:

$$\Omega \in \mathcal{A}, \quad (10.1)$$

$$A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}, \quad (10.2)$$

$$A_1, A_2, A_3, \dots \in \mathcal{A} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{A}. \quad (10.3)$$

(Man überlege sich, dass dann auch $\emptyset \in \mathcal{A}$ liegt, und dass z.B. der Durchschnitt von abzählbar vielen Mengen aus $\mathcal{A}$ wieder zu $\mathcal{A}$ gehört.)

Wir wollen nun zeigen, dass immer eine kleinste σ -Algebra existiert, die ein vorgegebenes System $\mathcal{F}$ von „einfachen“ Mengen enthält. Dazu brauchen wir

Lemma 10.2 Sei $I \neq \emptyset$ eine beliebige Indexmenge, und sei für alle $i \in I$ $\mathcal{A}_i$ eine σ -Algebra in Ω . Dann ist das System

$$\mathcal{A}_I := \bigcap_{i \in I} \mathcal{A}_i$$

aller Mengen $A \subset \Omega$, die für alle i zu $\mathcal{A}_i$ gehören, eine σ -Algebra.

Beweis: Ω gehört zu allen $\mathcal{A}_i$, also auch zu $\mathcal{A}_I$. Sei nun $A \in \mathcal{A}_I$. Dann gilt für alle i $A \in \mathcal{A}_i$ und also auch $A^c \in \mathcal{A}_i$, denn $\mathcal{A}_i$ ist σ -Algebra. Da i beliebig war, folgt $A^c \in \mathcal{A}_I$. Genauso zeigt man (10.3) für $\mathcal{A}_I$. $\square$

Satz 10.3 Sei $\Omega \neq \emptyset$ und $\mathcal{F}$ eine beliebige Familie von Teilmengen von Ω . Dann gibt es unter den σ -Algebren, die $\mathcal{F}$ enthalten, eine kleinste, nämlich die Familie $\mathcal{A}(\mathcal{F})$ aller Mengen A , die zu jeder σ -Algebra in Ω gehören, welche $\mathcal{F}$ enthält. Man nennt sie die von $\mathcal{F}$ erzeugte σ -Algebra.

Beweis: Es gibt mindestens eine σ -Algebra, die $\mathcal{F}$ enthält, nämlich die Familie $\mathcal{P}(\Omega)$ aller Teilmengen von Ω . Ist $\{\mathcal{A}_i : i \in I\}$ eine beliebige Indizierung der σ -Algebren, die $\mathcal{F}$ enthalten, so ist also $I \neq \emptyset$ und $\mathcal{A}(\mathcal{F}) = \mathcal{A}_I$. $\square$

Satz 10.3 ist ein reiner Existenzsatz. Nur in ganz speziellen Fällen, z.B. für endliches $\mathcal{F}$, lässt sich $\mathcal{A}(\mathcal{F})$ explizit beschreiben. Für uns ist vor allem ein anderer Fall interessant, nämlich der, in dem $\Omega = \mathbb{R}^n$ ist und $\mathcal{F}$ die Familie aller (z.B.) nach links halboffenen Intervalle. Dabei ist für $a = (a_1, \dots, a_n)$, $b = (b_1, \dots, b_n) \in \mathbb{R}^n$ mit $a \leq b$ (d.h. $a_i \leq b_i$ für alle i) das nach links halboffene Intervall $]a, b]$ definiert durch

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

Ist $\mathcal{F}$ die Familie aller dieser Intervalle, so heißt $\mathcal{B} := \mathcal{A}(\mathcal{F})$ die **borelsche σ -Algebra** in $\mathbb{R}^n$, und die zu $\mathcal{B}$ gehörigen Mengen heißen dann **borelsche Mengen**. Im Fall $\Omega = \mathbb{R}^n$ werden dies die „Ereignisse“ sein. Geht man statt von $\mathcal{F}$ vom System aller offenen oder vom System aller abgeschlossenen Intervalle im $\mathbb{R}^n$ aus, so erhält man das gleiche Mengensystem $\mathcal{B}$.

[Die Bezeichnung borelsche σ -Algebra wird allgemein in einem topologischen Raum Ω für die von der Familie der offenen Mengen erzeugte σ -Algebra benutzt. Im Fall des $\mathbb{R}^n$ ist es nicht schwer zu zeigen, dass diese mit $\mathcal{B}$ übereinstimmt.]

Da sich jede offene Teilmenge des $\mathbb{R}^n$ als abzählbare Vereinigung von Intervallen schreiben lässt, ist jede offene Menge in $\mathbb{R}^n$ borelsch, und da jede abgeschlossene Menge Komplement einer offenen ist, ist jede abgeschlossene Menge borelsch. Aber das ist erst der Anfang. Nicht nur alle geometrisch irgendwie vorstellbaren Mengen sind borelsch, sondern auch fast beliebige komplizierte, bei denen die Vorstellung versagt.

[Dass es überhaupt nicht borelsche Mengen im $\mathbb{R}^n$ gibt, zeigt man gewöhnlich mit Hilfe des Auswahlaxioms. Uns werden solche Mengen hier nicht begegnen.]

Zum Glück ist es nicht notwendig, alle borelschen Mengen zu kennen, oder die Wahrscheinlichkeiten $P(A)$ aller borelschen Mengen festzulegen. Die große Allgemeinheit dient vielmehr der Bequemlichkeit. Der springende Punkt ist der folgende:

Hat man die Wahrscheinlichkeiten $P(A)$ für alle A aus der obigen Familie $\mathcal{F}$ festgelegt, und genügen sie dort den Rechenregeln für Wahrscheinlichkeiten, so ist dadurch automatisch $P(B)$ für alle $B \in \mathcal{A}(\mathcal{F})$ eindeutig bestimmt, und zwar so, dass auch in $\mathcal{A}(\mathcal{F})$ die Rechenregeln weiter gelten.

Die Familie $\mathcal{F}$ ist also nur der Eingang zu einem großen Garten, den man nie verlässt solange man aus den dort vorgefundenen Ereignissen $A \subset \Omega$ neue nur mit abzählbaren mengentheoretischen Operationen bildet, und in dem die Gültigkeit der Rechenregeln gewährleistet ist. Niemand ist gezwungen, sich den ganzen Garten anzusehen. Tatsächlich bleibt man gewöhnlich sehr nah beim Eingang. Dass $P(B)$ für alle $B \in \mathcal{A}(\mathcal{F})$ in der angegebenen Weise bestimmt ist, ist ein Ergebnis der Maßtheorie, das wir hier nicht beweisen wollen.

In anderen Räumen als $\mathbb{R}^n$ benutzt man andere Familien $\mathcal{F}$. Wir definieren daher — bildlich gesprochen — einen Wahrscheinlichkeitsraum als einen Garten mit den genannten Eigenschaften ohne Bezug auf den Eingang.

Definition 10.4 Ein **messbarer Raum** ist ein Paar $(\Omega, \mathcal{A})$ bestehend aus einer nicht-leeren Menge Ω und einer σ -Algebra $\mathcal{A}$ von Teilmengen von Ω . Ein **Wahrscheinlichkeitsmaß** P ist eine auf $\mathcal{A}$ definierte Funktion mit Werten in $[0, 1]$, welche den folgenden drei Bedingungen genügt:

$$P(A) \geq 0 \quad \text{für alle } A \in \mathcal{A}; \quad (10.4)$$

$$P(\Omega) = 1; \quad (10.5)$$

P ist **σ -additiv**, d.h. für disjunkte $A_1, A_2, \dots \in \mathcal{A}$ ist

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i). \quad (10.6)$$

$(\Omega, \mathcal{A}, P)$ heißt dann **Wahrscheinlichkeitsraum**, P auch **Wahrscheinlichkeitsverteilung**. Teilmengen $A \subset \Omega$, die zu $\mathcal{A}$ gehören, heißen **Ereignisse**.

Die früher angegebenen Eigenschaften von Wahrscheinlichkeitsmaßen gelten weiter, nur dass wir von allen auftretenden Teilmengen annehmen müssen, dass sie zu $\mathcal{A}$ gehören. Der Fall der diskreten Wahrscheinlichkeitsräume ergibt sich hier als der Spezialfall, in dem Ω abzählbar und $\mathcal{A} = \mathcal{P}(\Omega)$ die Menge aller Teilmengen von Ω ist.

Zur Berechnung von Wahrscheinlichkeiten ist der folgende Satz oft nützlich.

Satz 10.5 *Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum. Ist $B_1 \subset B_2 \subset \dots$ eine wachsende Folge von Ereignissen und B deren Vereinigung, so gilt*

$$P(B) = \lim_{i \rightarrow \infty} P(B_i). \quad (10.7)$$

Ist $C_1 \supset C_2 \supset \dots$ eine fallende Folge von Ereignissen und C deren Durchschnitt, so ist

$$P(C) = \lim_{i \rightarrow \infty} P(C_i). \quad (10.8)$$

Beweis: Dies war für diskrete Wahrscheinlichkeitsräume schon in § 2.5 gezeigt worden, und der gleiche Beweis gilt hier weiter. $\square$

Wir bemerken, dass umgekehrt aus endlicher Additivität und (10.7) σ -Additivität folgt. Man braucht nur $B_n = \bigcup_{i=1}^n A_i$ zu setzen. Da (10.7) und (10.8) im Falle endlicher Additivität durch Übergang zu Komplementen äquivalent sind, folgt σ -Additivität auch aus endlicher Additivität und (10.8).

10.2 Beispiele von Verteilungen mit Dichten

Eindimensionale Dichten: Eine Funktion F auf $\mathbb{R}$ mit Werten in $[0, 1]$ heißt **Verteilungsfunktion**, wenn sie *rechtsstetig* und (nicht notwendig strikt) *monoton wachsend* ist, und wenn für $x \rightarrow -\infty$ $F(x) \rightarrow 0$, und für $x \rightarrow \infty$ $F(x) \rightarrow 1$ gilt.

Ist P ein Wahrscheinlichkeitsmaß auf $\mathbb{R}$, und setzt man $F(x) = P(]-\infty, x])$, so ist F Verteilungsfunktion. Die Monotonie ist klar. Ist nun $x_1 \geq x_2 \geq \dots$ eine von rechts gegen x strebende Folge, so ist $C =]-\infty, x]$ der Durchschnitt der $C_i =]-\infty, x_i]$ und also nach Satz 10.5 $F(x) = P(C) = \lim P(C_i) = \lim F(x_i)$. Für $x_i \rightarrow -\infty$ gilt $F(x_i) \rightarrow 0$, weil dann $C = \emptyset$ ist. Aus (10.7) folgt analog $F(x) \rightarrow 1$ für $x \rightarrow \infty$.

Man kann aber auch umgekehrt von einer beliebigen Verteilungsfunktion F ausgehen, und für Intervalle $]a, b]$ eine Wahrscheinlichkeit P durch $P(]a, b]) = F(b) - F(a)$ definieren. In der Maßtheorie zeigt man, dass es auf der σ -Algebra der borelschen Mengen von $\mathbb{R}$ ein eindeutig bestimmtes Wahrscheinlichkeitsmaß P gibt, das für die Intervalle $]a, b]$ die angegebenen Werte hat.

Der mit Abstand wichtigste Fall ist der, in dem F durch eine Dichte gegeben ist. Eine **Dichte** auf $\mathbb{R}$ ist eine nichtnegative Funktion f auf $\mathbb{R}$ mit

$$\int_{-\infty}^{\infty} f(x) dx = 1. \quad (10.9)$$

Das Integral soll dabei wohldefiniert sein. Wir benötigen hier nur den Fall, in dem f stetig bis auf höchstens endlich viele Sprungstellen ist, so dass man mit dem Riemann-Integral auskommt. Ist f eine Dichte, so wird durch

$$F(x) = \int_{-\infty}^x f(t) dt \quad (10.10)$$

eine Verteilungsfunktion definiert. Die Wahrscheinlichkeit von Intervallen ist dann also durch

$$P([a, b]) = \int_a^b f(t) dt \quad (10.11)$$

gegeben. Ist (a_n) eine wachsende Folge mit $a_n < b$, die gegen b strebt, so ist $\{b\}$ der Durchschnitt der $[a_n, b]$, und die über diese Intervalle erstreckten Integrale streben gegen 0. Also ist $P(\{b\}) = 0$ für alle b . Kein Ergebnis $b \in \mathbb{R}$ hat also positive Wahrscheinlichkeit.

Ist F eine beliebige stetige Verteilungsfunktion auf $\mathbb{R}$, die auf dem Komplement einer endlichen oder leeren Menge C stetig differenzierbar ist, so wird durch

$$f(x) = \frac{d}{dx} F(x) \quad (x \in \mathbb{R} \setminus C) \quad (10.12)$$

eine Dichte f zu F definiert. (Auf C kann man f beliebig festsetzen.)

Ist der Stichprobenraum Ω nur ein Teilintervall von $\mathbb{R}$, so geht man ganz analog vor. Als σ -Algebra wählt man dann das System der borelschen Teilmengen von Ω .

Spezialfälle: (a) **Gleichverteilung in einem Intervall $[a, b]$:** Dies ist die Verteilung mit der Dichte f , die in $[a, b]$ den Wert $1/(b-a)$ und außerhalb von $[a, b]$ den Wert 0 annimmt. Die Wahrscheinlichkeit eines Teilintervalls I von $[a, b]$ ist also proportional zur Länge von I .

Die Gleichverteilung ist ein nahe liegendes Modell für Reste, die man erhält, wenn man breit streuende Größen um die Vielfachen einer festen Zahl reduziert. Ein beliebtes Bild ist das des Zufallsrades, auf dessen Peripherie eine Skala von 0 bis 2π angebracht ist. Nach einer schwungvollen Drehung bleibt es so stehen, dass ein starrer Zeiger auf eine zufällige Stelle ω der Peripherie zeigt. Als Modell würde man $\Omega = [0, 2\pi[$ mit der Gleichverteilung nehmen. Werden Messungen reellwertiger Größen auf die jeweils nächstgelegene ganze Zahl hin auf- bzw. abgerundet, so betrachtet man gewöhnlich den Abrundungsfehler als gleichverteilt in $[-1/2, +1/2]$. Kommt man zu einer zufälligen Zeit an eine geschlossene Schranke, die insgesamt 2 Minuten geschlossen ist, so ist es nahe liegend anzunehmen, die Wartezeit bis zur Öffnung der Schranke sei gleichverteilt in $[0, 2]$.

Wichtig ist auch, dass man die Gleichverteilung leicht in andere Verteilungen transformieren kann und umgekehrt. Darauf gehen wir später ein.

(b) **Exponentialverteilung:** Für jedes $\lambda > 0$ ist durch

$$f_\lambda(x) = \lambda e^{-\lambda x} \quad (x \geq 0)$$

(und $f_\lambda(x) = 0$ für $x < 0$) eine Dichte definiert. Sie heißt Dichte der Exponentialverteilung mit Parameter λ und ist ein Analogon zur geometrischen Verteilung für den Fall stetiger Zeit; vgl. Bild 10.1.

Ereignen sich gewisse Vorkommnisse, etwa Meteoriteneinschläge in einem bestimmten Gebiet, Flugzeugabstürze oder radioaktive Zerfallsvorgänge unabhängig voneinander mit zeitlich konstanter „Rate“, so eignet sich die Exponentialverteilung als Verteilung der Wartezeit bis zum ersten Eintreten eines solchen Vorkommnisses von einem festen Zeitpunkt an gerechnet. (Beim Studium von Poisson-Prozessen in § 18 gehen wir darauf noch genauer ein.)

Charakteristische Eigenschaft der Exponentialverteilung ist ihre Gedächtnislosigkeit. Wir nennen eine Verteilung P auf dem offenen Intervall $]0, \infty[$ **gedächtnislos**, wenn für alle positiven t und x mit $P(]t, \infty[) > 0$

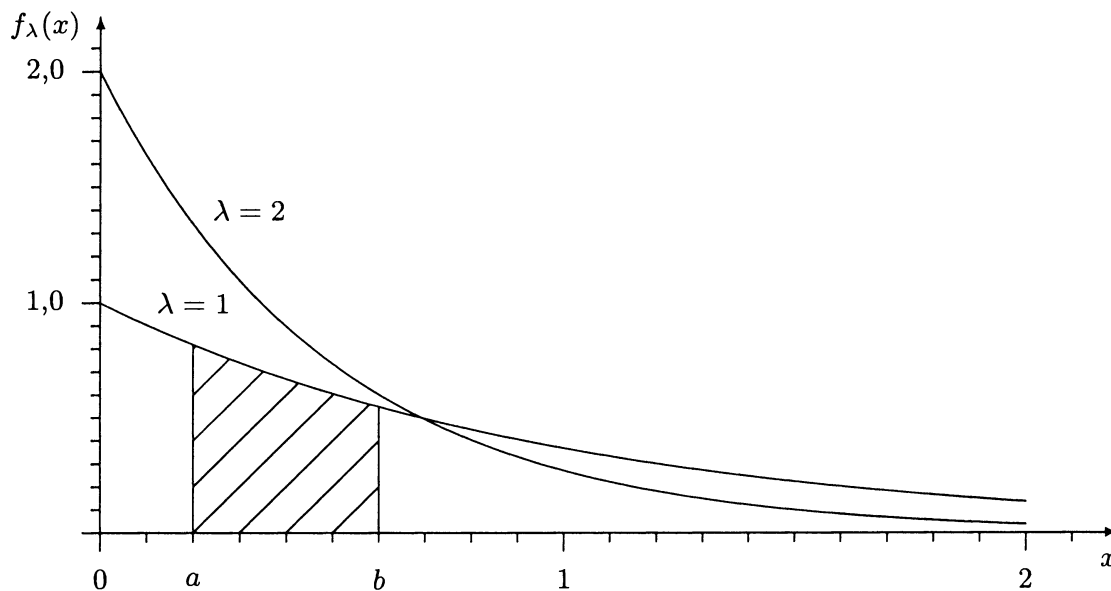


Bild 10.1 Dichten der Exponentialverteilungen für $\lambda = 1$ und $\lambda = 2$. Die schraffierte Fläche ist im Fall $\lambda = 1$ die dem Intervall $[a, b]$ zugeordnete Wahrscheinlichkeit.

$$P(\cdot | x+t, \infty[\mid \cdot | t, \infty[) = P(\cdot | x, \infty[) \quad (10.13)$$

ist.

(Ist P z.B. die Verteilung der Wartezeit bis zum ersten Meteoriteneinschlag, so bedeutet dies: Ist der erste Einschlag bis zum Zeitpunkt t noch nicht beobachtet worden, so ist die Wahrscheinlichkeit, dass man noch mindestens x Zeiteinheiten darauf warten muss, genauso groß wie zur Zeit 0.) Ist F die Verteilungsfunktion zu P und $G(x) = 1 - F(x)$, so ist (10.13) gleich bedeutend damit, dass G die Funktionalgleichung

$$\frac{G(x+t)}{G(t)} = G(x) \quad (10.14)$$

erfüllt. Für die Exponentialverteilung ist $G(x) = e^{-\lambda x}$. Also ist die Exponentialverteilung gedächtnislos.

Ist nun P eine beliebige gedächtnislose Verteilung auf $]0, \infty[$, so ist für hinreichend kleines t $G(t) > 0$, denn $G(1/n) = P(\cdot | 1/n, \infty[)$ strebt wegen (10.8) gegen $P(\cdot | 0, \infty[) = 1$. Aus (10.14) folgt dann leicht $G(s) = G(1)^s$ für rationales $s > 0$ und aus Monotoniegründen für alle $s > 0$. Wäre $G(1) = 1$, so wäre $F(s) = 0$ für alle $s > 0$, was $F(s) \rightarrow 1$ ($s \rightarrow \infty$) widerspricht. Also muss $0 < G(1) < 1$ gelten. Damit ist $G(s) = e^{-\lambda s}$ für ein $\lambda > 0$. Jede gedächtnislose Verteilung auf $]0, \infty[$ ist also eine Exponentialverteilung.

(c) **Normalverteilung:** Als (eindimensionale) Normalverteilung mit Erwartungswert μ und Varianz σ^2 , oder kurz als $N(\mu, \sigma^2)$ -Verteilung, bezeichnet man die Verteilung P mit der Dichte

$$\varphi_{\mu, \sigma^2}(x) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right). \quad (10.15)$$

(Dass μ ein Erwartungswert und σ^2 eine Varianz ist, können wir natürlich erst zeigen, wenn diese Begriffe für Verteilungen mit Dichten definiert sind.) Ist φ die in § 5 benutzte gaußsche Glockenkurve, so ist $\varphi_{0,1}(x) = \varphi(x)$ und $\varphi_{\mu, \sigma^2}(x) = (1/\sigma) \varphi((x - \mu)/\sigma)$. Mit Hilfe der Substitution $y = (x - \mu)/\sigma$ folgt daher

$$\int_a^b \varphi_{\mu, \sigma^2}(x) dx = \int_{a^*}^{b^*} \varphi(y) dy = \Phi(b^*) - \Phi(a^*)$$

mit $a^* = (a - \mu)/\sigma$ und $b^* = (b - \mu)/\sigma$.

Die $N(\mu, \sigma^2)$ -Verteilung dient oft als Approximation der Verteilung zufälliger Größen wie Kopfumfang, Ernteertrag, Niederschlagsmenge, die von vielen unabhängigen Größen beeinflusst sind. Natürlich sollte man sich davon überzeugen, dass im speziellen Fall diese Größen wirklich eine Verteilung mit glockenförmiger Dichte haben, und nicht nur aus Bequemlichkeit die Normalverteilung benutzen.

(d) **Weibull-Verteilung***: Hat die Verteilung der Lebensdauer eines Lebewesens, eines Geräts oder einer Maschine eine Dichte f auf $\mathbb{R}^+$, so kann man

$$r(t) = \frac{f(t)}{P(]t, \infty[)} = \frac{f(t)}{\int_t^\infty f(x) dx}$$

als **Ausfallrate** bezeichnen. Ist f stetig, so ist

$$r(t) = \lim_{\varepsilon \rightarrow 0+0} \frac{P(]0, t + \varepsilon[\mid]t, \infty[)}{\varepsilon}.$$

Aus der Gedächtnislosigkeit der Exponentialverteilung folgt, dass die Exponentialverteilung eine zeitunabhängige Ausfallrate hat. Die Verteilung auf $\mathbb{R}^+$ mit Dichte

$$f(x) = \alpha \beta x^{\beta-1} \exp(-\alpha x^\beta) \quad (x > 0)$$

nennt man Weibull-Verteilung mit Parameter (α, β) . Die Ausfallrate ist

$$r(t) = \alpha \beta x^{\beta-1}.$$

Für $\beta = 1$ ist die Weibull-Verteilung gerade die Exponentialverteilung mit Parameter $\lambda = \alpha$. Für $\beta > 1$ ist die Ausfallrate wachsend, für $\beta < 1$ fallend.

(e) **Von-Mises-Verteilung***: Für die Beschreibung der Verteilung zufälliger Richtungsdaten (z.B. Windrichtungen, Abflugrichtungen in einem Taubenflugexperiment) und für Häufigkeitsverteilungen zyklischer Art (z.B. Niederschlagsmengen im Jahresverlauf) benötigt man auch andere Verteilungen als die Gleichverteilung. Eine Verteilung, bei der eine Richtung $\vartheta_0 \in [0, 2\pi[$ mit maximaler Dichte bevorzugt wird, und die ebenfalls symmetrisch zu ϑ_0 eine glockenähnliche Gestalt hat, ist die Von-Mises-Verteilung. Sie wird durch ϑ_0 und einen Konzentrationsparameter κ festgelegt und hat in $\vartheta \in [0, 2\pi[$ die Dichte

$$g(\vartheta \mid \vartheta_0, \kappa) = \frac{1}{2\pi I_0(\kappa)} \exp(\kappa \cos(\vartheta - \vartheta_0)) \quad \text{mit} \quad I_0(\kappa) = \sum_{r=0}^{\infty} \frac{(\kappa/2)^{2r}}{(r!)^2}.$$

κ spielt die umgekehrte Rolle wie σ^2 bei der Normalverteilung. Große Werte von κ bedeuten, dass die Verteilung eng um ϑ_0 herum konzentriert ist. Bild 10.2 zeigt den Verlauf der Dichte in $[-\pi, \pi]$ für $\vartheta_0 = 0$ und drei Werte von κ .

Die Statistik von Richtungsdaten ist ein Sonderkapitel, das in den meisten Statistik-Büchern nicht erwähnt wird und auch hier nicht weiter verfolgt wird. Eine ausführliche Darstellung hat Mardia (1972) gegeben.

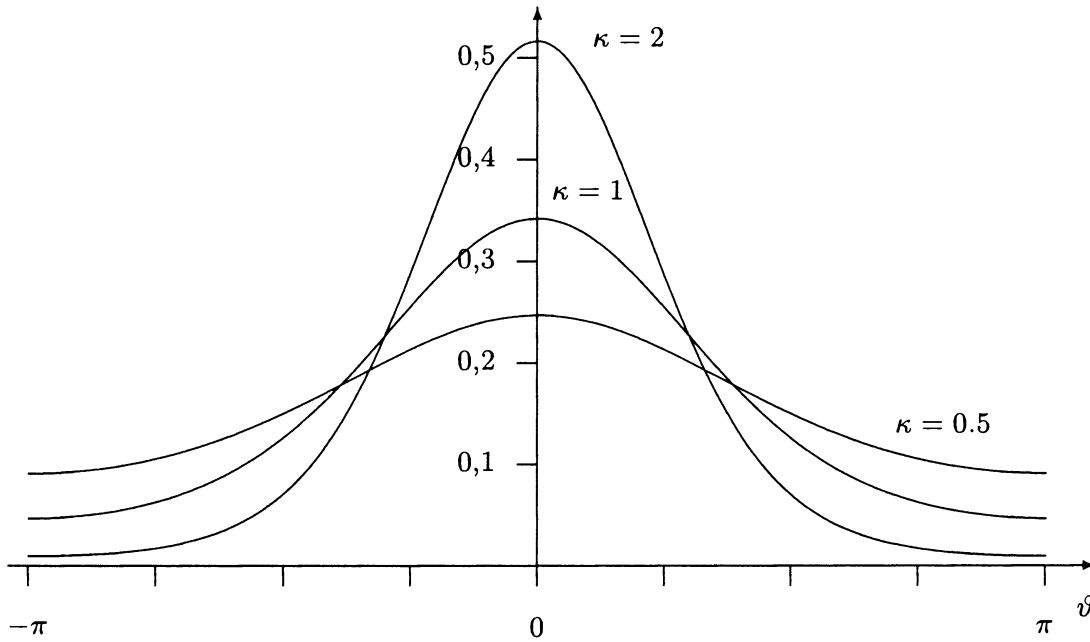


Bild 10.2 Dichte der Von-Mises-Verteilung

Weitere wichtige Verteilungen mit Dichten (χ^2 -Verteilung, t -Verteilung, F -Verteilung, Gamma-Verteilung) werden wir zusammen mit ihren statistischen Anwendungen besprechen. Eine reichhaltige Sammlung von Verteilungen mit Diskussion der wichtigsten Eigenschaften haben Johnson und Kotz (1970) zusammengestellt.

Dichten in $\mathbb{R}^n$: Eine Dichte in $\mathbb{R}^n$ ist eine nichtnegative integrierbare Funktion f auf $\mathbb{R}^n$ mit

$$\int_{\mathbb{R}^n} f(x_1, \dots, x_n) dx_1 \dots dx_n = 1. \quad (10.16)$$

Wieder sollen die Integrale dabei wohldefiniert sein, z.B. f stetig. Für $a = (a_1, \dots, a_n)$, $b = (b_1, \dots, b_n)$ betrachten wir wieder die Intervalle $]a, b]$. Wir setzen

$$P(]a, b]) = \int_{]a, b]} f(x) dx = \int_{a_1}^{b_1} \dots \int_{a_n}^{b_n} f(x_1, \dots, x_n) dx_n \dots dx_1, \quad (10.17)$$

wobei das in der Mitte stehende Integral einfach eine abkürzende Schreibweise für die rechte Seite ist. Es ergibt sich aus Ergebnissen der Maßtheorie, dass die Funktion P von der Familie $\mathcal{F}$ der Intervalle auf die von $\mathcal{F}$ erzeugte borelsche σ -Algebra $\mathcal{A}$ eindeutig fortgesetzt werden kann, und dass P auf $(\Omega, \mathcal{A})$ ein Wahrscheinlichkeitsmaß ist. Die Funktion

$$F(x) = P(\{y \in \mathbb{R}^n : y \leq x\})$$

auf $\mathbb{R}^n$ heißt wieder die zu P gehörige **Verteilungsfunktion**.

Beispiele: (a) **Gleichverteilung in einem Gebiet:** Sei G ein endliches Gebiet im $\mathbb{R}^2$ und $\lambda(G)$ seine Fläche. Das Wahrscheinlichkeitsmaß P mit der Dichte f , die in G den Wert $1/\lambda(G)$ und außerhalb von G den Wert 0 annimmt, heißt Gleichverteilung in G . Wenn wir sagen, dass ein Punkt zufällig in G gewählt wird, ist als Wahrscheinlichkeitsmaß für dieses Experiment die Gleichverteilung gemeint.

(b) **Produktdichten:** Sind n Teilexperimente durch Dichten f_i beschrieben, so verwendet man

$$f(x_1, \dots, x_n) = f_1(x_1) f_2(x_2) \dots f_n(x_n) \quad (10.18)$$

als Dichte für die Verteilung in $\mathbb{R}^n$, die die unabhängige Hintereinanderausführung der Teilexperimente beschreibt; vgl. Satz 11.7.

Anhang*

Grundlagenfragen: Noch zu Beginn des 20. Jahrhunderts waren die Grundlagen der Wahrscheinlichkeitstheorie ungeklärt. Selbst der laplacesche Begriff war nicht eindeutig, denn er ließ offen, was gleichmögliche Fälle waren. Z.B. sprach Markow noch 1912, wenn aus einer Urne mit unbekannten Zahlen von weißen, schwarzen, roten und blauen Kugeln gezogen wurde, davon, dass alle Farben gleichmöglich seien. Für ihn war die Wahrscheinlichkeit, eine weiße Kugel zu ziehen, $1/4$. Er unterschied also nicht die Unsicherheit, die mit der zufälligen Ziehung zusammenhing, von der, die durch mangelnde Kenntnis der Zahlen der Kugeln verursacht war.

Eine gedankliche Wende deutet sich in einem Artikel des Göttinger Privatdozenten Georg Bohlmann an, der in einem einleitenden Abschnitt eines 1901 erschienenen Enzyklopädieartikels über Versicherungsmathematik eine erste Axiomatik der Wahrscheinlichkeitstheorie vorschlug. Für ihn waren Wahrscheinlichkeiten den Ereignissen E zugeordnete Zahlen $P(E)$ zwischen 0 und 1 mit den Eigenschaften: (i) das sichere Ereignis hat Wahrscheinlichkeit 1, das unmögliche Wahrscheinlichkeit 0; (ii) falls sich E_1 und E_2 ausschließen, hat das Ereignis, dass E_1 oder E_2 eintritt, die Wahrscheinlichkeit $P(E_1) + P(E_2)$. (Ein drittes von Bohlmann angegebenes Axiom könnte man als Definition der bedingten Wahrscheinlichkeit werten.) Bohlmann konnte den Begriff Ereignis noch nicht mathematisch ausdrücken.

Bohlmann war durch Hilberts „Grundlagen der Geometrie“ (1899) inspiriert, in denen Hilbert Punkte, Geraden und Ebenen durch Angabe von Axiomen „definierte“. Am 8. August 1900 stellte Hilbert mit Hinweis auf Bohlmann als sechstes seiner berühmten 23 Probleme die Aufgabe der Axiomatisierung „der Gebiete der Physik, in denen Mathematik schon heute eine hervorragende Rolle spielt; dies sind in erster Linie die Wahrscheinlichkeitsrechnung und die Mechanik“. Dies lässt vermuten, dass Hilbert Bohlmanns Ansatz einer axiomatischen Definition des Wahrscheinlichkeitsbegriffs für im Prinzip richtig, aber im Detail unausgereift hielt. Hilberts Schüler U. Broggi beschrieb 1907 Ereignisse als Teilmengen einer abstrakten Menge.

Der schwedische Mathematiker Wiman verwandte schon 1901 Maße; allerdings nur in der geometrischen Wahrscheinlichkeitstheorie, wo schon lange Wahrscheinlichkeiten durch Längen und Flächen beschrieben worden waren.

Von Mises entwarf 1919 eine Axiomatisierung der Wahrscheinlichkeitsrechnung mittels abstrakter relativer Häufigkeiten. Sein Ansatz regte fruchtbare Entwicklungen an, war aber lange unpräzise und setzte sich nicht durch.

Wiener, Paley und Zygmund arbeiteten mit Wahrscheinlichkeiten wie mit Maßen. Kolmogorow gab dann 1933 die heute übliche Definition von Wahrscheinlichkeiten an. Dass sein Buch „Grundbegriffe der Wahrscheinlichkeitsrechnung“ der entscheidende Durchbruch zur Entwicklung der Wahrscheinlichkeitstheorie als mathematischer Disziplin war, lag aber wohl weniger daran, dass nun die Axiome explizit und klar ausgesprochen waren, sondern vor allem daran, dass es ihm gelang, mit maßtheoretischen Mitteln die theoretischen Grundlagen der Theorie der stochastischen Prozesse zu legen. [Mehr zur historischen Entwicklung findet man bei Schneider (1988) und Krengel (1990).]

Warum nun ist es zweckmäßig im Fall allgemeiner Stichprobenräume Ω die Familie der Teilmengen, die als Ereignisse zugelassen werden, einzuschränken? Der folgende Satz von Banach und Kuratowski (1929) zeigt, dass nur diskrete Wahrscheinlichkeitsmaße auf $\mathbb{R}$ existieren, wenn man als Definitionsbereich die Familie $\mathcal{P}(\mathbb{R})$ aller Teilmengen von $\mathbb{R}$ zulässt:

Satz 10.6 *Unter der Voraussetzung der Gültigkeit der (mit den Grundaxiomen der Mengenlehre konsistenten) Kontinuumshypothese gibt es keine auf ganz $\mathcal{P}(\mathbb{R})$ definierte σ -additive Funktion P mit $P(\mathbb{R}) = 1$, die jeder aus einem einzigen Punkt bestehenden Menge $\{\omega\}$ den Wert $P(\{\omega\}) = 0$ zuordnet.*

Für Wahrscheinlichkeitsmaße P , die durch Dichten gegeben sind, ist aber $P(\{\omega\}) = 0$ für alle $\omega \in \mathbb{R}$. Sie lassen sich also nicht σ -additiv auf ganz $\mathcal{P}(\mathbb{R})$ definieren.

Satz 10.6 zeigt, dass man mindestens eine der Eigenschaften von Wahrscheinlichkeiten aufgeben müsste, wollte man darauf bestehen, dass $\mathcal{P}(\Omega)$ der Definitionsbereich von P sein soll. Dafür käme wohl allenfalls die σ -Additivität in Frage. Aber auch das wäre äußerst problematisch. Z.B. ist eine nicht-negative additive Mengenfunktion P auf der Familie $\mathcal{P}(\Omega)$ aller Teilmengen A von $\Omega = [0, 1[$ nicht

eindeutig bestimmt, wenn man nur festlegt, dass für Intervalle $I \subset [0, 1[$ $P(I)$ die Länge des Intervalls sein soll. Das würde es schwer machen, P überhaupt festzulegen. Wollte man eine Gleichverteilung P auf $\Omega = [0, 1]^n$ definieren, so würde man natürlich verlangen, dass kongruente Teilmengen A, B die gleiche Wahrscheinlichkeit $P(A) = P(B)$ haben. Für $n \geq 3$ existieren nach einem Ergebnis von Hausdorff solche auf ganz $\mathcal{P}(\Omega)$ definierten P nicht mehr, selbst wenn man statt σ -Additivität nur Additivität fordert.

Historischer Exkurs; Bertrands Paradox: Als die Wahrscheinlichkeitstheorie noch nicht rigoros entwickelt war, bestand ein Einwand gegen die Verwendung von Wahrscheinlichkeiten bei überabzählbaren Ergebnismengen Ω in der (scheinbaren) Widersprüchlichkeit verschiedener Lösungsansätze für die folgende Aufgabe:

Eine Sehne wird zufällig in einem Kreis mit Radius r um einen Punkt O gezogen. Wie groß ist die Wahrscheinlichkeit p , dass ihre Länge s größer ist als die Seitenlänge d eines einbeschriebenen gleichseitigen Dreiecks? (Bild 10.3)

1. *Ansatz:* Aus Symmetriegründen können wir ohne Beschränkung der Allgemeinheit annehmen einer der Schnittpunkte der Sehne mit dem Kreis sei gerade die Spitze des Dreiecks. (Sonst würde man eben das Dreieck etwas drehen.) Nun gilt $s > d$ genau dann, wenn der zweite Schnittpunkt der Sehne mit dem Kreis auf dem der Spitze gegenüberliegenden Kreissegment liegt. Da die drei Kreissegmente zwischen den Dreiecksseiten gleich groß sind, ist $p = 1/3$.

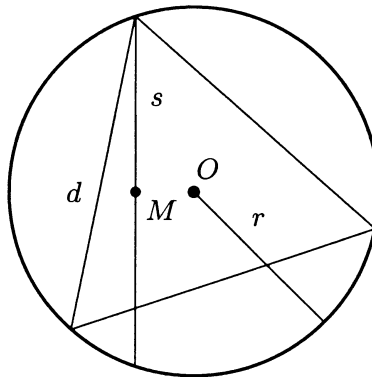


Bild 10.3 Zu Bertrands Paradox

2. *Ansatz:* $s > d$ gilt genau dann, wenn der Mittelpunkt M der Sehne von O einen Abstand $< r/2$ hat. Die Richtung der Sehne spielt dabei keine Rolle. Wir können die Sehne also in Richtung der y -Achse wählen. Dann liegt M auf der x -Achse zwischen $-r$ und $+r$, und $s > d$ gilt genau dann, wenn M zwischen $-r/2$ und $r/2$ liegt. Es folgt $p = 1/2$.

3. *Ansatz:* Wir wählen M zufällig im Innern des Kreises. $s > d$ gilt genau dann, wenn M innerhalb des Kreises mit Radius $r/2$ um O liegt. Da dessen Fläche $1/4$ der Fläche des Kreises mit Radius r ist, ist $p = 1/4$.

Der Widerspruch ist nur scheinbar, weil das Wort „zufällig“ in der Aufgabenstellung das Wahrscheinlichkeitsmaß noch nicht eindeutig festlegt. In den drei Ansätzen geht man von drei verschiedenen Zufallsmechanismen aus, und die Aufgabe ist erst dann eindeutig gestellt, wenn gesagt wird, welcher dieser Zufallsmechanismen für die Wahl der Sehne benutzt werden soll.

Durch mangelnde Angabe von P lässt sich ebenso gut ein Paradoxon für diskrete Experimente angeben. Fragt man z.B. nach der Wahrscheinlichkeit dafür, dass zwei nacheinander gezogene Kugeln aus einer Urne mit drei weißen und drei schwarzen Kugeln die gleiche Farbe haben, so ergeben sich verschiedene Wahrscheinlichkeiten — je nachdem ob mit oder ohne Rücklegen gezogen wird.

Bei wohldefinierter Angabe des zu benutzenden Wahrscheinlichkeitsmaßes ist die obige Aufgabe eine typische Fragestellung aus der geometrischen Wahrscheinlichkeitstheorie. In dieser beschäftigt man sich mit zufällig gewählten Punkten, Geraden, Ebenen, Kreisen usw., und fragt nach der Verteilung geometrisch interessanter Größen. Anwendungen treten z.B. in der Astronomie, Biologie und Medizin auf, aber einige elementare Aufgaben können auch schon den Schulunterricht bereichern. Aus Platzgründen verweisen wir für eine erste Einführung auf Pfanzagl (1988) und für weiter gehende Studien auf Stoyan-Kendall-Mecke (1987).

Die Anfangszifferverteilung: Wenn man eine Liste mit Börsenkursen liest, eine Tabelle von Naturkonstanten anschaut oder Listen mit Flusslängen oder spezifischen Wärmen ansieht, kann man die Feststellung machen, dass sehr viel mehr Zahlen mit einer 1 beginnen, als beispielsweise mit einer 6. Diese Beobachtung machte 1881 auch der Astronom und Mathematiker Simon Newcomb, dem aufgefallen

war, dass die ersten Seiten von Logarithmentafeln viel abgegriffener waren, als die späteren. Er gab im American Journal of Mathematics eine heuristische Begründung dafür an, dass die Anfangsziffer k die Wahrscheinlichkeit $\log_{10}(1+k^{-1})$ hat ($k = 1, \dots, 9$). Für $k = 1$ ergibt sich der Wert 0,301, für $k = 2$ der Wert 0,176, usw. bis zum Wert 0,046 für $k = 9$. Frank Benford entdeckte 57 Jahre später das Gesetz neu und untermauerte es mit über 20 000 Daten aus den verschiedensten Tabellen. Da man die newcombsche Arbeit übersah, erhielt die Verteilung den Namen „**benfordsches Gesetz**“.

Es gab viele Versuche, das Gesetz mathematisch zu begründen, aber bis in die neuere Zeit war keiner überzeugend. Z.B. konvergiert die relative Häufigkeit h_n der Anfangsziffer 1 unter den ersten n natürlichen Zahlen nicht. Das Gesetz sollte skaleninvariant sein, denn es sollte nicht davon abhängen, ob man Längen in Yards oder Metern misst. Aber es gibt kein skaleninvariantes Wahrscheinlichkeitsmaß auf der σ -Algebra der borelschen Teilmengen der Menge $\mathbb{R}^+$ der positiven reellen Zahlen.

Die Fixierung auf die übliche σ -Algebra verhinderte offenbar lange einen unvoreingenommenen Zugang. T. P. Hill machte 1995 den natürlichen Vorschlag, eine andere σ -Algebra zu betrachten. Sei $D_m(x)$ für $x \in \mathbb{R}^+$ die m -te signifikante Ziffer in der Dezimalbruchdarstellung von x . Also für $x = 0,0721$, aber auch für $x = 72,1$ sei $D_1(x) = 7$, $D_2(x) = 2$, $D_3(x) = 1$. Die **Mantissen- σ -Algebra** $\mathcal{M}$ sei die von den Mengen $\{x \in \mathbb{R}^+ : D_1(x) = d_1, D_2(x) = d_2, \dots, D_j(x) = d_j\}$ mit $j \in \mathbb{N}$, $d_1 \in \{1, 2, \dots, 9\}$, $d_2, \dots, d_j \in \{0, 1, \dots, 9\}$ erzeugte σ -Algebra. Eine Menge $M \subset \mathbb{R}^+$ gehört zu $\mathcal{M}$, wenn eine borelsche Menge $B \subset [1, 10[$ existiert, für die M die Vereinigung der Mengen $10^n \cdot B$ mit $n \in \mathbb{Z}$ ist. (Dabei ist $\alpha B = \{\alpha x : x \in B\}$.)

Für $M \in \mathcal{M}$ und $\alpha > 0$ gilt auch $\alpha M \in \mathcal{M}$. Ein Wahrscheinlichkeitsmaß P auf $\mathcal{M}$ heißt skaleninvariant, wenn für alle $M \in \mathcal{M}$ und alle $\alpha > 0$ die Gleichung $P(\alpha M) = P(M)$ gilt. Hill bemerkte, dass es genau ein skaleninvariantes Wahrscheinlichkeitsmaß P auf $\mathcal{M}$ gibt. Für $B = [1, t)$ mit $1 \leq t \leq 10$ und das daraus wie oben gebildete M muss $P(M) = \log_{10} t$ sein. Dies entspricht genau dem **newcomb-benfordschen Gesetz**. Hill gab eine weitere (tiefer liegende) Charakterisierung mittels „Basis-Invarianz“ an. Schließlich bewies er noch einen Grenzwertsatz für Stichproben, die mittels *zufällig* gewählter Verteilungsfunktionen erhoben werden. Als Grenzverteilung ergab sich wieder das gleiche Gesetz. (Literatur: Hill (1995)). Die Verteilung hat inzwischen interessante Anwendungen z.B. in der Steuerfahndung gefunden.

Aufgaben

1. Man bestimme Konstanten c_1, c_2 so, dass die folgenden Funktionen Dichten sind:

$$f_1(x) = \frac{c_1}{\alpha^2 + (x - \beta)^2} \quad (\alpha > 0; \beta, x \in \mathbb{R}); \quad f_2(x) = \begin{cases} 0 & \text{für } x \leq 1, \\ c_2 x^{-r} & \text{für } x > 1. \end{cases}$$

Dabei sei $r > 1$.

2. Man zeige, dass jede Verteilungsfunktion auf $\mathbb{R}$ höchstens an abzählbar vielen Stellen unstetig ist.
3. Man zeige, dass jede offene Menge in $\mathbb{R}^n$ sich als abzählbare Vereinigung von n -dimensionalen halboffenen Intervallen $[a(k), b(k)[$ darstellen lässt. (L)
4. Man zeige, dass nicht jede auf $\mathbb{R}^2$ stetige Funktion $F(x, y)$, die in jeder Koordinate monoton wachsend ist, und für die $F(0, 0) = 0$, $F(1, 1) = 1$ und $0 \leq F(x, y) \leq 1$ gilt, eine Verteilungsfunktion ist. (L)
5. Für eine unbekannte stetige Dichte $f(x, y)$ im $\mathbb{R}^2$ gelte

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

Wie kann man bei Kenntnis von F die Dichte bestimmen? (Vgl. (10.12))

6. In Bertrands Paradox sei

$$\Omega = \{M = (m_1, m_2) \in \mathbb{R}^2 : m_1^2 + m_2^2 \leq r^2\}.$$

Dem dritten Ansatz entspricht das Wahrscheinlichkeitsmaß P_3 mit konstanter Dichte auf Ω . Seien $(\rho(M), \phi(M))$ mit $-r < \rho(M) < r$ und $0 \leq \phi(M) < \pi$ die Polarkoordinaten von M . Dem zweiten Ansatz entspricht der Zufallsmechanismus, bei dem $\rho(M)$ und $\phi(M)$ unabhängig voneinander und mit Gleichverteilung in $] -r, +r[$ und $[0, \pi[$ gewählt werden. Welche Dichte f_2 beschreibt das zugehörige Wahrscheinlichkeitsmaß P_2 ?

7. Seien $A_1, A_2, \dots$ borelsche Mengen in $\mathbb{R}$, und sei B die Menge aller $\omega \in \mathbb{R}$, die für jedes n in n aufeinander folgenden A_i vorkommen. Man zeige, dass B borelsch ist.

8. In einer Kreisscheibe mit Radius 1 wird ein Punkt zufällig gewählt (Gleichverteilung). Bestimmen Sie die Dichte der Verteilung seines Abstandes vom Mittelpunkt M des Kreises. (L)
9. Sei $\beta > 1$. Bestimmen Sie die Ausfallrate der Verteilung mit Dichte $f_\beta(x) = x^{-\beta}/(1 - \beta)$ für $x \geq 1$ (und $= 0$ für $x < 1$).
10. Zeigen Sie, dass es eine stetige Verteilungsfunktion auf $\mathbb{R}$ mit $F(0) = 0$ und $F(1) = 1$ gibt, die auf abzählbar vielen offenen Teilintervallen von $[0, 1]$, deren Längen sich zu 1 summieren, konstant ist. (Das zugehörige P hat keine Dichte, und dennoch gilt $P(\{\omega\}) = 0$ für alle ω .) (L)

§ 11 Zufallsvariable und ihre Momente

11.1 Messbare Funktionen

Im diskreten Fall hatten wir *jede* Abbildung X von Ω in $\mathbb{R}$ Zufallsvariable genannt. Für allgemeines Ω ist das nicht zweckmäßig. Wir wollen z.B. von der Wahrscheinlichkeit sprechen können, dass $X \leq 7$ ist. Dazu muss $\{X \leq 7\}$ ein Ereignis sein, also zu der σ -Algebra gehören, auf der P definiert ist.

Am besten betrachten wir gleich den Fall allgemeiner Wertebereiche. Ist f eine beliebige Abbildung einer Menge Ω in eine Menge Ω' , und $A' \subset \Omega'$, so bezeichne $f^{-1}(A')$ die Menge $\{\omega \in \Omega : f(\omega) \in A'\}$ aller Urbilder von Punkten in A' .

Definition 11.1 Sind $(\Omega, \mathcal{A})$ und $(\Omega', \mathcal{A}')$ messbare Räume, so nennen wir eine Abbildung f von Ω in Ω' **messbar**, wenn für alle $A' \in \mathcal{A}'$

$$f^{-1}(A') \in \mathcal{A} \quad (11.1)$$

gilt. Man nennt f dann auch **messbare Funktion** auf Ω mit Werten in Ω' . Eine **Zufallsvariable** ist eine auf dem Stichprobenraum Ω eines Wahrscheinlichkeitsraums definierte messbare Funktion.

Vor allem sind natürlich reellwertige und vektorwertige Zufallsvariable von Interesse. Dann ist also $\Omega' = \mathbb{R}$ bzw. $= \mathbb{R}^n$. Die σ -Algebra $\mathcal{A}'$ sei dann immer die der borelschen Mengen.

Zum Glück braucht man zum Nachweis der Messbarkeit von f die Bedingung (11.1) nicht für alle $A' \in \mathcal{A}'$ zu prüfen. Ist nämlich $\mathcal{F}'$ ein **Erzeugendensystem** von $\mathcal{A}'$, d.h. eine Familie von Teilmengen von Ω' für welche $\mathcal{A}'$ die erzeugte σ -Algebra $\mathcal{A}(\mathcal{F}')$ ist, und gilt (11.1) für alle $A' \in \mathcal{F}'$, so gilt (11.1) sogar für alle $A' \in \mathcal{A}'$. Das folgt aus:

Lemma 11.2 Ist f eine Abbildung von Ω in Ω' und $\mathcal{A}$ eine σ -Algebra in Ω , so ist die Familie

$$\mathcal{A}'_f := \{A' \subset \Omega' : f^{-1}(A') \in \mathcal{A}\}$$

eine σ -Algebra in Ω' .

Beweis: Die Mengenabbildung f^{-1} ist mit allen mengentheoretischen Operationen vertauschbar. Z.B. gilt

$$f^{-1}\left(\bigcup_{i=1}^{\infty} A'_i\right) = \bigcup_{i=1}^{\infty} f^{-1}(A'_i). \quad (11.2)$$

Für Mengen $A'_i \in \mathcal{A}'_f$ gehört $f^{-1}(A'_i)$ zu $\mathcal{A}$ und daher auch die Vereinigung dieser Mengen. Wegen (11.2) gehört daher die Vereinigung der A'_i zu $\mathcal{A}'_f$. Genauso argumentiert man für das Komplement. Wegen $f^{-1}(\Omega') = \Omega \in \mathcal{A}$ gilt auch $\Omega' \in \mathcal{A}'_f$. $\square$

Da die Familie $\mathcal{F}'$ der Intervalle $] -\infty, x]$ (und auch die der entsprechenden offenen Intervalle) Erzeugendensysteme der borelschen σ -Algebra in $\mathbb{R}$ sind, ist eine reellwertige Funktion auf einem Wahrscheinlichkeitsraum genau dann Zufallsvariable, wenn alle Mengen $\{X \leq x\}$ zu $\mathcal{A}$ gehören (und auch genau dann wenn $\{X < x\} \in \mathcal{A}$ für alle x gilt).

Jede stetige Abbildung f von $\mathbb{R}^m$ in $\mathbb{R}^n$ ist messbar, denn für offenes $B \subset \mathbb{R}^n$ ist dann $f^{-1}(B)$ offen und damit borelsch. Auch monoton wachsende oder fallende Funktionen f auf $\mathbb{R}$ sind offenbar messbar, denn für alle x ist dann $f^{-1}(]-\infty, x])$ ein Intervall. Ein anderes Beispiel messbarer Funktionen sind die Indikatorfunktionen 1_A von Mengen $A \in \mathcal{A}$.

Wir wollen uns jetzt überlegen, dass man mit messbaren Funktionen so ziemlich alles machen kann, ohne den Bereich der messbaren Funktionen zu verlassen.

Lemma 11.3 *Sind $X : \Omega \rightarrow \Omega'$ und $g : \Omega' \rightarrow \Omega''$ messbar (bzgl. $\mathcal{A}, \mathcal{A}', \mathcal{A}''$), so ist $Y = g \circ X$ messbar.*

Beweis: Für $B'' \in \mathcal{A}''$ ist $Y^{-1}(B'') = X^{-1}(g^{-1}(B'')) \in \mathcal{A}$. $\square$

Lemma 11.4 *Sind $X_1, \dots, X_n$ reellwertige messbare Funktionen auf Ω , so wird durch $X(\omega) = (X_1(\omega), \dots, X_n(\omega))$ eine $\mathbb{R}^n$ -wertige messbare Funktion X definiert (und umgekehrt).*

Beweis: Für $a = (a_1, \dots, a_n)$ und $b = (b_1, \dots, b_n)$ ist

$$X^{-1}(]a, b]) = \bigcap_{i=1}^n X_i^{-1}(]a_i, b_i])$$

und die Intervalle erzeugen die borelsche σ -Algebra. Die Umkehrung folgt aus

$$X_i^{-1}(]a_i, b_i]) = X^{-1}(D_i)$$

mit $D_i = \{(x_1, \dots, x_n) : a_i < x_i \leq b_i\}$, da D_i borelsch ist. $\square$

Diese beiden Lemmata gelten analog auch für stetige Funktionen, aber der folgende Satz enthält auch Aussagen für unendliche Folgen, die nicht analog für stetige Funktionen gelten:

Satz 11.5 *Sind $X_1, X_2, \dots$ reellwertige messbare Funktionen und $\alpha_1, \alpha_2, \dots \in \mathbb{R}$, so sind auch die Funktionen*

$$\begin{aligned} \alpha_1 X_1 + \dots + \alpha_n X_n, & \quad X_1 X_2 \dots X_n, \\ \sup\{X_i, i \geq 1\}, & \quad \inf\{X_i, i \geq 1\}, \\ \limsup X_i, & \quad \liminf X_i \end{aligned}$$

messbar (mit Wertebereich $\bar{\mathbb{R}} = [-\infty, +\infty]$).

Beweis: Setzt man $X = (X_1, \dots, X_n)$ und $g(x_1, \dots, x_n) = \alpha_1 x_1 + \dots + \alpha_n x_n$, so ist $\alpha_1 X_1 + \dots + \alpha_n X_n = g \circ X$. g ist stetig und daher messbar. Die Behauptung folgt also aus den beiden Lemmata. Ebenso folgt die Messbarkeit des Produkts.

Die Messbarkeit abzählbarer Suprema folgt aus

$$\left\{ \sup_i X_i \leq x \right\} = \bigcap_i \{X_i \leq x\},$$

und die abzählbarer Infima aus

$$\left\{ \inf_i X_i < x \right\} = \bigcup_i \{X_i < x\}.$$

Schließlich ist noch $\limsup X_n = \inf_k (\sup_{i \geq k} X_i)$, usw. $\square$

11.2 Verteilungen von Zufallsvariablen

Für eine Zufallsvariable X nennen wir das durch

$$P_X(A') = P(X \in A') \quad (11.3)$$

auf der σ -Algebra $\mathcal{A}'$ im Wertebereich Ω' definierte Wahrscheinlichkeitsmaß P_X die Verteilung von X . Ist X reellwertig, so lässt sich die Verteilung eindeutig durch die **Verteilungsfunktion**

$$F_X(x) = P(X \leq x) = P_X([-\infty, x])$$

beschreiben. Hat die Verteilung von X eine Dichte f , so sagt man auch verkürzt, X habe die Dichte f .

Wenn man aus einer Zufallsvariablen X eine andere berechnet, z.B. $Y = X^2$, stellt sich die Frage, wie man aus der Verteilung von X die von Y berechnet. Ist $X \geq 0$, so ist für $y \geq 0$

$$F_Y(y) = P(X^2 \leq y) = P(X \leq \sqrt{y}) = F_X(\sqrt{y}).$$

Ist die Verteilung von X durch eine Dichte f gegeben, so erhält man mit der Substitution $z = x^2$

$$F_Y(y) = \int_0^{\sqrt{y}} f(x) dx = \int_0^y f(\sqrt{z}) \cdot \frac{1}{2\sqrt{z}} dz.$$

Also hat die Verteilung von Y die Dichte

$$g(y) = \frac{f(\sqrt{y})}{2\sqrt{y}}.$$

Ganz genauso kann man vorgehen, wenn allgemeiner $Y = \varphi(X)$ ist, und φ eine Funktion ist, die eine bis auf Sprungstellen stetige und überall *strikt* positive (oder überall strikt negative) Ableitung $\varphi'(x)$ im Wertebereich von X hat. Man erhält, wenn φ^{-1} die Umkehrfunktion von φ bezeichnet,

$$g(y) = \frac{f(\varphi^{-1}(y))}{|\varphi'(\varphi^{-1}(y))|}. \quad (11.4)$$

Das leuchtet, wie man in Bild 11.1 sieht, auch anschaulich ein.

Einem Zuwachs von x um dx entspricht ein Zuwachs $dy = \varphi'(x) dx$, und für infinitesimale Zuwächse müssen die Flächen $f(x) dx$ und $g(y) dy$ gleich sein, weil sie der gleichen Wahrscheinlichkeit

$$P(X \in [x, x + dx]) = P(Y \in [y, y + dy])$$

entsprechen. (Ist $\varphi'(x) < 0$, so ist dy negativ, aber die zugehörige Fläche ist $|g(y) dy|$.)

Es lohnt, zwei häufig auftretende Spezialfälle festzuhalten:

Satz 11.6 Ist $Y = X + a$ und hat X die Dichte f , so hat Y die Dichte

$$g(y) = f(y - a).$$

Ist $Y = cX$ und $c \neq 0$, so gilt

$$g(y) = \frac{1}{|c|} f\left(\frac{y}{c}\right).$$

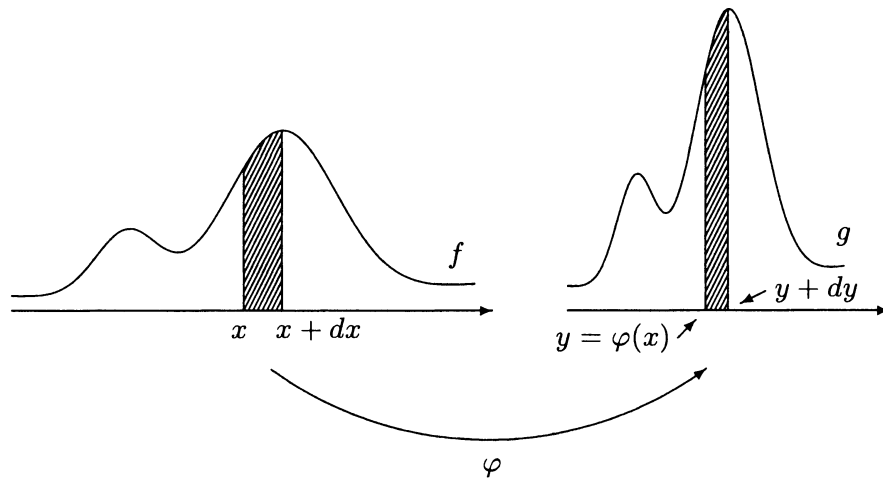


Bild 11.1 Transformation von Dichten

Wenn $\varphi'(x)$ nicht im ganzen Wertebereich von X das gleiche Vorzeichen hat, kann man gewöhnlich den Wertebereich von X in Teilintervalle aufteilen, in denen die Formel (11.4) anwendbar ist. Hat z.B. X positive und negative Werte und ist $Y = X^2$, so ist

$$P(Y \leq y) = P(Y \leq y, X \geq 0) + P(Y \leq y, X < 0),$$

und man erhält

$$g(y) = \frac{f(\sqrt{y})}{2\sqrt{y}} + \frac{f(-\sqrt{y})}{2\sqrt{y}}.$$

11.3 Unabhängigkeit

Seien $(\Omega_i, \mathcal{A}_i)$ messbare Räume ($i \in I$). Eine Familie von Zufallsvariablen X_i auf $(\Omega, \mathcal{A}, P)$ mit Wertebereichen Ω_i heißt unabhängig, wenn für alle $A_i \in \mathcal{A}_i$ die Ereignisse $\{X_i \in A_i\}$ (voneinander) unabhängig sind. (Für Ereignisse verwenden wir den alten im diskreten Fall definierten Begriff.)

Zum Glück braucht man die Unabhängigkeit nicht für jede Wahl von $A_i \in \mathcal{A}_i$ zu prüfen. Sind z.B. alle X_i reellwertige Zufallsvariable, so genügt es, zu zeigen, dass für alle $x_i \in \mathbb{R}$ die Ereignisse $\{X_i \leq x_i\}$ unabhängig sind.

[Dies folgt aus einem (relativ einfachen) Ergebnis der Maßtheorie. Man kann zeigen: Sind $\mathcal{F}_i$ ($i \in I$) unabhängige Teilfamilien von $\mathcal{A}$ (d.h. sind für jede Wahl von $F_i \in \mathcal{F}_i$ die Ereignisse F_i unabhängig), und enthält jedes $\mathcal{F}_i$ mit je zwei Mengen auch deren Durchschnitt, so sind sogar die von den $\mathcal{F}_i$ erzeugten σ -Algebren $\mathcal{A}(\mathcal{F}_i)$ unabhängig.]

Satz 11.7 (i) Sind $X_1, \dots, X_n$ unabhängige reellwertige Zufallsvariable, und hat X_i die Dichte f_i , so hat $X = (X_1, \dots, X_n)$ die Dichte f mit

$$f(x_1, \dots, x_n) = f_1(x_1) \dots f_n(x_n). \quad (11.5)$$

(ii) Umgekehrt gilt: Hat X die Dichte f , so sind die X_i unabhängig mit Dichten f_i .

Beweis: (i) Sei Q das Wahrscheinlichkeitsmaß mit Dichte f in $\mathbb{R}^n$. Für $a, b \in \mathbb{R}^n$ mit $a \leq b$ gilt dann

$$\begin{aligned}
Q([a, b]) &= \int_{a_1}^{b_1} \dots \int_{a_n}^{b_n} f_1(x_1) \dots f_n(x_n) dx_1 \dots dx_n \\
&= \prod_{i=1}^n \int_{a_i}^{b_i} f_i(x_i) dx_i \stackrel{(*)}{=} \prod_{i=1}^n P(X_i \in]a_i, b_i]) \\
&= P(X_i \in]a_i, b_i] \text{ für alle } i) = P_X([a, b]).
\end{aligned}$$

Da ein Wahrscheinlichkeitsmaß durch die Werte auf den Intervallen eindeutig bestimmt ist, ist $Q = P_X$.

(ii) Hat X die Dichte f , so ist $Q = P_X$. Man stellt nun die letzte Zeile der Rechnung an den Anfang und rechnet wie oben. Setzt man zunächst bei beliebigem festen j für alle $i \neq j$ $a_i = -\infty$ und $b_i = +\infty$, so folgt aus der Rechnung vor (*), dass P_{X_j} die Dichte f_j hat. Damit gilt auch die Gleichheit (*). lässt man jetzt wieder beliebige reelle $a_i \leq b_i$ zu, so folgt die Unabhängigkeit der X_i . $\square$

Die Existenz von Wahrscheinlichkeitsräumen $(\Omega, \mathcal{A}, P)$, auf denen eine Folge $X_1, X_2, \dots$ von unabhängigen Zufallsvariablen mit vorgegebenen Verteilungen P_i definiert werden kann, ist ein Ergebnis der Maßtheorie, auf das wir hier nicht eingehen wollen.

Faltungen: Seien X_1, X_2 unabhängige reellwertige Zufallsvariable, deren Verteilungen Dichten f_1, f_2 haben. Was ist die Verteilung von $S = X_1 + X_2$? Ist $B = \{(x_1, x_2) : x_1 + x_2 \leq s\}$, so ist

$$P(S \leq s) = P_{(X_1, X_2)}(B) = \iint_B f_1(x_1) f_2(x_2) dx_1 dx_2.$$

Mit der Substitution $u = x_1 + x_2, v = x_2$ erhalten wir

$$= \int_{-\infty}^s \left(\int_{-\infty}^{\infty} f_1(u-v) f_2(v) dv \right) du.$$

Wir setzen

$$(f_1 * f_2)(u) = \int_{-\infty}^{\infty} f_1(u-v) f_2(v) dv. \quad (11.6)$$

$f_1 * f_2$ heißt die **Faltung** von f_1 und f_2 . Aus unserer Rechnung folgt

Satz 11.8 Sind X_1, X_2 unabhängige Zufallsvariable mit Dichten f_1, f_2 , so hat $X_1 + X_2$ die Dichte $f_1 * f_2$.

Sind die Dichten f_1 und f_2 auf der negativen Halbachse $= 0$, so tragen in (11.6) nur die Werte u, v mit $v \geq 0$ und $u-v \geq 0$ zum Wert des Integrals bei. Dann ist also

$$(f_1 * f_2)(u) = \int_0^u f_1(u-v) f_2(v) dv. \quad (11.7)$$

Als Anwendungsbeispiel beweisen wir den später noch wichtigen

Satz 11.9 Sind X_1, X_2 unabhängig, und ist X_i $N(\mu_i, \sigma_i^2)$ -verteilt, so ist $X_1 + X_2$ $N(\mu, \sigma^2)$ -verteilt mit $\mu = \mu_1 + \mu_2$ und $\sigma^2 = \sigma_1^2 + \sigma_2^2$.

Beweis: Aus $\varphi_{\mu, \sigma^2}(x) = \varphi_{0, \sigma^2}(x - \mu)$ und Satz 11.6 folgt, dass eine Zufallsvariable Y genau dann $N(\mu, \sigma^2)$ -verteilt ist, wenn $Y - \mu$ $N(0, \sigma^2)$ -verteilt ist. Daher können wir $\mu_1 = \mu_2 = 0$ annehmen. Nun ist

$$\varphi_{0, \sigma_1^2} * \varphi_{0, \sigma_2^2}(u) = \frac{1}{2\pi \sigma_1 \sigma_2} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2} \left\{ \frac{(u-v)^2}{\sigma_1^2} + \frac{v^2}{\sigma_2^2} \right\}\right) dv.$$

Macht man die Substitution

$$z = v \frac{\sigma}{\sigma_1 \sigma_2} - u \frac{\sigma_2}{\sigma \sigma_1},$$

so ist $dv = (\sigma_1 \sigma_2 / \sigma) dz$ und eine kleine Rechnung liefert $\{...\} = z^2 + u^2 / \sigma^2$. Damit erhalten wir für die obige Faltung den Wert

$$\frac{1}{2\pi \sigma_1 \sigma_2} \int_{-\infty}^{\infty} \exp\left(-\frac{z^2}{2}\right) \exp\left(-\frac{u^2}{2\sigma^2}\right) \frac{\sigma_1 \sigma_2}{\sigma} dz = \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{u^2}{2\sigma^2}\right) = \varphi_{0, \sigma^2}(u). \quad \square$$

11.4 Erwartungswerte

Ist X eine reellwertige Zufallsvariable auf $(\Omega, \mathcal{A}, P)$, die nur abzählbar viele Werte $\alpha_1, \alpha_2, \dots$ annimmt, so setzen wir wie im Fall diskreter Wahrscheinlichkeitsräume

$$EX = \sum_{i=1}^{\infty} \alpha_i P(X = \alpha_i).$$

(Wir sagen wieder, dass EX existiert, wenn diese Reihe absolut konvergiert.)

Sei $\mathcal{D}$ die Klasse der Zufallsvariablen mit nur abzählbar vielen Werten, deren Erwartungswert existiert. Dann gelten in $\mathcal{D}$ die alten Rechenregeln (Übung).

Für allgemeine reellwertige Zufallsvariable X definieren wir den Erwartungswert mittels einer Approximation. Für $k \in \mathbb{Z}$ und $n \in \mathbb{N}$ sei

$$A_{n,k} = \{k/n \leq X < (k+1)/n\}$$

und

$$X_n = \sum_{k=-\infty}^{\infty} \frac{k}{n} 1_{A_{n,k}}.$$

Dann ist $X_n \leq X < X_n + 1/n$ und $|X_n - X_m| \leq 1/n + 1/m$. Existiert daher EX_n für ein n , so existiert EX_n für alle n , und es gilt $|EX_n - EX_m| \leq 1/n + 1/m$. Wir sagen, dass EX existiert, wenn EX_1 existiert, und setzen

$$EX = \lim_{n \rightarrow \infty} EX_n.$$

Man verwendet für EX auch die Schreibweisen $E_P X$ und $\int X dP$, die deutlich machen, dass EX von P abhängt. Der so definierte Integralbegriff wird ausführlich in der Maßtheorie studiert. Dort wird gezeigt, dass die Aussagen, die wir über Erwartungswerte im diskreten Fall bewiesen haben, sich auf den allgemeinen Fall übertragen. Insbesondere gilt für Zufallsvariable X, Y , deren Erwartungswerte existieren, $E(X+Y) = EX + EY$. Sind X und Y außerdem unabhängig, so existiert auch $E(XY)$ und es ist $E(XY) = (EX)(EY)$. Ist $X \leq Y$, so gilt $EX \leq EY$.

[Beweis der Additivität: Es ist $(X + Y)_n \leq X + Y \leq X_n + 1/n + Y_n + 1/n \leq X + Y + 2/n \leq (X + Y)_n + 3/n$. Existieren EX und EY , so folgt unter Verwendung der Additivität in $\mathcal{D}$

$$|E((X + Y)_n) - (EX_n + EY_n)| \leq 2/n.$$

Für $n \rightarrow \infty$ gilt $EX_n \rightarrow EX$, $EY_n \rightarrow EY$. Damit gilt $E((X + Y)_n) \rightarrow EX + EY$ und also $E(X + Y) = EX + EY$.

Die Monotonie von EX ist klar. Den Beweis der Formel $E(XY) = (EX)(EY)$ für unabhängige X , Y verschieben wir auf die Übungen.]

Wir leiten nun eine Formel zur Berechnung von Erwartungswerten im Fall von Dichten her:

Satz 11.10 *Sei X eine reellwertige Zufallsvariable, deren Verteilung P_X eine bis auf endlich viele Sprungstellen stetige Dichte f hat, und sei g stetig auf $\mathbb{R}$. Dann existiert $E(g(X))$ genau dann, wenn $I := \int |g(x)| f(x) dx$ endlich ist, und in diesem Fall ist*

$$E(g(X)) = \int_{-\infty}^{\infty} g(x) f(x) dx. \quad (11.8)$$

Beweis: Zu jedem $\delta > 0$ existiert eine strikt monoton wachsende Folge $\{x_n, n \in \mathbb{Z}\}$ mit $x_n \rightarrow -\infty$ für $n \rightarrow -\infty$, $x_n \rightarrow +\infty$ für $n \rightarrow +\infty$, und mit

$$|g(x) - g(x_n)| < \delta \quad \text{für } x_n \leq x \leq x_{n+1}.$$

Sei $g_\delta(x) = g(x_n)$ für $x_n \leq x < x_{n+1}$. Dann ist

$$|g_\delta(x) - g(x)| \leq \delta \quad (11.9)$$

und

$$E(g_\delta(X)) = \sum_{n=-\infty}^{\infty} g(x_n) \int_{x_n}^{x_{n+1}} f(x) dx.$$

Die letzte Summe konvergiert genau dann absolut, wenn I endlich ist, und dann unterscheidet sie sich von $\int g(x) f(x) dx$ maximal um δ . Wegen (11.9) ist auch $|E(g(X)) - E(g_\delta(X))| \leq \delta$. Damit folgt die Behauptung. $\square$

Speziell gilt also für eine Zufallsvariable X mit Dichte f

$$EX = \int_{-\infty}^{+\infty} x f(x) dx, \quad (11.10)$$

und

$$E(X^2) = \int_{-\infty}^{+\infty} x^2 f(x) dx, \quad (11.11)$$

und diese Erwartungswerte existieren genau dann, wenn die Integrale absolut konvergieren. Existiert $E(X^2)$, so setzen wir wieder

$$\text{Var}(X) = E((X - EX)^2). \quad (11.12)$$

Die alten Regeln gelten dann auch in der allgemeineren Situation weiter. Insbesondere ist $\text{Var}(X) = E(X^2) - (EX)^2$.

Beispiel 11.11 (*Erwartungswert und Varianz der Normalverteilung* $N(\mu, \sigma^2)$) Sei X $N(\mu, \sigma^2)$ -verteilt. Dann hat nach den Transformationsformeln in Satz 11.6 die Zufallsvariable $Y = (X - \mu)/\sigma$ die Dichte φ der Standard-Normalverteilung.

Wegen $\varphi(-x) = \varphi(x)$ ist

$$\int_{-\infty}^0 x \varphi(x) dx = - \int_0^{\infty} x \varphi(x) dx,$$

und damit $EY = 0$. Es folgt $E(X - \mu) = \sigma EY = 0$, und also $EX = \mu$.

Wegen $EY = 0$ ist

$$\text{Var}(Y) = E(Y^2) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2} dx.$$

Durch partielle Integration erhält man

$$= \lim_{N \rightarrow \infty} \frac{1}{\sqrt{2\pi}} \left\{ \left[x e^{-x^2/2} \right]_{-N}^N - \int_{-N}^N -e^{-x^2/2} dx \right\} = 1.$$

Aus $\text{Var}(Y) = 1$ folgt $\text{Var}(X) = \text{Var}(X - \mu) = \sigma^2 \text{Var}(Y) = \sigma^2$. □

Wer die gerade eingeführten Begriffe jetzt erst einmal an Hand eines Beispiels vertiefen will, kann nun auch § 18 über den Poisson-Prozess lesen. Eine andere Möglichkeit ist es, sich jetzt den Anwendungen in der Statistik (§ 13 und § 14) zuzuwenden.

11.5 Mehrdimensionale Dichtetransformation und Normalverteilung*

Uns interessiert nun die Verallgemeinerung der Transformationsformel für Dichten aus § 10.2 auf den mehrdimensionalen Fall und ihre Anwendung auf Normalverteilungen.

Sei $X = (X_1, \dots, X_n)$ ein Zufallsvektor mit einer Dichte f in $\mathbb{R}^n$, d.h. für eine integrierbare, nichtnegative Funktion f auf $\mathbb{R}^n$ gelte

$$P(X_1 \leq b_1, \dots, X_n \leq b_n) = \int_{-\infty}^{b_1} \dots \int_{-\infty}^{b_n} f(x_1, \dots, x_n) dx_n \dots dx_1$$

für alle $b = (b_1, \dots, b_n) \in \mathbb{R}^n$. Für das Integral schreiben wir auch kürzer

$$\int_{]-\infty, b]} f(x) dx.$$

Sei φ eine stetig differenzierbare Abbildung von $\mathbb{R}^n$ in $\mathbb{R}^n$ und $Y = \varphi(X)$. Wie bestimmt man die Verteilung von Y ?

Sei zunächst φ bijektiv und in beiden Richtungen stetig differenzierbar auf einer offenen Menge $G \subset \mathbb{R}^n$, und sei G Träger der Verteilung von X , d.h. $P(X \in G) = 1$. Sind $\varphi_i(x)$ ($i = 1, \dots, n$) die Komponenten des Vektors $\varphi(x) \in \mathbb{R}^n$, so bezeichnet

$$\Delta(x) = \det \left(\left(\frac{\partial \varphi_i}{\partial x_k} \right)_{i,k=1,\dots,n} \right)$$

die Funktionaldeterminante von φ an der Stelle $x \in G$. Ist $y = \varphi(x)$, so ist $(\Delta(x))^{-1} = (\Delta(\varphi^{-1}(y)))^{-1}$ der Wert der Funktionaldeterminante der Umkehrabbildung φ^{-1} an der Stelle y . Die Menge $G' = \varphi(G)$ ist ebenfalls offen.

Nach dem Transformationssatz für mehrfache Integrale (siehe O. Forster (1981), Seite 17) gilt für alle stetigen Funktionen h' auf G' mit kompaktem Träger

$$\int_G h'(\varphi(x)) |\Delta(x)| dx = \int_{G'} h'(y) dy.$$

Wenden wir diesen Satz auf φ^{-1} an, so ergibt sich

$$\int_G h(x) dx = \int_{G'} h(\varphi^{-1}(y)) |\Delta(\varphi^{-1}(y))|^{-1} dy \quad (11.13)$$

für stetige Funktionen h auf G mit kompaktem Träger. Mit ein wenig lebesguescher Integrationstheorie sieht man, dass das gleiche Resultat sogar für integrierbare Funktionen h gilt; vgl. Forster (1981), S. 120. Ist nun $A' = [a', b']$ ein in G' enthaltenes Intervall, so ist

$$P(\varphi(X) \in A') = \int 1_{\varphi^{-1}A'}(x) f(x) dx.$$

Setzt man $h(x) = 1_{\varphi^{-1}A'}(x) \cdot f(x)$, so ist $h(\varphi^{-1}(y)) = 1_{A'}(y) \cdot f(\varphi^{-1}(y))$. Aus (11.13) folgt dann

$$P(\varphi(X) \in A') = \int_{A'} f(\varphi^{-1}(y)) |\Delta(\varphi^{-1}(y))|^{-1} dy.$$

Ist A' eine in $\mathbb{R}^n \setminus G'$ enthaltene messbare Menge, so ist $P(\varphi(X) \in A') = 0$. Setzen wir also

$$g(y) = \begin{cases} f(\varphi^{-1}(y)) |\Delta(\varphi^{-1}(y))|^{-1} & \text{für } y \in G', \\ 0 & \text{für } y \in \mathbb{R}^n \setminus G', \end{cases}$$

so ist $g(y)$ die Dichte von $\varphi(X)$.

Ein etwas allgemeineres Ergebnis benötigt man, wenn φ zwar nicht bijektiv ist, aber der Bildbereich von X sich in offene Teilmengen zerlegen lässt, auf denen φ bijektiv ist:

Satz 11.12 Sei $X = (X_1, \dots, X_n)$ ein Zufallsvektor, dessen Verteilung durch eine Dichte f gegeben ist. Es gebe k disjunkte offene Mengen $G_1, \dots, G_k$ derart, dass deren Vereinigung G Träger der Verteilung von X ist, und φ sei eine Abbildung von G in $\mathbb{R}^n$, deren Einschränkung φ^i auf G_i für $i = 1, \dots, k$ eine bijektive und in beide Richtungen stetig differenzierbare Abbildung auf $G'_i = \varphi(G_i)$ ist. Sei $\Delta_i(x)$ der Wert der Funktionaldeterminante von φ^i in $x \in G_i$ und

$$g_i(y) = \begin{cases} f((\varphi^i)^{-1}(y)) |\Delta_i((\varphi^i)^{-1}(y))|^{-1} & \text{für } y \in G'_i, \\ 0 & \text{für } y \in \mathbb{R}^n \setminus G'. \end{cases}$$

Dann ist $g = \sum_{i=1}^k g_i$ die Dichte von $\varphi(X)$.

Beweis: $P(\varphi(X) \in A')$ ist die Summe der Werte $P(\varphi(X) \in A', X \in G_i)$. $\square$

Die mehrdimensionale Normalverteilung: Uns interessieren nun vor allem affine Abbildungen und Normalverteilungen. Es ist bei diesem Thema gängiger Brauch, vor allem mit Spaltenvektoren zu arbeiten.

Ist A eine Matrix, so bezeichne tA die transponierte Matrix.

Seien $X_1, \dots, X_n$ unabhängige $N(0, 1)$ -verteilte Zufallsvariable. Dann ist nach Satz 11.7 die Dichte von $X = {}^t(X_1, \dots, X_n)$ durch das Produkt der Dichten der X_i gegeben, d.h. es ist

$$f(x) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2} \sum_{i=1}^n x_i^2\right)$$

für $x = {}^t(x_1, \dots, x_n) \in \mathbb{R}^n$. Wir sagen dann, dass X eine **n -dimensionale Standard-Normalverteilung** habe.

Ist φ von der Form

$$\varphi(x) = Ax + b,$$

wobei A eine $(n \times n)$ -Matrix und $b = {}^t(b_1, \dots, b_n)$ ist, so bezeichnen wir die Verteilung von $Y = \varphi(X)$ als (allgemeine) n -dimensionale Normalverteilung. Ist A nichtausgeartet, d.h. $\det(A) \neq 0$, so sprechen wir von einer **nichtausgearteten** n -dimensionalen Normalverteilung, anderenfalls von einer **ausgearteten** n -dimensionalen Normalverteilung.

Erwartungswerte von Zufallsvektoren und Zufallsmatrizen definieren wir komponentenweise. Insbesondere ist für $Y = {}^t(Y_1, \dots, Y_n)$

$$E(Y) = {}^t(E(Y_1), \dots, E(Y_n)).$$

Für das obige Y ergibt sich mit den Rechenregeln für Erwartungswerte sofort

$$E(Y) = A E(X) + b = b,$$

da alle $E(X_i) = 0$ sind. Für einen beliebigen Vektor $Z = {}^t(Z_1, \dots, Z_n)$ von Zufallsvariablen mit endlicher Varianz bezeichnen wir die Matrix

$$(\text{Cov}(Z_i, Z_k))_{i, k=1, \dots, n} = \left(E((Z_i - EZ_i) \cdot (Z_k - EZ_k)) \right)_{i, k=1, \dots, n}$$

der Kovarianzen als **Kovarianzmatrix** $\text{Cov}(Z, Z)$. Es ist

$$\text{Cov}(Z, Z) = E\left((Z - E(Z)) \cdot {}^t(Z - E(Z))\right).$$

Mit Hilfe von Satz 11.12 können wir nun zeigen, dass die Verteilung einer nichtausgearteten n -dimensionalen Normalverteilung eindeutig durch $C = \text{Cov}(Y, Y)$ und $b = E(Y)$ bestimmt ist:

Satz 11.13 *Ist $\det(A) \neq 0$, so hat die Verteilung von Y die Dichte*

$$g(y) = \frac{1}{|\det(C)|^{1/2} (2\pi)^{n/2}} \exp\left(-\frac{1}{2} {}^t(y - b) C^{-1} (y - b)\right).$$

Beweis: Wegen $C = A \cdot {}^tA$ ist $\det(C) = \det(A \cdot {}^tA) = \det(A)^2$. Die Funktionaldeterminante $\Delta(x)$ ist unabhängig von x und $= \det(A)$. Also ist $|\Delta| = \det(C)^{1/2}$.

Aus $y = Ax + b = \varphi(x)$ folgt $x = \varphi^{-1}(y) = A^{-1}(y - b)$. Also ist

$$\sum_{i=1}^n x_i^2 = {}^t x \cdot x = {}^t(y - b) {}^t A^{-1} \cdot A^{-1}(y - b) = {}^t(y - b) C^{-1}(y - b).$$

Der Vergleich mit der Dichte f ergibt die Behauptung. $\square$

Wir betrachten nun auch ausgeartete Normalverteilungen. Ist $\det(A) = 0$, so ist $\varphi(\mathbb{R}^n)$ ein echter affiner Unterraum von $\mathbb{R}^n$, so dass $\varphi(X)$ keine Dichte in $\mathbb{R}^n$ haben kann.

Sei $m = \text{Rg}(A)$. Ist $m = 0$, so ist $Y = b$. Interessanter ist der Fall $1 \leq m \leq n - 1$, den wir nun untersuchen. Zur Vereinfachung nehmen wir jetzt $b = 0$ an. Sonst kommt nur eine Translation hinzu.

Sei also $1 \leq m \leq n - 1$. Es gibt eine orthonormale Basis $c^1, \dots, c^n$ in $\mathbb{R}^n$ derart, dass die Vektoren $Ac^1, \dots, Ac^m$ linear unabhängig sind und $Ac^j = 0$ für $j = m + 1, \dots, n$ gilt. Sei

$$X = W_1 c^1 + \dots + W_n c^n,$$

d.h. für jedes ω seien $W_1(\omega), \dots, W_n(\omega)$ die Koordinaten von $X(\omega)$ in der neuen Basis. Die Verteilung von $W = (W_1, \dots, W_n)$ ist wieder eine n -dimensionale Standard-Normalverteilung, denn die Dichte von X hängt nur vom Abstand von 0 ab, und ist $x = w_1 c^1 + \dots + w_n c^n$, so ist

$$\sum_{i=1}^n x_i^2 = \sum_{i=1}^n w_i^2.$$

Ist H^m der von $c^1, \dots, c^m$ aufgespannte Raum und P_m die orthogonale Projektion von $\mathbb{R}^n$ auf H^m , so ist

$$P_m X = W_1 c^1 + \dots + W_m c^m$$

und $AX = A P_m X$. A bildet aber H^m bijektiv und linear auf den m -dimensionalen linearen Unterraum AH^m von $\mathbb{R}^n$ ab. Da $(W_1, \dots, W_m)$ eine m -dimensionale Standard-Normalverteilung hat, folgt, dass AX eine nichtausgeartete Normalverteilung auf AH^m hat.

Abbildungen von $\mathbb{R}^n$ in $\mathbb{R}^k$: Den Fall von allgemeinen $(k \times n)$ -Matrizen A können wir ohne Mühe auf den bereits betrachteten quadratischen Fall zurückführen. Ist $k > n$, so hat der Bildraum $A\mathbb{R}^n$ höchstens die Dimension n . Dies ermöglicht die Reduktion auf den quadratischen Fall durch Wahl einer geeigneten Basis in $\mathbb{R}^k$.

Wichtiger ist der Fall $k < n$. Ist dann $Y = {}^t(Y_1, \dots, Y_k) = AX$, so definieren wir $\tilde{Y}_i = Y_i$ für $i = 1, \dots, k$ und $\tilde{Y}_i = 0$ für $i = k + 1, \dots, n$. Die Abbildung $X \mapsto \tilde{Y}$ wird dann durch eine $(n \times n)$ -Matrix $\tilde{A}$ beschrieben, die aus A hervorgeht, indem wir unten $n - k$ Zeilen anfügen, die nur aus Nullen bestehen.

Ist außerdem $k = \text{Rg}(A) = m$, so besteht $\tilde{A}\mathbb{R}^n$ aus den Vektoren in $\mathbb{R}^n$, deren letzte $n - m$ Koordinaten = 0 sind. Sei $\tilde{H}^m$ wie oben, aber mit $\tilde{A}$ gebildet. $\tilde{A}\tilde{H}^m$ hat Dimension m und ist in $\tilde{A}\mathbb{R}^n$ enthalten; also ist $\tilde{A}\tilde{H}^m = \tilde{A}\mathbb{R}^n$. $\tilde{A}X$ hat also eine nichtausgeartete Normalverteilung in $\tilde{A}\mathbb{R}^n$. Dies bedeutet, dass AX eine nichtausgeartete Normalverteilung in $\mathbb{R}^m$ hat.

Folgerung 11.14 (a) Ist $Y = {}^t(Y_1, \dots, Y_n)$ normalverteilt, so sind alle Marginalverteilungen von Y Normalverteilungen.

(b) Ist $Y = {}^t(Y_1, \dots, Y_n)$ normalverteilt, so ist $Y_1 + \dots + Y_n$ normalverteilt.

Beweis: Wir können wieder $b = 0$ annehmen. Sei für $1 \leq j_1 < \dots < j_r \leq n$ $Y' = (Y_{j_1}, \dots, Y_{j_r})$, und sei $Y'' = Y_1 + \dots + Y_n$. Die Abbildungen $X \mapsto Y'$ und $X \mapsto Y''$ sind durch die Hintereinanderausführung der linearen Abbildungen gegeben, die die Abbildungen $X \mapsto Y$ und $Y \mapsto Y'$ bzw. $Y \mapsto Y''$ beschreiben. $\square$

Aufgaben

1. Seien X, Y unabhängig und auf $[a, b]$ mit $0 < a < b$ gleichverteilt. Man berechne die Verteilung von $X^2 Y$.
2. Die gemeinsame Verteilung von (X_1, X_2) sei eine Gleichverteilung in $\{(x_1, x_2) : x_1^2 + x_2^2 \leq 1\}$. Man bestimme die Verteilung von $X_1 + X_2$. (L)
3. Sei U eine in $[0, 1]$ gleichverteilte Zufallsvariable und F eine Verteilungsfunktion. Sei $F^{-1}(u) = \inf\{x : F(x) \geq u\}$. Man zeige, dass $F^{-1}(U)$ eine Zufallsvariable mit Verteilungsfunktion F ist.
4. Sei X eine reellwertige Zufallsvariable und $F(x) = P(X \leq x)$. Man zeige, dass $F(X)$ genau dann in $[0, 1]$ gleichverteilt ist, wenn F stetig ist.
5. Sei X eine in $[0, 2\pi]$ gleichverteilte Zufallsvariable, $U = \cos X$ und $V = \sin X$. Man berechne:
 - a) $P(-r \leq U \leq r)$ ($0 \leq r \leq 1$);
 - b) $P((U - 1)^2 + V^2 \leq r^2)$ ($r \geq 0$);
 - c) $P(|U| \leq |V|)$.
6. Sei X exponentiell verteilt mit Parameter $\lambda > 0$. Man zeige: $EX = 1/\lambda$ und $\text{Var}(X) = 1/\lambda^2$.
7. Ein Stab der Länge 1 wird zufällig in zwei Hälften gebrochen, so dass die Länge der rechten Hälfte auf dem Einheitsintervall gleichverteilt ist. Danach wird ebenso zufällig die längere der beiden Hälften in zwei Stücke gebrochen. Wie groß ist die Wahrscheinlichkeit, dass sich aus den entstandenen drei Stäben ein Dreieck bilden lässt?
8. Seien U, V zwei unabhängig voneinander mit Gleichverteilung in $[0, 1]$ gewählte Punkte. Bestimmen Sie die Verteilung des Abstandes $|U - V|$, sowie deren Erwartungswert und Varianz. (L)
9. Seien A und B zwei unabhängig voneinander in einer Kreisscheibe vom Radius 1 um den Punkt M gemäß der Gleichverteilung gewählte Punkte. Sei X der Abstand von M und A , und Y der Abstand von B zur Diagonalen durch A und M . Bestimmen Sie die Dichte von Y . Sind X und Y unabhängig? Bestimmen Sie den Erwartungswert der Fläche F des Dreiecks mit den Ecken M, A und B . (Aufgabe 8 aus § 10 kann verwendet werden.) (L)
10. In einem Friseursalon arbeiten 5 Angestellte, die pro Haarschnitt 10 Minuten Zeit benötigen. Ein Kunde, der den Salon betritt, sieht, dass alle 5 Angestellten arbeiten und noch weitere 4 Kunden warten. Wie groß ist der Erwartungswert seiner Wartezeit, wenn der Zeitpunkt X_i , zu dem der i -te Angestellte mit dem gegenwärtig bedienten Kunden fertig ist, gleichverteilt in $[0, 10]$ ist und die X_i unabhängig sind.
11. Zeigen Sie, dass für unabhängige reellwertige Zufallsvariable X, Y , deren Erwartungswerte existieren, auch $E(XY)$ existiert und $= (EX)(EY)$ ist. (Anleitung: Betrachten Sie zuerst den Fall, dass X und Y beschränkt und ≥ 1 sind.)
- 12.* $X_1, X_2, \dots$ seien diskrete Zufallsvariable, deren Verteilung folgendermaßen bestimmt wird: Zuerst wird der Wert λ einer Zufallsvariablen Λ mit Dichte $f(\lambda)$ beobachtet. Dann seien bei festem Wert von λ die Zufallsvariablen $X_1, X_2, \dots$ unabhängig mit Verteilung $P(x | \lambda)$. Es sei also

$$P(\Lambda \in [a, b], X_1 = x_1, \dots, X_n = x_n) = \int_a^b f(\lambda) P(x_1 | \lambda) \dots P(x_n | \lambda) d\lambda.$$

Es könnten etwa $X_1, X_2, \dots$ die Schadenshöhen einer Versicherung in den Jahren 1, 2, ... sein. Deren Verteilung hängt von einem Parameter λ ab, welcher eine apriori-Verteilung mit Dichte $f(\lambda)$ hat. Man könnte $E(X_1)$ als Basiswert einer Versicherungsprämie für das erste Jahr ansehen, zu dem Kosten- und Gewinnzuschläge zu addieren wären. $E(X_{n+1} | X_1 = x_1, \dots, X_n = x_n)$ wäre der Basiswert für das $(n + 1)$ -te Jahr. (Die bedingte Erwartung sei die Erwartung unter der bedingten Wahrscheinlichkeit.) Sei hier speziell $P(x | \lambda) = e^{-\lambda} \lambda^x / (x!)$ für $x = 0, 1, \dots$ und f Dichte einer Γ -Verteilung mit Parametern $\alpha = a_0$, $\gamma = n_0 + 1$, vgl. Anhang zu § 12.

- (a) Zeigen Sie, dass die bedingte Verteilung von Λ , gegeben $X_1 = x_1, \dots, X_n = x_n$ eine Γ -Verteilung mit veränderten Parameterwerten ist.
- (b) Bestimmen Sie $E(X_{n+1} \mid X_1 = x_1, \dots, X_n = x_n)$.

[Literatur: Gerber (1995). Der springende Punkt ist, dass für dieses Paar von Verteilungen (und mehrere weitere) die neue Prämie eine konvexe Kombination der Ausgangsprämie $E(X_1)$ und des Mittelwerts der bisherigen Schadenshöhen ist.]

§ 12 Grenzwertsätze*

In diesem Abschnitt wollen wir eine Verschärfung des schwachen Gesetzes der großen Zahlen herleiten und die Normalapproximation der Binomialverteilung verallgemeinern.

12.1 Das starke Gesetz der großen Zahlen

Zunächst stellen wir fest, dass die tschebyschewsche Ungleichung (Satz 3.15) auch für beliebige reellwertige Zufallsvariable mit endlicher Varianz gilt. Die Voraussetzung, dass der Wahrscheinlichkeitsraum diskret sei, erübrigt sich, nachdem EX und $\text{Var}(X)$ nun auch allgemein definiert sind. Das Gleiche gilt für das schwache Gesetz der großen Zahlen (Satz 3.16).

Wir sagen, dass eine Folge (Y_n) von reellwertigen Zufallsvariablen **stochastisch** gegen eine Zufallsvariable Y **konvergiere**, wenn für alle $\varepsilon > 0$

$$P(|Y_n - Y| \geq \varepsilon) \rightarrow 0 \quad (12.1)$$

gilt. Das schwache Gesetz der großen Zahlen besagt also, dass für eine Folge (X_n) von unabhängigen Zufallsvariablen mit gleichem Erwartungswert und beschränkten Varianzen die Folge der Zufallsvariablen

$$\frac{1}{n} (X_1 + \dots + X_n) \quad (12.2)$$

stochastisch gegen die konstante Zufallsvariable EX_1 konvergiert. Stochastische Konvergenz von (Y_n) gegen Y besagt, dass Y_n für großes n mit großer Wahrscheinlichkeit nahe bei Y liegt, aber nicht, dass auch nur für ein einziges $\omega \in \Omega$ die Folge $(Y_n(\omega))$ gegen $Y(\omega)$ konvergiert.

Man betrachtet daher noch einen weiteren Konvergenzbegriff: Wir sagen, dass eine Folge (Y_n) von Zufallsvariablen **fast sicher gegen Y konvergiere**, wenn

$$P(\{\omega \in \Omega : \lim_{n \rightarrow \infty} Y_n(\omega) = Y(\omega)\}) = 1 \quad (12.3)$$

ist.

Satz 12.1 *Fast sichere Konvergenz impliziert stochastische Konvergenz, aber nicht umgekehrt.*

Beweis: Sei $\varepsilon > 0$. Die Mengen

$$B_N = \{\omega \in \Omega : |Y_n(\omega) - Y(\omega)| < \varepsilon \text{ für alle } n \geq N\}$$

bilden eine aufsteigende Folge, und ihre Vereinigung B enthält die Menge

$$A := \{\omega : \lim Y_n(\omega) = Y(\omega)\}.$$

Wegen $P(A) = 1$ ist also $P(B) = 1$, und daraus folgt nach Satz 10.5 $P(B_N) \rightarrow 1$. Also gilt

$$P(|Y_N - Y| \geq \varepsilon) \leq P(B_N^c) \rightarrow 0.$$

Um zu sehen, dass nicht umgekehrt aus stochastischer Konvergenz fast sichere Konvergenz folgt, betrachten wir $\Omega = [0, 1[$ mit der Gleichverteilung P , also mit $P([a, b]) = b - a$. Sei

$$c_n = 1 + \frac{1}{2} + \dots + \frac{1}{n} \quad \text{und} \quad A_n = \{\omega \in [0, 1[: \omega \text{ liegt mod } 1 \text{ in } [c_{n-1}, c_n]\}.$$

Die Folge $(Y_n) = (1_{A_n})$ konvergiert stochastisch gegen 0, denn für $0 < \varepsilon < 1$ gilt

$$P(|Y_n - 0| \geq \varepsilon) = P(A_n) = 1/n \rightarrow 0.$$

Andererseits ist es nicht schwer zu prüfen, dass für jedes $\omega \in \Omega$ und für jedes N Zahlen $m, n \geq N$ mit $\omega \in A_m$ und $\omega \in A_n^c$ existieren. Also gilt für jedes ω

$$\liminf Y_n(\omega) = 0 \quad \text{und} \quad \limsup Y_n(\omega) = 1. \quad \square$$

Wir wollen nun zeigen, dass die Folge in (12.2) fast sicher gegen EX_1 konvergiert. Als Vorbereitung benötigen wir das folgende einfache Lemma, das in der Wahrscheinlichkeitstheorie sehr häufig angewendet wird.

Lemma 12.2 (Borel-Cantelli) *Für eine Folge $A_1, A_2, \dots$ von Ereignissen sei*

$$A^* = \{\omega \in \Omega : \omega \in A_k \text{ für unendlich viele } k\}.$$

(i) *Gilt $\sum_{k=1}^{\infty} P(A_k) < \infty$, so ist $P(A^*) = 0$.*

(ii) *Sind die A_k unabhängig, und ist $\sum_{k=1}^{\infty} P(A_k) = \infty$, so ist $P(A^*) = 1$.*

Beweis: (i) $\omega \in \Omega$ gehört zu A^* , wenn für jedes n ein (von ω abhängiges) $k \geq n$ mit $\omega \in A_k$ existiert. Also ist

$$A^* = \bigcap_{n=1}^{\infty} \bigcup_{k \geq n} A_k. \quad (12.4)$$

Für jedes n ist

$$P(A^*) \leq P\left(\bigcup_{k=n}^{\infty} A_k\right) \leq \sum_{k=n}^{\infty} P(A_k),$$

und die rechte Seite strebt wegen der Konvergenz der Summe der $P(A_k)$ gegen 0.

(ii) Für $0 \leq \alpha_i \leq 1$ gilt $\log(1 - \alpha_i) \leq -\alpha_i$. Damit folgt für $n < N$

$$\log\left(\prod_{k=n}^N (1 - \alpha_k)\right) \leq -\sum_{k=n}^N \alpha_k.$$

Wegen der Unabhängigkeit der Ereignisse gilt dann

$$P\left(\bigcap_{k=n}^N A_k^c\right) = \prod_{k=n}^N (1 - P(A_k)) \leq \exp\left(-\sum_{k=n}^N P(A_k)\right).$$

Bei festem n strebt die rechte Seite für $N \rightarrow \infty$ gegen 0. Damit folgt nach (10.7)

$$P\left(\bigcap_{k=n}^{\infty} A_k^c\right) = 0.$$

Wegen der σ -Additivität von P hat dann auch

$$(A^*)^c = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k^c$$

die Wahrscheinlichkeit 0, und damit gilt $P(A^*) = 1$. $\square$

Wir sagen, dass eine Aussage für **fast alle** ω (oder **fast sicher**, oder **fast überall**) gelte, wenn die Menge B der ω , für die sie nicht gilt Wahrscheinlichkeit $P(B) = 0$ hat. Konvergiert die Summe der $P(A_k)$, so gehört also fast jedes ω nur zu endlich vielen A_k ; divergiert sie, und sind die A_k unabhängig, so gehört fast jedes ω zu unendlich vielen A_k .

Beispiel 12.3 Wir betrachten das folgende Gedankenexperiment. Aus einer unendlichen Folge von Urnen wird je eine Kugel gezogen. Enthält die n -te Urne eine weiße und $n - 1$ schwarze Kugeln, und ist A_n das Ereignis, dass die aus der n -ten Urne gezogene Kugel weiß ist, so ist $P(A_n) = 1/n$. Da die Ziehungen unabhängig sind und die Summe der Zahlen $1/n$ divergiert, ist $P(A^*) = 1$. Fast sicher werden unendlich viele weiße Kugeln gezogen.

Enthält die n -te Urne eine weiße Kugel und $n^2 - 1$ schwarze, so werden fast sicher nur endlich viele weiße Kugeln gezogen. Es ist unmöglich, die Zahlen der weißen und schwarzen Kugeln in den Urnen so zu wählen, dass A^* eine echt zwischen 0 und 1 liegende Wahrscheinlichkeit hat.

Da fast sichere Konvergenz eine stärkere Aussage ist als stochastische Konvergenz, bezeichnet man einen Satz über stochastische Konvergenz von Mittelwerten oft als schwaches Gesetz der großen Zahlen, und einen Satz über fast sichere Konvergenz solcher Mittelwerte als **starkes Gesetz der großen Zahlen**. Wir begnügen uns hier mit dem Fall beschränkter Varianzen, der von Rajchman stammt. Die übliche Annahme der Unabhängigkeit ist darin durch die schwächere Annahme der Unkorreliertheit ersetzt, ohne dass dadurch der Beweis erschwert wird.

Satz 12.4 Sei $X_1, X_2, \dots$ eine Folge von reellwertigen, unkorrelierten Zufallsvariablen und

$$\text{Var}(X_i) \leq M < \infty \quad \text{für alle } i.$$

Dann konvergiert die durch

$$Z_n = \frac{1}{n} \sum_{i=1}^n (X_i - EX_i)$$

definierte Folge fast sicher gegen 0.

Beweis: (1) Im ersten Beweisschritt zeigen wir, dass $(Z_{n^2})_{n \in \mathbb{N}}$ fast sicher gegen 0 konvergiert. Ohne Beschränkung der Allgemeinheit können wir $EX_i = 0$ annehmen. Da die X_i unkorreliert sind, ist für $i \neq j$ $\text{Cov}(X_i, X_j) = 0$, und also nach den (auch für nicht diskrete Wahrscheinlichkeitsräume gültigen) Rechenregeln in Satz 3.13

$$\text{Var}(Z_{n^2}) = \text{Var}\left(\frac{1}{n^2} (X_1 + \dots + X_{n^2})\right) = \frac{1}{n^4} \sum_{i=1}^{n^2} \text{Var}(X_i) \leq \frac{M}{n^2}.$$

Nach Tschebyschew gilt für alle $\varepsilon > 0$

$$P(|Z_{n^2}| \geq \varepsilon) \leq \frac{1}{\varepsilon^2} \text{Var}(Z_{n^2}) \leq \frac{M}{\varepsilon^2 n^2}.$$

Ist $A_n = \{|Z_{n^2}| \geq \varepsilon\}$, so ist also die Summe der $P(A_n)$ konvergent. Aus dem Lemma von Borel-Cantelli folgt nun, dass fast jedes ω nur zu endlich vielen A_n gehört. Setzt man $\varepsilon = 1/k$, so ergibt sich, dass

$$E_k := \{\omega : |Z_{n^2}(\omega)| \geq 1/k \text{ für unendlich viele } n\}$$

die Wahrscheinlichkeit $P(E_k) = 0$ hat. Damit hat auch die Vereinigung E der E_k Wahrscheinlichkeit 0. Für $\omega \in E^c$ gibt es zu jedem k nur endlich viele n mit $|Z_{n^2}(\omega)| \geq 1/k$. Für diese ω gilt also $\lim Z_{n^2}(\omega) = 0$.

(2) Für $m \in \mathbb{N}$ sei nun $n = n(m)$ die natürliche Zahl mit $n^2 \leq m < (n+1)^2$. Wir wollen Z_m mit Z_{n^2} vergleichen. Sei

$$S_k = \sum_{i=1}^k X_i.$$

Nach den Rechenregeln für Varianzen ist

$$\text{Var}(S_m - S_{n^2}) = \sum_{i=n^2+1}^m \text{Var}(X_i) \leq M(m - n^2).$$

Nach Tschebyschew folgt für $\varepsilon > 0$

$$P(|S_m - S_{n^2}| \geq \varepsilon n^2) \leq \frac{M}{\varepsilon^2 n^4} (m - n^2).$$

Summieren wir über m , so erhalten wir

$$\begin{aligned} \sum_{m=1}^{\infty} P\left(\frac{1}{n(m)^2} |S_m - S_{n(m)^2}| \geq \varepsilon\right) &\leq \frac{M}{\varepsilon^2} \sum_{n=1}^{\infty} \sum_{m=n^2}^{(n+1)^2-1} \frac{m - n^2}{n^4} \\ &= \frac{M}{\varepsilon^2} \sum_{n=1}^{\infty} \frac{1}{n^4} (1 + 2 + \dots + 2n) \\ &= \frac{M}{\varepsilon^2} \sum_{n=1}^{\infty} \frac{(2n)(2n+1)}{2n^4} < \infty. \end{aligned}$$

Nach Borel-Cantelli gilt dann für fast alle ω und für alle hinreichend großen m , etwa für $m \geq m_\varepsilon(\omega)$,

$$\frac{1}{n(m)^2} |S_m(\omega) - S_{n(m)^2}(\omega)| < \varepsilon.$$

Ferner ist nach dem ersten Beweisschritt für fast alle ω und für hinreichend großes m

$$\left| \frac{1}{n(m)^2} S_{n(m)^2}(\omega) \right| < \varepsilon.$$

Für fast alle ω gilt also beides. Für diese ω ist dann von einem hinreichend großen m an $|S_m(\omega)/n(m)^2| < 2\varepsilon$. Wegen $m > n(m)^2$ ist dann aber auch $|Z_m(\omega)| = |S_m(\omega)/m| < 2\varepsilon$.

Setzt man wieder $\varepsilon = 1/k$, so folgt, dass für jedes k

$$P(\{\omega : |Z_m(\omega)| > 2/k \text{ für unendlich viele } m\}) = 0$$

ist. Wie oben ergibt sich daraus, dass (Z_m) fast sicher gegen 0 strebt. $\square$

Bemerkung: Aus Satz 12.4 können wir folgern, dass in einer unendlichen Folge $X_1, X_2, \dots$ von Bernoulli-Versuchen mit Erfolgswahrscheinlichkeit p die Wahrscheinlichkeit 1 ist, dass die relative Häufigkeit $h_n = n^{-1}(X_1 + \dots + X_n)$ der Erfolge gegen p konvergiert.

Ist etwa $X_i = 1$, wenn im i -ten Wurf eines Würfels eine 6 kommt und $= 0$ sonst, so ist $Z_n = h_n - 1/6$. Das schwache Gesetz besagt $P(|h_n - 1/6| > \varepsilon) \rightarrow 0$. Wäre nur dies richtig, so wäre es immer noch denkbar, dass für fast jedes ω beliebig große n mit $h_n(\omega) \approx 1$ existieren. (Mit Hilfe von Borel-Cantelli überlegt man sich leicht, dass fast jede Folge von Würfeln beliebig lange Sechsenfolgen enthält.) Nach dem starken Gesetz der großen Zahlen wissen wir aber, dass fast sicher $h_n \rightarrow 1/6$ gilt. Fast sicher werden sehr lange Sechsenfolgen also erst so spät auftreten, dass sie die relativen Häufigkeiten nicht mehr stark beeinflussen.

12.2 Normale Zahlen*

Emile Borel, der 1909 das erste starke Gesetz der großen Zahlen, nämlich das für Bernoulli-Folgen, angegeben hat (vgl. Anhang), hat eine amüsante Anwendung auf die reine Mathematik gegeben.

Eine Zahl $\omega \in [0, 1[$ heißt **normal**, wenn in der Dezimalbruchentwicklung

$$\omega = 0, x_1 x_2 x_3 \dots$$

jede endliche Ziffernfolge $a = (a_1, \dots, a_k)$ mit der relativen Häufigkeit 10^{-k} vorkommt, d.h. wenn für alle $k \in \mathbb{N}$ und alle $a \in \{0, 1, \dots, 9\}^k$

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{h=1}^n 1_{\{a\}}(x_h, x_{h+1}, \dots, x_{h+k-1}) = 10^{-k} \quad (12.5)$$

ist.

Im langfristigen Mittel soll also z.B. jede zehnte Ziffer x_h eine sechs sein, für durchschnittlich jedes tausendste h soll $x_h x_{h+1} x_{h+2}$ der Ziffernblock 377 sein, usw. Es scheint hoffnungslos schwierig zu sein, für Zahlen wie $\pi - 3$, $e - 2$ oder $1/\sqrt{2}$ zu prüfen, ob sie normal sind. Man kann aber einige normale Zahlen ad hoc konstruieren. Die bekannteste ist die *champernownsche Zahl*

$$0,0123456789000102 \dots 09101112 \dots 99000001002 \dots,$$

in der zuerst die Ziffern $0, \dots, 9$, dann alle Ziffernpaare, dann alle Zifferntripel, ... der Größe nach geordnet erscheinen. (Der Beweis der Normalität dieser Zahl ist nicht trivial.)

Der folgende Satz von Borel zeigt, dass normale Zahlen im Überfluss existieren. Da er nicht nur für Entwicklungen zur Basis 10 sondern für Entwicklungen zu einer beliebigen Basis $b \geq 2$ gilt, folgt aus ihm auch die Existenz von Zahlen, die simultan in allen Entwicklungen normal sind.

Satz 12.5 Sei P die Gleichverteilung auf $\Omega = [0, 1[$. Dann ist fast jede Zahl $\omega \in \Omega$ normal.

Beweis: Sei $X_i(\omega) = x_i$ die i -te Stelle in der Dezimalbruchentwicklung von ω . (Die abzählbar vielen ω , deren Dezimalbruchentwicklung nicht eindeutig ist, können wir ignorieren, da die Menge dieser ω ohnehin die Wahrscheinlichkeit 0 hat.)

Da für jedes $m \geq 1$ und jedes $b = (b_1, \dots, b_m) \in \{0, \dots, 9\}^m$ die Menge $\{X_1 = b_1, X_2 = b_2, \dots, X_m = b_m\}$ ein Intervall der Länge 10^{-m} ist, ist

$$P(X_1 = b_1, \dots, X_m = b_m) = 10^{-m} = P(X_1 = b_1) \dots P(X_m = b_m).$$

Daher sind die X_i unabhängige Zufallsvariable, deren Verteilung die Gleichverteilung auf $\{0, 1, \dots, 9\}$ ist. Damit folgt nach dem starken Gesetz der großen Zahlen sofort, dass jeder „Block“ a der Länge $k = 1$ für fast alle ω die relative Häufigkeit $1/10$ hat. Sei nun $k > 1$ und a fest. Wir setzen für $0 < j \leq k$ und $i \geq 0$

$$Y_i^{(j)}(\omega) = 1_{\{X_{ik+j}=a_1, \dots, X_{ik+j+k-1}=a_k\}}(\omega).$$

($Y_i^{(j)}(\omega)$ ist also genau dann 1, wenn $(x_{ik+j}, \dots, x_{ik+j+k-1}) = a$ ist.)

Für festes j hängt $Y_i^{(j)}$ nur von den Koordinaten mit Index in $I_i = \{ik + j, \dots, ik + j + k - 1\}$ ab. Da $I_1, I_2, \dots$ disjunkt sind, sind bei festem j die Zufallsvariablen $Y_1^{(j)}, Y_2^{(j)}, \dots$ unabhängig.

Es ist

$$P(Y_i^{(j)} = 1) = P(X_{ik+j} = a_1, \dots, X_{ik+j+k-1} = a_k) = 10^{-k}.$$

Also ist $E(Y_i^{(j)}) = 10^{-k}$. Nach dem starken Gesetz der großen Zahl folgt nun für fast alle ω , d.h. für alle ω außerhalb einer Menge $B_j(a)$ mit $P(B_j(a)) = 0$

$$\lim_{m \rightarrow \infty} \frac{1}{m} \sum_{i=0}^{m-1} Y_i^{(j)}(\omega) = 10^{-k}. \quad (12.6)$$

Für eine feste Blocklänge k gibt es nur endlich viele a und nur endlich viele j . Insgesamt gibt es also nur abzählbar viele $B_j(a)$ und ihre Vereinigung B hat Wahrscheinlichkeit $P(B) = 0$. Für $\omega \in B^c$ gilt (12.6) für alle a und für $j = 1, \dots, k$. (12.6) besagt für $\omega = 0, x_1 x_2 x_3 \dots$, dass

$$\frac{1}{m} \sum_{i=0}^{m-1} 1_{\{a\}}(x_{ik+j}, \dots, x_{ik+j+k-1}) \rightarrow 10^{-k}$$

gilt. Dies bedeutet, dass die relativen Häufigkeiten, mit denen in der Folge $(x_1, x_2, \dots)$ der Block a an den Stellen $(x_h, x_{h+1}, \dots, x_{h+k-1})$ vorkommt, gegen 10^{-k} konvergiert, wenn wir nur die h mit $h \equiv j \pmod{k}$ betrachten. Da aber j beliebig war, muss dann auch die relative Häufigkeit entlang der Folge aller h gegen 10^{-k} konvergieren. $\square$

12.3 Der Zentrale Grenzwertsatz

Der Zentrale Grenzwertsatz ist die Verallgemeinerung des Satzes von de Moivre-Laplace auf allgemeinere Verteilungen. Der gängigste Beweis verwendet **charakteristische Funktionen** $E(\exp(itX))$ von Zufallsvariablen, die auch ein fundamentales Werkzeug für die Bearbeitung vieler anderer Probleme sind, ihren natürlichen Platz aber erst in einem weiter fortgeschrittenen Abschnitt der Wahrscheinlichkeitstheorie finden.

Hier soll ein Beweis von Kersting dargestellt werden, der mit elementaren Hilfsmitteln auskommt.

Für Verteilungsfunktionen F, G verwenden wir die Metrik

$$d(F, G) = \sup_x |F(x) - G(x)|.$$

Sind X, Y Zufallsvariable mit Verteilungsfunktionen $F(x) = P(X \leq x)$ und $G(x) = P(Y \leq x)$, so schreiben wir statt $d(F, G)$ auch $d(X, Y)$ oder $d(X, G)$.

Lemma 12.6 *Ist $E(Y^2) \leq \eta$, so ist*

$$d(X + Y, \Phi) \leq d(X, \Phi) + 2\eta^{1/3}. \quad (12.7)$$

Beweis: Die maximale Steigung der Verteilungsfunktion Φ der Standard-Normalverteilung ist $\Phi'(0) = \varphi(0) = 1/\sqrt{2\pi} \leq 1$. Also ist für alle x und alle $\delta > 0$

$$|\Phi(x \pm \delta) - \Phi(x)| \leq \delta.$$

Ist $X \leq x - \delta$, so ist entweder $Y > \delta$ oder $X + Y \leq x$. Ist $X + Y \leq x$, so ist entweder $X \leq x + \delta$ oder $Y < -\delta$. Daher gilt

$$P(X \leq x - \delta) - P(Y > \delta) \leq P(X + Y \leq x) \leq P(X \leq x + \delta) + P(Y < -\delta).$$

Daraus folgt

$$\begin{aligned} P(X \leq x - \delta) - \Phi(x - \delta) - \delta - P(Y > \delta) &\leq P(X + Y \leq x) - \Phi(x) \\ &\leq P(X \leq x + \delta) - \Phi(x + \delta) \\ &\quad + \delta + P(Y < -\delta) \end{aligned}$$

und damit

$$d(X + Y, \Phi) \leq d(X, \Phi) + \delta + P(|Y| > \delta).$$

Setzt man $\delta = \eta^{1/3}$, so gilt wegen $\eta \geq E(Y^2) \geq \delta^2 P(|Y| > \delta)$ die Abschätzung $P(|Y| > \delta) \leq \eta^{1/3}$, und daraus folgt (12.7). $\square$

Wir beweisen den Zentralen Grenzwertsatz zuerst unter Zusatzbedingungen:

Satz 12.7 *Sind $X_1, X_2, \dots$ unabhängige Zufallsvariable mit endlich vielen Werten, die für alle m die Bedingungen*

$$EX_m = 0, \quad \text{Var}(X_m) = 1 \quad (12.8)$$

und

$$|X_m| \leq B < \infty \quad (12.9)$$

erfüllen, und ist $S_n = X_1 + \dots + X_n$ und $S_n^* = S_n/\sqrt{n}$, so gilt

$$d(S_n^*, \Phi) \rightarrow 0.$$

Beweis: Sei Y_0 eine von den X_i unabhängige $N(0, 1)$ -verteilte Zufallsvariable. (Auf einem geeigneten Wahrscheinlichkeitsraum lassen sich Y_0 und die X_i definieren.) Sei N zunächst eine fest gewählte natürliche Zahl. Wir setzen für $n \geq N$

$$Z_n = \sqrt{\frac{N}{n}} Y_0 + \frac{1}{\sqrt{n}} (S_n - S_N).$$

Sei $F_n(x) = P(Z_n \leq x)$. Es gilt

$$Z_{n+1} = \sqrt{\frac{n}{n+1}} Z_n + \frac{1}{\sqrt{n+1}} X_{n+1}.$$

Sind $x_1, x_2, \dots, x_k$ die möglichen Werte von X_{n+1} und $p_1, \dots, p_k$ die zugehörigen Wahrscheinlichkeiten, so ist — wegen der Unabhängigkeit der letzten beiden Summanden —

$$\begin{aligned} F_{n+1}(x) &= \sum_{i=1}^k p_i P\left(\sqrt{\frac{n}{n+1}} Z_n \leq x - \frac{1}{\sqrt{n+1}} x_i\right) \\ &= \sum_{i=1}^k p_i F_n\left(\sqrt{\frac{n+1}{n}} x - \frac{1}{\sqrt{n}} x_i\right) \\ &= E\left(F_n\left(\sqrt{\frac{n+1}{n}} x - \frac{1}{\sqrt{n}} X_{n+1}\right)\right). \end{aligned}$$

Sei $\alpha = \sqrt{(n+1)/n}$, $\beta = 1/\sqrt{n}$. Wir erhalten

$$|F_{n+1}(x) - \Phi(x)| \leq |E(F_n(\alpha x - \beta X_{n+1}) - \Phi(\alpha x - \beta X_{n+1}))| + |E(\Phi(\alpha x - \beta X_{n+1}) - \Phi(x))|$$

und damit

$$d(F_{n+1}, \Phi) \leq d(F_n, \Phi) + \sup_x |E(\Phi(\alpha x - \beta X_{n+1}) - \Phi(x))|. \quad (12.10)$$

Die Taylorentwicklung von Φ liefert

$$\Phi(\alpha x - \beta x_i) = \Phi(\alpha x) - \beta x_i \Phi'(\alpha x) + \frac{1}{2} (\beta x_i)^2 \Phi''(\alpha x) - \frac{1}{6} (\beta x_i)^3 \Phi'''(\alpha x - \theta_i(x) \beta x_i)$$

mit $|\theta_i(x)| \leq 1$.

Für nicht notwendig explizit erwähnte Terme $r(x, n)$, für die $r(x, n)/g(n)$ beschränkt ist, schreiben wir wie üblich $O(g(n))$. Z.B. können wir den letzten Term durch $O(n^{-3/2})$ abschätzen, denn $\Phi''' = \varphi''$ ist beschränkt, die $|x_i|$ sind wegen (12.9) durch B beschränkt, und es ist $\beta = n^{-1/2}$.

Es ist $\alpha = 1 + 1/(2n) + O(n^{-2})$. Φ und alle Ableitungen von Φ sowie $x \Phi'(x)$ und $x^2 \Phi''(x)$ sind beschränkt. Die Taylorentwicklung von $\Phi(\alpha x)$ liefert

$$\begin{aligned} \Phi(\alpha x) &= \Phi(x) + \left(\frac{1}{2n} + O\left(\frac{1}{n^2}\right)\right) x \Phi'(x) \\ &\quad + \left(\frac{1}{2n} + O\left(\frac{1}{n^2}\right)\right) \frac{x^2}{2!} \Phi''\left(x + \theta x \left(\frac{1}{2n} + O\left(\frac{1}{n^2}\right)\right)\right) \end{aligned}$$

mit $0 \leq |\theta| \leq 1$. Damit erhalten wir

$$\Phi(\alpha x) = \Phi(x) + \frac{x}{2n} \Phi'(x) + O\left(\frac{1}{n^2}\right).$$

Analog zeigt man

$$\Phi''(\alpha x) = \Phi''(x) + \frac{x}{2n} \Phi'''(x) + O\left(\frac{1}{n^2}\right).$$

Bildet man nun $E(\Phi(\alpha x - \beta X_{n+1})) = \sum_{i=1}^k \Phi(\alpha x - \beta x_i) p_i$ und berücksichtigt, dass aus (12.8)

$$\sum x_i p_i = 0 \quad \text{und} \quad \sum x_i^2 p_i = 1$$

folgt, so ergibt sich

$$E(\Phi(\alpha x - \beta X_{n+1})) = \Phi(x) + \frac{x}{2n} \Phi'(x) + \frac{1}{2n} \Phi''(x) + O\left(\frac{1}{n^{3/2}}\right). \quad (12.11)$$

Der Schlüssel zu dem Beweis ist nun die Identität $\Phi''(x) = -x \Phi'(x)$, die man durch Differenzieren von $\varphi(x) = \Phi'(x)$ verifiziert. In (12.11) heben sich also die Glieder mit Φ' und mit Φ'' gegenseitig weg, so dass aus (12.10) für geeignetes von n und N unabhängiges $K > 0$

$$|d(F_{n+1}, \Phi) - d(F_n, \Phi)| \leq K n^{-3/2}$$

folgt. Wegen $F_N = \Phi$ ist $d(F_N, \Phi) = 0$ und also für alle $n \geq N$

$$d(Z_n, \Phi) = d(F_n, \Phi) \leq K \sum_{i=N}^{\infty} \frac{1}{i^{3/2}}.$$

Wegen der Konvergenz der Reihe auf der rechten Seite können wir bei vorgegebenem $\varepsilon > 0$ N so groß wählen, dass $d(Z_n, \Phi) < \varepsilon/2$ für alle $n \geq N$ gilt. Nun gilt aber bei festem N für $n \rightarrow \infty$

$$E((S_n^* - Z_n)^2) = E\left(\left(\sqrt{\frac{N}{n}} Y_0 - \frac{1}{\sqrt{n}} S_N\right)^2\right) \rightarrow 0.$$

Nach Lemma 12.6 folgt daher für hinreichend großes n

$$d(S_n^*, \Phi) < \varepsilon.$$

□

Den Zentralen Grenzwertsatz für identisch verteilte Zufallsvariable erhalten wir nun durch Approximation.

Satz 12.8 Sind $X_1, X_2, \dots$ unabhängige identisch verteilte Zufallsvariable mit $0 < \sigma^2 := \text{Var}(X_i) < \infty$, und schreibt man

$$S_n = X_1 + \dots + X_n \quad \text{und} \quad S_n^* = \frac{S_n - n EX_1}{\sigma \sqrt{n}},$$

so gilt

$$d(S_n^*, \Phi) \rightarrow 0.$$

Beweis: Sei $f_m(x) = k/m$ für $k/m \leq x < (k+1)/m$ mit $-m^2 \leq k \leq m^2$, und $= 0$ sonst. Sei ferner $Y_{m,i} = f_m(X_i)$, $\mu_m = EY_{m,i}$ und $\sigma_m^2 = E(Y_{m,i} - \mu_m)^2$. Es ist nicht schwer zu zeigen, dass $\mu_m \rightarrow \mu = EX_1$, $\sigma_m^2 \rightarrow \sigma^2$ und

$$E((Y_{m,i} - X_i)^2) \rightarrow 0$$

gilt.

Die Zufallsvariablen $X_{m,i} = (Y_{m,i} - \mu_m)/\sigma_m$ erfüllen für festes m die Voraussetzungen von Satz 12.7. Definiert man also

$$S_{m,n}^* = \frac{1}{\sqrt{n}} (X_{m,1} + \dots + X_{m,n}),$$

so gilt $d(S_{m,n}^*, \Phi) \rightarrow 0$ ($n \rightarrow \infty$). Es ist

$$S_n^* - S_{m,n}^* = \left(1 - \frac{\sigma_m}{\sigma}\right) S_{m,n}^* + \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^n [(X_i - \mu) - (Y_{m,i} - \mu_m)].$$

Bezeichnet man die beiden Summanden auf der rechten Seite mit $A_{m,n}$ und $B_{m,n}$, so ist $E(A_{m,n}^2) = (1 - \sigma_m/\sigma)^2$ und

$$E(B_{m,n}^2) = \frac{1}{\sigma^2} \text{Var}(X_1 - Y_{m,1}) \leq \frac{1}{\sigma^2} E((X_1 - Y_{m,1})^2),$$

denn die Terme [...] sind unabhängig, identisch verteilt und haben Erwartung 0. Für großes m ist daher $E((S_n^* - S_{m,n}^*)^2)$ klein. Mit Hilfe von Lemma 12.6 folgt nun die Behauptung. $\square$

[Die Gleichheit der Verteilungen der X_i ist in Satz 12.8 nicht wesentlich. Man kann ganz ähnlich argumentieren, wenn nur gefordert wird, die X_i seien unabhängig mit $EX_i = 0$, $\text{Var}(X_i) \rightarrow \sigma^2 > 0$ und $\sup_i E(X_i^2 1_{\{|X_i| \geq B\}}) \rightarrow 0$ für $B \rightarrow \infty$. Die Methode lässt sich sogar in viel allgemeineren Situationen anwenden.]

Der Zentrale Grenzwertsatz ist von fundamentaler Bedeutung. In unzähligen wahrscheinlichkeitstheoretischen und statistischen Problemen lässt sich die Verteilung der interessierenden Zufallsvariablen entweder nicht oder nur mit enormem Aufwand bestimmen. In sehr vielen Fällen lassen sich aber asymptotische Aussagen machen, die für die Praxis völlig befriedigend sind und oft durch Tabellen für kleine Beobachtungszahlen ergänzt werden können. Sehr oft spielen Varianten oder Verschärfungen des Zentralen Grenzwertsatzes eine Schlüsselrolle beim Beweis solcher asymptotischen Aussagen.

Anhang*

Anmerkungen zum Gesetz der großen Zahlen: Die Aussage des Gesetzes der großen Zahlen gilt für identisch verteilte Zufallsvariable bereits unter der schwächeren Annahme der Endlichkeit des Erwartungswerts.

Satz 12.9 (Kolmogorow) *Seien $X_1, X_2, \dots$ unabhängige reellwertige Zufallsvariable mit den gleichen Verteilungen. Existiert EX_1 , so konvergiert $Z_n = n^{-1}(X_1 + \dots + X_n)$ fast sicher gegen EX_1 . Für die fast sichere Konvergenz von Z_n (gegen einen reellwertigen Grenzwert) ist die Existenz von EX_1 auch notwendig.*

(Z_n) konvergiert auch schon fast sicher, wenn nur EX_1 existiert und $X_1, X_2, \dots$ eine **stationäre** Verteilung hat, d.h. für jedes n die gemeinsame Verteilung von $X_1, \dots, X_n$ gleich der von $X_2, \dots, X_{n+1}$ ist (**Ergodensatz**).

Borel hat versucht, das starke Gesetz der großen Zahlen für Bernoulli-Folgen (X_i) mit Hilfe der Normalapproximation zu beweisen, dann aber so gerechnet, als ob S_n^* für großes n *exakt* normalverteilt wäre. Diese Lücke lässt sich nur durch genauere Abschätzung des Fehlerterms in der Normalapproximation schließen. Den ersten vollständigen Beweis des starken Gesetzes der großen Zahlen für Bernoulli-Folgen hat F. Hausdorff 1914 in seinem Buch „Grundzüge der Mengenlehre“ gegeben. Hausdorff bezeichnete den Satz als „Satz von Borel“.

Anmerkungen zur Normalapproximation: Eine Abschätzung des Fehlers in der Normalapproximation gibt

Satz 12.10 (Berry-Esseen) Seien $X_1, X_2, \dots$ unabhängig und identisch verteilt. Ist $0 < \sigma^2 = \text{Var}(X_1) < \infty$ und $\gamma = E(|X_1 - EX_1|^3) < \infty$, so gilt $d(S_n^*, \Phi) \leq 0,8 \gamma / (\sigma^3 \sqrt{n})$.

(Siehe z.B. Feller (1966) oder Gänssler-Stute (1977). Die ursprüngliche Konstante war 7,59 statt 0,8. Diese war mehrfach verbessert worden. Den Wert 0,8 hat van Beek (1972) erhalten.)

Eine andere Richtung der Verschärfung sind asymptotische Entwicklungen (**Edgeworth-Entwicklungen**). Man erhält genauere Approximationen als $\Phi(x)$ durch Berücksichtigung der höheren Momente; siehe Feller (1966).

Es gibt heute eine hoch entwickelte Theorie der Grenzverteilungen von Folgen von Zufallsvariablen auch im abhängigen Fall. Der Name „Zentraler Grenzwertsatz“ wird gewöhnlich gebraucht, wenn die Grenzverteilung Φ ist. Die ersten rigorosen über den Satz von de Moivre-Laplace hinausgehenden Resultate verdanken wir Ljapounow (1900).

Die hier verwendete Methode von Kersting wurde von ihm nur skizziert, in publizierter Form erst 1978. (Sie diente ihm zur Lösung eines anderen Problems.) Denker (1975) hat in einem Skript den Beweis ausgearbeitet. Wir haben uns an Pflug (1983) angelehnt, aber auch ein Argument von Rösler (1987) aus einem anderen elementaren Beweis verwandt.

Aufgaben

1. Für Primzahlen p sei $A_p = \{n \in \mathbb{N} : p \text{ teilt } n\}$. Man zeige: Es gibt kein Wahrscheinlichkeitsmaß P auf $\Omega = \mathbb{N}$, für welches die A_p unabhängig sind und $P(A_p) = 1/p$ gilt. (Hinweis: $\sum_p \text{prim } 1/p = \infty$.) (L)
2. $X_1, X_2, \dots$ seien unabhängige Zufallsvariable, und X_k sei exponentiell verteilt mit Parameter $\sqrt{k}$. Zeigen Sie, dass $n^{-1}(X_1 + \dots + X_n)$ fast sicher gegen 0 konvergiert.
3. Beweisen Sie die Äquivalenz der folgenden Aussagen:
a) $X_n \rightarrow X$ fast sicher.
b) Für jedes $\varepsilon > 0$ gilt $P(\sup_{k \geq n} |X_k - X| > \varepsilon) \rightarrow 0$.
4. Man zeige, dass in Beispiel 3.17 $X_n \rightarrow 0$ fast sicher gilt.
5. (X_n) sei eine Folge von reellwertigen Zufallsvariablen mit $X_k \leq X_{k+1}$ für alle k . Man zeige: Die Folge (X_n) konvergiert genau dann fast sicher gegen eine reellwertige Zufallsvariable X , wenn sie stochastisch gegen X konvergiert. (L)
6. (X_n) sei eine Folge von unabhängigen Zufallsvariablen mit der Dichte $f(x) = 1/(\pi(1+x^2))$. Gibt es eine Konstante $a \in \mathbb{R}$ derart, dass $n^{-1}(X_1 + \dots + X_n)$ stochastisch gegen a konvergiert?
7. Sei A_n das Ereignis, dass zwischen dem 2^n -ten Versuch (einschließlich) und dem 2^{n+1} -ten Versuch (ausschließlich) einer unendlichen Folge von Bernoulli-Versuchen mit Erfolgswahrscheinlichkeit $0 < p < 1$ mindestens n Erfolge in ununterbrochener Reihenfolge auftreten. Man bestimme mit Hilfe des Borel-Cantelli-Lemmas die Wahrscheinlichkeit dafür, dass unendlich viele A_n eintreten. (Man betrachte die Fälle $p < 1/2$, $p = 1/2$, $p > 1/2$ separat.)
8. Man zeige, dass die Aussage von Aufgabe 4 in § 5 ein Spezialfall des Zentralen Grenzwertsatzes ist. (Der Einfachheit halber sei $\lambda \in \mathbb{N}$.)
9. Sei X eine Zufallsvariable mit stetiger Verteilungsfunktion und sei (Y_n) eine Folge von Zufallsvariablen, die stochastisch gegen 0 strebt. Man zeige: $d(X + Y_n, X) \rightarrow 0$. (L)
10. Sei $X_1, X_2, \dots$ eine Folge von unabhängigen Zufallsvariablen mit Verteilungsfunktion F .

$$F_n(t, \omega) = \frac{1}{n} \sum_{i=1}^n 1_{]-\infty, t]} \circ X_i(\omega)$$

ist die relative Häufigkeit der $X_i(\omega)$ mit Werten $\leq t$ unter den ersten n . (F_n heißt **empirische Verteilungsfunktion**.) Man zeige: Für fast alle ω gilt $F_n(t, \omega) \rightarrow F(t)$.

11. Es seien $X_1, X_2, \dots$ unabhängig und identisch verteilt mit $EX_i = 0$ und $E(|X_i|^4) < \infty$. Zeigen Sie, dass für $n \rightarrow \infty$

$$\frac{1}{\sqrt{n} \log n} \sum_{i=1}^n X_i$$

stochastisch gegen 0 konvergiert. (L)

§ 13 Schätzverfahren und Fehlerrechnung

Eine einigermaßen vollständige Einführung in die wichtigsten statistischen Methoden soll hier nicht versucht werden. Wir wollen nur exemplarisch einige davon vorstellen. Dies ist relativ leicht, wenn man nur rezeptartig die Verfahren beschreibt. Bei den meisten Verfahren ist es dagegen schwierig nachzuweisen, dass sie in einem geeigneten Sinn optimal sind. Hier wollen wir einen Zwischenweg beschreiten und zunächst einige Schätzmethoden, dann — im nächsten Paragraphen — einige gebräuchliche Tests, aus allgemeinen Überlegungen heraus motivieren. Dafür ist die Maximum-Likelihood-Methode besonders geeignet.

13.1 Maximum-Likelihood-Schätzungen bei Dichten

Es werde eine Zufallsvariable X mit Werten in $\mathbb{R}^n$ beobachtet. Die Verteilung von X hänge von einem unbekannten Parameter $\vartheta \in \Theta$ ab. Wir nennen sie P_ϑ . Hat P_ϑ die Dichte $f(\cdot | \vartheta)$, so ist für jedes $x \in \mathbb{R}^n$ $P_\vartheta(x) = 0$. Man kann also nicht, wie im diskreten Fall, aus der Betrachtung von $P_\vartheta(x)$ Schätzer ableiten. Es ist zweckmäßig, die Likelihood-Funktion nun dadurch zu definieren, dass wir $f(x | \vartheta)$ an die Stelle von $P_\vartheta(x)$ setzen. (Ist $f(x | \vartheta)$ in x stetig und U_x eine kleine Umgebung von x mit Länge (bzw. Fläche, Volumen) $\lambda(U_x)$, so ist $P_\vartheta(U_x) \approx f(x | \vartheta) \lambda(U_x)$. Der Wert von ϑ , welcher $f(x | \vartheta)$ maximiert, wird also (bis auf kleine Terme) auch $P_\vartheta(U_x)$ maximieren.) Als **Likelihood-Funktion** bezeichnen wir daher jetzt die Funktion

$$\vartheta \mapsto L_x(\vartheta) = f(x | \vartheta).$$

Beispiel 13.1 (*Maximum-Likelihood-Schätzer für die Normalverteilung*) Seien $X_1, X_2, \dots, X_n$ unabhängig und $N(\mu, \sigma^2)$ -verteilt. Dann ist $\vartheta = (\mu, \sigma^2)$. Die Dichte der Verteilung von X_i ist

$$f_i(x_i | \vartheta) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right).$$

Also hat $X = (X_1, \dots, X_n)$ an der Stelle $x = (x_1, \dots, x_n)$ die Dichte

$$f(x | \vartheta) = \prod_{i=1}^n f_i(x_i | \vartheta) = \left(\frac{1}{\sigma \sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right). \quad (13.1)$$

Der Maximum-Likelihood-Schätzer $\hat{\vartheta} = \hat{\vartheta}(x)$ ist wieder der $L_x(\vartheta) = f(x | \vartheta)$ maximierende Parameterwert. Wir bestimmen ihn, indem wir das Maximum von $\mathcal{L}_x(\vartheta) = \log L_x(\vartheta)$ aufsuchen. Es ist

$$\mathcal{L}_x(\vartheta) = -n \log(\sigma \sqrt{2\pi}) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2.$$

Wir betrachten drei Fälle:

I. μ ist unbekannt, aber $\sigma^2 = \sigma_0^2$ ist bekannt. Dann ist

$$\Theta = \{(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 = \sigma_0^2\}.$$

Durch Differenziation finden wir, dass

$$\frac{d}{d\mu} \mathcal{L}_x(\vartheta) = 0 \quad \text{genau dann gilt, wenn} \quad \sum_{i=1}^n (x_i - \mu) = 0$$

ist. Daraus ergibt sich der Maximum-Likelihood-Schätzer

$$\hat{\mu} = \bar{x} := \frac{1}{n} \sum_{i=1}^n x_i.$$

Durch Bildung der zweiten Ableitung stellt man fest, dass wirklich ein Maximum in $\hat{\mu}$ vorliegt.

- II. $\mu = \mu_0$ ist bekannt und $\sigma^2 > 0$ unbekannt. Dann ist $\Theta = \{(\mu, \sigma^2) : \mu = \mu_0, \sigma^2 > 0\}$.
Aus

$$\frac{d}{d\sigma} \mathcal{L}_x(\vartheta) = 0 \quad \text{leitet man} \quad -\frac{n}{\sigma} - \frac{1}{2} \sum_{i=1}^n (x_i - \mu_0)^2 (-2)/\sigma^3 = 0$$

her. Daraus ergibt sich für σ^2 der Maximum-Likelihood-Schätzer

$$\tilde{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu_0)^2.$$

- III. μ und σ^2 sind beide unbekannt. Die Gleichungen

$$\frac{d}{d\mu} \mathcal{L}_x(\vartheta) = 0 \quad \text{und} \quad \frac{d}{d\sigma} \mathcal{L}_x(\vartheta) = 0$$

liefern die Maximum-Likelihood-Schätzer $\bar{x}$ für μ und

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

für σ^2 . (Durch Betrachtung der Funktionaldeterminante sieht man, dass ein Maximum vorliegt.)

In jedem der drei Fälle sind die angegebenen Schätzer konsistent; vgl. Abschnitt 4.6. Die Schätzer $\hat{\mu}$ und $\tilde{\sigma}^2$ sind auch erwartungstreu. Statt $\hat{\sigma}^2$ verwendet man

$$s^2 = s^2(x) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (13.2)$$

als Schätzer von σ^2 im Fall III. (Da die Rechenregeln für Erwartungswerte nun auch für Verteilungen mit Dichten zur Verfügung stehen, ist, wie in Abschnitt 4.3, s^2 erwartungstreue Schätzung von σ^2 , während $\hat{\sigma}^2$ nicht erwartungstreu ist.)

[$\hat{\sigma}^2$ ist die **empirische Varianz**, d.h. die Varianz der Verteilung, die durch die empirische Verteilungsfunktion beschrieben wird. Für jedes t ist der Wert der empirischen Verteilungsfunktion (vgl. § 12, Aufg. 10) an der Stelle t erwartungstreu Schätzung von $F(t) = P(X_i \leq t)$.]

Wie erwähnt wird das Modell der Normalverteilung häufig auf Messungen von Gewichten, Längen, Stromstärken usw. angewandt. Sind dann $x_1, \dots, x_n$ die Ergebnisse von n Messungen, so geht man davon aus, dass sie die „Realisierungen“ $x_i = X_i(\omega)$ von Zufallsvariablen seien, die wegen zufälliger Einflüsse um einen wahren zu messenden Wert schwanken. Dieser wird dann durch $\bar{x}$ geschätzt. Die unbekannte Standardabweichung σ wird durch $s = \sqrt{s^2(x)}$ geschätzt. Sie dient als Maß für die Genauigkeit der Messungen.

13.2 Konfidenzintervalle

Es ist bemerkenswert einfach, Konfidenzintervalle für μ anzugeben.

Betrachten wir zunächst den Fall I, in dem μ unbekannt und $\sigma^2 > 0$ bekannt war. Sei dann P_μ das Wahrscheinlichkeitsmaß auf Ω , unter dem $X_1, \dots, X_n$ unabhängig und $N(\mu, \sigma^2)$ -verteilt sind. Wir wissen bereits, dass $X_1 + \dots + X_n$ normalverteilt ist (Satz 11.9). Aus Satz 11.6 folgt, dass für normalverteiltes Y und beliebige $c \neq 0$ und $a \in \mathbb{R}$ die Zufallsvariable $cY + a$ normalverteilt ist. Sei wieder $\bar{X} = n^{-1}(X_1 + \dots + X_n)$. Dann ist

$$Z(X) = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}$$

normalverteilt. Nach den Rechenregeln für Erwartungswerte und Varianzen ergibt sich leicht, dass $Z(X)$ Erwartungswert 0 und Varianz 1 hat, also $N(0, 1)$ -verteilt ist.

Wegen $\Phi(1,96) \geq 0,975$ ist also

$$P_\mu(|Z(X)| \leq 1,96) \geq 0,95.$$

Sei

$$C(x) = \{\mu \in \mathbb{R} : |\bar{x} - \mu| \leq 1,96 \sigma / \sqrt{n}\}.$$

Dann ist $\mu \in C(X)$ äquivalent mit $|Z(X)| \leq 1,96$. Also gilt für alle μ

$$P_\mu(\mu \in C(X)) \geq 0,95.$$

Das bedeutet, dass $C(x)$ ein Konfidenzintervall zum Niveau 0,95 zur Beobachtung $x = (x_1, \dots, x_n)$ ist. Erst recht ist dann

$$[\bar{x} - 2\sigma/\sqrt{n}, \bar{x} + 2\sigma/\sqrt{n}]$$

ein Konfidenzintervall für μ . Gewöhnlich ist natürlich σ^2 unbekannt und wird durch s^2 geschätzt. Man gibt dann für große n

$$[\bar{x} - 2s/\sqrt{n}, \bar{x} + 2s/\sqrt{n}]$$

als Konfidenzintervall an, weil $s^2 \approx \sigma^2$ ist. Für kleine n sollte zur Konstruktion der Konfidenzintervalle die im nächsten Paragraphen diskutierte t -Verteilung benutzt werden. Der Faktor 2 wird dann durch von n abhängige andere Faktoren ersetzt, z.B. für $n = 10$ durch 2,26, für $n = 20$ durch 2,1, und für $n = 30$ durch 2,05. Erst etwa ab $n = 60$ ist der Faktor wirklich ≤ 2 , aber so genau nehmen es wohl die Wenigsten.

Fehlergrenzen: Ist ein unbekannter Wert μ n -mal gemessen worden, und sind $x_1, \dots, x_n$ die Messwerte, so gibt man in der Praxis $\bar{x} \pm s/\sqrt{n}$ als Schätzwert und Fehlergrenzen an. Dahinter steckt genau die obige Überlegung und der Ansatz, die Messungen seien um μ herum normalverteilt. Nehmen wir wieder an, n sei so groß, dass s als gute Schätzung von σ betrachtet werden kann, so ist $[\bar{x} - s/\sqrt{n}, \bar{x} + s/\sqrt{n}]$ ein Konfidenzintervall zum Niveau $2\Phi(1) - 1 \geq 2 \cdot 0,84 - 1 = 0,68$. Für kleines n gibt man das gleiche Intervall an; das Niveau ist dann etwas niedriger.

13.3 Das Fehlerfortpflanzungsgesetz*

Es kommt oft vor, dass eine Größe nicht direkt gemessen wird, sondern aus verschiedenen gemessenen Größen berechnet wird. Wir nehmen an, es lägen Messungen $x_1, \dots, x_m$ einer Größe μ_x vor und Messungen $y_1, \dots, y_n$ einer Größe μ_y . Wir interessieren uns für $f(\mu_x, \mu_y)$, wobei f eine uns bekannte Funktion ist, also z.B. für

$$\psi := f(\mu_x, \mu_y) = \mu_x / \mu_y.$$

Ist $\bar{x} = m^{-1} (x_1 + \dots + x_m)$ und $\bar{y} = n^{-1} (y_1 + \dots + y_n)$, so liegt es nahe,

$$\hat{\psi} = f(\bar{x}, \bar{y})$$

als Schätzung von ψ zu wählen. Wie aber sollen wir den Fehler abschätzen?

Das Fehlerfortpflanzungsgesetz ist eine approximative Methode zur Beantwortung dieser Frage.

Wir betrachten $x_1, \dots, x_m$ als die beobachteten Werte von $N(\mu_x, \sigma_x^2)$ -verteilten Zufallsvariablen $X_1, \dots, X_m$ und $y_1, \dots, y_n$ als die beobachteten Werte von $N(\mu_y, \sigma_y^2)$ -verteilten Zufallsvariablen $Y_1, \dots, Y_n$. All diese Zufallsvariablen seien unabhängig. $f(x, y)$ sei in einer Umgebung von (μ_x, μ_y) zweimal stetig differenzierbar, und diese sei so groß, dass sie $(\bar{X}, \bar{Y})$ „fast mit Wahrscheinlichkeit 1“ enthält.

Sind $\bar{X} - \mu_x$ und $\bar{Y} - \mu_y$ mit großer Wahrscheinlichkeit klein, also σ_x und σ_y klein oder m und n groß, so können wir von einer Taylor-Entwicklung

$$\psi = f(\bar{X}, \bar{Y}) + \left(\frac{\partial f}{\partial x} \right) (\mu_x - \bar{X}) + \left(\frac{\partial f}{\partial y} \right) (\mu_y - \bar{Y}) + \text{kleine Terme}$$

ausgehen, wobei die partiellen Ableitungen in $(\bar{X}, \bar{Y})$ zu wählen sind. Wir vernachlässigen jetzt die kleinen Terme, die $(\mu_x - \bar{X})^2$, $(\mu_y - \bar{Y})^2$ oder $(\mu_x - \bar{X})(\mu_y - \bar{Y})$ als Faktor enthalten, und nehmen zusätzlich an, $(\bar{X}, \bar{Y})$ läge so nahe bei (μ_x, μ_y) , dass wir die partiellen Ableitungen in dieser Umgebung durch Konstanten α_x , α_y ersetzen können. Dann ist

$$\psi - f(\bar{X}, \bar{Y}) \approx \alpha_x (\mu_x - \bar{X}) + \alpha_y (\mu_y - \bar{Y}).$$

Nach Satz 11.9 ist die rechte Seite normalverteilt mit Erwartung 0 und Varianz $\alpha_x^2 \sigma_x^2 / m + \alpha_y^2 \sigma_y^2 / n$. Nun sind

$$s_x^2 = \frac{1}{m-1} \sum_{i=1}^m (x_i - \bar{x})^2 \quad \text{und} \quad s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$$

die Schätzungen von σ_x^2 und σ_y^2 . Also sind

$$\Delta_x = \frac{s_x}{\sqrt{m}} \quad \text{und} \quad \Delta_y = \frac{s_y}{\sqrt{n}}$$

die Schätzungen der Standardabweichungen von $\bar{X} - \mu_x$ und $\bar{Y} - \mu_y$. Wir erhalten so

$$\Delta = \sqrt{\left(\frac{\partial f}{\partial x} \right)^2 \Delta_x^2 + \left(\frac{\partial f}{\partial y} \right)^2 \Delta_y^2}$$

als Schätzung der Standardabweichung von $\hat{\psi} - \psi$. Sind die obigen Approximationen hinreichend genau, so ist

$$[\hat{\psi} - \Delta, \hat{\psi} + \Delta]$$

ein 68%-Konfidenzintervall für ψ . Dies ist das **gaußsche Fehlerfortpflanzungsgesetz**, das natürlich analog für Größen gilt, die aus mehr als zwei gemessenen Größen berechnet werden. Die obige Überlegung ist nach mathematischen Standards schlampig, aber sie lässt sich mit mehr Aufwand durchaus präzisieren. Exakt lässt sie sich als Grenzwertsatz für $m \rightarrow \infty, n \rightarrow \infty$ formulieren. (Vgl. Übungsaufgabe 5.)

Beispiel 13.2 Sei $f(x, y) = x/y$, also $\psi = \mu_x/\mu_y$ zu schätzen. Wegen

$$\frac{\partial f}{\partial x} = \frac{1}{y} \quad \text{und} \quad \frac{\partial f}{\partial y} = -\frac{x}{y^2}$$

ist

$$\Delta^2 = \frac{1}{\bar{y}^2} \Delta_x^2 + \frac{\bar{x}^2}{\bar{y}^4} \Delta_y^2.$$

Man schreibt dann $\psi = \bar{x}/\bar{y} \pm \Delta$. (Für dieses spezielle ψ lässt sich auch ein exaktes Konfidenzintervall angeben. Es stimmt asymptotisch mit dem angegebenen überein.)

13.4 Die Methode der kleinsten Quadrate

Oft stellt sich das Problem, eine Gerade, Parabel oder eine andere „einfache“ Funktion einer gegebenen Menge von Messwerten anzupassen. Z.B. kann in Abhängigkeit von einer Größe x eine Größe y gemessen worden sein, und nun liegen n Messpunkte $(x_1, y_1), \dots, (x_n, y_n)$ vor. Wenn diese Punkte relativ gut auf einer Geraden liegen, könnte man einen linearen Zusammenhang der beobachteten Größen vermuten, der nur durch Messfehler z_i gestört ist. Dann wäre

$$y_i = \alpha + \beta x_i + z_i \quad (i = 1, \dots, n). \quad (13.3)$$

In anderen Fällen könnte etwa aus Naturgesetzen oder aus der Versuchsanordnung folgen, dass ein Ansatz

$$y_i = \alpha + \beta x_i + \gamma x_i^2 + z_i$$

geboten scheint, in dem nur noch α, β und γ unbekannt sind.

Allgemeiner nehmen wir an, $\vartheta_1, \dots, \vartheta_p$ seien unbekannte Parameter, und für bekannte Funktionen φ_i sei

$$\xi_i = \varphi_i(\vartheta_1, \dots, \vartheta_p) \quad (i = 1, \dots, n) \quad (13.4)$$

der wahre zu messende Wert bei der i -ten Messung und $y_i = \xi_i + z_i$ der tatsächlich beobachtete Wert, also z_i der Messfehler. Im Beispiel der Geraden wäre $\vartheta_1 = \alpha, \vartheta_2 = \beta$ und $\varphi_i(\vartheta_1, \vartheta_2) = \vartheta_1 + \vartheta_2 x_i$.

Man fragt, welche Parameterwerte am besten zu den y_i passen. Die Methode der kleinsten Quadrate schreibt vor ϑ_k so zu bestimmen, dass

$$Q := \sum_{i=1}^n (y_i - \xi_i)^2 \quad (13.5)$$

minimal wird. Dies ist als ad hoc Ansatz ohne jede Statistik formulierbar und wird auch oft so angewandt. Nehmen wir nun an, die z_i seien die Realisierungen, d.h. die Werte von unabhängigen $N(0, \sigma^2)$ -verteilten Zufallsvariablen Z_i , so ist $y = (y_1, \dots, y_n)$ die Realisierung von $Y = (Y_1, \dots, Y_n)$ mit $Y_i = \xi_i + Z_i$. Die Y_i sind unabhängig und $N(\xi_i, \sigma^2)$ -verteilt. Also hat Y in y die Dichte

$$\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \xi_i)^2\right).$$

Diese Dichte ist maximal, wenn Q minimal ist. Unter den erwähnten Annahmen sind also die nach der Methode der kleinsten Quadrate bestimmten ϑ_k gerade Maximum-Likelihood-Schätzer.

Beispiel 13.3 (Regressionsgerade) Suchen wir den Messwerten $(x_1, y_1), \dots, (x_n, y_n)$ eine Gerade möglichst gut anzupassen, so müssen wir Schätzwerte $\hat{\alpha}, \hat{\beta}$ für α und β in (13.3) bestimmen. (Man beachte, dass die x_i hier bekannte *feste* Werte sind, während die y_i Realisierungen von Zufallsvariablen sind. Wir weichen hier von unserer Gewohnheit ab, die beobachteten Werte der Zufallsvariablen, auf die wir die Schätzungen stützen, mit x zu bezeichnen, um nicht die übliche Lage der x -Achse und der y -Achse vertauschen zu müssen.) Nun ist also

$$Q = \sum_{i=1}^n (y_i - (\alpha + \beta x_i))^2.$$

Um Q zu minimieren, setzen wir die partiellen Ableitungen von Q nach α und β gleich Null und erhalten die sog. **Normalgleichungen**

$$\sum_{i=1}^n (y_i - \alpha - \beta x_i) = 0 \quad \text{und} \quad \sum_{i=1}^n x_i (y_i - \alpha - \beta x_i) = 0. \quad (13.6)$$

Sei wie üblich $\bar{x} = n^{-1} (x_1 + \dots + x_n)$, $\bar{y} = n^{-1} (y_1 + \dots + y_n)$, und sei

$$\overline{xx} := n^{-1} (x_1^2 + \dots + x_n^2), \quad \overline{xy} := n^{-1} (x_1 y_1 + \dots + x_n y_n).$$

Dann lässt sich (13.6) in der Form

$$\bar{y} - \alpha - \beta \bar{x} = 0, \quad \overline{xy} - \alpha \bar{x} - \beta \overline{xx} = 0$$

schreiben, und dies liefert die Schätzungen

$$\hat{\beta} = \frac{\overline{xy} - \bar{x} \bar{y}}{\overline{xx} - \bar{x} \bar{x}} \quad \text{und} \quad \hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}.$$

Die Gerade $y = \hat{\alpha} + \hat{\beta} x$ nennt man **Regressionsgerade**. Oft schreibt man $\hat{\beta}$ in etwas anderer Form. Sei

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad \text{und} \quad s_{xx} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

Dann ist $\hat{\beta} = s_{xy}/s_{xx}$. Man nennt $\hat{\beta}$ den **empirischen Regressionskoeffizienten**. $\hat{\alpha}$ und $\hat{\beta}$ sind erwartungstreue Schätzer. Auf Verteilungsaussagen gehen wir erst in § 14 ein.

Wir haben nicht angenommen, dass die x_i alle verschieden sind. Aber es müssen natürlich mindestens zwei verschiedene x_i vorkommen. Sonst ist der Nenner in der Formel für $\hat{\beta}$ Null. Sind einige x_i untereinander gleich, so kann man die zu den gleichen x -Werten gehörigen y -Werte in Gruppen zusammenfassen. In modifizierter Beschreibung des Modells sind dann $x_1, \dots, x_n$ die verschiedenen x -Werte und y_{ij} ($j = 1, \dots, n_i$) die n_i an der Stelle x_i gemessenen y -Werte. Dies hat rechentechnische Vorteile, aber es gibt

noch einen wichtigeren Grund so vorzugehen. Oben hatten wir angenommen, die Varianzen aller y -Beobachtungen seien gleich. Man kann nun aus den y_{ij} die Varianz bei festem x -Wert x_i schätzen. Ergibt sich dabei eine nennenswerte Abhängigkeit von der Stelle x_i , so ist die obige Auswertung nicht unverändert anwendbar. (Man kann modifizierte Verfahren angeben, wenn man weiß, wie die Varianz der Y_{ij} von x_i abhängt, also z.B. im Fall $\text{Var}(Y_{ij}) = \sigma^2 x_i$. Aber dies führt hier zu weit.)

Nicht frei variierende Parameterwerte: Wenn in (13.4) die Parameterwerte $\vartheta_1, \dots, \vartheta_p$ frei über eine offene Teilmenge des $\mathbb{R}^p$ variieren, kann man das Minimum von Q in (13.5) gewöhnlich bestimmen, indem die partiellen Ableitungen von Q nach ϑ_i gleich Null gesetzt werden. Oft gibt es aber Beziehungen unter den ϑ_i . Sind z.B. a, b, c Messwerte für die drei Winkel α, β, γ in einem Dreieck, so gilt (im Bogenmaß) $\alpha + \beta + \gamma = \pi$. Betrachtet man $\vartheta_1 = \alpha, \vartheta_2 = \beta, \vartheta_3 = \gamma$ als die unbekannten Parameter, so variieren sie also nicht frei im $\mathbb{R}^3$. Man kann dies auf verschiedene Weise berücksichtigen. Eine Möglichkeit ist natürlich, nur zwei ϑ_i zu verwenden. Dann ist $y_1 = a, y_2 = b, y_3 = c, \vartheta_1 = \alpha, \vartheta_2 = \beta, \xi_1 = \vartheta_1, \xi_2 = \vartheta_2, \xi_3 = \pi - \vartheta_1 - \vartheta_2$, also

$$Q = (a - \alpha)^2 + (b - \beta)^2 + (c - (\pi - \alpha - \beta))^2$$

nach α und β zu differenzieren usw. Eine zweite Möglichkeit ist es, drei Parameterwerte zu verwenden, also $\vartheta_1 = \alpha, \vartheta_2 = \beta, \vartheta_3 = \gamma$, aber dann muss man in $Q = (a - \alpha)^2 + (b - \beta)^2 + (c - \gamma)^2$ das Minimum unter der *Nebenbedingung* $\alpha + \beta + \gamma = \pi$ suchen. Dieses zweite Verfahren ist (jedenfalls bei komplizierteren Schätzproblemen) oft vorzuziehen. Maxima und Minima unter Nebenbedingungen werden mit der Methode der Lagrangeschen Multiplikatoren bestimmt.

13.5 Median, Ausreißer und Robuste Schätzer*

Wenn in einem Schweizer Bergdorf fünf zugezogene Multimillionäre und fünfzig Einheimische leben, ist es für die Einheimischen wenig befriedigend, wenn man ihnen erklärt, das durchschnittliche Einkommen in diesem Ort sei doch recht hoch. Der Erwartungswert einer Verteilung ist manchmal keine gute Maßzahl für das „Zentrum“ einer Verteilung. Man betrachtet daher auch andere Maßzahlen. Die bekannteste davon ist der **Median**.

Ist Z eine reellwertige Zufallsvariable, so heißt jede Zahl μ_m mit $P(Z \geq \mu_m) \geq 1/2$ und $P(Z \leq \mu_m) \geq 1/2$ ein Median von Z (oder von der Verteilung von Z). μ_m ist durch diese Bedingung nicht notwendig eindeutig bestimmt. Gibt es ein Intervall $[a, b]$ mit $P(Z \geq b) = 1/2$ und $P(Z \leq a) = 1/2$, so ist jede Zahl zwischen a und b ein Median. Dies ist aber auch der einzige Fall ohne Eindeutigkeit. Ist Z **symmetrisch** um eine Zahl c **verteilt**, d.h. gilt $P(Z \geq c + t) = P(Z \leq c - t)$ für alle $t \geq 0$, so ist c ein Median. Für symmetrische Verteilungen mit endlicher Erwartung ist, wie man leicht prüft, $EZ = c$.

Für n Messwerte $x_1, \dots, x_n \in \mathbb{R}$ definieren wir den empirischen Median $\hat{\mu}_m$ folgendermaßen: Sind $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ die der Größe nach geordneten x_i , so sei für ungerades n

$$\hat{\mu}_m = \hat{\mu}_m(x_1, \dots, x_n) = x_{((n+1)/2)}$$

das $x_{(k)}$ mit dem mittleren Index. Ist n gerade, so gibt es keinen mittleren Index. Dann sei $\hat{\mu}_m(x)$ der Mittelwert der beiden $x_{(k)}$ in der Mitte der aufsteigenden Folge, also

$$\hat{\mu}_m = \frac{1}{2} (x_{(n/2)} + x_{((n/2)+1)}) .$$

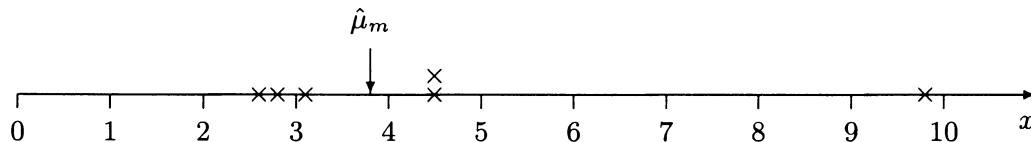
Beispiel 13.4 Ist

$$x_1 = 3,1, \quad x_2 = 4,5, \quad x_3 = 4,5, \quad x_4 = 2,8, \quad x_5 = 2,6, \quad x_6 = 9,8,$$

so ist $x_{(1)}$ der kleinste dieser Werte, also $x_{(1)} = 2,6$. $x_{(2)}$ ist der nächstgrößere, usw. Mehrfach auftretende Werte werden dabei mehrfach berücksichtigt. Man erhält

$$x_{(1)} = 2,6, \quad x_{(2)} = 2,8, \quad x_{(3)} = 3,1, \quad x_{(4)} = 4,5, \quad x_{(5)} = 4,5, \quad x_{(6)} = 9,8.$$

Der Median ist $\hat{\mu}_m = (3,1 + 4,5)/2 = 3,8$ (Bild 13.1).

**Bild 13.1** Median

Die Folge $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ nennt man die **Ordnungsstatistik** zu den beobachteten Werten $x_1, \dots, x_n$.

Für eine Folge von Zufallsvariablen $X_1, \dots, X_n$ sei $X_{(1)}(\omega) \leq X_{(2)}(\omega) \leq \dots \leq X_{(n)}(\omega)$ die aufsteigend geordnete Folge der $X_1(\omega), \dots, X_n(\omega)$. (Man beachte, dass die Umordnung von ω abhängt.)

Aus dem starken Gesetz der großen Zahlen lässt sich ableiten: Ist Z eine reellwertige Zufallsvariable mit eindeutigem Median μ_m und sind $Z_1, Z_2, \dots$ unabhängige Zufallsvariable mit der gleichen Verteilung wie Z , so gilt für $n \rightarrow \infty$

$$\hat{\mu}_m(Z_1, \dots, Z_n) \rightarrow \mu_m \quad \text{fast sicher.}$$

Wir verschieben dies auf die Übungen (Aufgabe 6). Man kann daher den empirischen Median als Schätzer des Medians, und gegebenenfalls des Symmetriezentrums einer Verteilung mit eindeutigem Median verwenden.

Ausreißer: Wenn man eine Reihe von Messwerten wie in Bild 13.1 registriert hat, besteht oft zu Recht die Vermutung, dass ein Messwert, der sich relativ weit von den übrigen entfernt, durch eine Störung zustande gekommen ist und dass er das Ergebnis verfälschen würde, wenn man ihn nicht streicht. Z.B. kann in Bild 13.1 der ganz rechts liegende Wert ein solcher „Ausreißer“ sein. Wenn man den Mittelwert $\bar{x} = n^{-1}(x_1 + \dots + x_n)$ als Schätzer benutzt, stört solch ein Ausreißer relativ stark. Für den Median $\hat{\mu}_m$ spielt er eine geringere Rolle. Man sagt, dass $\hat{\mu}_m$ ein **robusterer Schätzer** als $\bar{x}$ sei.

Man könnte deshalb daran denken, einen Parameter μ , für den n Messwerte $x_1, \dots, x_n$ vorliegen, stets durch den empirischen Median zu schätzen. Aber auch dies hat Nachteile. Man kann sich z.B. überlegen, dass der empirische Median relativ langsam konvergiert, so dass die Schätzung von μ mittels des Medians gewöhnlich unnötig ungenau ist.

Es sind daher — zunächst mehr oder weniger intuitiv — andere Methoden vorgeschlagen worden, wie man Schätzer bilden soll, die den Einfluss von Ausreißern klein halten und dabei möglichst genau sind. Bekannt ist z.B. das **getrimmte Mittel** $\bar{x}_\alpha$. Sei $0 \leq \alpha < 1/2$ und $[n\alpha]$ die größte ganze Zahl $\leq n\alpha$. Man lässt die $[n\alpha]$ kleinsten und die $[n\alpha]$ größten Werte weg und bildet aus den verbleibenden Werten den Mittelwert $\bar{x}_\alpha$, d.h. man bildet

$$\bar{x}_\alpha = (n - 2[n\alpha])^{-1} (x_{([n\alpha]+1)} + x_{([n\alpha]+2)} + \dots + x_{(n-[n\alpha])}).$$

α wird so gewählt, dass man davon ausgehen kann, dass in der betreffenden Situation gewöhnlich nicht mehr als $2\alpha \cdot 100\%$ der Beobachtungen Ausreißer sind. Üblich sind Werte mit $0,1 \leq \alpha \leq 0,2$.

Andere robuste Schätzer sind das **Winsorisierte Mittel** w_α und das **Lehmann-Hodges-Mittel**. w_α wird gebildet, indem die $[n\alpha]$ kleinsten und die $[n\alpha]$ größten Werte nicht weggelassen werden, sondern nur zum nächstgelegenen der übrigen Werte hin verschoben werden, bevor gemittelt wird. Das Lehmann-Hodges-Mittel ist der empirische Median der Werte $(x_i + x_j)/2$ mit $i \neq j$. Es gibt inzwischen nicht nur umfangreiche theoretische Studien zu diesem Problemkreis, sondern auch Simulationen, mit denen man feststellt, welche dieser Schätzer in welchen Situationen vorteilhaft sind; ferner hat man robuste Schätzer mit Hilfe von historischen Datensätzen verglichen. (Vgl. Bickel-Doksum (1977), Huber (1981), Andrews u.a. (1972), Stigler (1977).)

Fehlergrenzen für das getrimmte Mittel: Um wie bisher von Fehlergrenzen sprechen zu können, müssen wir natürlich Modellannahmen formulieren, die nun zulassen sollen, dass die Messfehler mit nicht zu vernachlässigender Wahrscheinlichkeit auch relativ groß ausfallen können. Seien die Messfehler Z_i unabhängig und mit einer um 0 symmetrischen Dichte $f(x) = f(-x)$ verteilt. Wir beobachten $X_i = \mu + Z_i$. Sei zunächst f bekannt, und seien $q_\alpha, q_{1-\alpha}$ die sog. α - und $(1 - \alpha)$ -Quantile zu f , d.h. $P(Z_i \leq q_\alpha) = \alpha$ und $P(Z_i \leq q_{1-\alpha}) = 1 - \alpha$. (Wegen der Symmetrie von f ist $q_{1-\alpha} = -q_\alpha$.) Ist $\bar{X}_\alpha(\omega)$ das getrimmte Mittel zu $X_1(\omega), \dots, X_n(\omega)$ (das natürlich von n abhängt), so gilt nach Bickel (1965):

$$P(\sqrt{n}(\bar{X}_\alpha - \mu) \leq t) \rightarrow \Phi(\sigma_\alpha t) \quad \text{für } n \rightarrow \infty$$

mit

$$\sigma_\alpha^2 = \frac{2}{(1 - 2\alpha)^2} \left\{ \int_0^{q_{1-\alpha}} t^2 f(t) dt + \alpha q_{1-\alpha}^2 \right\}.$$

Nun ist f und damit σ_α^2 natürlich nicht bekannt, aber σ_α^2 lässt sich schätzen durch die **Winsorisierte Stichprobenvarianz**

$$\hat{\sigma}_\alpha^2 = \frac{1}{(1 - 2\alpha)^2 n} \left\{ \sum_{i=[n\alpha]+1}^{n-[n\alpha]} (X_{(i)} - \bar{X}_\alpha)^2 + [n\alpha] ((X_{([n\alpha]+1)} - \bar{X}_\alpha)^2 + (X_{(n-[n\alpha])} - \bar{X}_\alpha)^2) \right\}.$$

(Vgl. Jaeckel (1971)). $T = \sqrt{n}(\bar{X}_\alpha - \mu)/\hat{\sigma}_\alpha$ ist asymptotisch $N(0, 1)$ -verteilt, d.h. für $n \rightarrow \infty$ gilt $P(T \leq t) \rightarrow \Phi(t)$. Will man μ robust schätzen, so sollte man also $\bar{x}_\alpha \pm \hat{\sigma}_\alpha$ statt $\bar{x} \pm s$ angeben. Ist in Wirklichkeit f die Dichte einer Normalverteilung (gibt es also keine Ausreißer), so verliert man auf diese Weise nicht viel an Präzision.

Anhang*

Historisches: A. M. Legendre hat 1805 die Methode der kleinsten Quadrate als Erster beschrieben. C. F. Gauß löste einen Prioritätsstreit aus, als er 1809 in seinem astronomischen Hauptwerk „Theoria motus corporum coelestium“ erwähnte, er habe die Methode seit 1795 benutzt. Dafür gibt es tatsächlich in einem 1799 publizierten Brief von Gauß Anhaltspunkte, die allerdings nicht schlüssig sind. (Vgl. Stigler (1981).)

Man kann aber wohl in jedem Fall Legendre und Gauß die Väter der Methode nennen, denn Gauß hat die Methode wahrscheinlichkeitstheoretisch begründet und entscheidend entwickelt. 1809 verwandte Gauß ein Maximum-Likelihood-Argument. In seiner Schrift „Theoria combinationis observationum erroribus minimis obnoxiae“ (1819–1822) hat Gauß dann die Fehlertheorie geschaffen. Seine hier gegebene Begründung der Methode der kleinsten Quadrate, der er später stets den Vorzug gab, lief darauf hinaus, dass der mittlere quadratische Fehler (§ 4) möglichst klein gehalten werden solle. Dies wollen wir etwas präzisieren, da es auch heute noch wichtig ist. Wir verwenden die Bezeichnungen aus Abschnitt 13.4.

Satz von Gauß-Markow: Man spricht von einem **linearen Modell**, wenn die ξ_i von der Form $\xi_i = c_{i1} \vartheta_1 + \dots + c_{ip} \vartheta_p$ sind. Die Z_i seien unabhängige Messfehler mit $EZ_i = 0$ und $\text{Var}(Z_i) = \sigma^2$, aber nicht notwendig normalverteilt. Ein Wert $\hat{\vartheta}$ von $\vartheta = (\vartheta_1, \dots, \vartheta_p)$, für den $\sum_{i=1}^n (y_i - \xi_i)^2$ minimal ist, heißt **Kleinste-Quadrate-Schätzer** von ϑ . (Solch ein $\hat{\vartheta}$ existiert stets; es braucht aber nicht eindeutig bestimmt zu sein.)

Sei ϑ_j oder allgemeiner $g(\vartheta) = g_1 \vartheta_1 + \dots + g_p \vartheta_p$ zu schätzen. Ein Schätzer der Form $T(y) = t_1 y_1 + \dots + t_n y_n$ heißt **linear**. Der Satz von Gauß-Markow besagt:

Wenn es überhaupt einen linearen erwartungstreuen Schätzer von $g(\vartheta)$ gibt (was von (c_{ij}) abhängt), ist $g(\hat{\vartheta})$ der lineare erwartungstreue Schätzer mit kleinster Varianz. (Er ist auch dann eindeutig bestimmt, wenn $\hat{\vartheta}$ es nicht ist.)

Mehr zur Methode der kleinsten Quadrate findet man z.B. bei van der Waerden (1971) und Linnik (1961).

Ausblick: In neuerer Zeit werden neben den robusten Methoden intensiv statistische Methoden studiert, die mittels Computersimulationen von Stichproben aus einem festen Datensatz Konfidenzbereiche festlegen, sog. „**Bootstrap-Methoden**“. Wir wollen die Grundidee wenigstens andeuten:

Seien $X_1, X_2, \dots, X_n$ unabhängige identisch verteilte Zufallsvariable mit unbekannter Verteilungsfunktion F und endlicher Varianz $\sigma^2 > 0$. Um die Abhängigkeit der Wahrscheinlichkeiten von F anzuzeigen, bezeichnen wir das Wahrscheinlichkeitsmaß mit P_F .

Nehmen wir einmal an, wir suchten ein Konfidenzintervall für $\mu = E_F(X_i)$ zu vorgegebenem Niveau $1 - \alpha$. Legen wir den Mittelwert $\hat{\mu} = n^{-1} (X_1 + \dots + X_n)$ als Schätzer für μ zugrunde, so liegt es nahe, ein Intervall der Form $[\hat{\mu} - \delta, \hat{\mu} + \delta]$ zu verwenden, wobei $\delta > 0$ so zu bestimmen ist, dass $P_F(|\hat{\mu} - \mu| \leq \delta) \geq 1 - \alpha$ für alle F gilt. δ darf vom beobachteten Vektor $X = (X_1, \dots, X_n)$ abhängen, aber nicht von F .

Ist n nicht zu klein, so wird die durch $F_n(t) = (1/n) \cdot (\text{Anzahl der } i \text{ mit } X_i \leq t)$ definierte empirische Verteilungsfunktion F_n sehr wahrscheinlich nahe bei F liegen. Dies lässt vermuten, dass man gute Approximationen für Wahrscheinlichkeiten $P_F(\cdot)$ erhalten kann, wenn man F durch F_n ersetzt. Da F_n nur von X und nicht von F abhängt, wäre damit viel gewonnen.

Natürlich müsste das sich auf F_n stützende Experiment das Ursprungsexperiment simulieren. Seien daher $X_1^*, \dots, X_n^*$ bei gegebenem X unabhängige Zufallsvariable mit Verteilungsfunktion F_n . Diese Voraussetzung ist erfüllt, wenn $X_1^*, \dots, X_n^*$ die Ergebnisse von n Ziehungen mit Zurücklegen aus einer Urne mit n Kugeln sind, auf denen die Zahlen $X_1, \dots, X_n$ stehen. Der Erwartungswert der X_i^* bei gegebenem X ist $\hat{\mu}$. Ist $\hat{\mu}^*$ der Mittelwert der X_i^* , so sollte gelten

$$P_{F_n}(\hat{\mu}^* - \hat{\mu} \leq t) \approx P_F(\hat{\mu} - \mu \leq t).$$

Tatsächlich strebt die Differenz der beiden Seiten fast sicher gleichmäßig in t gegen 0. (Die Konvergenz lässt sich noch durch Standardisieren verbessern.) Der Wert der linken Seite lässt sich leicht mit einem Computer approximativ bestimmen, indem man das Simulationsexperiment bei festem X sehr oft unabhängig wiederholt und die relative Häufigkeit auszählt, mit der das Ereignis $\hat{\mu}^* - \hat{\mu} \leq t$ auftritt. Man spricht daher von **resampling**, weil man aus einer festen Stichprobe immer wieder neue Stichproben zieht. Das Wort „bootstrap“ bedeutet Stiefelschlaufe. Man zieht sich an der Stiefelschlaufe F_n aus dem Sumpf der mangelnden Kenntnis von F wie Münchhausen an seinen Haaren.

Ausführlicher und mit Begründung ist die Idee bei Falk (1987) dargestellt. Diaconis und Efron (1983) haben eine — so weit wie möglich — allgemein verständliche Darstellung der von Efron (1979) vorgeschlagenen Bootstrap-Methoden gegeben. Eine elementare anwendungsbezogene Einführung in diese Thematik liefert das Buch von Efron und Tibshirani (1993).

Simulationen und Monte-Carlo-Methoden spielen natürlich auch in anderen Teilgebieten der Stochastik eine große Rolle; siehe z.B. Rubinstein (1981).

Ein anderer moderner Zweig der Statistik ist die explorative Datenanalyse, in der es nicht um mathematische Bestätigungen sondern darum geht, wie man aus unübersichtlichen Daten Zusammenhänge und Trends erkennen lernt; siehe Tukey (1977).

Aufgaben

1. Sei Z eine reellwertige integrierbare Zufallsvariable mit Dichte f und Median μ_m . Man zeige: Für alle $a \in \mathbb{R}$ ist $E|Z - a| \geq E|Z - \mu_m|$ und Gleichheit gilt genau dann, wenn μ_m ein Median von Z ist.
2. Je 10 Messungen der Winkel α, β, γ eines Dreiecks ergeben folgende Mittelwerte: $\bar{\alpha} = 61,3^\circ$, $\bar{\beta} = 44,6^\circ$, $\bar{\gamma} = 73,4^\circ$. Die Winkelsumme ist auf Grund der Messfehler nicht 180° . Man bestimme mit Hilfe der Methode der kleinsten Quadrate Schätzungen $\hat{\alpha}, \hat{\beta}, \hat{\gamma}$ mit Winkelsumme 180° . (L)
3. Wir nehmen an, dass Knaben- und Mädchengeburten gleich wahrscheinlich seien, und bezeichnen die Wahrscheinlichkeit, dass ein Knabe rot-grün-blind ist, mit $1 - \vartheta$. Nach einem genetischen Modell ist dann $(1 - \vartheta)^2$ die Wahrscheinlichkeit, dass ein Mädchen rot-grün-blind ist. Es sind 8870 Männer und 9140 Frauen untersucht worden, und davon waren 670 Männer und 40 Frauen rot-grün-blind. Man bestimme eine Maximum-Likelihood-Schätzung für ϑ .
4. Sei $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ die Ordnungsstatistik zu n auf dem Einheitsintervall gleichverteilten unabhängigen Zufallsvariablen $X_1, \dots, X_n$. Man zeige: $E(X_{(k)}) = k/(n+1)$ für $k = 1, \dots, n$. (Hinweis: Um $P(X_{(k)} \leq t)$ zu bestimmen, kann man die Beobachtung von X_i als Erfolg werten, wenn $X_i \leq t$ gilt.) (L)
5. (X_n) sei eine Folge von unabhängigen $b_{1,p}$ -verteilten Zufallsvariablen; $0 < p < 1$. Sei $S_n = X_1 + \dots + X_n$. f sei eine auf dem Einheitsintervall definierte und dort zweimal stetig differenzierbare reellwertige Funktion. Man zeige: Ist $f'(p) \neq 0$, so gilt für alle $x_1 < x_2$

$$\lim_{n \rightarrow \infty} P \left(x_1 \leq \frac{\sqrt{n}(f(S_n/n) - f(p))}{f'(p) \sqrt{p(1-p)}} \leq x_2 \right) = \Phi(x_2) - \Phi(x_1).$$

6. $X_1, X_2, \dots$ seien unabhängig mit Dichte $f(x) = \pi^{-1} (1 + x^2)^{-1}$. Man zeige, dass der empirische Median fast sicher gegen 0 konvergiert. (Man verwende Aufgabe 10 aus § 12.)
7. Sei mit den Bezeichnungen aus § 13.4 $Y_i = \xi_i + Z_i$. Nun aber seien die Z_i $N(0, w_i \sigma^2)$ -verteilt mit bekannten $w_i > 0$. Wie bestimmt man Maximum-Likelihood-Schätzer für $\vartheta_1, \dots, \vartheta_p$? Wenden Sie dies auf Beispiel 13.3 an. (L)
8. Gibt es Kleinste-Quadrate-Schätzer im linearen Regressionsmodell, wenn nur Parameterwerte $\alpha > 0, \beta > 0$ möglich sind?
9. Bestimmt man Schätzer der Parameter ϑ_k dadurch, dass (mit den Bezeichnungen aus § 13.4) $Q^* = \sum |y_i - \xi_i|$ minimiert wird, so spricht man von der Methode der kleinsten absoluten Werte. Zeigen Sie, dass diese Methode gerade die Maximum-Likelihood-Methode ist, wenn die Z_i eine Dichte $f(z) = \exp(-|z|)/2$ haben.

§ 14 Einige wichtige Testverfahren

Erinnern wir uns, dass in einem Testproblem eine — meist vektorwertige — Zufallsvariable X beobachtet wird, deren Verteilung P_ϑ einer Familie $\{P_\vartheta : \vartheta \in \Theta\}$ angehört, der Menge der im Modell in Betracht gezogenen Verteilungen. Θ ist die disjunkte Vereinigung zweier nichtleerer Mengen H und K , der Hypothese und der Alternative, und es soll aufgrund des beobachteten Wertes x von X entschieden werden, ob ϑ zu H oder zu K gehört.

Wir bilden den **Likelihood-Quotienten**

$$q(x) = \frac{\sup\{L_x(\vartheta) : \vartheta \in K\}}{\sup\{L_x(\vartheta) : \vartheta \in H\}}. \quad (14.1)$$

Ein **Test** φ ist nun eine *messbare* Abbildung des Wertebereichs $\mathcal{X}$ von X in $[0, 1]$. Wird x beobachtet, so besagt $\varphi(x) = 1$ wieder, dass die Hypothese verworfen werden soll; im Fall $\varphi(x) = 0$ soll sie angenommen werden, und im Fall $0 < \varphi(x) < 1$ soll ein zusätzliches Zufallsexperiment mit Wahrscheinlichkeit $\varphi(x)$ zur Verwerfung führen.

φ heißt **Likelihood-Quotienten-Test**, wenn für geeignetes $c \in \mathbb{R}$ im Fall $q(x) > c$ die Hypothese verworfen wird und sie im Fall $q(x) < c$ angenommen wird, also wenn

$$\{x : q(x) > c\} \subset \{x : \varphi(x) = 1\} \quad \text{und} \quad \{x : q(x) < c\} \subset \{x : \varphi(x) = 0\} \quad (14.2)$$

gilt.

Nach dem Neyman-Pearson-Lemma, das sich mit dem gleichen Beweis auf die Situation mit Dichten überträgt, existiert im Fall einfacher Hypothesen und Alternativen zu $0 \leq \alpha \leq 1$ ein Likelihood-Quotienten-Test zum Niveau α , und er ist schärfster Test zu seinem Niveau.

Für zusammengesetzte (d.h. nicht einfache) Hypothesen und Alternativen ist es nicht mehr allgemein richtig, dass ein Likelihood-Quotienten-Test schärfster Test zu seinem Niveau ist. Zahlreiche wichtige Tests lassen sich aber als Likelihood-Quotienten-Tests darstellen. Dass sie in geeignetem Sinn optimal sind, lässt sich oft zeigen, aber die dazu erforderliche Theorie kann in diesem Rahmen nicht behandelt werden. Wir wollen einige häufig verwendete Tests durch Betrachtung der Likelihood-Quotienten ableiten und die dabei benötigten Verteilungen studieren.

14.1 Der t -Test

$X_1, \dots, X_n$ seien unabhängig $N(\mu, \sigma^2)$ -verteilt mit unbekanntem (μ, σ^2) . Für ein gegebenes festes μ_0 sei zu testen, ob $\mu = \mu_0$ oder $\mu \neq \mu_0$ gilt. Dann ist $\vartheta = (\mu, \sigma^2)$,

$$H = \{(\mu, \sigma^2) : \mu = \mu_0, \sigma^2 > 0\}, \quad K = \{(\mu, \sigma^2) : \mu \neq \mu_0, \sigma^2 > 0\}$$

und $\Theta = H \cup K$. Ist $f(x|\vartheta)$ die in (13.1) angegebene Dichte der Verteilung von $X = (X_1, \dots, X_n)$, so ist aus Stetigkeitsgründen, da K in Θ dicht ist,

$$\sup\{f(x|\vartheta) : \vartheta \in K\} = \sup\{f(x|\vartheta) : \vartheta \in \Theta\}.$$

Die Parameterwerte (μ, σ^2) , die bei festem x die Likelihood-Funktion $L_x(\vartheta) = f(x|\vartheta)$ maximieren, hatten wir schon am Anfang von § 13 bestimmt. Nach Fall III wurde das Supremum über *alle* $\vartheta \in \Theta$ im Punkt $(\bar{x}, \hat{\sigma}^2)$ angenommen, d.h. es gilt für $x = (x_1, \dots, x_n)$

$$\sup\{L_x(\vartheta) : \vartheta \in K\} = f(x | (\bar{x}, \hat{\sigma}^2)).$$

Ebenso ergibt sich aus Fall II

$$\sup\{L_x(\vartheta) : \vartheta \in H\} = f(x | (\mu_0, \tilde{\sigma}^2)).$$

Nun ist

$$f(x | (\bar{x}, \hat{\sigma}^2)) = \left(\frac{1}{\sqrt{2\pi} \hat{\sigma}}\right)^n \exp\left(-\frac{1}{2\hat{\sigma}^2} \sum_{i=1}^n (x_i - \bar{x})^2\right) = \left(\frac{1}{\sqrt{2\pi} \hat{\sigma}}\right)^n \exp\left(-\frac{n}{2}\right),$$

und ebenso

$$f(x | (\mu_0, \tilde{\sigma}^2)) = \left(\frac{1}{\sqrt{2\pi} \tilde{\sigma}}\right)^n \exp\left(-\frac{n}{2}\right).$$

Daher ist

$$q(x) = \left(\frac{\tilde{\sigma}}{\hat{\sigma}}\right)^n.$$

Ist φ Likelihood-Quotienten-Test, so gilt für geeignetes c die Bedingung (14.2). Für geeignetes $c' = c^{2/n}$ gilt also

$$\varphi(x) = 1 \quad \text{auf} \quad \{\tilde{\sigma}^2 > c' \hat{\sigma}^2\} \quad \text{und} \quad \varphi(x) = 0 \quad \text{auf} \quad \{\tilde{\sigma}^2 < c' \hat{\sigma}^2\}. \quad (14.3)$$

Es ist

$$\frac{\tilde{\sigma}^2}{\hat{\sigma}^2} = \frac{\sum_{k=1}^n (x_k - \mu_0 - \bar{x} + \bar{x})^2}{\sum_{k=1}^n (x_k - \bar{x})^2} = \dots = 1 + \frac{(\bar{x} - \mu_0)^2}{\hat{\sigma}^2}. \quad (14.4)$$

Wir setzen

$$T(x) = \frac{\sqrt{n}(\bar{x} - \mu_0)}{s(x)} \quad \text{mit} \quad s(x) = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}. \quad (14.5)$$

Dann ist der letzte Summand in (14.4) ein Vielfaches von $|T(x)|^2$. Also ist für geeignetes $t \in \mathbb{R}$ die Bedingung $q(x) > c$ äquivalent zu $|T(x)| > t$.

Für jeden Likelihood-Quotienten-Test φ gibt es demnach ein t mit

$$\varphi(x) = \begin{cases} 1, & \text{falls } |T(x)| > t, \\ 0, & \text{falls } |T(x)| < t. \end{cases} \quad (14.6)$$

Für x mit $|T(x)| = t$ setzen wir $\varphi(x) = 0$. (Wir können hier willkürlich vorgehen, weil die Verteilung von $T(X)$ unter jedem P_ϑ eine Dichte hat, so dass $P_\vartheta(|T(X)| = t) = 0$ gilt. Für $\vartheta \in H$ werden wir das gleich sehen.)

Um zu vorgegebenem Signifikanzniveau $\alpha > 0$ die Zahl t zu finden, für die φ gerade Niveau α hat, muss man die Verteilung von $T(X)$ unter der Hypothese kennen.

Die Zufallsvariablen $Y_i = (X_i - \mu_0)/\sigma$ sind unter der Hypothese unabhängig und $N(0, 1)$ -verteilt. Ist $Y = (Y_1, \dots, Y_n)$, so ist, wie man leicht nachrechnet, $s(X) = \sigma s(Y)$ und

$$T(X) = \frac{\bar{Y} \sqrt{n}}{s(Y)}. \quad (14.7)$$

Insbesondere ist die Verteilung von $T(X)$ unabhängig von μ_0 und σ^2 . Besonders letzteres ist wichtig, da wir σ^2 ja nicht kennen. Die Verteilung von Y hat die Dichte

$$g(y_1, \dots, y_n) = \left(\frac{1}{\sqrt{2\pi}} \right)^n \exp \left(-\frac{1}{2} \sum_{i=1}^n y_i^2 \right).$$

g ist invariant unter orthogonalen Transformationen. Ist also A eine orthogonale $n \times n$ -Matrix, und $Z = (Z_1, \dots, Z_n)$ der Zufallsvektor YA , so gilt für jedes n -dimensionale Rechteck $[a, b[$

$$\begin{aligned} P(Z \in [a, b[) &= P(Y \in [a, b[A^{-1}) = \int_{[a, b[A^{-1}} g(y_1, \dots, y_n) dy_1 \dots dy_n \\ &= \int_{[a, b[} g(y_1, \dots, y_n) dy_1 \dots dy_n = P(Y \in [a, b[). \end{aligned}$$

Z hat also die gleiche Verteilung wie Y , d.h. $Z_1, \dots, Z_n$ sind unabhängig $N(0, 1)$ -verteilt.

Sei nun A eine orthogonale Matrix, die den in Richtung der Diagonalen weisenden Einheitsvektor $d = (1/\sqrt{n}) (1, 1, \dots, 1)$ auf den Vektor $(1, 0, 0, \dots, 0)$ abbildet.

Die Projektion yP von $y = (y_1, \dots, y_n)$ auf die Diagonale ist der Vektor $(\bar{y}, \bar{y}, \dots, \bar{y})$, denn aus Symmetriegründen wird durch P jeder der Basisvektoren $(0, \dots, 0, 1, 0, \dots)$ auf den gleichen Vektor $(\rho, \rho, \dots, \rho)$ abgebildet, und aus

$$(1, 1, \dots, 1) = (1, 1, \dots, 1) P = n (\rho, \dots, \rho)$$

folgt $\rho = 1/n$.

A bildet den zur Diagonale orthogonalen Vektor $y - yP$ auf einen Vektor ab, der orthogonal zu $(1, 0, \dots, 0)$ ist, also als erste Koordinate 0 hat. Daher hat $z = yA = yPA + (y - yP)A$ dieselbe erste Koordinate z_1 wie yPA , und dies ist $\sqrt{n} \bar{y}$. Damit ist $Z_1 = \sqrt{n} \bar{Y}$.

Bezeichnet $\|\cdot\|$ den euklidischen Abstand im $\mathbb{R}^n$, so ist für jedes ω $\|Y(\omega)\| = \|Y(\omega)A\| = \|Z(\omega)\|$. Damit erhalten wir

$$\begin{aligned} (n-1) s^2(Y) &= \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i^2 - 2\bar{Y} Y_i + \bar{Y}^2) \\ &= \sum_{i=1}^n Y_i^2 - n\bar{Y}^2 = \|Z\|^2 - Z_1^2 = Z_2^2 + \dots + Z_n^2. \end{aligned}$$

Da Z_1 von $Z_2^2 + \dots + Z_n^2$ unabhängig ist, haben wir gezeigt, dass $\bar{Y}$ von $s^2(Y)$ unabhängig ist, was wohl einigermaßen überraschend ist, weil $\bar{Y}$ in der Formel für $s^2(Y)$ vorkommt. (Wir haben entscheidend Eigenschaften der Normalverteilung ausgenutzt. Es lässt sich zeigen, dass für nichtkonstante unabhängige Y_i mit *anderen* Verteilungen $\bar{Y}$ und $s^2(Y)$ stets abhängig sind und die Verteilung von Y nicht rotationsinvariant ist.)

In (14.7) sind also auf der rechten Seite Zähler und Nenner unabhängig. Wir benötigen nun noch die Verteilung von $(n-1) s^2(Y)$, also die von $Z_2^2 + \dots + Z_n^2$.

Definition 14.1 Die Verteilung der Summe der Quadrate von k unabhängigen $N(0, 1)$ -verteilten Zufallsvariablen nennen wir eine χ_k^2 -Verteilung oder ausführlicher eine χ^2 -Verteilung mit k Freiheitsgraden.

Die Berechnung der Dichte g_k dieser Verteilung findet der Leser im Anhang zu diesem Paragraphen. Es ist

$$g_k(x) = \frac{1}{2^{k/2} \Gamma(k/2)} x^{(k/2)-1} e^{-x/2} \quad (x > 0) \quad (14.8)$$

und natürlich $g_k(x) = 0$ für $x \leq 0$. In Bild 14.1 ist der Verlauf der Dichten g_k für einige Werte k angegeben. Nach Definition 14.1 hat $Z_2^2 + \dots + Z_n^2$ eine χ_{n-1}^2 -Verteilung.

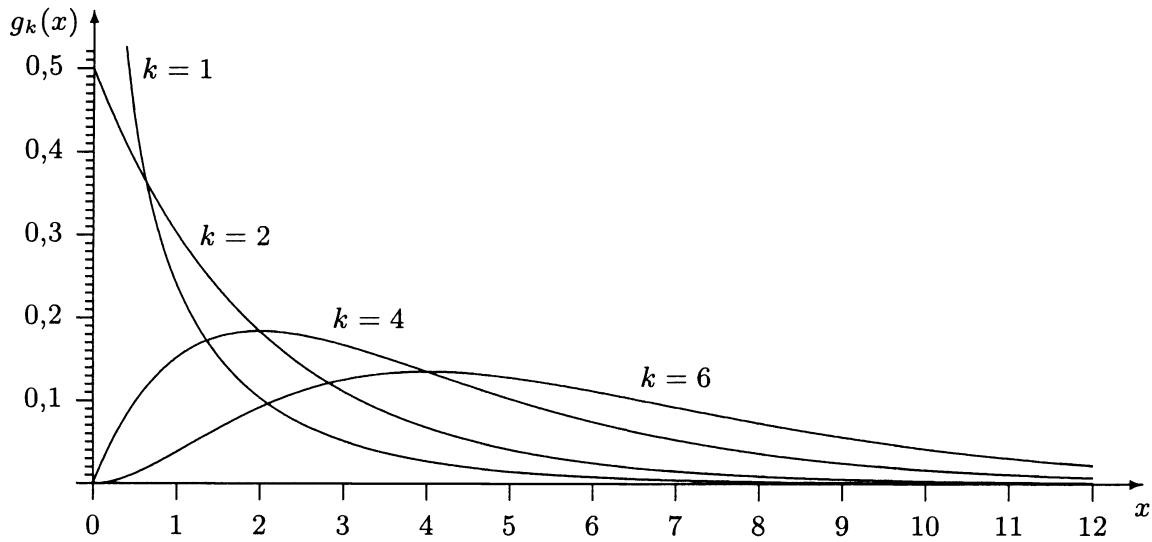


Bild 14.1 Dichte $g_k(x)$ der χ_k^2 -Verteilung für $k = 1, 2, 4, 6$

Definition 14.2 Sind W und U_k unabhängige Zufallsvariable, und ist W $N(0, 1)$ -verteilt und U_k χ_k^2 -verteilt, so nennt man die Verteilung von

$$T_k = \frac{W}{\sqrt{U_k/k}} \quad (14.9)$$

eine t_k -Verteilung oder ausführlicher eine **t-Verteilung** mit k Freiheitsgraden.

Unsere oben durchgeführten Überlegungen führen nun unmittelbar zu

Satz 14.3 Sind $X_1, \dots, X_n$ unabhängige $N(\mu_0, \sigma^2)$ -verteilte Zufallsvariable und ist $T(X)$ durch (14.5) mit $x_i = X_i(\omega)$ definiert, so ist $T(X)$ t_{n-1} -verteilt.

Beweis: Es gilt (14.7). Setzt man also $W = \sqrt{n} \bar{Y} = Z_1$ und $U_{n-1} = Z_2^2 + \dots + Z_n^2$, so ist $T(X) = T_{n-1}$. $\square$

Die Berechnung der Dichte h_k von T_k führen wir wieder im Anhang zu diesem Paragraphen durch. Es ist

$$h_k(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right) \Gamma\left(\frac{1}{2}\right) \sqrt{k}} \left(1 + \frac{x^2}{k}\right)^{-(k+1)/2}. \quad (14.10)$$

Für $k \rightarrow \infty$ gilt $h_k(x) \rightarrow \varphi(x)$, was nicht verwunderlich ist, da für $n \rightarrow \infty$ $s^2(X) \rightarrow \sigma^2$ gilt und $\sqrt{n}(\bar{X} - \mu_0)/\sigma$ $N(0, 1)$ -verteilt ist. $h_k(x)$ ist in der Nähe vom Nullpunkt $x = 0$ etwas kleiner als $\varphi(x)$ und für große $|x|$ etwas größer.

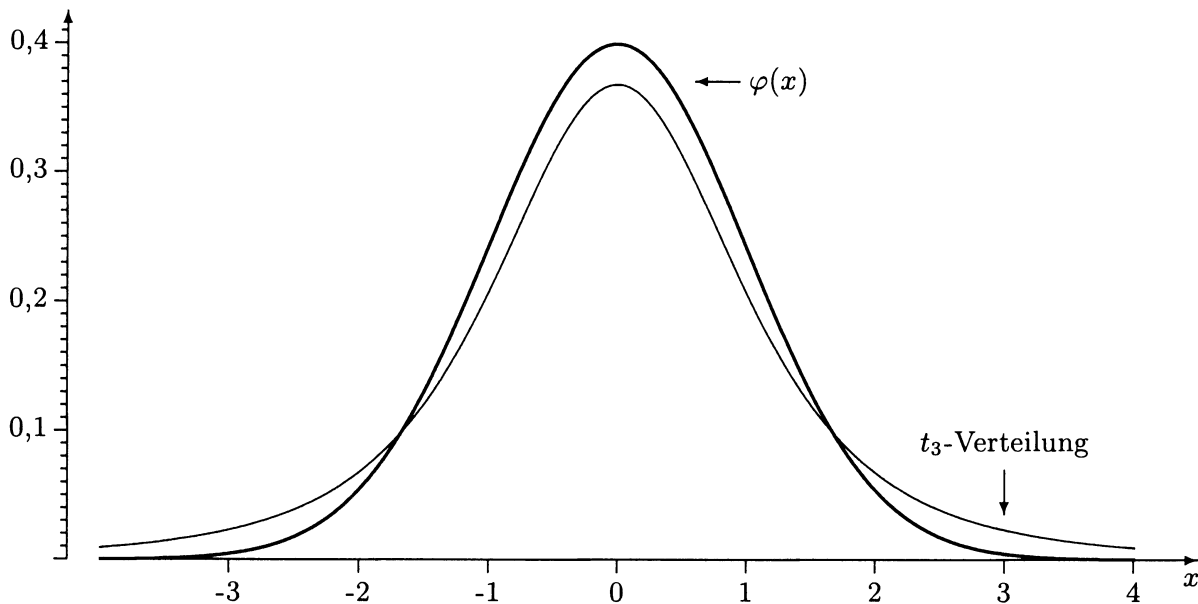


Bild 14.2 Vergleich der Dichte h_3 der t_3 -Verteilung mit der Dichte φ der $N(0, 1)$ -Verteilung

Die t_1 -Verteilung heißt auch **Standard-Cauchy-Verteilung**. Sie hat die Dichte $h_1(x) = 1/(\pi(1 + x^2))$. Die allgemeine t_k -Verteilung stammt von dem britischen Statistiker W. S. Gosset, der unter dem Pseudonym „Student“ publizierte. Die t -Verteilung heißt daher auch **studentsche Verteilung**.

Anwendung auf das Testproblem: Kehren wir zu dem am Anfang dieses Abschnitts beschriebenen Testproblem zurück. Für den Test (14.6) ist die Verwerfungswahrscheinlichkeit unter der Hypothese $E(\varphi(X)) = P(|T(X)| > t)$.

Für t_k -verteiltes T nennt man die Zahl $t_{k,\beta}$ mit

$$P(T \leq t_{k,\beta}) = \beta$$

das (untere) β -Quantil der t_k -Verteilung. (Allgemein heißt für eine Verteilungsfunktion F die Zahl

$$F^{-1}(\beta) := \inf\{t : F(t) \geq \beta\}$$

das (untere) β -Quantil zu F .)

Um einen Test zum Niveau α zu erhalten, bestimmt man aus Tabellen der t_{n-1} -Verteilung die Zahl $t = t_{n-1, 1-\alpha/2}$. Wegen der Symmetrie der t_{n-1} -Verteilung um den Nullpunkt ist dann $P(|T(X)| > t) = \alpha$. Der Test (14.6) hat dann also gerade Niveau α . (Für großes k , etwa für $k > 25$, ist die t_k -Verteilung fast identisch mit der $N(0, 1)$ -Verteilung, so dass man dann die Tabelle für Φ verwendet.)

Man nennt den Test (14.6) einen **zweiseitigen t-Test**, da sowohl für besonders kleine wie für besonders große Werte von $T(x)$ die Hypothese verworfen wird. Daneben gibt es noch **einseitige t-Tests**. Man verwendet einen solchen Test, wenn man für

unabhängige $X_1, \dots, X_n$ die Hypothese, dass die X_i $N(\mu, \sigma^2)$ -verteilt sind mit $\mu \leq \mu_0$, gegen die Alternative testet, dass die X_i $N(\mu, \sigma^2)$ -verteilt sind mit $\mu > \mu_0$. Ist $\mu > \mu_0$, so wird $\bar{X} - \mu_0$ größere Werte annehmen als im Fall $\mu = \mu_0$. Man verwirft die Hypothese dann also nur für große $T(x)$, d.h. man wählt einen Test der Form

$$\varphi(x) = 1 \quad \text{für } T(x) > t \quad \text{und} \quad \varphi(x) = 0 \quad \text{für } T(x) \leq t.$$

Damit er das Niveau α hat, muss man $t = t_{n-1, 1-\alpha}$ wählen.

Wir haben uns nun noch nicht an das in § 6 propagierte Prinzip gehalten, auch die Macht des Tests φ zu betrachten. Dazu braucht man die Verteilung von $T(X)$ auch unter der Alternative, die sog. **nichtzentrale t-Verteilung**. Ihre Berechnung würde aber hier zu weit führen. Man findet sie in den ausführlichen Werken über Statistik, etwa bei Witting (1985), und natürlich gibt es auch dafür Tabellen.

Häufig stellt sich das Problem des qualitativen Vergleichs von zwei Methoden oder z.B. des Vergleichs von zwei Düngemitteln. Man hat dann zwei Reihen von Messungen $X_1, \dots, X_m$ (1. Methode) und $Y_1, \dots, Y_n$ (2. Methode), die alle unabhängig seien. Man nimmt dann häufig an, $X_1, \dots, X_m$ seien $N(\mu_1, \sigma_1^2)$ -verteilt und $Y_1, \dots, Y_n$ seien $N(\mu_2, \sigma_2^2)$ -verteilt. Getestet wird z.B. die Hypothese $\mu_1 = \mu_2$ gegen $\mu_1 \neq \mu_2$. Der beste Test hierfür hängt bei festem (σ_1, σ_2) stark vom Verhältnis $\sigma_1 : \sigma_2$ ab. Bei unbekanntem $\sigma_1 : \sigma_2$ führt dies zu Schwierigkeiten, dem „**Behrens-Fisher-Problem**“ (siehe z.B. Pfanzagl (1974)). Im einfachsten Fall $\sigma_1 = \sigma_2$ kann man ähnlich wie oben vorgehen.

Als Schätzung der Varianz verwendet man

$$s^2 := s^2(X, Y) = \frac{1}{m+n-2} \left(\sum_{i=1}^m (X_i - \bar{X})^2 + \sum_{j=1}^n (Y_j - \bar{Y})^2 \right).$$

Unter der Hypothese $\mu_1 = \mu_2$ hat die Teststatistik

$$T(X, Y) = \frac{\bar{X} - \bar{Y}}{s \sqrt{\frac{1}{m} + \frac{1}{n}}}$$

eine t_{m+n-2} -Verteilung. Die Hypothese wird verworfen, wenn

$$|T(X, Y)| > t_{m+n-2, 1-\alpha/2}$$

gilt.

Wenn wie oben die X_i und die Y_j in voneinander getrennten Versuchen ermittelt werden, spricht man von einem **Zweistichprobenproblem mit unverbundenen Stichproben**. Manchmal ist es möglich und zweckmäßig, die X_i und Y_i als paarweise Vergleiche zu beobachten, also z.B. (X_1, Y_1) an Versuchsperson 1, (X_2, Y_2) an Versuchsperson 2 usw. (Bekannt ist das Beispiel des Testens der Qualität von Schuhsohlen. Jede Versuchsperson trägt simultan eine Schuhsohle der Sorte 1 an einem Fuß und der Sorte 2 am anderen.) Dann spricht man von **verbundenen Stichproben**. In diesem Fall bildet man die Differenzen $D_i = X_i - Y_i$ und reduziert das Problem damit auf ein Einstichprobenproblem.

14.2 Einfache Varianzanalyse*

Will man mehrere Behandlungsmethoden oder Düngemittel miteinander vergleichen und prüfen, ob ein nennenswerter Unterschied zwischen ihnen besteht, so wendet man jede

der Methoden mehrmals an und misst die Größe, an der man interessiert ist, etwa den Ertrag an Weizen. Seien $x_{i,1}, x_{i,2}, \dots, x_{i,n_i}$ die Messergebnisse mit der i -ten Methode.

Nehmen wir an, dies seien Realisierungen von unabhängigen Zufallsvariablen $X_{i,j}$ ($i = 1, \dots, r$; $j = 1, \dots, n_i$), und $X_{i,j}$ sei $N(\mu_i, \sigma^2)$ -verteilt mit unbekanntem, aber für alle i gleichen σ^2 . Dann wäre die Hypothese $\mu_1 = \mu_2 = \dots = \mu_r$ gegen die Alternative, dass mindestens zwei der μ_i verschieden sind, zu testen.

Mit dem obigen Zweistichproben- t -Test kann man die Hypothese $\mu_1 = \mu_2$ testen. Man könnte nun daran denken, alle Hypothesen $\mu_i = \mu_j$ ($i \neq j$) nacheinander zu testen. Das ist aber nicht sinnvoll: Wenn man für festes $i \neq j$ die Hypothese $\mu_i = \mu_j$ zum Niveau α testet, und sie trifft zu, macht man mit Wahrscheinlichkeit α einen Fehler. Macht man nun mehrere solche Vergleiche, so ist die Wahrscheinlichkeit, wenigstens bei einem dieser Vergleiche einen Fehler zu machen, relativ groß. Würde man, um dies auszugleichen, α sehr klein wählen, so würde wiederum ein tatsächlich vorhandener Unterschied zwischen μ_1 und μ_2 zu selten aufgedeckt werden.

Wir suchen daher nach einem Test, der simultan auf die möglichen Unterschiede anspricht. Hier soll nur die Grundidee diskutiert werden. Im Bild 14.2 sind die Messergebnisse $x_{1,1}, x_{1,2}, \dots, x_{1,n_1}$ der „ersten Klasse“ als Kreuze an den Stellen $(j, x_{1,j})$ eingetragen, dann — durch einen senkrechten Strich getrennt — Kreuze für die Messwerte $x_{2,1}, \dots, x_{2,n_2}$ der zweiten Klasse, usw. Es gibt hier $r = 3$ Klassen.

Die Streuung in jeder der Klassen ist von der Größenordnung σ . Sind nun zwei oder mehr der μ_i verschieden, so wird dadurch die aus *allen* $x_{i,j}$ berechnete Streuung zusätzlich vergrößert. Man wird daher die Hypothese verwerfen, wenn die Gesamtstreuung verglichen mit der Schätzung für σ , die man durch Vergleich der $x_{i,j}$ mit den Mittelwerten der zugehörigen Klassen erhält, groß ist.

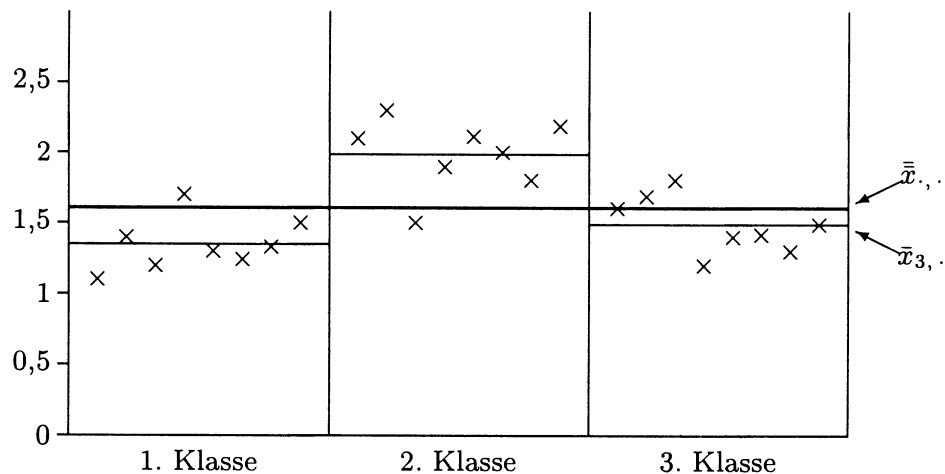


Bild 14.3 Mittelwerte der Klassen und Gesamtmittelwert

Sei

$$N = \sum_{i=1}^r n_i \quad \text{die Gesamtzahl der Beobachtungen,}$$

$$\bar{X}_{i,\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{i,j} \quad \text{der Mittelwert der } i\text{-ten Klasse}$$

und

$$\bar{\bar{X}}_{.,.} = \frac{1}{N} \sum_{i,j} X_{i,j} \quad \text{der Gesamtmittelwert.}$$

(Der Punkt steht für einen Index, über den man summiert. Der Querstrich deutet Mittelung an.) Es ist

$$\begin{aligned} Q &:= \sum_{i,j} (X_{i,j} - \bar{\bar{X}}_{.,.})^2 = \sum_{i,j} (X_{i,j} - \bar{X}_{i,.} + \bar{X}_{i,.} - \bar{\bar{X}}_{.,.})^2 \\ &= \sum_{i,j} (X_{i,j} - \bar{X}_{i,.})^2 + 2 \sum_{i,j} (X_{i,j} - \bar{X}_{i,.}) (\bar{X}_{i,.} - \bar{\bar{X}}_{.,.}) + \sum_{i,j} (\bar{X}_{i,.} - \bar{\bar{X}}_{.,.})^2 \\ &= \sum_{i,j} (X_{i,j} - \bar{X}_{i,.})^2 + \sum_i n_i (\bar{X}_{i,.} - \bar{\bar{X}}_{.,.})^2 =: Q_1 + Q_2. \end{aligned}$$

Man nennt dies die Zerlegung der „**Gesamtvarianz**“ Q in Q_1 , die „**Varianz innerhalb der Klassen**“, und Q_2 , die „**Varianz zwischen den Klassen**“. Wie bei der Herleitung der Verteilung von $T(X)$ im letzten Abschnitt zeigt man, dass $\sigma^{-2} Q_1$ eine χ^2_{N-r} -Verteilung hat, und dass Q_1 und Q_2 unabhängig sind. Unter der Hypothese $\mu_1 = \mu_2 = \dots = \mu_r$ hat ferner $\sigma^{-2} Q_2$ eine χ^2_{r-1} -Verteilung.

Ist U χ^2_m -verteilt und V χ^2_n -verteilt und sind U und V unabhängig, so nennt man die Verteilung von $(U/m)/(V/n)$ eine $F_{m,n}$ -Verteilung oder eine **F-Verteilung mit m und n Freiheitsgraden**. Die Dichte $g_{m,n}$ dieser Verteilung lässt sich auf genau die gleiche Weise ausrechnen, wie die der t -Verteilung.

Es ist $g_{m,n}(x) = 0$ für $x \leq 0$ und

$$g_{m,n}(x) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right) \Gamma\left(\frac{n}{2}\right)} m^{m/2} n^{n/2} \frac{x^{(m/2)-1}}{(n+mx)^{(m+n)/2}}$$

für $x > 0$.

Aus den oben angegebenen Aussagen über die Verteilungen von $\sigma^{-2} Q_1$ und $\sigma^{-2} Q_2$ und der Definition der $F_{m,n}$ -Verteilung folgt unmittelbar, dass die Teststatistik

$$F = \frac{Q_2/(r-1)}{Q_1/(N-r)}$$

unter der Hypothese eine $F_{r-1, N-r}$ -Verteilung hat. Sind die μ_i nicht alle gleich, so wird Q_2 groß sein. Man verwirft daher die Hypothese, wenn F das $(1-\alpha)$ -Quantil der $F_{r-1, N-r}$ -Verteilung übertrifft, das man Tabellen entnimmt. Da die Teststatistik eine F -Verteilung hat, spricht man von einem F -Test. Der Buchstabe F soll an R. A. Fisher erinnern, der die Varianzanalyse entwickelt hat. Die Verteilung von F unter der Alternative ist eine so genannte nichtzentrale F -Verteilung.

14.3 χ^2 -Tests

Angenommen, man vermutet für die Vererbung eines bestimmten Merkmals — etwa der Blütenfarbe — einen bestimmten Vererbungsmechanismus, bei dessen Vorliegen die drei möglichen Ausprägungen des Merkmals — etwa: rotblühend, rosablühend, weißblühend — mit den Wahrscheinlichkeiten $1/4$, $1/2$, $1/4$ auftreten. Unter $n = 320$ Nachkommen

würde man die Ausprägungen also etwa mit den Häufigkeiten 80, 160 und 80 erwarten. Tatsächlich werde in einem Experiment die erste Ausprägung $x_1 = 102$ mal, die zweite $x_2 = 156$ mal und die dritte $x_3 = 62$ mal beobachtet. Sprechen diese Zahlen gegen die vermutete Theorie oder sind sie als zufällige Abweichungen von den „erwarteten“ Zahlen 80, 160, 80 interpretierbar?

Dies ist ein Spezialfall des folgenden Problems: Es werden n unabhängige, untereinander gleiche Teilerperimente ausgeführt. Diese haben $r \geq 2$ mögliche Ausgänge und der i -te Ausgang hat Wahrscheinlichkeit p_i . Der Parameter $\vartheta = (p_1, \dots, p_r)$ ist unbekannt. Wir nehmen an, dass alle p_i positiv sind. Für einen gegebenen Wahrscheinlichkeitsvektor $\pi = (\pi_1, \dots, \pi_r)$ ist zu testen, ob $\vartheta = \pi$ ist.

Die Wahrscheinlichkeit, die Häufigkeiten $x_1, \dots, x_r$ zu beobachten, ist nach der Multinomialverteilung

$$P_{\vartheta}(x) = \binom{n}{x_1, \dots, x_r} p_1^{x_1} p_2^{x_2} \dots p_r^{x_r}.$$

Der Vektor $x = (x_1, \dots, x_r)$ der Anzahlen muss natürlich die Bedingung $x_1 + \dots + x_r = n$ erfüllen. Wir sind wieder im diskreten Fall. Die Likelihoodfunktion ist $L_x(\vartheta) = P_{\vartheta}(x)$. Bei der Ermittlung der Maximum-Likelihood-Schätzung $\hat{p}_i(x)$ der p_i ist die Nebenbedingung $p_1 + \dots + p_r = 1$ zu beachten. Die Methode der lagrangeschen Multiplikatoren führt zu dem Gleichungssystem

$$\frac{\partial \log L_x(\vartheta)}{\partial p_1} = \frac{\partial \log L_x(\vartheta)}{\partial p_2} = \dots = \frac{\partial \log L_x(\vartheta)}{\partial p_r},$$

und aus diesem ergibt sich leicht der Maximum-Likelihood-Schätzer $\hat{p}_i(x) = x_i/n$, den wir im Bernoulli-Fall ($r = 2$) ja schon kennen.

Es ist nun leicht, sich zu überlegen, wie der exakte Likelihood-Quotienten-Test der Hypothese $H = \{\pi\}$ gegen die Alternative $K = \{\vartheta : \vartheta \neq \pi\}$ aussieht; aber dieser ist wenig praktikabel. Die Verteilung des Likelihood-Quotienten hängt von zu vielen Parametern ab, um effektiv mit Tabellen arbeiten zu können.

Da Rückschlüsse auf das Vorliegen einer bestimmten Verteilung π nur möglich sind, wenn n hinreichend groß ist, arbeitet man mit Approximationen. Für die Multinomialverteilung gibt es eine Approximation, die die Normalapproximation der Binomialverteilung verallgemeinert.

Sei

$$\mathcal{X}(n) = \left\{ x = (x_1, \dots, x_r) \in \mathbb{Z}^r : x_i \geq 0, \sum_{i=1}^r x_i = n \right\}$$

die Menge der möglichen Ergebnisse bei n Teilversuchen. Wir betrachten Folgen $x(n) = (x_1(n), x_2(n), \dots, x_r(n)) \in \mathcal{X}(n)$, schreiben aber zur Abkürzung nur x_i statt $x_i(n)$.

Satz 14.4 *Ist $C < \infty$, so gilt gleichmäßig für alle Folgen $x(n)$ mit $|x_i - np_i| \leq C\sqrt{n}$*

$$P_{\vartheta}(x(n)) \sim \left(\frac{1}{\sqrt{2\pi n}} \right)^{r-1} \frac{1}{\sqrt{p_1 p_2 \dots p_r}} \exp \left(-\frac{1}{2} \sum_{i=1}^r \frac{(x_i - np_i)^2}{np_i} \right). \quad (14.11)$$

Der Beweis ist analog zu dem von Satz 5.2. (Vgl. van der Waerden (1971), Morgenstern (1968). Genauere Approximationen finden sich bei Dinges-Rost (1982).)

Für den Likelihood-Quotienten $q(x)$ ergibt sich mit $\hat{\vartheta} = (\hat{p}_1, \dots, \hat{p}_r)$ wegen $x_i - n\hat{p}_i = 0$ die Approximation

$$q(x) = \frac{P_{\hat{\vartheta}}(x)}{P_{\pi}(x)} \approx \left(\frac{\hat{p}_1}{\pi_1} \cdots \frac{\hat{p}_r}{\pi_r} \right)^{-1/2} \exp \left(\frac{1}{2} \sum_{i=1}^r \frac{(x_i - n\pi_i)^2}{n\pi_i} \right). \quad (14.12)$$

Sei

$$V^2(x) = \sum_{i=1}^r \frac{(x_i - n\pi_i)^2}{n\pi_i}.$$

Gilt die Hypothese $\vartheta = \pi$, so ist nach dem Gesetz der großen Zahlen für großes n mit Wahrscheinlichkeit nahe bei 1 $\hat{p}_i = x_i/n$ fast gleich π_i . Der erste Faktor auf der rechten Seite von (14.12) ist dann nahe bei 1. Wir können deshalb als Approximation des Likelihood-Quotienten-Tests den Test nehmen, der die Hypothese ablehnt, wenn $V^2(x)$ eine vom gewählten Niveau α abhängige Schranke c_α übertrifft.

Entscheidend ist nun die Entdeckung von Karl Pearson aus dem Jahr 1900, dass die Verteilung von V^2 unter der Hypothese für große n durch eine χ^2_{r-1} -Verteilung approximiert werden kann. Man benötigt daher nicht für verschiedene π und n verschiedene Tabellen, wenn man mit dieser Approximation statt mit der exakten Verteilung arbeitet. (Karl Pearson war der Vater des uns schon bekannten Egon Pearson.)

Satz 14.5 * Ist G_k die Verteilungsfunktion der χ^2_k -Verteilung, so gilt für $n \rightarrow \infty$

$$P_{\pi}(V^2 \leq u) \rightarrow G_{r-1}(u).$$

Beweis: Ist $x \in \mathcal{X}(n)$ und $y_i = (x_i - n\pi_i)/\sqrt{n\pi_i}$, so ist

$$\sum_{i=1}^n y_i \sqrt{n\pi_i} = \sum_{i=1}^n (x_i - n\pi_i) = n - n = 0.$$

Die zu den verschiedenen $x \in \mathcal{X}(n)$ gehörigen Vektoren $y = (y_1, \dots, y_r)$ liegen also alle auf der $(r-1)$ -dimensionalen Hyperebene durch den Koordinatenursprung 0, die orthogonal zu

$$b = (\sqrt{\pi_1}, \sqrt{\pi_2}, \dots, \sqrt{\pi_r})$$

ist. Ist $\mathcal{X}'(n)$ die Menge der $x \in \mathbb{Z}^r$ mit $\sum x_i = n$ (ohne die Nebenbedingung $x_i \geq 0$), so bilden die zugehörigen y auf der Hyperebene ein diskretes Gitter $D = D(n)$. Erhöht man in x , um zu einem benachbarten Gitterpunkt zu gelangen, die i -te Koordinate für ein $i \leq r-1$ um 1, so muss man eine andere — etwa die letzte — um 1 vermindern. Ist daher für $i \leq r-1$ $a^i = (a_1^i, \dots, a_r^i)$ der Vektor mit

$$a_i^i = \frac{1}{\sqrt{n\pi_i}}, \quad a_j^i = 0 \quad (i \neq j \neq r) \quad \text{und} \quad a_r^i = -\frac{1}{\sqrt{n\pi_r}},$$

und ist $y(0)$ z.B. das zu $x = (0, 0, \dots, 0, n)$ gehörige y , so ist

$$D = \left\{ y(0) + \sum_{i=1}^{r-1} k_i a^i : k_i \in \mathbb{Z} \right\}.$$

Da b die Länge 1 hat und auf dem Gitter senkrecht steht, ist die Fläche F eines Gitterelements das Volumen des von $a^1, a^2, \dots, a^{r-1}, b$ aufgespannten Parallelotops. Also ist F der Absolutbetrag der Determinante der Matrix M mit den Zeilenvektoren $a^1, \dots, a^{r-1}, b$. Addiert man die mit $\sqrt{\pi_i/\pi_r}$ multiplizierte i -te Spalte von M für $i \leq r-1$ zur letzten, so ergibt sich

$$\begin{aligned}
& \det \begin{pmatrix} \frac{1}{\sqrt{n \pi_1}} & 0 & \dots & 0 & -\frac{1}{\sqrt{n \pi_r}} \\ 0 & \frac{1}{\sqrt{n \pi_2}} & \dots & 0 & -\frac{1}{\sqrt{n \pi_r}} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & \dots & \frac{1}{\sqrt{n \pi_{r-1}}} & -\frac{1}{\sqrt{n \pi_r}} \\ \sqrt{\pi_1} & \dots & \dots & \sqrt{\pi_{r-1}} & \sqrt{\pi_r} \end{pmatrix} \\
&= \det \begin{pmatrix} \frac{1}{\sqrt{n \pi_1}} & 0 & \dots & 0 & 0 \\ 0 & \frac{1}{\sqrt{n \pi_2}} & \dots & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & \dots & \frac{1}{\sqrt{n \pi_{r-1}}} & 0 \\ \sqrt{\pi_1} & \dots & \dots & \sqrt{\pi_{r-1}} & \eta \end{pmatrix}
\end{aligned}$$

mit $\eta = \sqrt{\pi_r} + \pi_1/\sqrt{\pi_r} + \dots + \pi_{r-1}/\sqrt{\pi_r} = 1/\sqrt{\pi_r}$. Es folgt

$$F = \frac{1}{n^{(r-1)/2} \sqrt{\pi_1 \pi_2 \dots \pi_r}}.$$

Ist $V^2(x) \leq u$, so ist $|x_i - n\pi_i| \leq \sqrt{n\pi_i u}$. Wegen $\pi_i > 0$ gilt dann $x_i \geq 0$ für alle hinreichend großen n , und für $p_i = \pi_i$ ist die in Satz 14.4 angegebene Approximation für $n \rightarrow \infty$ gleichmäßig in $\{x : V^2(x) \leq u\}$ anwendbar.

$V^2(x)$ ist das Quadrat $\|y\|^2$ der euklidischen Länge von y . Die rechte Seite von (14.11) hat daher die Form

$$\frac{F}{(2\pi)^{(r-1)/2}} \exp\left(-\frac{\|y\|^2}{2}\right),$$

und für $n \rightarrow \infty$ gilt

$$P_\pi(V^2 \leq u) \sim \sum_{\{y \in D : \|y\|^2 \leq u\}} \frac{F}{(2\pi)^{(r-1)/2}} \exp\left(-\frac{\|y\|^2}{2}\right). \quad (14.13)$$

Nun rotieren wir das Koordinatensystem noch so, dass b die Koordinaten $(0, 0, \dots, 0, 1)$ erhält. Sind $(z_1, \dots, z_r)$ die neuen Koordinaten von y , so ist (wegen der Orthogonalität von y und b) $z_r = 0$. Aus der rechten Seite von (14.13) wird wegen $\|y\|^2 = \|z\|^2$

$$\sum_{\{z : z_r=0, \|z\|^2 \leq u\}} \frac{F}{(2\pi)^{(r-1)/2}} \exp\left(-\frac{z_1^2 + \dots + z_{r-1}^2}{2}\right). \quad (14.14)$$

F ist auch im rotierten Koordinatensystem die Fläche eines Gitterelements. Daher ist (14.14) eine Riemann-Summe, die das Integral

$$\int \dots \int_{Q(u)} \varphi(z_1) \varphi(z_2) \dots \varphi(z_{r-1}) dz_1 \dots dz_{r-1}$$

über den Bereich

$$Q(u) = \left\{ (z_1, \dots, z_{r-1}) \in \mathbb{R}^{r-1} : \sum_{i=1}^{r-1} z_i^2 \leq u \right\}$$

approximiert. Das Integral ist aber gerade die Wahrscheinlichkeit $G_{r-1}(u)$, dass die Summe der Quadrate von $r-1$ unabhängigen $N(0, 1)$ -verteilten Zufallsvariablen $\leq u$ ist. $\square$

Zur Ausführung des χ^2 -Tests: Man bestimmt aus einer Tabelle der χ_{r-1}^2 -Verteilung das $(1-\alpha)$ -Quantil, also den Wert $\chi_{r-1, 1-\alpha}^2$ mit $G_{r-1}(\chi_{r-1, 1-\alpha}^2) = 1-\alpha$. Die Hypothese wird verworfen wenn $V^2(x) > \chi_{r-1, 1-\alpha}^2$ ist.

Van der Waerden (1971) gibt als Faustregel an, dass die Approximationen hinreichend gut sind, wenn für alle i $n\pi_i \geq 2$ (oder ≥ 3) ist. Ist r relativ groß, etwa $r \geq 10$, so soll auch bei $n\pi_i \geq 1$ die Approximation schon sehr gut sein.

Für die praktische Berechnung benutzt man meist die durch Umformung der Definitionsgleichung herleitbare Formel

$$V^2(x) = \frac{1}{n} \sum_{i=1}^r \frac{x_i^2}{\pi_i} - n.$$

Für $\vartheta \neq \pi$ hat V^2 unter P_ϑ asymptotisch eine „nichtzentrale“ χ_{r-1}^2 -Verteilung.

Test auf Unabhängigkeit*: Die χ^2 -Methode lässt sich auf den Fall verallgemeinern, in dem die p_i in funktional bekannter Form und hinreichend glatt von einem unbekannten Parameter abhängen, der mehrdimensional sein darf. Ein wichtiger Spezialfall davon führt zum Test auf Unabhängigkeit mit Hilfe von **Kontingenztafeln**. Dabei werden an n voneinander unabhängigen Objekten oder Personen je zwei Merkmale beobachtet, z.B. Konfession und Abiturnote in Mathematik. Das erste Merkmal trete in r , das zweite in s „Ausprägungen“ auf. Z.B. wären {evangelisch, katholisch, jüdisch, sonstige} vier Ausprägungen. Wir bezeichnen die Ausprägungen des ersten Merkmals mit $1, \dots, r$ die des zweiten mit $1, \dots, s$. Die aus den Beobachtungen resultierende Kontingenztafel ist die $r \times s$ -Matrix der Anzahlen $n_{j,k}$ der Objekte, für die das erste Merkmal die Ausprägung j , das zweite die Ausprägung k hat.

Für jedes Objekt sei die Wahrscheinlichkeit ρ_j , dass es das erste Merkmal in der j -ten Ausprägung hat. Entsprechend sei τ_k die Wahrscheinlichkeit, mit der das zweite Merkmal die Ausprägung k hat. Ist $\vartheta_{j,k}$ die Wahrscheinlichkeit der Merkmalskombination (j, k) , so lässt sich die Hypothese der Unabhängigkeit der Merkmale durch die Bedingung beschreiben, dass $\vartheta_{j,k}$ von der Form $\rho_j \tau_k$ ist. Die $n_{j,k}$ spielen demnach die Rolle, die in den obigen Überlegungen zum Test der Hypothese $\vartheta = \pi$ die x_i gespielt haben. Allerdings müssen die unbekannten ρ_j und τ_k nun noch geschätzt werden, während π_i oben bekannt war.

Sei

$$n_{j, \cdot} = \sum_{k=1}^s n_{j,k} \quad \text{and} \quad n_{\cdot, k} = \sum_{j=1}^r n_{j,k}.$$

$\hat{\rho}_j = n_{j,\cdot}/n$ und $\hat{\tau}_k = n_{\cdot,k}/n$ sind die uns bereits bekannten Maximum-Likelihood-Schätzer für ρ_j und τ_k . Ersetzt man nun in der Definitionsgleichung von $V^2(x)$ die x_i durch $n_{j,k}$ und π_i durch $\hat{\rho}_j \hat{\tau}_k$, so kommt man auf die Teststatistik

$$\tilde{V}^2 = \sum_{j=1}^r \sum_{k=1}^s \frac{(n_{j,k} - n_{j,\cdot} n_{\cdot,k}/n)^2}{n_{j,\cdot} n_{\cdot,k}/n}.$$

Die Zahl der beobachteten Werte ist jetzt rs statt r . Man könnte also vermuten, dass $\tilde{V}^2$ unter der Hypothese für $n \rightarrow \infty$ asymptotisch eine χ^2_{rs-1} -Verteilung haben müsste, wenn die $n_{j,k}$ als Zufallsvariable betrachtet werden. Das wäre richtig, wenn die ρ_j und τ_k bekannt wären, und wir $\rho_j \tau_k$ statt $\hat{\rho}_j \hat{\tau}_k$ eingesetzt hätten. Die genaue Herleitung der Grenzverteilung zeigt jedoch, dass man mit jedem geschätzten frei variierenden Parameter einen Freiheitsgrad verliert, in diesem Fall $r-1 + s-1$. Für $n \rightarrow \infty$ hat daher $\tilde{V}^2$ eine $\chi^2_{(r-1)(s-1)}$ -Verteilung. (Im Prinzip kennen wir diese Verringerung der Zahl der Freiheitsgrade bereits aus einem anderen Beispiel: Sind $X_1, \dots, X_n$ unabhängig $N(\mu, 1)$ -verteilt, so ist $\sum (X_i - \mu)^2$ χ^2_n -verteilt, aber $\sum (X_i - \bar{X})^2$ ist χ^2_{n-1} -verteilt.)

Der χ^2 -Test für Kontingenztafeln verwirft also die Hypothese der Unabhängigkeit der Merkmale, wenn $\tilde{V}^2 > \chi^2_{(r-1)(s-1), 1-\alpha}$ ist.

Für $r = s = 2$ und kleine n wird statt des χ^2 -Tests der **exakte Test von Fisher** verwendet, der nur die hypergeometrische Verteilung benutzt (s. Aufgaben 7 und 8).

[Bei den ASW-Experimenten hatten wir einen anderen Test auf Unabhängigkeit empfohlen. Das lag daran, dass wir dort die Macht des Tests für spezielle Alternativen groß machen wollten.]

Die mathematische Theorie der χ^2 -Tests ist bei C. R. Rao (1973) gut dargestellt.

14.4 Nichtparametrische Tests

Bei den meisten hier besprochenen Tests sind wir von einer Normalverteilung ausgegangen. Was macht man nun, wenn diese Voraussetzung verletzt ist oder man die Form der Verteilung nicht hinreichend gut kennt? Hierfür sind Tests entwickelt worden, die nur noch die gegenseitigen Größenbeziehungen der beobachteten Werte berücksichtigen und dadurch von *speziellen* Verteilungsannahmen unabhängig sind. Man nennt sie nichtparametrische Tests.

Oft ist eine numerische Bewertung von Beobachtungen problematisch, während man größer und kleiner noch unterscheiden kann. Man denke etwa an Intelligenzleistungen. Dann liegt es nahe, nichtparametrische Tests zu verwenden.

Nichtparametrische Verfahren sind in der Praxis von großer Wichtigkeit. Wir können aber in diesem Rahmen nur einige Grundbegriffe erläutern und Beispiele geben.

Der einfachste nichtparametrische Test ist wohl der **Zeichentest**. Er wird verwendet, um zu testen, ob der Median einer Verteilung einen bestimmten Wert annimmt. Sind $D_1, \dots, D_n$ unabhängige identisch verteilte Zufallsvariable und ist μ ein Median der Verteilung der D_i und $P(D_i = \mu) = 0$, so gilt

$$P(D_i > \mu) = \frac{1}{2} = P(D_i < \mu).$$

Ist dann S_n die Anzahl der $i \leq n$ mit $D_i > \mu$, so ist S_n $b_{n,p}$ -verteilt mit $p = 1/2$. Der Zeichentest verwendet S_n als Teststatistik für das Testen der Hypothese, dass μ Median der Verteilung der D_i ist.

Eine typische Anwendung des Zeichentests ist das Testen verbundener Stichproben. Werden etwa an n Testpersonen zwei blutdrucksenkende Mittel A und B getestet, und ist X_i ein Maß für den Erfolg des Mittels A an Person i und Y_i die entsprechende Größe für Mittel B , so kann man den Zeichentest auf $D_i = X_i - Y_i$ anwenden. Sind beide Mittel gleich gut, so sollte der Median μ von D_i gleich 0 sein. Hierfür hatten wir in § 14.1 auch den t -Test in Betracht gezogen, aber dessen Anwendbarkeit hängt davon ab, ob die D_i normalverteilt sind.

Viele nichtparametrische Verfahren stützen sich auf die sog. Ränge von Zufallsvariablen. Diesen Begriff wollen wir als Nächstes studieren.

Die Angabe eines Datenvektors $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ lässt sich in zwei Teilinformationen aufspalten, indem man zuerst mitteilt, welche Werte wie oft in x vorkommen und indem man zweitens deren Anordnung in x spezifiziert. Die erste Angabe haben wir bereits in § 13.5 gemacht: $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ war die zu x gehörige Ordnungsstatistik. Die $x_{(i)}$ sind die Koordinaten von x in aufsteigender Reihenfolge, wobei jeder Wert in der aufsteigenden Folge $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ so oft vorkommt wie in x .

Die Anordnung beschreibt man mittels der **Ränge**. Das geht am einfachsten, wenn alle Werte x_i verschieden sind. Dem kleinsten x_i gibt man dann den Rang 1, dem nächstgrößeren den Rang 2, usw. Formal definiert man den Rang r_i von x_i in x durch

$$r_i = \text{card}\{j \leq n : x_j \leq x_i\}. \quad (14.15)$$

Ist etwa

$$x_1 = 2,5, \quad x_2 = 7,8, \quad x_3 = 1,2, \quad x_4 = 1,9,$$

so ist

$$x_{(1)} = 1,2, \quad x_{(2)} = 1,9, \quad x_{(3)} = 2,5, \quad x_{(4)} = 7,8$$

und

$$r_1 = 3, \quad r_2 = 4, \quad r_3 = 1, \quad r_4 = 2.$$

x_1 ist nämlich der drittkleinste der Werte, usw. Offenbar gilt

$$x_i = x_{(r_i)} \quad (i = 1, \dots, n) \quad (14.16)$$

und auch diese Identität können wir zur Definition der Ränge verwenden.

Sind nicht alle x_i verschieden, so spricht man von **Bindungen** (engl.: ties). Ergänzt man etwa den obigen Vektor $(x_1, \dots, x_4)$ durch $x_5 = 2,5$, so gibt es eine Bindung zwischen x_1 und x_5 . Dann ist die neue Ordnungsstatistik

$$x_{(1)} = 1,2, \quad x_{(2)} = 1,9, \quad x_{(3)} = 2,5, \quad x_{(4)} = 2,5, \quad x_{(5)} = 7,8.$$

(14.16) ergibt keine eindeutige Definition von r_i mehr. Diese Gleichung ist für $i = 1$ sowohl mit $r_1 = 3$ als auch mit $r_1 = 4$ erfüllt. Es ist also „strittig“, ob x_1 den Rang 3 oder 4 erhalten soll, und auch für x_5 sind diese beiden Ränge mit (14.16) verträglich. Es gibt nun mehrere verschiedene Vorschläge, mit diesem Problem umzugehen:

- (a) Mittels (14.15) gibt man jedem x_i den höchsten möglichen Rang, so wie man manchmal mehrere Goldmedaillen vergibt. (Manchmal entscheidet man sich auch für den niedrigsten möglichen Rang.)
- (b) Man vergibt den Mittelwert der strittigen Ränge an alle potenziellen Kandidaten für diese Ränge.

(c) Man verteilt die strittigen Ränge *zufällig* unter die potenziellen Kandidaten.

Jede dieser Methoden hat, wie man sich leicht überlegt, die Eigenschaft, dass aus der Ordnungsstatistik und den Rängen der Vektor $(x_1, \dots, x_n)$ rekonstruierbar ist.

Der Einfachheit halber wollen wir hier nur den bindungsfreien Fall betrachten. Sind $X_1, \dots, X_n$ Zufallsvariable, so ist für jedes $\omega \in \Omega$ $X_1(\omega), \dots, X_n(\omega)$ eine endliche Folge von reellen Zahlen. $R_i(\omega)$ sei dann der Rang von $X_i(\omega)$ in dieser Folge. Die Abhängigkeit von ω wird gewöhnlich nicht extra erwähnt. Man sagt, dass R_i der Rang von X_i in $(X_1, \dots, X_n)$ ist. Diese Ränge spielen eine wichtige Rolle bei vielen Tests. Wir betrachten als Beispiel den **Wilcoxon-Zweistichprobentest**, der auch **Rangsummentest** genannt wird.

Beobachtet werden unabhängige Zufallsvariable $X_1, \dots, X_m, Y_1, \dots, Y_n$. Alle X_i sollen die gleiche Verteilungsfunktion $F(x) = P(X_i \leq x)$ haben, alle Y_j die gleiche Verteilungsfunktion $G(x) = P(Y_j \leq x)$. Z.B. können die Werte X_i die Daten für den Benzinverbrauch pro km bei m Fahrten mit Benzinsorte A sein und die Y_j die Daten für n vergleichbare Fahrten mit Sorte B . Ein anderes Beispiel ist die Beobachtung von Ernteerträgen bei Verwendung verschiedener Düngemittel. Wir nehmen an, dass F und G stetig sind, also für jedes $x \in \mathbb{R}$ die Bedingung $P(X_i = x) = P(Y_j = x) = 0$ für alle i, j gilt. Wir wollen uns kurz überlegen, dass dies wegen der geforderten Unabhängigkeit zur Folge hat, dass Bindungen nur mit Wahrscheinlichkeit 0 auftreten:

Da F und G stetig sind, gibt es zu jedem $\varepsilon > 0$ eine Zerlegung von $\mathbb{R}$ in disjunkte Teilintervalle $I_1, \dots, I_K$ mit der Eigenschaft $P(X_1 \in I_k) < \varepsilon$ und $P(Y_1 \in I_k) < \varepsilon$ für alle $k = 1, \dots, K$. Dann ist

$$\begin{aligned} P(X_1 = Y_1) &\leq \sum_{k=1}^K P(X_1 \in I_k, Y_1 \in I_k) \\ &= \sum_{k=1}^K P(X_1 \in I_k) P(Y_1 \in I_k) \leq \varepsilon \sum_{k=1}^K P(X_1 \in I_k) = \varepsilon. \end{aligned}$$

Da $\varepsilon > 0$ beliebig klein gewählt werden kann, folgt $P(X_1 = Y_1) = 0$. Genauso zeigt man, dass bei den übrigen Paaren von Zufallsvariablen Bindungen nur mit Wahrscheinlichkeit 0 auftreten.

Es sei die Hypothese zu testen, dass $F = G$ gilt. Die Alternative sei $F \neq G$, aber wir nehmen an, dass wir besonders daran interessiert seien, die Hypothese zu verwerfen, wenn die Y_j eher größere Werte annehmen, als die X_i . (Z.B. könnte Sorte B minimal billiger sein als Sorte A , so dass man sie im Fall $F = G$ vorziehen würde, aber natürlich auch, wenn der Verbrauch für Sorte B geringer ist. Das Verwerfen der Hypothese würde einen Wechsel zu Sorte A bedeuten.)

Das bedeutet: Wir versuchen nicht die Macht des Tests für alle $F \neq G$ groß zu machen, sondern nur für solche $F \neq G$ für die $P(Y_j > t) \geq P(X_i > t)$ für alle t gilt. Oft schränkt man die Alternative noch weiter ein.

Die kombinierte, aus allen Beobachtungen bestehende Stichprobe wird durch den Zufallsvektor $(Z_1, \dots, Z_{n+m})$ mit

$$Z_i = X_i \quad \text{für } 1 \leq i \leq m \quad \text{und} \quad Z_{m+j} = Y_j \quad \text{für } 1 \leq j \leq n$$

beschrieben. Wenn die Hypothese $F = G$ gilt, ist die Folge $Z_1, Z_2, \dots, Z_N$ mit $N = m + n$ eine Folge unabhängiger identisch verteilter Zufallsvariablen. Jede Permutation dieser Zufallsvariablen hat dann die gleiche Verteilung. Dies bedeutet, dass dann für jede Permutation π von $\{1, \dots, N\}$ das Ereignis

$$A_\pi = \{\omega \in \Omega : Z_{\pi(1)}(\omega) < Z_{\pi(2)}(\omega) < \dots < Z_{\pi(N)}(\omega)\}$$

die gleiche Wahrscheinlichkeit hat. Da Bindungen nur mit Wahrscheinlichkeit 0 auftreten, ist die Summe aller $P(A_\pi)$ gleich 1. Gilt $F = G$, so ist also

$$P(A_\pi) = \frac{1}{N!} \quad (14.17)$$

für alle π .

Ist $F \neq G$ und sind die Y_j tendenziell eher größer als die X_i , so werden die kleineren der Z -Werte eher von den X -Beobachtungen herrühren, die größeren eher von den Y -Beobachtungen. Die Ränge der X -Beobachtungen in der kombinierten Stichprobe werden dann eher kleiner sein. F. Wilcoxon hat 1945 vorgeschlagen, die Summe W der Ränge aller X -Beobachtungen in der kombinierten Stichprobe als Teststatistik zu verwenden. Trägt man etwa in Bild 14.4 die beobachteten Werte x_i der X_i als Kreuze auf der Koordinatenachse ein und die beobachteten Werte y_j der Y_j als dicke Punkte, so ist der beobachtete Wert w von W die Zahl 15, denn den x_i kommen die Ränge 1, 3, 5, 6 zu. (Es spielt keine Rolle welches x_i zu welchem Kreuz gehört.)

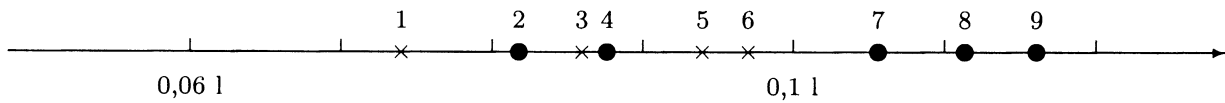


Bild 14.4 Ränge in zwei Stichproben

Da kleine Werte von W gegen die Hypothese sprechen, wird man bei gegebenem Niveau $\alpha > 0$ eine Verwerfungsschranke $c = c(\alpha, m, n)$ so bestimmen, dass c die größte Zahl mit $P(W \leq c) \leq \alpha$ ist, und die Hypothese im Fall $W \leq c$ verwerfen. Die Werte von $c(\alpha, m, n)$ findet man für nicht zu große Werte von m und n in Tabellen, aber wir wollen uns trotzdem überlegen, wie man sie bestimmen kann.

Dazu ordnen wir den Beobachtungsergebnissen Folgen von m Symbolen x und n Symbolen y so zu, dass die x -Symbole gerade an den Stellen stehen, die den Rängen der X_i entsprechen. Der Beobachtung in Bild 14.4 entspricht also

$$xyxyxyxyy.$$

Jeweils $m!n!$ Permutationen π ergeben die gleiche x - y -Folge. Also hat unter der Hypothese jede x - y -Folge die Wahrscheinlichkeit $m!n!/N! = \binom{N}{m}^{-1}$.

Der kleinste mögliche Wert von W ist $m(m+1)/2$. Er ergibt sich, wenn $X_i < Y_j$ für alle i und j gilt, also nur für die Folge $xx \dots xy y \dots y$. Daher ist

$$P\left(W = \frac{m(m+1)}{2}\right) = \binom{N}{m}^{-1}.$$

Sei $k(w | m, n)$ die Anzahl der x - y -Folgen, für die die Summe der x -Ränge w ist, und in denen x genau m -mal und y genau n -mal vorkommt. Dann ist unter der Hypothese $F = G$

$$P(W = w) = k(w | m, n) \binom{N}{m}^{-1}.$$

Unterscheidet man die Fälle, in denen das letzte Symbol ein x bzw. ein y ist, so erhält man die Rekursionsgleichung

$$k(w | m, n) = k(w | m, n - 1) + k(w - n - m | m - 1, n).$$

Die Anfangsbedingungen für die Rekursion sind ziemlich offensichtlich. Wir wissen schon, dass

$$k(m(m+1)/2 | m, n) = 1 \quad \text{und} \quad k(w | m, n) = 0 \quad \text{für} \quad w < m(m+1)/2$$

gilt. Ferner gilt offenbar

$$k(w | m, n) = \begin{cases} 1 & \text{für } w = 0, m = 0, n \in \mathbb{N}, \\ 0 & \text{für } w \neq 0, m = 0, n \in \mathbb{N}, \\ 0 & \text{für } w \neq m(m+1)/2, m \in \mathbb{N} \text{ und } n = 0. \end{cases}$$

Die einzige Eigenschaft von F , die wir benutzt haben, war die Stetigkeit. Während also z.B. bei der Ableitung des „parametrischen“ t -Tests benötigt wurde, dass die Verteilungsfunktion auf ganz bestimmte Weise von den Parametern abhing, ist der Rangsummentest sehr allgemein anwendbar. Wir erwähnen noch, dass selbst die Stetigkeit von F nicht benötigt wird, wenn man strittige Ränge im Fall von Bindungen zufällig vergibt.

Sind m und n groß, so greifen Tabellen nicht mehr. Man benutzt dann, dass sich die Verteilung von W nach Normierung durch eine Standard-Normalverteilung approximieren lässt. Dazu brauchen wir den Erwartungswert und die Varianz von W . Sei U die Anzahl der Paare (i, j) mit $X_i > Y_j$. Dann ist, wenn wir Bindungen wie bisher ausschließen, $W = m(m+1)/2 + U$. Der Wert $W = m(m+1)/2$ ergibt sich nämlich gerade für den Fall, dass alle x_i kleiner als alle y_j sind. Und vertauschen wir in der zugeordneten x - y -Folge ein x mit einem direkt rechts neben ihm stehenden y , so erhöhen sich sowohl der Wert von U als auch der Wert von W um genau 1. (Die Teststatistik U ist 1946 unabhängig von Wilcoxons Arbeit von Mann-Whitney vorgeschlagen worden.) Sei $I_{i,j}$ die Indikatorfunktion des Ereignisses $\{X_i > Y_j\}$. Unter der Hypothese gilt $E(I_{i,j}) = 1/2$. U ist die Summe aller $I_{i,j}$. Daher ist $E(U) = mn/2$ und

$$E(W) = \frac{m}{2} (m + n + 1).$$

Mit Hilfe der gleichen Summendarstellung von U verifiziert man durch Berechnung der Kovarianzen leicht

$$\text{Var}(W) = \text{Var}(U) = \frac{mn}{12} (n + m + 1).$$

Es lässt sich zeigen, dass die Verteilung von

$$W^* = \frac{W - E(W)}{\sqrt{\text{Var}(W)}}$$

gegen $N(0, 1)$ strebt, wenn m und n gegen ∞ streben; siehe z.B. Bickel und Doksum (1977), Satz 9.1.1. Daher braucht man für große m und n keine Tabellen der Verteilung von W .

Ist man daran interessiert, die Hypothese $F = G$ nicht nur dann zu verwerfen, wenn die Y_j tendenziell größer sind als die X_i , sondern auch dann, wenn sie tendenziell kleiner sind, so wendet man den Rangsummentest auch zweiseitig an.

F. Wilcoxon hat noch einen weiteren Test vorgeschlagen, der in ähnlichen Situationen wie der Zeichentest angewandt wird, aber, im Gegensatz zu letzterem, sehr großen und sehr kleinen Werten mehr Gewicht verleiht als Werten, die nahe bei 0 liegen. Gemeint ist der **Vorzeichenrangtest**.

Wir betrachten wieder den bindungsfreien Fall. Seien $D_1, D_2, \dots, D_n$ unabhängig und identisch verteilt. Sei $T_i = 1$, wenn $D_i > 0$ ist und $= 0$ sonst. Ist die Verteilung der D_i *symmetrisch* bzgl. 0, so bedeutet dies, dass für alle $0 \leq a < b$ die Gleichung

$$P(-b < D_i \leq -a) = P(a \leq D_i < b)$$

gilt. Dann ist

$$P(a \leq |D_i| < b, T_i = 1) = \frac{1}{2} P(a \leq |D_i| < b).$$

Für $a \rightarrow 0$ und $b \rightarrow \infty$ erhalten wir $P(T_i = 1) = 1/2$ und damit die Unabhängigkeit von $|D_i|$ und T_i .

Sei R_i^+ der Rang von $|D_i|$ in $|D_1|, |D_2|, \dots, |D_n|$. Die Verteilung von $(R_1^+, \dots, R_n^+)$ ist die Gleichverteilung auf der Menge der Permutationen von $\{1, \dots, n\}$.

Die Vorzeichen-Rang-Statistik von Wilcoxon ist definiert durch

$$W^+ = \sum_{i=1}^n T_i R_i^+.$$

Sind die D_i symmetrisch bzgl. 0 verteilt, so ist die Verteilung von W^+ durch die hier gemachten Angaben über die Verteilungen der T_i und R_i^+ eindeutig bestimmt. Große Werte von W^+ sprechen dafür, dass die D_i mehr positive große Werte annehmen, als stark negative Werte.

Eine ausführliche über weite Strecken elementar gehaltene Darstellung nichtparametrischer Methoden mit zahlreichen Beispielen aus der Praxis findet sich bei Lehmann (1975). Als relativ leicht zugängliche Einführung in die mathematischen Methoden eignet sich Randles-Wolfe (1979). Einen rascheren Überblick für Leser mit Grundkenntnissen in Mathematischer Statistik bieten Bickel-Doksum (1977).

Anhang

Die Gammaverteilung und verwandte statistische Verteilungen: Als Gammafunktion bezeichnet man bekanntlich die für $t > 0$ durch

$$\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx$$

definierte Funktion. Durch partielle Integration ergibt sich

$$\Gamma(t+1) = t\Gamma(t) \quad (t > 0).$$

Da offenbar $\Gamma(1) = 1$ ist, folgt induktiv $\Gamma(n) = (n-1)!$. Sei für $\alpha > 0$ und $\nu > 0$

$$f_{\alpha, \nu}(x) = \frac{1}{\Gamma(\nu)} \alpha^\nu x^{\nu-1} e^{-\alpha x} \quad (x > 0) \quad (14.18)$$

(und $= 0$ für $x \leq 0$). Für $\alpha = 1$ ist $f_{\alpha, \nu}$ offenbar eine Dichte. Hat X die Dichte $f_{1, \nu}(x)$, so hat nach Satz 11.6 X/α die Dichte $\alpha f_{1, \nu}(\alpha x)$ und dies ist gerade $f_{\alpha, \nu}(x)$. Also ist für jedes $\alpha > 0$ und $\nu > 0$ durch (14.18) eine Dichte gegeben. Man nennt sie die Dichte der **Gammaverteilung** mit Parametern α, ν . Man rechnet leicht nach, dass

$$EX = \nu \quad \text{und} \quad \text{Var}(X) = \nu \quad (14.19)$$

gilt, so dass Erwartungswert und Varianz zur Dichte $f_{\alpha, \nu}$ die Werte ν/α und ν/α^2 haben. Der folgende Satz ist sehr wichtig.

Satz 14.6 Die Faltung $f_{\alpha, \nu} * f_{\alpha, \mu}$ ist $f_{\alpha, \nu+\mu}$. (In anderen Worten: Die Summe von zwei unabhängigen $f_{\alpha, \mu}$ - und $f_{\alpha, \nu}$ -verteilten Zufallsvariablen ist $f_{\alpha, \mu+\nu}$ -verteilt.)

Beweis: Wir können $\alpha = 1$ annehmen. Nach (11.7) hat die Faltung in x die Form

$$\frac{1}{\Gamma(\mu)\Gamma(\nu)} \int_0^x (x-y)^{\mu-1} e^{-(x-y)} y^{\nu-1} e^{-y} dy.$$

Mit der Substitution $y = x \cdot t$ wird daraus

$$\begin{aligned} &= \frac{e^{-x}}{\Gamma(\mu)\Gamma(\nu)} \int_0^1 x^{\mu-1} (1-t)^{\mu-1} x^{\nu-1} t^{\nu-1} x dt \\ &= \frac{\Gamma(\mu+\nu)}{\Gamma(\mu)\Gamma(\nu)} f_{1, \mu+\nu}(x) \int_0^1 (1-t)^{\mu-1} t^{\nu-1} dt. \end{aligned}$$

Das rechts stehende Integral bezeichnet man als **Beta-Funktion** $B(\nu, \mu)$. Da die Faltung eine Dichte ist, muss

$$B(\nu, \mu) = \frac{\Gamma(\mu)\Gamma(\nu)}{\Gamma(\mu+\nu)}$$

gelten, so dass die Faltung $= f_{1, \mu+\nu}(x)$ ist. □

Hat eine Zufallsvariable X eine $N(0, 1)$ -Verteilung, so ist nach (11.5) die Dichte von X^2 durch

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

gegeben. Da dies bis auf die Normierungskonstante mit $f_{1/2, 1/2}(x)$ übereinstimmt und beides Dichten sind, müssen auch die Normierungskonstanten $1/\sqrt{2\pi}$ und $\alpha^\nu/\Gamma(\nu) = 1/(\sqrt{2}\Gamma(1/2))$ übereinstimmen. Also ist $\Gamma(1/2) = \sqrt{\pi}$, und X^2 hat die Dichte $f_{1/2, 1/2}(x)$. Nach Satz 14.6 ist dann die χ_k^2 -Verteilung die Gammaverteilung mit Parametern $\alpha = 1/2$, $\nu = k/2$.

Um die t -Verteilung zu bestimmen, benötigen wir

Lemma 14.7 Sind X, Y unabhängige Zufallsvariable mit Dichten f, g und ist $Y > 0$, so hat X/Y die Dichte

$$h(t) = \int_0^\infty f(ty) g(y) y dy.$$

Beweis: Es ist

$$P(X/Y \leq u) = \iint_B f(x) g(y) dx dy$$

mit $B = \{(x, y) : x \leq yu\}$. Bei der Substitution $t = x/y$ ist $dx = y dt$, und wir erhalten

$$= \int_{-\infty}^u \int_0^\infty f(ty) g(y) y dy dt = \int_{-\infty}^u h(t) dt. \quad \square$$

Da nun in Definition 14.2 U_k die Dichte $g_k(x)$ hat, hat nach Satz 11.6 U_k/k die Dichte $k g_k(kx)$. Aus (11.4) folgt, dass die Wurzel aus einer nichtnegativen Zufallsvariablen mit Dichte $d(x)$ die Dichte $2x d(x^2)$ hat. Also ist die Dichte von $Y = \sqrt{U_k/k}$ von der Form $g(y) = k g_k(ky^2) 2y$. Da die Dichte von W in Definition 14.2 $\varphi(x)$ ist, folgt nach Lemma 14.7 für die Dichte h_k von T_k

$$h_k(t) = \int_0^\infty \varphi(ty) k g_k(ky^2) 2y^2 dy.$$

Mit Hilfe von Formel (14.8), der Substitution $s = y^2(t^2 + k)/2$, der Identität $\Gamma(1/2) = \sqrt{\pi}$ und der Definition der Gammafunktion ergibt sich hieraus nach etwas Rechnung (14.10).

Aufgaben

1. Aus den Messungen der Wärmeleitfähigkeit von 15 Glasfaserplatten einer bestimmten Stärke ergab sich der Mittelwert $\bar{x} = 17,1$ und der Wert $s^2 = 0,36$. Unter der Annahme, die Messwerte seien unabhängig und normalverteilt gebe man ein Konfidenzintervall zum Niveau $\alpha = 0,1$ für den Erwartungswert μ der Wärmeleitfähigkeit an.
2. Sei $\{C(x) : x \in \mathcal{X}\}$ eine Familie von Konfidenzbereichen zum Niveau $1 - \alpha$ für $g(\vartheta)$. Man zeige, dass der Test mit Verwerfungsbereich $R = \{x : g_0 \notin C(x)\}$ ein Test zum Niveau α der Hypothese $g(\vartheta) = g_0$ ist. lässt sich auch umgekehrt den Verwerfungsbereichen zu einer Familie von nichtrandomisierten Tests eine Familie von Konfidenzbereichen zuordnen? (L)
3. Seien X_1, X_2 unabhängig $N(\mu, \sigma^2)$ -verteilt und $T = \bar{X} \sqrt{2}/s$. Man zeige, dass $|T|$ für $\mu = 0$ die gleiche Verteilung hat wie $1/|T|$. (Anleitung: $(\bar{x}, \bar{x})$ ist die Projektion von (x_1, x_2) auf die Diagonale. Hinweis: Diese Aufgabe illustriert die Bedeutung der Alternative. Nur vom Standpunkt der Hypothese aus betrachtet ist der Test mit Verwerfungsbereich $R = \{x : |T(x)| > 1\}$ ebenso gut wie der mit Verwerfungsbereich R^c , der genau die umgekehrte Entscheidung liefert.) (L)
4. Es wird vermutet, dass bei Pferderennen auf einer kreisförmigen Rennbahn die Startposition einen Einfluss auf die Gewinnchancen hat. In 144 Rennen hatten die Sieger die Startpositionen 1, 2, ..., 8 mit den folgenden Häufigkeiten: 29, 19, 18, 25, 17, 10, 15, 11. Man teste die Hypothese, dass alle Positionen die gleiche Siegwahrscheinlichkeit besitzen zum Niveau 0.05.
5. Seien $X_1, \dots, X_{10}$ unabhängig $N(\mu, 1)$ -verteilt. Man bestimme und zeichne die Gütefunktion des Tests, der die Hypothese $\mu = 0$ verwirft, wenn $|\bar{X}| > 1,96/\sqrt{10}$ ist.
6. In § 14.4 bestimme man $P(W = m(m+1)/2 + 2)$ für $m = n = 10$.
7. In einer (2×2) -Kontingenztafel sei $N_{j,k}$ die Häufigkeit der Merkmalskombination (j, k) bei n Beobachtungen. Der Vektor $(N_{1,1}, N_{1,2}, N_{2,1}, N_{2,2})$ ist also multinomialverteilt mit Parametern n und $(\vartheta_{1,1}, \vartheta_{1,2}, \vartheta_{2,1}, \vartheta_{2,2})$. Man zeige, dass bei gegebenem $\{N_{1,\cdot} = n_{1,\cdot}\}$ die Zufallsvariablen $N_{1,1}$ und $N_{2,1}$ unabhängig und binomialverteilt sind mit Parametern $n_{1,\cdot}$ und $p_1 = \vartheta_{1,1}/(\vartheta_{1,1} + \vartheta_{1,2})$, bzw. $n_{2,\cdot} = n - n_{1,\cdot}$ und $p_2 = \vartheta_{2,1}/(\vartheta_{2,1} + \vartheta_{2,2})$.
8. (Fortsetzung). Seien $N_{1,1}$ und $N_{2,1}$ unabhängig und binomialverteilt mit Parametern $n_{1,\cdot}$ und p_1 , bzw. $n_{2,\cdot}$ und p_2 . Man zeige: Ist $p_1 = p_2$, so ist

$$P(N_{1,1} = n_{1,1} \mid N_{1,\cdot} = n_{1,\cdot}) = \frac{\binom{n_{1,\cdot}}{n_{1,1}} \binom{n_{2,\cdot}}{n_{2,1} - n_{1,1}}}{\binom{n}{n_{1,1}}}.$$

(Anwendung: Medikament 1 soll mit Medikament 2 verglichen werden. Im Paar (j, k) sei j die Nummer des Medikaments. Im Fall eines Erfolges sei $k = 1$, sonst $k = 2$. Die Hypothese, dass Medikament 2 mindestens so gut wie Medikament 1 ist, kann durch $p_1 \leq p_2$ beschrieben werden. Sie wird bei fester erster Spaltensumme $n_{\cdot,1}$ für große $n_{1,1}$ verworfen, wobei die Schranke aus der obigen hypergeometrischen Verteilung bestimmt wird. Aus der vorigen Aufgabe ergibt sich, dass man diesen Test, den exakten Test von Fisher, auch anwenden kann, um die Unabhängigkeit der Merkmale zu prüfen. Dann gibt die hypergeometrische Verteilung den Wert von $P(N_{1,1} = n_{1,1} \mid N_{1,\cdot} = n_{1,\cdot}, N_{\cdot,1} = n_{\cdot,1})$ unter der Hypothese an.)

9. Zu einem unbekannten Zeitpunkt θ wird ein Prozess mit einer zufälligen Dauer T ausgelöst, die exponentiell mit Dichte $f(x) = \lambda \exp(-\lambda x) 1_{[0,\infty)}$ verteilt ist. λ sei bekannt. Nur $S = \theta + T$ kann beobachtet werden. Beschreiben Sie einen schärfsten Test zum Niveau $\alpha = 0,05$ für die Hypothese $\theta \leq 3$ gegen die Alternative $\theta > 3$.
10. Die Zufallsvariablen X_1, X_2 seien unabhängig; X_1 sei $N(\mu, 1)$ -verteilt, X_2 sei $N(\mu, 4)$ -verteilt. Man bestimme den schärfsten Test der Hypothese $\mu = 0$ gegen die Alternative $\mu = 2$ zum Niveau $\alpha = 0,1$, sowie die Wahrscheinlichkeit eines Fehlers 2. Art zu diesem Test.

Kapitel III Markowsche Ketten

Zentrales Thema der Wahrscheinlichkeitstheorie ist das Studium von stochastischen Prozessen, d.h. von Familien von Zufallsvariablen, die meist die zeitliche, gelegentlich die räumliche, Entwicklung eines Zufallsgeschehens beschreiben. Neben den Folgen von unabhängigen Zufallsvariablen, die bisher im Vordergrund unseres Interesses standen, ist eine Klasse von Prozessen besonders wichtig, die man markowsche Ketten oder Markow-Ketten nennt. Sie sind durch eine spezielle übersichtliche Form der Abhängigkeit der Variablen charakterisiert.

§ 15 Die markowsche Eigenschaft

15.1 Definition und Beispiele

Definition 15.1 Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum, T eine beliebige nichtleere Indexmenge, und $(I, \mathcal{I})$ ein messbarer Raum. Eine Familie $\{X_t, t \in T\}$ von Zufallsvariablen mit Werten in I heißt **stochastischer Prozess** mit **Parameterbereich** T und **Zustandsraum** I .

Hier sei I stets abzählbar, d.h. endlich oder abzählbar unendlich. $\mathcal{I}$ sei die Potenzmenge von I . Wir brauchen also $\mathcal{I}$ nicht mehr zu erwähnen. Weiter soll stets $T = \mathbb{Z}^+ = \{0, 1, 2, \dots\}$ sein.

Definition 15.2 Eine **markowsche Kette** ist ein stochastischer Prozess $\{X_n, n \in \mathbb{Z}^+\}$ mit abzählbarem Zustandsraum I , der die folgende **markowsche Eigenschaft** besitzt:

Für alle $n \in \mathbb{Z}^+$ und für alle $i_0, \dots, i_{n+1} \in I$ mit

$$P(X_0 = i_0, \dots, X_n = i_n) > 0 \quad (15.1)$$

ist

$$P(X_{n+1} = i_{n+1} \mid X_0 = i_0, \dots, X_n = i_n) = P(X_{n+1} = i_{n+1} \mid X_n = i_n). \quad (15.2)$$

Wir interpretieren X_n als Zustand eines Systems zur Zeit n . Der Prozess hat die markowsche Eigenschaft, wenn die Wahrscheinlichkeit, zur Zeit $n+1$ in einen beliebigen Zustand zu gelangen, nur vom Zustand zur Zeit n und von n abhängt, aber nicht davon, in welchen Zuständen das System früher war. Das wird an Beispielen noch klarer werden.

Die Bedingung (15.1) soll nur garantieren, dass die bedingten Wahrscheinlichkeiten wohldefiniert sind. Wir werden solche Bedingungen nicht immer hinschreiben.

Um eine modifizierte Formulierung der markowschen Eigenschaft herzuleiten, benutzen wir:

Lemma 15.3 Sind $C_1, C_2, \dots$ disjunkte Ereignisse mit Vereinigung C , und sind die bedingten Wahrscheinlichkeiten $P(A \mid B \cap C_k) = p$ unabhängig von k , so ist

$$p = P(A \mid B \cap C).$$

Dies ist Satz 2.2 (iv) für die bedingte Wahrscheinlichkeit $P(\cdot | B)$.

Satz 15.4 Die markowsche Eigenschaft ist äquivalent dazu, dass für jedes $n \in \mathbb{Z}^+$ und alle $i_n, i_{n+1} \in I$ gilt: Für alle $i_0, \dots, i_{n-1} \in I$ mit (15.1) sind die bedingten Wahrscheinlichkeiten $P(X_{n+1} = i_{n+1} | X_0 = i_0, \dots, X_n = i_n)$ gleich.

Beweis: Die eine Richtung ist an (15.2) ablesbar, die andere folgt aus Lemma 15.3 mit $A = \{X_{n+1} = i_{n+1}\}$ und $B = \{X_n = i_n\}$, wenn $C_1, C_2, \dots$ eine Durchnummerierung der Ereignisse $\{X_0 = i_0, \dots, X_{n-1} = i_{n-1}\}$ ist. $\square$

Beispiele: (1) *Summen unabhängiger Zufallsvariabler:* Sind $X_0, Y_1, Y_2, \dots$ unabhängige Zufallsvariable mit Werten in $\mathbb{Z}^d = I$, so ist die durch $X_n = X_0 + Y_1 + \dots + Y_n$ definierte Folge eine markowsche Kette. Mit Hilfe der Unabhängigkeit folgt nämlich

$$\begin{aligned} P(X_{n+1} = i_{n+1} | X_0 = i_0, \dots, X_n = i_n) \\ &= \frac{P(X_0 = i_0, \dots, X_{n+1} = i_{n+1})}{P(X_0 = i_0, \dots, X_n = i_n)} \\ &= \frac{P(X_0 = i_0, Y_1 = i_1 - i_0, \dots, Y_{n+1} = i_{n+1} - i_n)}{P(X_0 = i_0, Y_1 = i_1 - i_0, \dots, Y_n = i_n - i_{n-1})} \\ &= P(Y_{n+1} = i_{n+1} - i_n). \end{aligned}$$

(2) *Einfaches Warteschlangenmodell:* Seien $0, 1, 2, \dots$ die Zeitpunkte, an denen ein Skilift, der pro Zeiteinheit eine Person befördern kann, abfährt. Zwischen den Zeitpunkten n und $n+1$ kommen Y_n neue Skifahrer an. Die Y_n seien unabhängig. Die Länge X_n der Warteschlange unmittelbar vor der Abfahrt zur Zeit n bestimmt sich rekursiv durch

$$X_n = \max(0, X_{n-1} - 1) + Y_{n-1} \quad (n \geq 1).$$

$X_0 = i_0$ sei eine bekannte Zahl, z.B. $X_0 = 0$.

Da Y_n unabhängig von $Y_0, \dots, Y_{n-1}$ ist, ist Y_n auch unabhängig von $(X_0, \dots, X_n)$, denn die X_i mit $i \leq n$ sind Funktionen von $(Y_1, \dots, Y_{n-1})$. Wir erhalten für $i_n \geq 1$

$$\begin{aligned} P(X_{n+1} = i_{n+1}, X_n = i_n, \dots) \\ &= P(Y_n = i_{n+1} - i_n + 1, X_n = i_n, X_{n-1} = i_{n-1}, \dots) \\ &= P(Y_n = i_{n+1} - i_n + 1) P(X_n = i_n, X_{n-1} = i_{n-1}, \dots) \end{aligned}$$

und somit

$$P(X_{n+1} = i_{n+1} | X_n = i_n, X_{n-1} = i_{n-1}, \dots) = P(Y_n = i_{n+1} - i_n + 1).$$

Ebenso sieht man, dass im Fall $i_n = 0$

$$P(X_{n+1} = i_{n+1} | X_n = i_n, X_{n-1} = i_{n-1}, \dots) = P(Y_n = i_{n+1})$$

ist. Die bedingten Wahrscheinlichkeiten sind also unabhängig von $i_0, \dots, i_{n-1}$.

(3) *Lagerhaltung:* Am n -ten Tag kommen Y_n Kunden in ein Geschäft, die ein bestimmtes Gerät kaufen wollen. Im Geschäft sind am n -ten Tag bei Ladenschluss noch X_n Geräte vorrätig. Ist $X_n \leq 2$, so werden aus dem Vorstadtlager 5 Exemplare über Nacht herangeschafft. Ist $X_n \geq 3$, so unterbleibt dies. Die Y_n seien unabhängig. Dann ist

$$X_{n+1} = \begin{cases} (X_n + 5 - Y_{n+1})^+, & \text{falls } X_n \leq 2, \\ (X_n - Y_{n+1})^+, & \text{falls } X_n \geq 3, \end{cases}$$

wobei $a^+ = \max(a, 0)$ ist. Die Folge $X_0, X_1, \dots$ ist eine markowsche Kette. Der Beweis ist analog zu dem im vorigen Beispiel.

Später folgen weitere Beispiele, aber um ein Gefühl für die markowsche Eigenschaft zu bekommen, ist es nützlich, sich auch Beispiele zu überlegen, in denen sie verletzt ist:

(a) Wir betrachten eine Personenfolge, in der die $(n+1)$ -te Person Kind der n -ten Person ist. Sei $X_n = 1$, wenn an der n -ten Person ein latentes Erbmerkmal in Erscheinung tritt, und $= 0$ sonst. $X_1, X_2, \dots$ ist *keine* markowsche Kette: Wenn $X_n = 0$ ist, kann die Erbanlage vorhanden sein, ohne in Erscheinung zu treten. Ist auch $X_{n-1} = 0$, so ist dies weniger wahrscheinlich als im Fall $X_{n-1} = 1$. Ist Z_n der *Genotyp* der n -ten Person, so ist hier (Z_n) markowsche Kette, aber nicht das Erscheinungsbild X_n . Da X_n sich als Funktion von Z_n beschreiben lässt, zeigt dieses Beispiel auch, dass Funktionen $f(Z_n)$ einer markowschen Kette nicht notwendig markowsche Ketten sind.

(b) Kommen beim Lagerhaltungsbeispiel nicht bediente Kunden am nächsten Tag wieder, so ist die Folge (Y_i) nicht unabhängig und $X_0, X_1, \dots$ keine markowsche Kette. Die markowsche Eigenschaft geht auch verloren, wenn die Lieferung zusätzlicher Geräte aus dem Lager zwei Tage dauert.

Die große Bedeutung der markowschen Ketten für zahlreiche Anwendungen legt die Vermutung nahe, dass die Theorie durch Anwendungsfragen motiviert wurde. Tatsächlich entwickelte Markow zwischen 1906 und 1908 diesen Abhängigkeitsbegriff jedoch bei der Untersuchung der Frage, ob Unabhängigkeit eine notwendige Bedingung für das Gesetz der großen Zahlen und für die Normalapproximation ist. Er gab nur ein „reales“ Beispiel an, die Abfolge der Zustände „Konsonant“ und „Vokal“ in Romanen wie Puschkins „Eugen Onegin“ (für die die Beschreibung durch eine markowsche Kette nur eine erste Approximation darstellt).

15.2 Einfache Folgerungen aus der markowschen Eigenschaft

Ist $X_0, X_1, \dots$ eine markowsche Kette, so gilt für alle n und alle $i_0, \dots, i_n \in I$

$$\begin{aligned} P(X_0 = i_0, \dots, X_n = i_n) \\ = P(X_0 = i_0) P(X_1 = i_1 | X_0 = i_0) \dots P(X_n = i_n | X_{n-1} = i_{n-1}). \end{aligned} \quad (15.3)$$

Dies folgt unmittelbar aus Formel (2.4) und der markowschen Eigenschaft.

Häufig benötigt man die markowsche Eigenschaft in einer formal schärferen Form, nämlich so wie sie im folgenden Satz beschrieben wird:

Satz 15.5 *Sei $0 < n < N$. Ist $X_0, X_1, \dots$ eine markowsche Kette, so gilt für alle $i_n \in I$ und alle Teilmengen $E \subset I^n$, $F \subset I^{N-n}$*

$$\begin{aligned} P((X_{n+1}, \dots, X_N) \in F | X_n = i_n, (X_0, \dots, X_{n-1}) \in E) \\ = P((X_{n+1}, \dots, X_N) \in F | X_n = i_n). \end{aligned} \quad (15.4)$$

Beweis: Da man auf beiden Seiten die σ -Additivität ausnutzen kann, genügt es anzunehmen, F bestehe aus nur einem Element $(i_{n+1}, \dots, i_N)$ von I^{N-n} . Sei $p_k(j|i) = P(X_{k+1} = j | X_k = i)$. Für beliebige $i_0, \dots, i_{n-1}$ ist nach (15.3)

$$\begin{aligned}
P((X_{n+1}, \dots, X_N) \in F \mid X_n = i_n, (X_0, \dots, X_{n-1}) = (i_0, \dots, i_{n-1})) \\
&= \frac{P(X_0 = i_0, \dots, X_N = i_N)}{P(X_0 = i_0, \dots, X_n = i_n)} \\
&= \frac{P(X_0 = i_0) p_0(i_1 \mid i_0) p_1(i_2 \mid i_1) \dots p_{N-1}(i_N \mid i_{N-1})}{P(X_0 = i_0) p_0(i_1 \mid i_0) p_1(i_2 \mid i_1) \dots p_{n-1}(i_n \mid i_{n-1})} \\
&= p_n(i_{n+1} \mid i_n) p_{n+1}(i_{n+2} \mid i_{n+1}) \dots p_{N-1}(i_N \mid i_{N-1}) =: p.
\end{aligned}$$

Da p von $(i_0, \dots, i_{n-1})$ unabhängig ist, ist nach Lemma 15.3 für beliebige disjunkte Vereinigungen C von Mengen der Form $\{(X_0, \dots, X_{n-1}) = (i_0, \dots, i_{n-1})\}$

$$P((X_{n+1}, \dots, X_N) \in F \mid \{X_n = i_n\} \cap C) = p.$$

Setzt man erst $C = \{(X_0, \dots, X_{n-1}) \in E\}$ und dann $C = \Omega$, so sieht man, dass die linke und die rechte Seite von (15.4) beide $= p$ sind. $\square$

Hin und wieder stößt man auf den Fehler, dass in Formel (15.4) auch der Zustand i_n durch eine nichtleere Menge von Zuständen ersetzt wird. Eine solche Verallgemeinerung von (15.4) gilt jedoch nicht.

Satz 15.6 (Chapman-Kolmogorow-Gleichung) *Ist $X_0, X_1, \dots$ eine markowsche Kette und $k < m < n$, so gilt für alle $h, j \in I$*

$$P(X_n = j \mid X_k = h) = \sum_{i \in I} P(X_m = i \mid X_k = h) P(X_n = j \mid X_m = i). \quad (15.5)$$

Beweis: Es ist

$$\begin{aligned}
P(X_k = h, X_n = j) &= \sum_i P(X_k = h, X_m = i, X_n = j) \\
&= \sum_i P(X_k = h, X_m = i) P(X_n = j \mid X_k = h, X_m = i).
\end{aligned}$$

Nach (15.4) kann man im letzten Faktor die Bedingung $\{X_k = h\}$ weglassen. Man braucht nun nur noch durch $P(X_k = h)$ zu dividieren. $\square$

15.3 Stationäre Übergangswahrscheinlichkeiten

Bisher haben wir — ohne dadurch mehr Arbeit zu haben — zugelassen, dass die Wahrscheinlichkeit eines Übergangs von Zustand i in Zustand j vom gerade betrachteten Zeitpunkt abhängt. Besonderes Interesse verdient der Fall, dass der Zeitpunkt keine Rolle spielt.

Definition 15.7 Eine markowsche Kette heißt **homogen** oder Kette mit **stationären Übergangswahrscheinlichkeiten**, wenn für alle $i, j \in I$ $P(X_{n+1} = j \mid X_n = i) =: p_{ij}$ unabhängig von n ist.

$\mathbb{P} = (p_{ij})$ ist dann eine **stochastische Matrix**, d.h. es gilt

$$p_{ij} \geq 0 \quad (i, j \in I) \quad \text{und} \quad \sum_{j \in I} p_{ij} = 1 \quad (i \in I). \quad (15.6)$$

Von jetzt an betrachten wir nur homogene markowsche Ketten. Das Wahrscheinlichkeitsmaß $\pi = (\pi_i, i \in I)$ mit $\pi_i = P(X_0 = i)$ heißt **Startverteilung**. Kennt man π und die Matrix (p_{ij}) der Übergangswahrscheinlichkeiten, so sind die gemeinsamen Verteilungen durch die Formel

$$P(X_0 = i_0, \dots, X_n = i_n) = \pi_{i_0} p_{i_0 i_1} p_{i_1 i_2} \cdots p_{i_{n-1} i_n}, \quad (15.7)$$

die ein Spezialfall von (15.3) ist, eindeutig bestimmt. Ist π ein beliebiges Wahrscheinlichkeitsmaß auf I und $\mathbb{P}$ eine beliebig vorgegebene stochastische Matrix, so existiert eine markowsche Kette mit $\mathbb{P}$ als Matrix der Übergangswahrscheinlichkeiten und Startverteilung π . Auf den Beweis dieses Existenzsatzes gehen wir hier nicht ein.

Die obigen Beispiele von markowschen Ketten sind genau dann homogen, wenn die Verteilung der Y_n unabhängig von n ist (Übung). Für das Warteschlangenbeispiel ergibt die oben durchgeführte Überlegung die Übergangswahrscheinlichkeiten

$$p_{0j} = P(Y_n = j) \quad \text{und} \quad p_{ij} = P(Y_n = j - i + 1) \quad \text{für} \quad i \geq 1.$$

Ist $P(Y_1 = j) = p_j$, so hat die Übergangsmatrix also die Form

$$\mathbb{P} = (p_{ij})_{i,j \geq 0} = \begin{pmatrix} p_0 & p_1 & p_2 & p_3 & \cdots \\ p_0 & p_1 & p_2 & p_3 & \cdots \\ 0 & p_0 & p_1 & p_2 & \cdots \\ 0 & 0 & p_0 & p_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Auch die in § 7 besprochenen Bienaymé-Galton-Watson-Prozesse sind homogene markowsche Ketten (Übung).

Andere Beispiele lassen sich am besten durch Angabe der Übergangsmatrix $\mathbb{P}$ konstruieren:

Sei etwa $I = \{0, 1, \dots, b\}$, $b \geq 1$, $0 < p < 1$, $p_{00} = 1$, $p_{bb} = 1$, und für $0 < i < b$ sei $p_{i,i+1} = p$ und $p_{i,i-1} = 1 - p = q$. Eine markowsche Kette mit diesen Übergangswahrscheinlichkeiten heißt **Irrfahrt mit absorbierendem Rand**. Man könnte dabei an die folgende Situation denken: Peter und Paul spielen ein Spiel, bei dem Peter jede gespielte Runde mit Wahrscheinlichkeit p gewinnt. Wer eine Runde gewinnt erhält vom anderen Spieler eine Mark. Es wird so lange gespielt bis einer der Spieler pleite ist. Beide Spieler zusammen haben b Mark. X_n gibt an, wie viel Mark Peter nach n Runden hat. (Werden nur $m < n$ Runden gespielt, so sei $X_n = X_m$.) Die Verteilung dieser markowschen Kette liegt fest, sobald die Startverteilung oder der Startzustand angegeben ist.

Wir denken uns nun immer die Matrix (p_{ij}) der Übergangswahrscheinlichkeiten gegeben. Es wird manchmal nützlich sein, mehrere Startverteilungen zu betrachten. Sei P_i das Wahrscheinlichkeitsmaß, das die Verteilung des Prozesses beschreibt, wenn man in i startet; also

$$P_i(X_0 = i_0, X_1 = i_1, \dots, X_n = i_n) = \delta_{i i_0} p_{i_0 i_1} p_{i_1 i_2} \cdots p_{i_{n-1} i_n}, \quad (15.8)$$

wobei δ_{ij} das Kroneckersymbol ist. Für ein durch eine allgemeine Startverteilung π wie in (15.7) definiertes P ist dann

$$P_i(X_0 = i_0, \dots, X_n = i_n) = P(X_0 = i_0, \dots, X_n = i_n | X_0 = i), \quad (15.9)$$

sofern nur die rechte Seite wohldefiniert ist, also $\pi_i > 0$ gilt. Für die homogene markowsche Kette nennt man

$$p_{ij}^{(m)} := P(X_{n+m} = j \mid X_n = i)$$

die **m -Schritt-Übergangswahrscheinlichkeit** von i nach j . Dass sie nicht von n abhängt, ist für $m = 1$ Definition der Homogenität. Für $m \geq 2$ ergibt es sich induktiv mittels der Chapman-Kolmogorow-Gleichung. Diese lässt sich nun in der Form

$$p_{hj}^{(t+s)} = \sum_i p_{hi}^{(t)} p_{ij}^{(s)} \quad (15.10)$$

schreiben. Wir können (15.10) auch für $t = 0$ und $s = 0$ als erfüllt ansehen, indem wir $p_{ij}^{(0)} = \delta_{ij}$ setzen. Aus (15.10) und $p_{ij}^{(1)} = p_{ij}$ folgt durch Induktion, dass die Matrix $(p_{ij}^{(m)})$ gerade die m -te Potenz $\mathbb{P}^m$ der Matrix $\mathbb{P} = (p_{ij})$ ist.

Wir kommen nun zu einer ersten Anwendung.

15.4 Absorptionswahrscheinlichkeiten

Sei $J \subset I$ eine Teilmenge des Zustandsraums einer homogenen markowschen Kette und

$$T(\omega) = \inf\{n \geq 0 : X_n(\omega) \in J\}$$

der Zeitpunkt, zu dem zum ersten Mal ein Zustand aus J angenommen wird. Man interpretiert den Verlauf des Prozesses auch oft als *Bewegung* von Zustand zu Zustand und nennt $T(\omega)$ die Zeit des ersten Eintreffens in J . Falls kein $n \geq 0$ mit $X_n(\omega) \in J$ existiert, setzen wir $T(\omega) = \infty$. Uns interessieren nun die Wahrscheinlichkeiten

$$\alpha_i = P_i(T < \infty)$$

bei Start in i je nach J zu gelangen. Für $i \in J$ ist offenbar $\alpha_i = 1$.

Setzt man für $k \geq 0$ $Y_k = X_{k+1}$, so hat wegen der Homogenität der Kette $X_0, X_1, \dots$ die Kette $Y_0, Y_1, \dots$ die gleichen Übergangswahrscheinlichkeiten. Für alle j gilt daher

$$\begin{aligned} P(\exists n \text{ mit } 0 \leq n \leq N \text{ und } Y_n \in J \mid Y_0 = j) \\ = P(\exists n \text{ mit } 0 \leq n \leq N \text{ und } X_n \in J \mid X_0 = j) = P_j(T \leq N). \end{aligned}$$

Für $i \in J^c$ erhalten wir

$$\begin{aligned} P_i(T \leq N+1) &= P_i(\exists n \text{ mit } 1 \leq n \leq N+1 \text{ und } X_n \in J) \\ &= \sum_{j \in I} P_i(X_1 = j) P(\exists n \text{ mit } 1 \leq n \leq N+1 \text{ und } X_n \in J \mid X_1 = j) \\ &= \sum_{j \in I} p_{ij} P(\exists n \text{ mit } 0 \leq n \leq N \text{ und } Y_n \in J \mid Y_0 = j) \\ &= \sum_{j \in I} p_{ij} P_j(T \leq N). \end{aligned}$$

Für $N \rightarrow \infty$ gilt $P_j(T \leq N) \rightarrow \alpha_j$, und wir erhalten

$$\alpha_i = \sum_{j \in I} p_{ij} \alpha_j \quad \text{für } i \in J^c \quad \text{und} \quad \alpha_i = 1 \quad \text{für } i \in J. \quad (15.11)$$

(Es ist bequem die Gleichungen für $i \in J^c$ mit den Randbedingungen für $i \in J$ zusammenzufassen.)

In manchen Fällen lassen sich die α_i mit Hilfe von (15.11) explizit bestimmen. Betrachten wir etwa die oben beschriebene Irrfahrt mit absorbierendem Rand und $J = \{0\}$. Dann ist α_i die Wahrscheinlichkeit, dass Peter all sein Geld verliert, wenn sein Startkapital i Mark ist. Man spricht von der Wahrscheinlichkeit des Ruins von Peter bei Start in i .

(15.11) besagt, dass $\alpha_0 = 1$ und

$$\alpha_i = p \alpha_{i+1} + q \alpha_{i-1} \quad (0 < i < b) \quad (15.12)$$

gilt. Wegen $p_{bb} = 1$ und $b \geq 1$ erreicht man von b startend nie 0. Daher gilt die *zusätzliche* Randbedingung $\alpha_b = 0$. Ist $p = q = 1/2$, so lässt sich (15.12) für beliebige reelle Zahlen c, d durch einen Ansatz $\alpha_i = c + di$ lösen. Die Konstanten c, d bestimmen sich aus den Randbedingungen.

Man erhält so die Lösung

$$\alpha_i = 1 - i/b \quad (i = 0, \dots, b). \quad (15.13)$$

Ist $0 < p < 1$ und $p \neq q$, so führt ein Ansatz

$$\alpha_i = c + d \left(\frac{q}{p} \right)^i \quad (15.14)$$

zum Ziel. Man bestimmt c und d wieder aus den Randbedingungen und erhält die Lösung

$$\alpha_i = \frac{(q/p)^b - (q/p)^i}{(q/p)^b - 1} \quad (i = 0, \dots, b). \quad (15.15)$$

Um sicher zu sein, die richtige Lösung zu haben, müssen wir uns noch überlegen, dass die Lösung unter den Randbedingungen $\alpha_0 = 1, \alpha_b = 0$ eindeutig ist. Betrachten wir den Fall $p \neq q$. Ist (α_i) eine beliebige Lösung, so können wir c und d so bestimmen, dass (15.14) für $i = 0$ und $i = 1$ gilt. Aus (15.12) folgt dann (15.14) für $i = 2$, und dann sukzessive für $i = 3, \dots, b$. Damit hat jede Lösung die Form (15.14), und die Lösung ist eindeutig. Genauso verfährt man im Fall $p = q$.

Wir haben hier außer der Randbedingung $\alpha_i = 1$ ($i \in J$) aus (15.11) noch die zusätzliche Randbedingung $\alpha_b = 0$ benutzt. Sonst hätte es noch die Lösung $\alpha_i \equiv 1$ gegeben. Wir sehen, dass (15.11) alleine die Lösung nicht eindeutig bestimmt.

Im folgenden Abschnitt soll unter anderem gezeigt werden, dass $\alpha_i = P_i(T < \infty)$ stets die minimale Lösung von (15.11) mit $0 \leq \alpha_i \leq 1$ ist.

15.5 Absorptionsverteilungen*

Wir wollen nun noch genauer wissen, mit welchen Wahrscheinlichkeiten die Absorption in den einzelnen Punkten von J erfolgt, d.h. wir interessieren uns für

$$P_i(T < \infty, X_T = j).$$

Als Vorbereitung betrachten wir den Operator $\mathbb{P}$, der beschränkten Funktionen f auf I die Funktion $\mathbb{P}f$ mit

$$(\mathbb{P}f)(i) = \sum_k p_{ik} f(k) \quad (15.16)$$

zuordnet. (Wenn wir f als Spaltenvektor schreiben, entspricht (15.16) der üblichen Matrizenmultiplikation.) Wegen $\mathbb{P}^m = (p_{ik}^{(m)})$ ist — wenn E_i die zu P_i gehörige Erwartung bezeichnet —

$$(\mathbb{P}^m f)(i) = \sum_{k \in I} f(k) P_i(X_m = k) = E_i(f(X_m)).$$

Stellt man sich vor, dass $f(k)$ eine Auszahlung ist, die man im Zustand k erhalten kann, so ist $(\mathbb{P}^m f)(i)$ die erwartete Auszahlung, wenn man in i startet und zur Zeit m ausgezahlt wird.

Es ist nützlich, neben der Kette (X_n) noch den Prozess (Z_n) zu betrachten, der genau wie der (X_n) -Prozess verläuft, solange kein Zustand aus J angenommen wird, und der vom Zeitpunkt T an im gleichen Zustand verbleibt. Sei also $T \wedge n = \min(T, n)$ und $Z_n = X_{T \wedge n}$. Es liegt auf der Hand, dass $Z_0, Z_1, \dots$ eine markowsche Kette ist, deren Übergangsmatrix $Q = (q_{ij})$ durch

$$q_{ij} = p_{ij} \quad (i \in J^c), \quad q_{ii} = 1 \quad (i \in J), \quad q_{ij} = 0 \quad (i \in J, i \neq j)$$

gegeben ist. Man verifiziert dies leicht, indem man für alle Folgen $i_0, \dots, i_n \in I$

$$P_i(Z_0 = i_0, \dots, Z_n = i_n) = \delta_{ii_0} q_{i_0 i_1} q_{i_1 i_2} \cdots q_{i_{n-1} i_n}$$

nachprüft. Ist nun $f = 1_{\{j\}}$ für ein $j \in J$, so ist

$$(Q^n f)(i) = q_{ij}^{(n)} = P_i(Z_n = j) = P_i(T \leq n, X_T = j).$$

Der Limes dieser Folge für $n \rightarrow \infty$ ist die gesuchte Wahrscheinlichkeit.

Wir wollen nun den Limes auf andere Weise charakterisieren, und damit die Antwort erst praktikabel machen. Dabei schränken wir uns nicht auf den Fall $f = 1_{\{j\}}$ ein. Wir schicken ein elementares Lemma voraus:

Lemma 15.8 *Ist (f_m) eine konvergente Folge von gleichmäßig beschränkten Funktionen auf I und (p_k) eine Wahrscheinlichkeitsverteilung auf I , so gilt*

$$\sum_k p_k \lim_m f_m(k) = \lim_m \sum_k p_k f_m(k).$$

Beweis: Übung. (Endlich viele k machen schon $1 - \varepsilon$ der Gesamtwahrscheinlichkeit 1 aus.) $\square$

Neben den oben betrachteten Auszahlungen zur Zeit n interessieren wir uns nun für Auszahlungen zur Zeit T . Da $T(\omega) = \infty$ ist, wenn man nie nach J gelangt, und da $X_\infty(\omega)$ nicht definiert ist, setzen wir fest, dass für Punkte ω mit $T(\omega) = \infty$ stets $f(X_T(\omega)) = 0$ sein soll; wenn man nicht nach J gelangt, gibt es keine Auszahlung. Für beschränktes f auf I sei

$$(Hf)(i) = E_i(f(X_T))$$

die erwartete Auszahlung zur Zeit T bei Start in i . Für $f = 1_J$ ist $(Hf)(i)$ die im vorigen Abschnitt betrachtete Absorptionswahrscheinlichkeit $\alpha_i = P_i(T < \infty)$. Für $f = 1_{\{j\}}$ mit $j \in J$ ist $(Hf)(i) = P_i(T < \infty, X_T = j)$.

Satz 15.9 Sei $f \geq 0$ eine beschränkte Funktion auf I . Dann ist Hf die kleinste beschränkte nichtnegative Funktion h mit

$$Qh = h \quad \text{und} \quad h(j) = f(j) \quad \text{für} \quad j \in J. \quad (15.17)$$

Beweis: Für $f' = f 1_J$ ist $Hf' = Hf$. Ist $T(\omega) > n$, so ist $Z_n(\omega) = X_n(\omega) \in J^c$. Es folgt

$$(Q^n f')(i) = \sum_{k \in I} q_{ik}^{(n)} f'(k) = E_i(f'(Z_n)) = E_i(f(X_T) 1_{\{T \leq n\}}).$$

Die Folge der Mengen $\{T \leq n\}$ wächst, und ihre Vereinigung ist $\{T < \infty\}$. Daher ist die durch $f_n = Q^n f'$ definierte Folge wachsend, und ihr Limes f_∞ ist Hf . Wendet man Lemma 15.8 mit $p_k = q_{ik}$ an, so folgt

$$(Qf_\infty)(i) = \sum_k q_{ik} f_\infty(k) = \lim_n \sum_k q_{ik} (Q^n f')(k) = \lim_n (Q^{n+1} f')(i) = f_\infty(i).$$

Damit gilt $Q(Hf) = Hf$. Für $j \in J$ ist bei Start in j $T = 0$ und also

$$(Hf)(j) = E_j(f(X_0)) = f(j).$$

Die Funktion Hf erfüllt deshalb (15.17). Ist h eine beliebige nichtnegative, beschränkte Funktion, die (15.17) erfüllt, so gilt $Q^n h = h$ für alle n und $h \geq f'$. Daraus folgt

$$h(i) = (Q^n h)(i) = E_i(h(Z_n)) \geq E_i(f'(Z_n)).$$

Da die rechte Seite gegen $(Hf)(i)$ strebt, ist $Hf \leq h$. □

Dass Hf nicht die einzige Funktion zu sein braucht, die (15.17) erfüllt, haben wir schon an dem Beispiel von dem Spiel von Peter und Paul gesehen. Für $f = 1_J$ war $(Hf)(i) = \alpha_i$, aber $h \equiv 1$ war eine zweite Lösung.

Die mangelnde Eindeutigkeit der Lösung h von (15.17) hängt damit zusammen, dass J nicht mit Wahrscheinlichkeit 1 erreicht werden muss.

Bemerkung: Die Gleichung $Qh = h$ bedeutet, dass für alle $i \in J^c$

$$h(i) = \sum_k p_{ik} h(k)$$

ist. Eine Funktion h mit dieser Eigenschaft nennt man auf J^c **harmonisch** für P . Man sagt, dass Hf das **Dirichlet-Problem** löse, eine Funktion anzugeben, die auf dem „Rand“ J die durch eine Funktion f vorgegebenen Werte annimmt und in J^c harmonisch ist. Dass tatsächlich eine enge Beziehung zum klassischen Dirichlet-Problem aus der Potenzialtheorie besteht, lässt sich aber erst zeigen, wenn man Satz 15.9 auf markowsche Prozesse mit stetiger Zeit und Zustandsraum $\mathbb{R}^2$ ausdehnt. (Vgl. Dynkin-Jushkewitch (1969).)

Aufgaben

1. Sei $Y_0, Y_1, Y_2, \dots$ eine Bernoulli-verteilte Folge und $X_n = 2Y_n + Y_{n+1}$. Zeigen Sie, dass (X_n) eine markowsche Kette ist, und bestimmen Sie die Übergangswahrscheinlichkeiten. (L)
2. Eine Urne enthält $m \geq 2$ Kugeln, die mit $1, \dots, m$ nummeriert sind. Es wird fortlaufend mit Rücklegen gezogen. Sei X_n die Anzahl der verschiedenen Kugeln, die in den ersten n Ziehungen auftreten. Ist (X_n) eine markowsche Kette? Wenn ja, bestimme man die Übergangswahrscheinlichkeiten.
3. Sei $Y_1, Y_2, \dots$ unabhängig mit $P(Y_i = 1) = P(Y_i = -1) = 1/2$, und $S_n = Y_1 + \dots + Y_n$, $S_0 = 0$. Ferner sei $X_n = \max(S_0, S_1, \dots, S_n)$. Ist (X_n) eine markowsche Kette?

4. Ein stochastischer Prozess $X_0, X_1, \dots$ mit abzählbarem Zustandsraum I heißt **n -Schritt-markowsch**, wenn für alle $i_0, i_1, \dots \in I$ und alle $N \geq n$ gilt

$$\begin{aligned} P(X_{N+1} = i_{N+1} \mid X_N = i_N, \dots, X_0 = i_0) \\ = P(X_{N+1} = i_{N+1} \mid X_N = i_N, \dots, X_{N-n+1} = i_{N-n+1}). \end{aligned}$$

Sei $I' = I^n$ und $X'_k = (X_k, X_{k+1}, \dots, X_{k+n-1})$. Zeigen Sie, dass (X'_k) genau dann markowsch ist, wenn (X_k) n -Schritt-markowsch ist.

5. Sei (Y_n) eine Folge unabhängiger Zufallsvariablen mit $P(Y_n = 1) = p = 1 - P(Y_n = -1)$ ($0 < p < 1$, $n \geq 1$). Gibt es ein $k \geq 1$, für das die durch $X_n = Y_n Y_{n+1}$ definierte Folge k -Schritt-markowsch ist?
6. Wie ändern sich bei einer Irrfahrt mit absorbierendem Rand die Ruinwahrscheinlichkeiten bei Start in i , wenn der Spieleinsatz verkleinert wird, d.h. $I = \{0, \dots, b\}$ ersetzt wird durch $I' = \{0, 1/n, 2/n, \dots, b\}$? Für die neue Irrfahrt sei $p_{i/n, (i+1)/n} = p$, $p_{i/n, (i-1)/n} = q$ und wieder $p_{00} = p_{bb} = 1$. (L)
7. Spieler 1 und Spieler 2 spielen eine Variante der absorbierenden Irrfahrt, bei der ein Spieler, für den es schlecht steht, günstigere Gewinnwahrscheinlichkeiten für den nächsten Schritt erhält: Sei $I = \{0, 1, \dots, 5\}$, $p_{00} = 1$, $p_{55} = 1$, und für $1 \leq i \leq 4$ sei $p_{i, i+1} = 1 - i/5$, $p_{i, i-1} = i/5$. Man bestimme die Wahrscheinlichkeit der Absorption in 0 bei Start in i .
8. Eine markowsche Kette $X_0, X_1, \dots$ bewegt sich im Zustandsraum $\{0, 1, 2, 3\}$ nach den Übergangswahrscheinlichkeiten einer Irrfahrt mit absorbierendem Rand und $p = 1/2$. In den Zeitpunkten $n = 1, 2, \dots$ leuchtet jeweils mit Wahrscheinlichkeit w ein Lichtblitz auf und die Lage von X_n kann beobachtet werden. Bestimmen Sie in Abhängigkeit vom Startzustand $X_0 = i$ die Verteilung der Position U der Kette im Moment S der ersten Beobachtung. (L)

§ 16 Das Verhalten markowscher Ketten in langen Zeiträumen

Kennt man die Übergangswahrscheinlichkeiten (p_{ij}) einer markowschen Kette, so lassen sich Wahrscheinlichkeiten, die nur von einer kleinen Zahl von Übergängen abhängen, oft noch explizit ausrechnen. Der Rechenaufwand z.B. für die Berechnung der n -Schritt-Übergangswahrscheinlichkeiten kann aber für große n extrem hoch werden. Wir sind daher an Grenzwertsätzen für $n \rightarrow \infty$ interessiert.

Wir beginnen mit einem wichtigen Spezialfall, der uns den Weg weisen wird und für viele Anwendungen bereits ausreicht. Die Darstellung des allgemeinen Falls wird aber auch unabhängig lesbar sein.

16.1 Ketten mit endlich vielen Zuständen

Wir betrachten Ketten, für die man — nicht notwendig in nur einem Schritt — von jedem Zustand zu jedem anderen gelangen kann. Ferner schließen wir eine Art zyklische Bewegung aus. Der Zustandsraum soll nicht in echte Teilmengen $C_0, \dots, C_{d-1}$ zerlegbar sein, für die man von einem Zustand in C_ν in einem Schritt stets nur nach $C_{(\nu+1) \pmod d}$ gelangt. Eine solche Zerlegung ist etwa für die in Bild 16.1 dargestellte Kette möglich. Die Pfeile mögen die Übergänge andeuten, die in einem Schritt möglich sind. Man kann in diesem Fall $C_0 = \{1, 3\}$ und $C_1 = \{2, 4\}$ setzen.

Wir werden später sehen, dass sich diese Voraussetzungen für eine Kette mit endlich vielen Zuständen durch die im folgenden Satz beschreiben lassen:

Satz 16.1 *Hat für eine Kette mit endlich vielen Zuständen die L -Schritt-Übergangsmatrix $(p_{ij}^{(L)})$ für ein $L \geq 1$ nur strikt positive Elemente, so konvergieren die Übergangswahrscheinlichkeiten $p_{ij}^{(n)}$ für $n \rightarrow \infty$ (exponentiell schnell) gegen von i unabhängige Zahlen ρ_j . Der Vektor $\rho = (\rho_j)$ ist der eindeutig bestimmte Wahrscheinlichkeitsvektor, der das Gleichungssystem*

$$\rho_k = \sum_{j \in I} \rho_j p_{jk} \quad (k \in I) \quad (16.1)$$

löst.

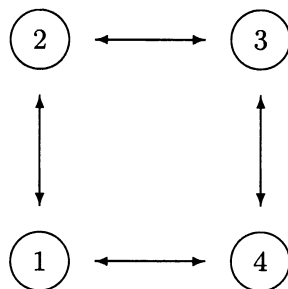


Bild 16.1

Bemerkung: Denken wir uns $\rho = (\rho_j)$ als Zeilenvektor geschrieben, so lässt sich (16.1) in Matrizenschreibweise als

$$\rho = \rho \mathbb{P} \quad \text{mit} \quad \mathbb{P} = (p_{ij}) \quad (16.2)$$

schreiben. Man nennt eine Wahrscheinlichkeitsverteilung ρ , die dieser Bedingung genügt, **invariant**.

Beweis von Satz 16.1: Wir setzen

$$m_j^{(n)} = \min_i p_{ij}^{(n)} \quad \text{und} \quad M_j^{(n)} = \max_i p_{ij}^{(n)}.$$

Dann gilt

$$m_j^{(n+1)} = \min_i \sum_{h \in I} p_{ih} p_{hj}^{(n)} \geq \min_i \sum_{h \in I} p_{ih} m_j^{(n)} = m_j^{(n)},$$

und genauso folgt $M_j^{(n+1)} \leq M_j^{(n)}$.

Da es nur endlich viele Paare (i, j) gibt, und alle $p_{ij}^{(L)}$ strikt positiv sind, gibt es ein $\delta > 0$ mit

$$p_{ij}^{(L)} \geq \delta > 0 \quad \text{für alle} \quad (i, j). \quad (16.3)$$

Bezeichnen wir für feste $h, i \in I$ die Summe über alle k , für die $p_{hk}^{(L)} \geq p_{ik}^{(L)}$ gilt, mit $\sum_{k+}$ und die Summe über die übrigen k mit $\sum_{k-}$, so ist

$$\sum_{k+} (p_{hk}^{(L)} - p_{ik}^{(L)}) + \sum_{k-} (p_{hk}^{(L)} - p_{ik}^{(L)}) = 1 - 1 = 0. \quad (16.4)$$

Ist nun, für festes n , h ein Zustand, für den $p_{hj}^{(n+L)}$ maximal ist, und i ein Zustand, für den $p_{ij}^{(n+L)}$ minimal ist, so ist

$$\begin{aligned} M_j^{(n+L)} - m_j^{(n+L)} &= p_{hj}^{(n+L)} - p_{ij}^{(n+L)} = \sum_k (p_{hk}^{(L)} - p_{ik}^{(L)}) p_{kj}^{(n)} \\ &\leq \sum_{k+} (p_{hk}^{(L)} - p_{ik}^{(L)}) M_j^{(n)} + \sum_{k-} (p_{hk}^{(L)} - p_{ik}^{(L)}) m_j^{(n)} \\ &\leq \sum_{k+} (p_{hk}^{(L)} - p_{ik}^{(L)}) (M_j^{(n)} - m_j^{(n)}) \quad (\text{wegen (16.4)}) \\ &\leq (1 - \delta) (M_j^{(n)} - m_j^{(n)}). \end{aligned}$$

Induktiv folgt nun

$$M_j^{(\nu L)} - m_j^{(\nu L)} \leq (1 - \delta)^\nu \quad (\nu \geq 0).$$

Da die Folge $(M_j^{(n)})_{n \geq 0}$ fallend und die Folge $(m_j^{(n)})_{n \geq 0}$ wachsend ist, ergibt sich hieraus für $\eta = (1 - \delta)^{1/L} < 1$ die Abschätzung

$$M_j^{(n)} - m_j^{(n)} \leq (1 - \delta)^{-1} \eta^n \quad (n \geq 0).$$

Ferner müssen die $m_j^{(n)}$ und die $M_j^{(n)}$ gegen Grenzwerte ρ_j mit

$$m_j^{(n)} \leq \rho_j \leq M_j^{(n)}$$

konvergieren. Da für alle $i \in I$ $p_{ij}^{(n)}$ zwischen diesen Grenzen liegt, gilt

$$|p_{ij}^{(n)} - \rho_j| \leq (1 - \delta)^{-1} \eta^n \quad (i \in I, n \geq 0). \quad (16.5)$$

Die erste der Aussagen ist damit bewiesen.

Durch Grenzübergang $n \rightarrow \infty$ folgt aus

$$p_{ik}^{(n+1)} = \sum_j p_{ij}^{(n)} p_{jk}$$

die Gleichung (16.1). Als Grenzwert von endlichdimensionalen Wahrscheinlichkeitsvektoren muss auch ρ Wahrscheinlichkeitsvektor sein. Ist nun $\tilde{\rho}$ ein beliebiger Wahrscheinlichkeitsvektor, für den $\tilde{\rho} = \tilde{\rho} \mathbb{P}$ gilt, so gilt auch für alle n $\tilde{\rho} = \tilde{\rho} \mathbb{P}^n$, bzw., ausführlich geschrieben,

$$\tilde{\rho}_k = \sum_j \tilde{\rho}_j p_{jk}^{(n)}.$$

Durch Grenzübergang folgt $\tilde{\rho}_k = \sum_j \tilde{\rho}_j \rho_k = \rho_k$. □

Beispiel 16.2 Die Übergangsmatrix

$$\mathbb{P} = \begin{pmatrix} 0 & 3/4 & 1/4 \\ 1/2 & 0 & 1/2 \\ 1 & 0 & 0 \end{pmatrix}$$

beschreibt die Kette in Bild 16.2. Mit Hilfe der Pfeile überlegt man sich leicht, dass man von jedem Zustand i in genau 4 Schritten zu jedem Zustand j kommen kann. Das bedeutet, dass $\mathbb{P}^4$ eine Matrix mit strikt positiven Elementen $p_{ij}^{(4)}$ ist, was man natürlich auch nachrechnen kann. Die Kette erfüllt also die Voraussetzungen des Satzes mit $L = 4$. Die $p_{ij}^{(n)}$ konvergieren also gegen Zahlen ρ_j . Aus den drei Gleichungen (16.1) und der Gleichung $\rho_1 + \rho_2 + \rho_3 = 1$ erhält man $\rho_1 = 8/19$, $\rho_2 = 6/19$ und $\rho_3 = 5/19$. Wo auch immer man startet, die Wahrscheinlichkeit, nach 100 Schritten im Zustand 1 zu sein, ist ziemlich genau $8/19$.

Wir wenden uns nun der allgemeineren Situation zu, wobei auch ganz neue Fragen auftreten werden. Wir werden auch für den Fall endlicher Zustandsräume etwas dazulernen.

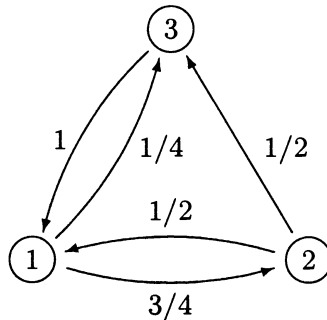


Bild 16.2

16.2 Kommunizierende Zustände und Periodizität

Haben in einer homogenen Kette mit Zustandsraum $I = \{1, 2, 3, 4\}$ nur die in Bild 16.3 durch Pfeile gekennzeichneten Übergänge $i \rightarrow j$ positive Wahrscheinlichkeit p_{ij} , so kann man nie von 1 nach 3 oder 4 gelangen.

Für das Studium der Wahrscheinlichkeiten $P(X_n = k | X_0 = 1)$ benötigt man also nur die p_{ij} mit $1 \leq i, j \leq 2$. Die folgenden Überlegungen dienen dazu, in diesem Sinn das Studium von $p_{ij}^{(n)}$ für große n zu vereinfachen.



Bild 16.3

Definition 16.3 Wir sagen, dass Zustand i in n Schritten zu Zustand j führe und schreiben dafür $i \rightsquigarrow j[n]$, wenn $p_{ij}^{(n)} > 0$ ist. Gibt es ein $n \geq 1$ mit $i \rightsquigarrow j[n]$, so sagen wir i führe zu j und schreiben $i \rightsquigarrow j$.

Aus der Chapman-Kolmogorow-Gleichung (15.10) folgt

$$p_{hj}^{(m+n)} \geq p_{hi}^{(m)} p_{ij}^{(n)}. \quad (16.6)$$

Daher ist die Relation $\rightsquigarrow$ transitiv, d.h. aus $h \rightsquigarrow i$ und $i \rightsquigarrow j$ folgt $h \rightsquigarrow j$. Aus (15.8) folgt, dass $i \rightsquigarrow j[n]$ zur Existenz einer Folge $i_0, i_1, \dots, i_n$ mit $i_0 = i$, $i_n = j$ und $p_{i_\nu, i_{\nu+1}} > 0$ äquivalent ist.

Wir sagen, i **kommuniziere** mit j , und schreiben dafür $i \longleftrightarrow j$, wenn $i \rightsquigarrow j$ und $j \rightsquigarrow i$ gilt. i heißt **wesentlich**, wenn jeder Zustand j , zu dem i führt, auch zurück zu i führt, also mit i kommuniziert. Ist i wesentlich, und führt i zu j , so ist auch j wesentlich. Wenn i nur zu Zuständen führt, von denen man nicht nach i zurückkommen kann, gilt nicht $i \rightsquigarrow i$. Aber auf der Teilmenge der wesentlichen Zustände ist $i \longleftrightarrow j$ offenbar eine Äquivalenzrelation.

In Bild 16.3 sind die Zustände 1, 2 wesentlich, aber nicht die Zustände 3, 4.

Für die Konvergenz der $p_{ij}^{(n)}$ ist noch eine weitere Eigenschaft der Zustände wichtig. Blicken wir noch einmal auf Bild 16.1. Von einem Zustand mit ungerader Nummer geht man stets zu einem Zustand mit gerader Nummer. Die möglichen Rückkehrzeiten in den Ausgangszustand sind daher alle durch 2 teilbar. So etwas müssen wir ausschließen.

Definition 16.4 Für einen Zustand i mit $i \rightsquigarrow i$ heißt der größte gemeinsame Teiler der potenziellen Rückkehrzeiten

$$d_i = \text{ggT}\{n \geq 1 : i \rightsquigarrow i[n]\}$$

die **Periode** von i . Gilt nicht $i \rightsquigarrow i$, so sei $d_i = \infty$. Zustände mit $d_i = 1$ heißen **aperiodisch**. Die Kette heißt aperiodisch, wenn alle Zustände aperiodisch sind, und **periodisch** mit Periode d , wenn alle $d_i = d \geq 2$ sind.

Satz 16.5 Kommuniziert i mit j , so ist $d_i = d_j$.

Beweis: Gilt $j \rightsquigarrow j[n]$ und sind k, m Zeitpunkte mit $i \rightsquigarrow j[k]$ und $j \rightsquigarrow i[m]$, so gilt $i \rightsquigarrow i[k+m]$ und $i \rightsquigarrow i[k+m+n]$. d_i teilt also $k+m$ und $k+m+n$ und also auch n . d_i ist also gemeinsamer Teiler aller n mit $j \rightsquigarrow j[n]$. Es folgt $d_i \leq d_j$. Aus Symmetriegründen muss auch $d_j \leq d_i$ gelten. $\square$

Sei nun $C(i)$ die Menge der mit i kommunizierenden Zustände. Sie ist genau dann nichtleer, wenn man von i nach i zurückkehren kann. Auf der Menge der Zustände, die zu sich selbst zurückführen, bilden die $C(i)$ die Äquivalenzklassen bezüglich der Äquivalenzrelation $\rightsquigarrow$, also gewissermaßen die Komponenten, in die sich die Kette zerlegen lässt. Die Kette in Bild 16.3 hat die Komponenten $C(1) = C(2) = \{1, 2\}$ und $C(3) = C(4) = \{3, 4\}$. Komponenten, die man verlassen kann (z.B. $C(3)$), bestehen aus nicht wesentlichen Zuständen.

Wir betrachten jetzt eine feste Komponente $C(i)$ und wollen uns überlegen, dass im Fall $d_i \geq 2$ die Menge $C(i)$ in zyklisch angeordnete Teilmengen zerfällt, von denen jede nur zur nächsten (und evtl. nach außen) führt. Wir benötigen ein Lemma aus der elementaren Zahlentheorie.

Lemma 16.6 *Ist d der größte gemeinsame Teiler von Zahlen $n_1, n_2, \dots \in \mathbb{N}$, so gibt es Zahlen K und L derart, dass sich jedes ld mit $l \geq L$ als Summe der Form*

$$ld = \sum_{k=1}^K c_k n_k \quad \text{mit } c_k \in \mathbb{N} \quad (16.7)$$

schreiben lässt.

Beweis: Wir können annehmen, dass d gleich 1 und der ggT einer endlichen Teilfolge $n_1, \dots, n_K$ sei. Nach dem Hauptsatz über den ggT (z.B. in Hasse (1964)) gibt es ganze Zahlen $a_1, \dots, a_K$ mit

$$a_1 n_1 + \dots + a_K n_K = 1.$$

Ist $a = \max(|a_1|, |a_2|, \dots, |a_K|)$ und $L = a n_1 (n_1 + \dots + n_K)$, so lässt sich jedes $l \geq L$ in der Form

$$l = a n_1 (n_1 + \dots + n_K) + i n_1 + r (a_1 n_1 + \dots + a_K n_K)$$

mit $i \geq 0$ und $0 \leq r < n_1$ schreiben. Darin ist die Summe der Koeffizienten für jedes n_k eine natürliche Zahl. $\square$

Satz 16.7 *Zu jedem $j \in C(i)$ gibt es eine eindeutig bestimmte Zahl r_j mit $0 \leq r_j < d_i$ derart, dass $i \rightsquigarrow j[n]$ nur für Zahlen $n \equiv r_j \pmod{d_i}$ gilt, aber für hinreichend großes $n \geq N(j)$ auch für alle diese n .*

Beweis: Es gibt ein k mit $j \rightsquigarrow i[k]$. Sind $m < n$ zwei natürliche Zahlen mit $i \rightsquigarrow j[m]$ und $i \rightsquigarrow j[n]$, so gilt $i \rightsquigarrow i[k+m]$ und $i \rightsquigarrow i[k+n]$. d_i teilt also $n - m$. Alle n mit $j \rightsquigarrow j[n]$ liegen also in der gleichen Restklasse $r_j \pmod{d_i}$.

Für ein $m \geq 0$ gilt also $i \rightsquigarrow j[md_i + r_j]$. Ist L die zu $d = d_i$ im vorigen Lemma bestimmte Zahl, so setzt man $N(j) = r_j + md_i + Ld_i$. Für jedes $n \equiv r_j \pmod{d_i}$ mit $n \geq N(j)$ gibt es dann ein $l \geq L$ mit $n = r_j + md_i + ld_i$. Aus $i \rightsquigarrow i[n_k]$ und (16.7) folgt $i \rightsquigarrow i[ld_i]$ und also $i \rightsquigarrow j[n]$. $\square$

Sei für $0 \leq r < d_i$

$$C_r(i) = \{j \in C(i) : r_j = r\}.$$

Gilt $p_{jk} > 0$ und kommuniziert k mit j , so kommuniziert k mit i und für großes l gilt $i \rightsquigarrow j [ld_i + r_j]$ und also $i \rightsquigarrow k [ld_i + r_j + 1]$. Also gilt $r_k = (r_j + 1) \pmod{d_i}$. Solange man $C(i)$ nicht verlässt, werden also die Mengen $C_r(i)$ wie in Bild 16.4 zyklisch durchlaufen.

Ist in einer aperiodischen Kette mit endlich vielen Zuständen, die alle kommunizieren, $n \geq N(j)$ für alle j , so sind alle n -Schritt-Übergangswahrscheinlichkeiten $p_{ij}^{(n)}$ strikt positiv. Umgekehrt müssen alle Zustände kommunizieren, wenn für ein $n \geq 1$ alle $p_{ij}^{(n)}$ strikt positiv sind. Ferner ist die Kette dann aperiodisch, denn erstens gilt $i \rightsquigarrow i[n]$, und zweitens $i \rightsquigarrow i[n+1]$, weil nämlich für mindestens ein j $i \rightsquigarrow j[1]$ und für alle j $j \rightsquigarrow i[n]$ gilt. Die Bedingung aus Satz 16.1 ist also für eine Kette mit endlich vielen Zuständen äquivalent zu der Bedingung, dass die Kette aperiodisch sei und alle Zustände kommunizieren. Das ist, wie ein Blick auf Bild 16.2 zeigt, oft leichter zu erkennen.

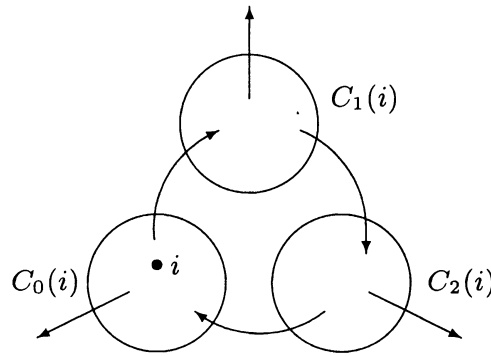


Bild 16.4 Zerlegung der Menge der mit einem Zustand i mit der Periode 3 kommunizierenden Zustände

16.3 Rekurrenz und Transienz

Das wahrscheinlichkeitstheoretische Verhalten einer markowschen Kette hängt stark davon ab, ob man zum Ausgangszustand mit Wahrscheinlichkeit 1 zurückkehrt oder nicht. Wir studieren daher als Nächstes das Rückkehrverhalten. Sei für $n \geq 1$

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

die Wahrscheinlichkeit, bei Start in i den Zustand j zum ersten Mal zur Zeit n zu besuchen, und sei $f_{ij}^{(0)} = 0$. Wir setzen

$$f_{ij}^* = \sum_{n=1}^{\infty} f_{ij}^{(n)} \quad \text{und} \quad p_{ij}^* = \sum_{n=1}^{\infty} p_{ij}^{(n)}.$$

Dann ist f_{ij}^* die Wahrscheinlichkeit, bei Start in i je nach j zu gelangen. p_{ij}^* ist die erwartete Anzahl der Besuche in j bei Start in i , denn

$$\begin{aligned} p_{ij}^* &= \sum_{n=1}^{\infty} p_{ij}^{(n)} = \sum_{n=1}^{\infty} E_i(1_{\{X_n=j\}}) = E_i\left(\sum_{n=1}^{\infty} 1_{\{X_n=j\}}\right) \\ &= E_i(\text{Anzahl } B_j \text{ der Besuche in } j \text{ zu Zeitpunkten } n \geq 1). \end{aligned}$$

Dabei bezeichnet E_i den Erwartungswert unter P_i .

Satz 16.8 *Es gilt $P_i(B_i \geq m) = (f_{ii}^*)^m$ ($m \geq 1$).*

Dies ist heuristisch leicht zu verstehen. Mit Wahrscheinlichkeit f_{ii}^* kehrt man nach i zurück. Wenn dies geschieht, gibt es eine zweite Rückkehr wieder mit der gleichen Wahrscheinlichkeit f_{ii}^* , denn die Übergangswahrscheinlichkeiten haben sich ja nicht geändert. Die Wahrscheinlichkeit, mindestens zweimal nach i zurückzukehren, ist also $(f_{ii}^*)^2$, und offenbar kann man so weiter argumentieren.

Dieses Argument rigoros zu machen erfordert allerdings etwas Arbeit. Die Schwierigkeit ist, dass wir hier die markowsche Eigenschaft nicht auf einen festen Zeitpunkt angewandt haben, sondern auf einen *zufälligen* Zeitpunkt, was im Allgemeinen falsch ist. Wir müssen uns überlegen, dass die Rückkehrzeiten zufällige Zeiten sind, für die unser Schluss erlaubt ist.

Eine Zufallsvariable $\tau : \Omega \rightarrow \{0, 1, 2, \dots, \infty\}$ heißt **Stoppzeit**, wenn für alle $n \geq 0$ das Ereignis $\{\omega : \tau(\omega) = n\}$ nur von $X_0, \dots, X_n$ abhängt. Wir erinnern uns: Dies bedeutet, dass für geeignetes $A \subset I^{n+1}$

$$\{\tau = n\} = \{(X_0, \dots, X_n) \in A\}$$

ist. Man kann dann zum Zeitpunkt n aufgrund der dann vorliegenden Beobachtungen $X_0, \dots, X_n$ schon entscheiden, ob $\tau = n$ gilt oder nicht.

Beweis von Satz 16.8: Sei

$$\tau_1(\omega) = \inf\{n \geq 1 : X_n(\omega) = i\}$$

und

$$\tau_{m+1}(\omega) = \inf\{n > \tau_m(\omega) : X_n(\omega) = i\}.$$

Als Infimum der leeren Menge $\emptyset$ wird dabei, wie üblich, der Wert ∞ festgesetzt. $\tau_m(\omega)$ ist der Zeitpunkt des m -ten Besuches in i , und wenn es keinen m -ten Besuch gibt, soll $\tau_m(\omega) = \infty$ sein.

Die τ_m sind Stoppzeiten. Ist nämlich A_{mn} die Menge der Folgen $(i_0, \dots, i_{n-1}) \in I^n$, für die $i_0 = i$ gilt und genau $m - 1$ weitere der Koordinaten $= i$ sind, so ist

$$\{X_0 = i, \tau_m = n\} = \{(X_0, \dots, X_{n-1}) \in A_{mn}, X_n = i\}.$$

Offenbar ist $\{\tau_m < \infty\} = \{B_i \geq m\}$. Da f_{ii}^* die Rückkehrwahrscheinlichkeit nach i ist, gilt für $m = 1$

$$P_i(\tau_m < \infty) = (f_{ii}^*)^m. \quad (16.8)$$

Sei $D_n^{n+k} = \{X_{n+1} \neq i, \dots, X_{n+k-1} \neq i, X_{n+k} = i\}$. Ist die Gleichung (16.8) für m bewiesen, so ist

$$\begin{aligned}
P_i(\tau_{m+1} < \infty) &= \sum_{k=1}^{\infty} \sum_{n=1}^{\infty} P_i(\tau_{m+1} - \tau_m = k, \tau_m = n) \\
&= \sum_{k=1}^{\infty} \sum_{n=1}^{\infty} P_i(\tau_{m+1} - \tau_m = k \mid \tau_m = n) P_i(\tau_m = n) \\
&= \sum_{k=1}^{\infty} \sum_{n=1}^{\infty} P_i(D_n^{n+k} \mid X_n = i, (X_0, \dots, X_{n-1}) \in A_{mn}) P_i(\tau_m = n) \\
&= \sum_{k=1}^{\infty} \sum_{n=1}^{\infty} P_i(D_n^{n+k} \mid X_n = i) P_i(\tau_m = n) \\
&= \sum_{k=1}^{\infty} \sum_{n=1}^{\infty} P_i(D_0^k \mid X_0 = i) P_i(\tau_m = n) \\
&= \sum_{k=1}^{\infty} f_{ii}^{(k)} P_i(\tau_m < \infty) = f_{ii}^* (f_{ii}^*)^m,
\end{aligned}$$

so dass (16.8) durch Induktion für alle m folgt. $\square$

Wir wollen nun ein paar Früchte ernten. Für $\{B_i = \infty\}$ schreiben wir auch

$$\{X_n = i \text{ } \infty \text{ oft}\}.$$

Definition 16.9 Ein Zustand i heißt **rekurrent**, wenn $P_i(X_n = i \text{ } \infty \text{ oft}) = 1$ ist. Andernfalls heißt i **transient**. Eine markowsche Kette heißt rekurrent (transient), wenn jeder Zustand rekurrent (transient) ist.

Satz 16.10 (Rekurrenzsatz) *Ein Zustand i ist genau dann rekurrent, wenn $f_{ii}^* = 1$ ist. Notwendig und hinreichend hierfür ist $p_{ii}^* = \infty$.*

Beweis: Ist $f_{ii}^* = 1$, so ist $P_i(B_i \geq m) = 1$ für alle m und daher $P_i(B_i = \infty) = 1$. Damit ist auch die Erwartung p_{ii}^* von B_i unter P_i unendlich. Gilt $f_{ii}^* < 1$, so ist

$$\sum_{m=1}^{\infty} P_i(B_i \geq m) = \sum_{m=1}^{\infty} (f_{ii}^*)^m < \infty.$$

Dann ist $P_i(B_i = \infty) = 0$. In der linken Summe kommt $P_i(B_i = k)$ für $m = 1, \dots, k$ vor, also k -mal. Daher ist sie die Erwartung p_{ii}^* von B_i unter P_i , und es gilt $p_{ii}^* < \infty$. $\square$

Das **Rekurrenzkriterium** $p_{ii}^* < \infty$ ist sehr nützlich, denn um $p_{ii}^* < \infty$ zu beweisen, braucht man nur *Abschätzungen* von Wahrscheinlichkeiten. Um $f_{ii}^* = 1$ direkt zu zeigen, müsste man dagegen unendlich viele $f_{ii}^{(n)}$ genau berechnen.

Die Anwendung von Satz 16.10 auf das Studium des Grenzverhaltens von $p_{ij}^{(n)}$ verschieben wir auf den nächsten Paragraphen. Bevor wir zu Beispielen kommen, halten wir aber noch eine nützliche Folgerung fest.

Folgerung 16.11 (a) *Alle mit einem rekurrenten Zustand kommunizierenden Zustände sind rekurrent.*

(b) *Ist i rekurrent, so gilt für alle j mit $i \rightsquigarrow j$ $f_{ji}^* = 1$. Insbesondere ist jeder rekurrente Zustand wesentlich.*

Beweis: (a) Gilt $p_{ij}^{(m)} > 0$ und $p_{ji}^{(k)} > 0$, so folgt aus $p_{ii}^* = \infty$ und

$$p_{jj}^{(k+n+m)} \geq p_{ji}^{(k)} p_{ii}^{(n)} p_{ij}^{(m)}$$

durch Summation $p_{jj}^* = \infty$, also die Rekurrenz von j .

(b) Wegen $i \rightsquigarrow j$ existiert ein m mit $p_{ij}^{(m)} > 0$. Da i rekurrent ist, ist

$$\begin{aligned} 1 &= P_i(\exists n > m : X_n = i) \\ &= \sum_{k \in I} P_i(\exists n > m : X_n = i, X_m = k) \\ &= \sum_{k \in I} P_i(X_m = k) P_i(\exists n > m : X_n = i \mid X_m = k) \\ &= \sum_{k \in I} p_{ik}^{(m)} P_k(\exists n > 0 : X_n = i) = \sum_{k \in I} p_{ik}^{(m)} f_{ki}^*. \end{aligned}$$

Wäre $f_{ij}^* < 1$, so wäre die letzte Summe < 1 . □

Beispiele: (1) *Irrfahrt auf $\mathbb{Z}$* : Seien $Y_1, Y_2, \dots$ unabhängige Zufallsvariable mit

$$P(Y_m = 1) = 1 - P(Y_m = -1) = p = 1 - q.$$

Sei $X_n = Y_1 + \dots + Y_n$. Dann ist $p_{00}^{(2n+1)} = 0$. Mit der stirlingschen Formel oder (5.2) erhält man

$$p_{00}^{(2n)} = \binom{2n}{n} (pq)^n \sim \frac{1}{\sqrt{\pi n}} (4pq)^n.$$

Für $p \neq q$ ist $4pq < 1$ und daher $p_{00}^* < \infty$. Die Kette ist für $p \neq q$ transient. Für $p = q = 1/2$ ist die Kette wegen $\sum n^{-1/2} = \infty$ rekurrent.

(2) *Irrfahrt auf $\mathbb{Z}^d$* : Sei $Y_n = (Y_{n1}, Y_{n2}, \dots, Y_{nd})$. Alle Y_{ni} seien unabhängig mit $P(Y_{ni} = 1) = P(Y_{ni} = -1) = 1/2$. Sei $X_n = Y_1 + \dots + Y_n$. Von $i = (i_1, \dots, i_d) \in \mathbb{Z}^d$ geht man also mit Wahrscheinlichkeit 2^{-d} zu jedem der Punkte $j = (j_1, \dots, j_d)$ mit $|i_\nu - j_\nu| = 1$ für $\nu = 1, \dots, d$. Von $(0, \dots, 0)$ kehrt man genau dann zum Zeitpunkt n nach $(0, \dots, 0)$ zurück, wenn jede der eindimensionalen Irrfahrten $X_{ni} = Y_{1i} + \dots + Y_{ni}$ zum Zeitpunkt n nach 0 zurückkehrt. Da diese unabhängig sind, ist

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

Für $d = 2$ ist diese Kette also wegen $\sum n^{-1} = \infty$ rekurrent; für $d \geq 3$ ist sie wegen $\sum n^{-3/2} < \infty$ transient.

Oft betrachtet man statt (X_n) die Irrfahrt $(\tilde{X}_n)$, bei der man von $i \in \mathbb{Z}^d$ mit der gleichen Wahrscheinlichkeit $1/(2d)$ zu den $2d$ „Nachbarn“ j geht, also zu den j mit $\sum |i_\nu - j_\nu| = 1$. Für $d = 2$ haben wir das alte Problem: Man drehe $\mathbb{Z}^d$ einfach um 45° . Auch $(\tilde{X}_n)$ ist also für $d = 2$ rekurrent. Für $d \geq 3$ ist die Bestimmung der Rückkehrwahrscheinlichkeiten für $(\tilde{X}_n)$ schwieriger. Man erhält wie für (X_n) Transienz.

(3) *Warteschlangen**: Sei wie in § 15 $Y_0, Y_1, \dots$ eine Folge von unabhängigen Zufallsvariablen mit $P(Y_n = k) = p_k$ ($k \geq 0$), und

$$X_{n+1} = \max(0, X_n - 1) + Y_n.$$

Die rekursive Definition der Kette legt es nahe, mit erzeugenden Funktionen zu arbeiten. Wir brauchen dabei zunächst nicht auf § 7 zurückzugreifen. Sei

$$g(t) = \sum_{k=0}^{\infty} p_k t^k \quad \text{und} \quad f_{ij}(t) = \sum_{n=1}^{\infty} f_{ij}^{(n)} t^n \quad (|t| \leq 1).$$

Satz 16.12 Für die Warteschlange (X_n) gilt:

- (a) $f_{k0}(t) = (f_{10}(t))^k = (f_{00}(t))^k \quad (k \geq 1);$
- (b) $f_{00}(t) = t g(f_{00}(t)).$

Beweis: (a) Ob man von $X_0 = 0$ oder von $X_0 = 1$ startet, macht keinen Unterschied, denn in beiden Fällen ist $X_1 = Y_1$. Daher ist $f_{00}^{(n)} = f_{10}^{(n)}$ für alle $n \geq 1$ und $f_{00}(t) = f_{10}(t)$. (a) gilt also für $k = 1$.

Für $k \geq 2$ zerlegen wir das Ereignis, dass der erste Besuch in 0 zur Zeit n erfolgt, nach dem Zeitpunkt m , zu dem die Länge der Warteschlange erstmals nur $k - 1$ ist:

$$\begin{aligned} f_{k0}^{(n)} &= P_k(X_n = 0, X_{n-1} > 0, \dots, X_1 > 0) \\ &= \sum_{m=1}^{n-1} P_k(X_n = 0, X_{n-1} > 0, \dots, X_{m+1} > 0, X_m = k-1, X_{m-1} \geq k, \dots, X_1 \geq k) \\ &= \sum_{m=1}^{n-1} P(X_n = 0, \dots, X_{m+1} > 0 \mid X_m = k-1) \cdot \\ &\quad \cdot P_k(X_m = k-1, X_{m-1} \geq k, \dots, X_1 \geq k) \\ &= \sum_{m=1}^{n-1} P_{k-1}(X_{n-m} = 0, X_{n-m-1} > 0, \dots, X_1 > 0) f_{k,k-1}^{(m)} \\ &= \sum_{m=1}^{n-1} f_{k-1,0}^{(n-m)} f_{k,k-1}^{(m)}. \end{aligned}$$

Ist $X_0 = k \geq 1$ und m der erste Zeitpunkt mit $X_m = k - 1$, so ist für $t = 1, \dots, m$ $X_t = X_{t-1} - 1 + Y_{t-1}$ und also $X_t = X_0 + (Y_0 + \dots + Y_{t-1}) - t$. Daher ist

$$\begin{aligned} &\{X_0 = k, X_1 \geq k, \dots, X_{m-1} \geq k, X_m = k-1\} \\ &= \{X_0 = k, (Y_0 + \dots + Y_t) \geq t \text{ für } 1 \leq t \leq m-1, Y_0 + \dots + Y_{m-1} = m-1\} \end{aligned}$$

und

$$f_{k,k-1}^{(m)} = P((Y_0 + \dots + Y_{t-1}) \geq t \text{ für } t \leq m-1, Y_0 + \dots + Y_{m-1} = m-1).$$

Da die rechte Seite unabhängig von k ist, ist $f_{k,k-1}^{(m)} = f_{1,0}^{(m)}$. Damit ist

$$f_{k0}^{(n)} = \sum_{m=1}^{n-1} f_{k-1,0}^{(n-m)} f_{10}^{(m)},$$

und dies ist äquivalent zu $f_{k0}(t) = f_{k-1,0}(t) f_{10}(t)$. Damit folgt (a) durch Induktion.

(b): Es ist $f_{00}^{(1)} = p_0$ und für $n \geq 2$

$$f_{00}^{(n)} = \sum_{k=1}^{\infty} P_0(X_1 = k, X_2 > 0, \dots, X_{n-1} > 0, X_n = 0) = \sum_{k=1}^{\infty} p_k f_{k0}^{(n-1)}.$$

Daraus folgt

$$\begin{aligned} f_{00}(t) &= p_0 t + \sum_{n=2}^{\infty} \sum_{k=1}^{\infty} p_k f_{k0}^{(n-1)} t^n = p_0 t + t \sum_{k=1}^{\infty} p_k f_{k0}(t) \\ &= p_0 t + t \sum_{k=1}^{\infty} p_k (f_{00}(t))^k = t g(f_{00}(t)). \end{aligned} \quad \square$$

Eine Rekurrenzbedingung für die Warteschlange*: Nur der Fall $p_0 > 0, p_0 + p_1 < 1$ ist von Interesse. Sonst kann die Warteschlange nicht kürzer, bzw. nicht länger werden. Offenbar ist

$$f_{00}(1) = f_{00}^*.$$

Nach Satz 16.12 (b) folgt, dass f_{00}^* Lösung der Gleichung

$$y = g(y)$$

ist. Diese Gleichung haben wir schon im Beweis von Satz 7.5 betrachtet. (Die Betrachtung ist unabhängig vom Rest von § 7 lesbar.)

Ist $g'(1) \leq 1$, so ist $y = 1$ die einzige Lösung, d.h. die Kette ist dann rekurrent.

$g'(1) = \sum k p_k = E(Y_1)$ ist die erwartete Zahl der Neankömmlinge pro Zeiteinheit. Ist $g'(1) > 1$, so gilt nach dem starken Gesetz der großen Zahlen $Y_1 + \dots + Y_n > n$ für alle hinreichend großen n . Dies schließt $X_n = 0$ aus. Daher ist die Kette im Fall $g'(1) > 1$ transient. Die Länge der Schlange strebt dann nach unendlich. Im Fall $g'(1) \leq 1$ sei $m_{00} = \sum n f_{00}^{(n)}$ die erwartete Dauer bis die Schlange wieder Länge 0 hat. Nach Satz 7.2 ist

$$m_{00} = \lim_{t \rightarrow 1} f'_{00}(t).$$

Nach der Kettenregel erhält man aus Satz 16.12 (b)

$$f'_{00}(t) = g(f_{00}(t)) + t g'(f_{00}(t)) f'_{00}(t)$$

und damit

$$m_{00} = \lim_{t \rightarrow 1} \frac{g(f_{00}(t))}{1 - t g'(f_{00}(t))} \begin{cases} = \infty, & \text{falls } g'(1) = 1, \\ < \infty, & \text{falls } g'(1) < 1. \end{cases}$$

Anhang*

Die Folge der Rückkehrzeiten: Kehren wir noch einmal zum Beweis von Satz 16.8 zurück. Sei i rekurrent. Dann sind mit Wahrscheinlichkeit 1 alle Rückkehrzeiten τ_m endlich.

Sei $\tau_0 = 0$ und $\sigma_m = \tau_m - \tau_{m-1}$ ($m \geq 1$) die für die m -te Rückkehr benötigte Zeit, vom Zeitpunkt des letzten Aufenthalts in i an gerechnet. Seien $s_1, s_2, \dots, s_m$ natürliche Zahlen und n ihre Summe. Ist A die Menge der Folgen $(i_0, \dots, i_{n-1})$, für welche $i_t = i$ für $t = 0, t = s_1, t = s_1 + s_2, \dots, t = s_1 + \dots + s_{m-1}$ und für keine weiteren t gilt, so ist

$$\{X_0 = i, \sigma_1 = s_1, \dots, \sigma_m = s_m\} = \{(X_0, \dots, X_{n-1}) \in A, X_n = i\}.$$

Für jedes $s \in \mathbb{N}$ gilt

$$\begin{aligned} P_i(\sigma_{m+1} = s \mid \sigma_1 = s_1, \dots, \sigma_m = s_m) \\ &= P_i(X_{n+s} = i, X_{n+s-1} \neq i, \dots, X_{n+1} \neq i \mid (X_0, \dots, X_{n-1}) \in A, X_n = i) \\ &= P(X_{n+s} = i, X_{n+s-1} \neq i, \dots, X_{n+1} \neq i \mid X_n = i) = f_{ii}^{(s)}. \end{aligned}$$

Da die bedingten Wahrscheinlichkeiten nicht von $s_1, \dots, s_m$ abhängen, ist σ_{m+1} unabhängig von $(\sigma_1, \dots, \sigma_m)$ und es gilt $P_i(\sigma_{m+1} = s) = f_{ii}^{(s)} = P_i(\sigma_1 = s)$. Wir haben damit den folgenden Satz bewiesen:

Satz 16.13 *Ist i rekurrent, so sind die σ_m ($m \geq 1$) unter P_i unabhängig und identisch verteilt.*

Eine Variante dieser Idee lässt sich nutzen, um viele Resultate über unabhängige identisch verteilte Zufallsvariable auf markowsche Ketten zu übertragen. (Vgl. Freedman (1971).)

Aufgaben

1. Sei eine markowsche Kette mit Zustandsraum $\{1, 2, \dots, 7\}$ durch die folgende Übergangsmatrix gegeben:

$$\begin{pmatrix} 0 & 0 & 1/3 & 0 & 1/2 & 1/6 & 0 \\ 0 & 0 & 0 & 2/5 & 0 & 0 & 3/5 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 2/3 & 0 & 0 & 0 & 1/3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1/4 & 0 & 3/4 & 0 & 0 & 0 \end{pmatrix}.$$

Zeichnen Sie den Übergangsgraphen, d.h. das System der Pfeile, die möglichen Übergängen entsprechen. Welche Zustände sind rekurrent, transient, periodisch?

2. Bestimmen Sie in Aufgabe 1 $\lim p_{ii}^{(n)}$ für die i , für die der Limes existiert. Zeigen Sie, dass es (mindestens) zwei verschiedene invariante Wahrscheinlichkeitsverteilungen gibt. (L)
3. $X_1, X_2, \dots$ sei eine Folge von Zufallsvariablen. S und T seien Stoppzeiten.
 - a) Welche der folgenden Zufallsvariablen sind dann ebenfalls Stoppzeiten:
 $S + T$, $S - T$ (falls $S > T$), $S - 3$ (falls $S > 3$), $S + 3$?
 - b) Ist $T(\omega) = \inf\{n \geq 1 : X_n(\omega) \geq \sup X_k(\omega) - 1\}$ eine Stoppzeit?
4. (**Starke markowsche Eigenschaft**). Man zeige: Ist $X_0, X_1, \dots$ eine homogene markowsche Kette und $T : \Omega \rightarrow \mathbb{Z}^+$ eine Stoppzeit mit $P(T < \infty) = 1$, so ist der Prozess $Y_0, Y_1, \dots$ mit $Y_k(\omega) = X_{T(\omega)+k}(\omega)$ eine markowsche Kette mit den gleichen Übergangswahrscheinlichkeiten. Ereignisse der Form $\{T = n\} \cap \{X_0 = i_0, \dots, X_{n-1} = i_{n-1}\}$ sind bei gegebenem $\{Y_0 = i\}$ unabhängig von $Y_0, Y_1, \dots$.
5. (Fortsetzung). Geben Sie ein Beispiel einer (inhomogenen) markowschen Kette $X_0, X_1, \dots$ und einer Stoppzeit T mit $P(T < \infty) = 1$ an, für die der Prozess (Y_k) aus Aufgabe 4 keine markowsche Kette ist. (L)
6. $Y_1, Y_2, \dots$ seien unabhängig mit $P(Y_i = 1) = p = 1 - P(Y_i = -1)$ ($0 < p < 1$). Zeigen Sie, dass $X_n = Y_1 Y_2 \dots Y_n$ eine markowsche Kette ist, und zeigen Sie ohne Rückgriff auf den Grenzwertsatz für markowsche Ketten, dass $\lim_{n \rightarrow \infty} P(X_n = j)$ existiert.
7. Sei $X_0, X_1, \dots$ (homogene) markowsche Kette. Zeigen Sie, dass für alle $m, n \geq 1$ und für alle $i_\nu \in I$

$$P(X_{n-1} = i_{n-1} \mid X_n = i_n, \dots, X_{n+m} = i_{n+m}) = P(X_{n-1} = i_{n-1} \mid X_n = i_n)$$

gilt, und dass $\tilde{p}_{ij} = P(X_{n-1} = j \mid X_n = i)$ unabhängig von n ist, wenn X_0 eine invariante Verteilung hat. (Die Kette zu $(\tilde{p}_{ij})$ heißt dann die **umgekehrte Kette**.)

8. Mit Hilfe von Satz 15.8 zeige man: Ist (p_{ik}) Übergangsmatrix einer Kette, bei der alle Zustände kommunizieren, so ist $k \in I$ (und damit die Kette) genau dann transient, wenn das Gleichungssystem

$$y_i = \sum_{j \in I} p_{ij} y_j \quad \text{für alle } i \neq k$$

eine beschränkte nichtkonstante Lösung hat. (Anleitung: Ist k transient, so sei

$$y_i = P_i(\exists n \geq 0 : X_n = k).$$

Existiert eine nichtkonstante Lösung, so kann man o.B.d.A. annehmen: $y_k = 1$ und $0 \leq y_i \leq 2$. Ist dann k rekurrent, so ist

$$1 = f_{ik}^* = (H1_{\{k\}})(i) \leq (Hy)(i) = y_i.)$$

9. Seien (X_n) und (Y_n) zwei aperiodische homogene Ketten, die den gleichen endlichen Zustandsraum I haben. Für beide Ketten mögen alle Zustände kommunizieren. Zeigen Sie:

$$P(\exists n \text{ mit } X_n = Y_n) = 1. \quad (\text{L})$$

§ 17 Der Erneuerungssatz

Wir können nun zur Frage der Konvergenz der Übergangswahrscheinlichkeiten $p_{ij}^{(n)}$ zurückkehren. Der Fall, in dem j transient ist, lässt sich nun ziemlich rasch abhaken. Im rekurrenten Fall benötigen wir noch einen „Erneuerungssatz“, der auch von eigenständigem Interesse ist. Die Idee der Erneuerung gehört zu den fruchtbarsten Ideen der Wahrscheinlichkeitstheorie.

17.1 Die Erneuerungsgleichung

Der Schlüssel zu den Resultaten in diesem Abschnitt ist

Satz 17.1 *Für homogene markowsche Ketten gilt*

$$p_{ij}^{(n)} = \sum_{m=1}^n f_{ij}^{(m)} p_{jj}^{(n-m)} \quad (n \geq 1). \quad (17.1)$$

Diese Formel ist ganz plausibel. Wenn man in i startet, muss man, um zur Zeit n in j zu sein, zu einem Zeitpunkt m mit $1 \leq m \leq n$ dort zum ersten Mal eingetroffen sein, und in den verbleibenden $n - m$ Schritten muss man von j nach j zurückgekehrt sein.

Beweis: Man zerlegt $\{X_n = j\}$ auf die angegebene Weise und benutzt $p_{jj}^{(0)} = 1$ und die markowsche Eigenschaft:

$$\begin{aligned} p_{ij}^{(n)} &= \sum_{m=1}^n P_i(X_1 \neq j, \dots, X_{m-1} \neq j, X_m = j, X_n = j) \\ &= \sum_{m=1}^n P_i(X_1 \neq j, \dots, X_{m-1} \neq j, X_m = j) P_i(X_n = j \mid X_m = j) \\ &= \sum_{m=1}^n f_{ij}^{(m)} p_{jj}^{(n-m)}. \end{aligned} \quad \square$$

Folgerung 17.2 *Für transientes j ist $p_{ij}^* < \infty$. Insbesondere gilt für transientes j*

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

Beweis: Es ist

$$p_{ij}^* = \sum_{n=1}^{\infty} \sum_{m=1}^n f_{ij}^{(m)} p_{jj}^{(n-m)} = \sum_{m=1}^{\infty} \sum_{k=0}^{\infty} f_{ij}^{(m)} p_{jj}^{(k)} = f_{ij}^* (1 + p_{jj}^*).$$

Da j transient ist, ist $p_{jj}^* < \infty$. □

Sei nun j rekurrent. Setzen wir in (17.1) $i = j$ und dann $u_n = p_{jj}^{(n)}$ und $f_m = f_{jj}^{(m)}$, so erhalten wir die **Erneuerungsgleichung**

$$u_n = \sum_{m=1}^n f_m u_{n-m} \quad (n \geq 1), \quad u_0 = 1. \quad (17.2)$$

Da j rekurrent ist, ist $(f_m)_{m \geq 1}$ ein Wahrscheinlichkeitsmaß auf $\mathbb{N}$. Die Erneuerungsgleichung tritt auch in Situationen auf, die auf den ersten Blick nichts mit markowschen Ketten zu tun haben. Der Name der Gleichung lässt sich durch das folgende Beispiel erklären:

Sei $T_1, T_2, \dots$ eine Folge unabhängiger, identisch verteilter Zufallsvariabler mit Werten in $\mathbb{N}$. Sei $f_m = P(T_i = m)$ und

$$u_n = P\left(\exists k \geq 0 \text{ mit } \sum_{i=1}^k T_i = n\right). \quad (17.3)$$

Da die leere Summe $= 0$ ist, ist dann $u_0 = 1$. Für $n \geq 1$ ist

$$\begin{aligned} u_n &= \sum_{m=1}^n P\left(T_1 = m, \exists k \geq 1 \text{ mit } \sum_{i=1}^k T_i = n\right) \\ &= P(T_1 = n) + \sum_{m=1}^{n-1} P\left(T_1 = m, \exists k \geq 1 \text{ mit } \sum_{i=2}^k T_i = n - m\right) \\ &= f_n + \sum_{m=1}^{n-1} f_m P\left(\exists k \geq 0 \text{ mit } \sum_{i=1}^k T_{i+1} = n - m\right) = \sum_{m=1}^n f_m u_{n-m}. \end{aligned}$$

Die letzte Gleichung gilt, weil $T_2, T_3, \dots$ die gleiche Verteilung hat wie $T_1, T_2, \dots$.

Wir stellen uns vor, T_1 sei die Lebensdauer einer Glühbirne. In dem Moment, in dem sie durchbrennt, wird sie durch eine zweite Glühbirne mit Lebensdauer T_2 ersetzt, usw. Die k -te Glühbirne muss dann zum Zeitpunkt $T_1 + \dots + T_k$ erneuert werden. u_n ist die Wahrscheinlichkeit, dass gerade im Zeitpunkt n eine Erneuerung stattfindet.

Ist (f_m) eine Wahrscheinlichkeitsverteilung auf $\mathbb{N}$ und gilt (17.2), so setzen wir

$$d_u = \text{ggT}\{n \geq 1 : u_n > 0\} \quad \text{und} \quad d_f = \text{ggT}\{m \geq 1 : f_m > 0\}.$$

Da für alle m mit $f_m > 0$ auch $u_m > 0$ gilt, ist d_u Teiler von d_f . Andererseits sind die Lebensdauern aller Lampen Vielfache von d_f . Eine Erneuerung gibt es also nur zu Zeitpunkten, die Vielfache von d_f sind, d.h. d_f ist auch Teiler von d_u . Daher ist $d_f = d_u$.

Satz 17.3 (Erneuerungssatz) *Sei $(f_m)_{m \geq 1}$ ein Wahrscheinlichkeitsmaß auf $\mathbb{N}$ und $(u_n)_{n \geq 0}$ durch (17.2) gegeben. Sei*

$$\mu = \sum_{m=1}^{\infty} m f_m$$

die Erwartung der Verteilung (f_m) . Ist $d_u = 1$, so konvergiert (u_n) gegen $1/\mu$. (Dabei soll $1/\infty = 0$ sein.)

Beweis: *Schritt 1:* Durch Induktion folgt aus (17.2) $0 \leq u_n \leq 1$. Sei $\lambda = \limsup u_n$ und (n_k) eine Folge mit $n_k \rightarrow \infty$ und $u_{n_k} \rightarrow \lambda$.

Für alle $m \geq 1$ ist

$$\begin{aligned}
\lambda &= \lim_k u_{n_k} = \lim \left(f_m u_{n_k-m} + \sum_{\substack{1 \leq s \leq n_k \\ s \neq m}} f_s u_{n_k-s} \right) \\
&\leq \liminf_k (f_m u_{n_k-m}) + \limsup_k \left(\sum \dots \right) \\
&\leq f_m \liminf_k u_{n_k-m} + \sum_{s \neq m} f_s \limsup_k u_{n_k-s} \leq f_m \liminf_k u_{n_k-m} + (1 - f_m) \lambda.
\end{aligned}$$

Ist $f_m > 0$, so folgt $\lambda \leq \liminf_k u_{n_k-m}$ und damit

$$\lim_k u_{n_k-m} = \lambda. \quad (17.4)$$

Wenden wir dieses Argument mehrfach an, so sehen wir, dass (17.4) für jedes m gilt, für das Zahlen $m_1, \dots, m_k$ mit $m = m_1 + \dots + m_k$ und $f_{m_\nu} > 0$ existieren. Nun ist $d_f = 1$. Nach Lemma 16.6 ist daher jedes hinreichend große m von dieser Form. (Da die m_ν nicht verschieden sein müssen, können wir auf die Faktoren c_ν verzichten.) Es gibt also ein M derart, dass (17.4) für $m \geq M$ gilt.

Schritt 2: Für $n \geq 0$ sei

$$r_n = \sum_{m=n+1}^{\infty} f_m.$$

Dann ist $r_0 = 1$ und

$$\mu = \sum_{m=1}^{\infty} m f_m = \sum_{m=1}^{\infty} f_m + \sum_{m=2}^{\infty} f_m + \dots = \sum_{n=0}^{\infty} r_n.$$

Aus $r_m - r_{m-1} = -f_m$ und der Erneuerungsgleichung folgt

$$r_0 u_n = u_n = - \sum_{m=1}^n (r_m - r_{m-1}) u_{n-m}$$

und, indem man die negativen Terme nach links bringt,

$$r_0 u_n + r_1 u_{n-1} + \dots + r_n u_0 = r_0 u_{n-1} + \dots + r_{n-1} u_0.$$

Nennt man in dieser Gleichung die linke Seite A_n , so ist die rechte A_{n-1} . Aus $A_0 = r_0 u_0 = 1$ erhalten wir damit für alle $n \geq 0$ $A_n = 1$, speziell

$$\sum_{\nu=0}^{n_k-M} r_\nu u_{n_k-M-\nu} = 1. \quad (17.5)$$

Schritt 3: Nun zeigen wir $\lambda = 1/\mu$. Ist $\mu < \infty$, so existiert zu $\varepsilon > 0$ ein N mit $r_0 + r_1 + \dots + r_N \geq \mu - \varepsilon$. Ist k so groß, dass $n_k - M \geq N$ gilt, so ist wegen (17.5)

$$1 \geq \sum_{\nu=0}^N r_\nu u_{n_k-(M+\nu)}, \quad (17.6)$$

und nach (17.4) folgt $1 \geq \lambda(r_0 + \dots + r_N) \geq \lambda(\mu - \varepsilon)$. Da $\varepsilon > 0$ beliebig war, ist $\lambda \leq 1/\mu$.

Andererseits ist wegen (17.5), $u_n \leq 1$ und $(r_{N+1} + r_{N+2} + \dots) \leq \varepsilon$

$$1 \leq \varepsilon + \sum_{\nu=0}^N r_{\nu} u_{n_k - (M+\nu)}.$$

Durch Grenzübergang $k \rightarrow \infty$ finden wir damit $1 \leq \varepsilon + \lambda \mu$ und also $\lambda \geq 1/\mu$. Wir haben somit $\lambda = 1/\mu$ im Fall $\mu < \infty$ gezeigt.

Im Fall $\mu = \infty$ gibt es zu jedem $C > 0$ ein N mit $r_0 + \dots + r_N > C$. Grenzübergang in (17.6) ergibt dann $1 \geq C\lambda$. Da C beliebig groß sein konnte, muss dann $\lambda = 0$ gelten.

Schritt 4: Im Fall $\mu = \infty$ ist der Beweis schon komplett. Im Fall $\mu < \infty$ muss noch $\liminf u_n = 1/\mu$ gezeigt werden. Dies geht völlig analog. Man setzt $\lambda' = \liminf u_n$, wählt eine Teilfolge (n_k) mit $u_{n_k} \rightarrow \lambda'$ usw. $\square$

Die Aussage des Erneuerungssatzes ist einleuchtend: Wenn μ die mittlere Lebensdauer der Glühbirnen ist, wird eine große Zahl n von Glühbirnen etwa bis zur Zeit $n\mu$ reichen. Auf den Zeitraum $n\mu$ kommen also n Erneuerungen, und damit gibt es im Mittel in jedem Zeitpunkt $1/\mu$ Erneuerungen.

17.2 Anwendung auf Übergangswahrscheinlichkeiten

Mit dem Erneuerungssatz lassen sich jetzt die gewünschten Konvergenzaussagen auch im rekurrenten Fall ableiten. Sei zur Abkürzung

$$m_{ii} = \sum_{n=1}^{\infty} n f_{ii}^{(n)}$$

der Erwartungswert der ersten Rückkehrzeit nach i bei Start in i .

Satz 17.4 *Ist i rekurrent und $j \in C_r(i)$, so ist für $n \not\equiv r \pmod{d_i}$ $p_{ij}^{(n)} = 0$ und es gilt:*

$$\lim_{n \rightarrow \infty} p_{ij}^{(nd_i+r)} = \frac{d_i}{m_{jj}}. \quad (17.7)$$

Beweis: Die erste Aussage folgt aus der Definition von $C_r(i)$ und ist nur der Vollständigkeit halber angeführt. Betrachten wir nun zuerst den Fall des Starts in j . Für alle n , die nicht Vielfache von $d_j = d_i$ sind, ist $f_{jj}^{(n)} = 0$ und $p_{jj}^{(n)} = 0$. Also ist

$$p_{jj}^{(nd_i)} = \sum_{m=1}^n f_{jj}^{(md_i)} p_{jj}^{((n-m)d_i)}.$$

Setzt man $u_n = p_{jj}^{(nd_i)}$ und $f_m = f_{jj}^{(md_i)}$, so gilt also die Erneuerungsgleichung und $d_u = 1$. Es ist

$$\mu = \sum_{m=1}^{\infty} m f_{jj}^{(md_i)} = \frac{m_{jj}}{d_i}.$$

Nach dem Erneuerungssatz konvergieren also die $p_{jj}^{(nd_i)}$ gegen d_i/m_{jj} . Die Behauptung folgt nun aus

$$p_{ij}^{(nd_i+r)} = \sum_{m=0}^n f_{ij}^{(md_i+r)} p_{jj}^{((n-m)d_i)}$$

und

$$\sum_{m=0}^{\infty} f_{ij}^{(md_i+r)} = 1. \quad \square$$

Wir sagen, I bestehe aus einer rekurrenten Klasse, wenn alle Zustände kommunizieren und rekurrent sind. In diesem Fall ergibt sich aus Satz 17.4 auch eine Aussage für beliebige Startverteilungen.

Folgerung 17.5 *Besteht I aus einer rekurrenten Klasse, so gilt im aperiodischen Fall ($d_i = 1$)*

$$P(X_n = j) \rightarrow \frac{1}{m_{jj}}$$

und im allgemeinen Fall

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^{N-1} P(X_n = j) = \frac{1}{m_{jj}}.$$

Beweis: Ist $d_i = 1$, so besagt Satz 17.4, dass $P_i(X_n = j) \rightarrow 1/m_{jj}$ gilt. Wegen $P = \sum \pi_i P_i$ folgt daraus die erste Behauptung.

Nun überlegt man sich: Ist (a_n) eine gegen a konvergente Folge und definiert man für ein $d > 1$ und $0 \leq r < d$ die Folge (b_m) durch

$$b_m = \begin{cases} 0, & \text{falls } m \not\equiv r \pmod{d}, \\ a_n, & \text{falls } m = nd + r, \end{cases}$$

so gilt $N^{-1}(b_1 + b_2 + \dots + b_N) \rightarrow a/d$. Wendet man dies mit $a_n = p_{ij}^{(nd_i+r)}$ und $d = d_i$ an, so folgt die Behauptung für $P = P_i$. Der allgemeine Fall folgt dann wieder aus $P = \sum \pi_i P_i$. $\square$

Definition 17.6 i heißt **positiver Zustand** oder ausführlicher **positiv rekurrenter Zustand**, wenn $\limsup p_{ii}^{(n)} > 0$ ist, und **Nullzustand** sonst. i heißt **nullrekurrent**, wenn i rekurrenter Nullzustand ist.

Transiente Zustände sind nach Folgerung 17.2 Nullzustände. Ähnlich wie in Folgerung 16.11 überlegt man sich leicht, dass Zustände, die mit einem positiven Zustand kommunizieren, positiv sind. Nach den obigen Resultaten gibt es drei in ihrem Rückkehrverhalten wesentlich verschiedene Arten von Zuständen:

- Ist i transient, so kehrt man von i startend mit positiver Wahrscheinlichkeit nicht nach i zurück. Die Wahrscheinlichkeit, dass man unendlich oft zurückkehrt, ist 0.
- Ist i nullrekurrent, so kehrt man mit Wahrscheinlichkeit 1 unendlich oft nach i zurück, aber der Erwartungswert m_{ii} der Rückkehrdauer ist ∞ .
- Ist i positiv rekurrent, so kehrt man von i startend mit Wahrscheinlichkeit 1 unendlich oft nach i zurück, und m_{ii} ist endlich.

Satz 17.7 *Ist I endlich, so ist jeder wesentliche Zustand positiv.*

Beweis: Sei i wesentlich. Wir können annehmen, dass I gleich $C(i)$ sei, da man von Zuständen in $C(i)$ ausgehend stets in $C(i)$ bleibt. Da für alle n $\sum_j p_{ij}^{(n)} = 1$ ist, muss ein j mit $\limsup p_{ij}^{(n)} > 0$ existieren. Für ein m ist $p_{ji}^{(m)} > 0$. Wegen $p_{ii}^{(n+m)} \geq p_{ij}^{(n)} p_{ji}^{(m)}$ ist dann i positiv. $\square$

17.3 Bestimmung der m_{ii}

Die Berechnung der m_{ii} aus der Definitionsgleichung ist bis auf wenige Ausnahmen unmöglich. Ein Vergleich mit dem zuerst studierten Fall endlicher Zustandsräume legt einen anderen Zugang nahe. Dort hatten wir den Grenzwert ρ_j der Folge $(p_{ij}^{(n)})_{n \geq 1}$ als Lösung des Gleichungssystems

$$\rho_k = \sum_{j \in I} \rho_j p_{jk} \quad (k \in I) \quad (17.8)$$

bestimmt. Ein Wahrscheinlichkeitsmaß $\rho = (\rho_j)$ auf I , das (17.8) erfüllt, heißt **invariant**. Wir wissen *jetzt*, dass der Grenzwert in Satz 16.1 den Wert $1/m_{jj}$ hat. Ist ρ die Lösung von (17.8), so ist also

$$m_{jj} = \frac{1}{\rho_j}. \quad (17.9)$$

Mit Hilfe des Erneuerungssatzes haben wir also einen einfachen Weg zur Berechnung von m_{jj} gefunden. Erfreulicherweise gilt (17.9) nicht nur unter den Bedingungen von Satz 16.1.

Satz 17.8 *Kommunizieren alle Zustände in I miteinander, so existiert ein invariantes Wahrscheinlichkeitsmaß ρ genau dann, wenn die Kette positiv rekurrent ist. Es ist dann eindeutig bestimmt, und es gilt (17.9).*

Beweis: Aus $\rho = \rho \mathbb{P}$ folgt wieder $\rho = \rho \mathbb{P}^n$, d.h. es gilt

$$\rho_j = \sum_i \rho_i p_{ij}^{(n)}. \quad (17.10)$$

Ist die Kette transient oder nullrekurrent, so gilt $p_{ij}^{(n)} \rightarrow 0$. Aus (17.9) folgt dann $\rho_j = 0$. Die Existenz eines invarianten Wahrscheinlichkeitsmaßes impliziert also die positive Rekurrenz der Kette.

Sei nun die Kette positiv rekurrent und $\rho_j = 1/m_{jj}$. Nach Folgerung 17.5 ist dann

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^{N-1} p_{ij}^{(n)} = \rho_j > 0. \quad (17.11)$$

Für endliches $J \subset I$ ist dann

$$\sum_{j \in J} \rho_j = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^{N-1} \sum_{j \in J} p_{ij}^{(n)} \leq 1.$$

Da J beliebig war, ist $\sum_{i \in I} \rho_i \leq 1$. Für endliches J gilt auch

$$\begin{aligned} \sum_{j \in J} \rho_j p_{jk} &= \sum_{j \in J} \left(\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^{N-1} p_{ij}^{(n)} \right) p_{jk} \\ &= \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^{N-1} \left(\sum_{j \in J} p_{ij}^{(n)} p_{jk} \right) \leq \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^{N-1} p_{ik}^{(n+1)} = \rho_k. \end{aligned}$$

Da auch hierin J beliebig war, folgt

$$\sum_{j \in I} \rho_j p_{jk} \leq \rho_k. \quad (17.12)$$

Wegen

$$\sum_k \sum_j \rho_j p_{jk} = \sum_j \sum_k \rho_j p_{jk} = \sum_j \rho_j$$

muss in (17.12) für alle k Gleichheit gelten, d.h. ρ ist invariant. Damit gilt auch wieder (17.10). Durch Mittelung über $n = 0, \dots, N-1$ erhalten wir

$$\rho_j = \sum_i \rho_i \left(\frac{1}{N} \sum_{n=0}^{N-1} p_{ij}^{(n)} \right). \quad (17.13)$$

Aus (17.11) ergibt sich damit $\rho_j = \sum_i \rho_i \rho_j$. Wegen $\rho_j > 0$ ist $\sum_i \rho_i = 1$. ρ ist also invariantes Wahrscheinlichkeitsmaß. Ist $\tilde{\rho}$ ein weiteres invariantes Wahrscheinlichkeitsmaß, so gilt (17.13) auch für $\tilde{\rho}$. Durch Grenzübergang $N \rightarrow \infty$ erhalten wir dann $\tilde{\rho}_j = \sum_i \tilde{\rho}_i \rho_j$, also $\rho = \tilde{\rho}$. $\square$

Satz 17.8 gibt natürlich indirekt auch über den nicht positiv rekurrenten Fall Auskunft: Existiert kein invariantes Wahrscheinlichkeitsmaß, so gilt $p_{ij}^{(n)} \rightarrow 0$.

Beispiele: (1) (*Ehrenfests Modell der Wärmebewegung*) Ein Behälter ist in zwei Teile I und II geteilt, zwischen denen wie in Bild 17.1 eine kleine Öffnung existiert. Insgesamt enthält der Behälter N Gasmoleküle. Durch die Wärmebewegung können Moleküle aus dem Bereich I in den Bereich II gelangen oder umgekehrt. Der Zustand des Systems ist durch die Anzahl i der Moleküle in Bereich I beschrieben. Für jedes Molekül soll es gleich wahrscheinlich sein, dass es als Nächstes durch die Öffnung fliegt.

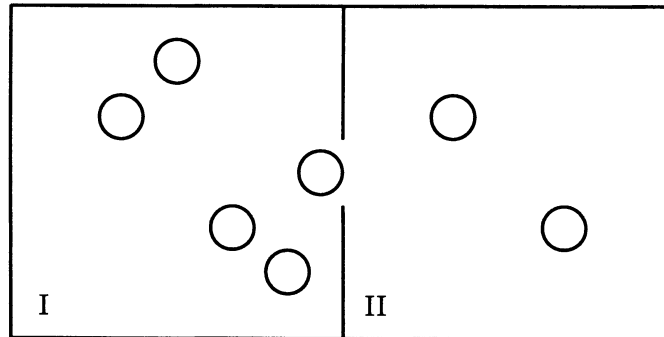


Bild 17.1 Ehrenfests Modell

Verwenden wir als Zeitskala die Zahl der Zustandswechsel, so gibt es also nur Übergänge $i \rightarrow i+1$ und $i \rightarrow i-1$, und es ist

$$p_{i,i+1} = \frac{N-i}{N}, \quad p_{i,i-1} = \frac{i}{N}.$$

In diesem Beispiel lässt sich die invariante Verteilung erraten: Nach einem langen Zeitraum dürfte für jedes der Gasmoleküle unabhängig von den anderen die Wahrscheinlichkeit, dass es im Bereich I ist, etwa $1/2$ sein. Die Wahrscheinlichkeit, dass j Moleküle im Bereich I sind, wäre dann ungefähr $\rho_j = \binom{N}{j} 2^{-N}$. Man rechnet leicht nach, dass das so definierte ρ tatsächlich invariant ist. Daher ist $m_{jj} = 2^N / \binom{N}{j}$.

Dies bedeutet: Beginnt man in Bereich I mit vollkommenem Vakuum und in Bereich II mit N Molekülen, so ist der Erwartungswert der Zahl der Zeiteinheiten, nach denen Bereich I wieder leer ist, gleich 2^N . Dies wächst mit N so schnell, dass bei tatsächlich auftretenden Molekülzahlen dieser Erwartungswert einen unvorstellbar hohen Wert besitzt. Zwar wird mit Wahrscheinlichkeit 1 der Bereich I irgendwann wieder leer sein, aber es lohnt nicht, darauf zu warten.

(Da die Kette die Periode $d_i = 2$ hat, konvergieren die Übergangswahrscheinlichkeiten nicht. Falls $|i - j|$ gerade ist, gilt $p_{ij}^{(2n+1)} = 0$ und $p_{ij}^{(2n)} \rightarrow 2\rho_j$. Bei ungeradem $|i - j|$ gilt $p_{ij}^{(2n)} = 0$ und $p_{ij}^{(2n+1)} \rightarrow 2\rho_j$.)

(2) (*Geburts- und Todesprozess*) Als **Geburts- und Todesprozess** bezeichnet man eine homogene markowsche Kette mit Zustandsraum $I = \{0, 1, 2, \dots\}$, für die man in einem Schritt von i nur zu den Zuständen $i - 1, i, i + 1$ kommen kann. Man stellt sich vor, dass der Zustand i die Zahl der Individuen einer Population beschreibt. Ein Übergang $i \rightarrow i - 1$ entspricht einem Todesfall, und ein Übergang $i \rightarrow i + 1$ einer Geburt.

Seien $p_i, q_i, r_i \geq 0$ mit $p_i + q_i + r_i = 1$, $q_0 = 0$ und $p_{i,i+1} = p_i$, $p_{i,i-1} = q_i$, $p_{ii} = r_i$. Hier betrachten wir der Einfachheit halber nur den Fall $r_i = 0$. (Der allgemeine Fall lässt sich ähnlich behandeln.)

Das Gleichungssystem (17.8) hat jetzt die Form

$$\begin{aligned}\rho_0 &= p_{10} \rho_1 = q_1 \rho_1, \\ \rho_j &= q_{j+1} \rho_{j+1} + p_{j-1} \rho_{j-1} \quad (j \geq 1).\end{aligned}$$

Wenn man ρ_0 kennt, kann man daraus die ρ_j ($j \geq 1$) sukzessive bestimmen. Man beginnt mit $\rho_1 = \rho_0 / q_1$. Dies setzt man in $\rho_1 = q_2 \rho_2 + p_0 \rho_0 = q_2 \rho_2 + \rho_0$ ein und erhält

$$\rho_2 = \frac{1}{q_2} \left(\frac{\rho_0}{q_1} - \rho_0 \right) = \frac{\rho_0 p_0 p_1}{q_1 q_2}.$$

So wird man auf die Lösung

$$\rho_j = \rho_0 \prod_{k=0}^{j-1} \frac{p_k}{q_{k+1}} \quad (17.14)$$

geführt, die man induktiv bestätigt. Wenn es eine Lösung gibt, muss sie die Form (17.14) haben. Aber im Allgemeinen braucht die so bestimmte Folge keine konvergente Summe zu haben. Dafür ist

$$\sigma := \sum_{j=1}^{\infty} \prod_{k=0}^{j-1} \frac{p_k}{q_{k+1}} < \infty \quad (17.15)$$

notwendig und hinreichend. Gilt (17.15), so definiert man $p_0 = 1/(1 + \sigma)$, damit $\rho = (\rho_j)$ ein Wahrscheinlichkeitsmaß ist, und hat damit die ρ_j bestimmt. Ist etwa für $k \geq 1$ $0 < p_k = p$ und $0 < q_k = 1 - p$ unabhängig von k , so prüft man nach, dass $p < q$ notwendig und hinreichend für (17.15) und damit für positive Rekurrenz ist.

(3) (*Warteschlangen*) Am Ende von § 16 hatten wir gezeigt, dass im Fall $p_0 > 0$, $p_0 + p_1 < 1$ der Erwartungswert m_{00} der Dauer bis zur Rückkehr nach 0 genau dann endlich ist, wenn $\sum k p_k < 1$ ist, also im Mittel weniger als ein Skifahrer pro Zeiteinheit zum Lift kommt. Dies ist auch notwendig und hinreichend für positive Rekurrenz. Man kann sich auch direkt überlegen, dass genau in diesem Fall ein invariantes Wahrscheinlichkeitsmaß ρ existiert.

Anmerkung: In der Erneuerungstheorie studiert man Prozesse, die zu bestimmten zufälligen Zeiten gewissermaßen neu beginnen, unabhängig vom Verlauf vor diesen Zeitpunkten. Sie brauchen nicht markowsch zu sein. Die Grundidee spielt schon in Arbeiten von Palm und Doeblin in den dreißiger und vierziger Jahren eine Rolle. Der hier bewiesene „arithmetische“ Fall des Erneuerungssatzes geht auf die Arbeit von Erdős, Feller und Pollard aus dem Jahr 1943 zurück, kann aber auch auf frühere Resultate von Kolmogorow zurückgeführt werden. Auch der Fall, in dem nicht mit Wahrscheinlichkeit 1 Erneuerungen eintreten (etwa die Rückkehr zu einem transienten Zustand), ist von Interesse. Man betrachtet dann unabhängige identisch verteilte Zufallsvariable $T_1, T_2, \dots$ mit Werten in $\mathbb{N} \cup \{\infty\}$ und lässt zu, dass die Summe $f = \sum_{i=1}^{\infty} f_i$ der Zahlen $f_i = P(T_1 = i)$ kleiner als 1 ist. Ist wieder $u_0 = 1$ und ist u_n für $n \geq 1$ rekursiv durch (17.2) definiert, so ist u_n wieder die Wahrscheinlichkeit, dass im Zeitpunkt n eine Erneuerung stattfindet. Erzeugende Funktionen sind äußerst nützlich bei diesen Untersuchungen, vgl. Aufg. 4. Studiert man Prozesse mit stetiger Zeit, so besteht der Wertebereich der T_i im Allgemeinen nicht mehr nur aus ganzen Zahlen. Ein Beispiel ist der Poisson-Prozess aus § 18.

Aufgaben

1. $(p_0, p_1, \dots)$ sei eine Wahrscheinlichkeitsverteilung auf $I = \{0, 1, 2, \dots\}$. Wir definieren eine Übergangsmatrix durch $p_{0j} = p_j$ ($j = 0, 1, \dots$), $p_{i, i-1} = 1$ ($i \geq 1$), und $p_{ij} = 0$ sonst. Unendlich viele p_j seien positiv. Unter welcher Bedingung existiert eine invariante Wahrscheinlichkeitsverteilung und welche Form hat sie?
2. Sei folgendes Zellteilungsmodell gegeben: Eine Zelle bestehe aus N Teilen, davon i der Sorte A , $N - i$ der Sorte B . Eine neue Zelle entsteht nach folgendem Mechanismus:
 - a) Jedes Teil verdoppelt sich; es gibt dann also $2i$ Teile der Sorte A und $2N - 2i$ Teile der Sorte B .
 - b) Aus diesen $2N$ Teilen werden N zufällig ausgewählt und bilden die neue Zelle.
 Man berechne die Wahrscheinlichkeit p_{ij} , dass die neue Zelle j Teile von Sorte A hat. Welche Zustände sind rekurrent bzw. transient? Berechnen Sie $\lim p_{ij}^{(n)}$.
3. Zeigen Sie für das ehrenfestsche Urnenmodell: Bei invariantem Startmaß gilt

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

(Diese Eigenschaft bzgl. Zeitumkehr nennt man **Reversibilität**.)

4. Seien $T_1, T_2, \dots$ unabhängige identisch verteilte Zufallsvariable mit Werten in $\mathbb{N} \cup \{\infty\}$. Ausgehend von $f_k = P(T_i = k)$ und $u_0 = 1$ sei u_n für $n \geq 1$ rekursiv durch (17.2) definiert. Sei $g_f(t) = \sum_{k=1}^{\infty} f_k t^k$ und $g_u(t) = \sum_{i=0}^{\infty} u_i t^i$ für $0 \leq t < 1$. Zeigen Sie: $g_u = 1/(1 - g_f)$. Zeigen Sie außerdem, dass $f = \sum_{k=1}^{\infty} f_k < 1$ genau dann gilt, wenn $u = \sum_{i=0}^{\infty} u_i$ endlich ist, und dass in diesem Fall $f = (u - 1)/u$ ist. (L)
5. Zu den Zeitpunkten $-2, -1, 0, 1, 2, \dots$ trifft jeweils mit Wahrscheinlichkeit p ein Teilchen auf ein Zählgerät. Dieses registriert einen Ausschlag, wenn ein Teilchen auftrifft und der Zähler unblockiert ist. Er ist blockiert, wenn in einem der letzten beiden Zeitpunkte ein Teilchen registriert wurde. Bestimmen Sie rekursiv die Wahrscheinlichkeiten w_n , dass zum Zeitpunkt n der Zähler unblockiert ist, und berechnen Sie $\lim w_n$. (L)

§ 18 Der Poisson-Prozess

Wir diskutieren nun eins der einfachsten Beispiele einer markowschen Kette mit stetiger Zeit, den Poisson-Prozess, der u.a. als Modell für die Beobachtung des radioaktiven Zerfalls dienen kann. Wir setzen nur § 10 und § 11, nicht aber die obigen Resultate über markowsche Ketten voraus.

18.1 Charakterisierung des Poisson-Prozesses

In Bild 18.1 sind die Zeitpunkte, an denen ein Geigerzähler eine radioaktive Emission registriert, als Kreuze auf der Zeitachse eingetragen.

Für ein Intervall I sei X_I die Anzahl der Emissionen im Zeitintervall I , und es sei $X_t = X_{]0, t]}$. Offenbar gilt nach dieser Interpretation:

(A0) *Die Zufallsvariablen X_t haben Werte in $\{0, 1, 2, \dots\} \cup \{\infty\}$, und X_t ist als Funktion von t monoton wachsend und rechtsstetig. Ferner ist $X_0 = 0$.*

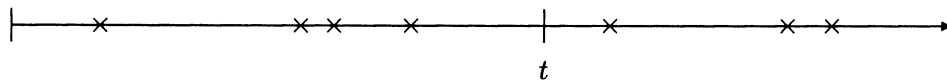


Bild 18.1 Zeitpunkte, an denen eine Emission registriert wird

Darüber hinaus gehen wir von den folgenden Eigenschaften dieses „Prozesses“ aus, die die experimentelle Erfahrung beschreiben.

- (A1) *Sind $I_1, I_2, \dots, I_r$ disjunkte Intervalle, so sind die Ereignisse $\{X_{I_i} = 0\}$ unabhängig.*
- (A2) *Sind I und I' gleich lange Intervalle, so gilt $P(X_I = 0) = P(X_{I'} = 0)$.*
- (A3) *Für alle endlichen Intervalle I ist $P(X_I < \infty) = 1$.*
- (A4) *Die Wahrscheinlichkeit dafür, dass 2 oder mehr Emissionen zum exakt gleichen Zeitpunkt auftreten, ist gleich 0.*

Wir nennen eine Familie $(X_t, t \geq 0)$, die diesen Bedingungen genügt, einen **Poisson-Prozess**.

Verschieben wir für einen Moment die Formalisierung des Modells. Zuerst wollen wir die Verteilung von X_t bestimmen. Sei $p(t) = P(X_t \geq 1)$, $q(t) = 1 - p(t)$ und $\mathbf{q} = q(1)$. Da X_t monoton wächst, ist auch $p(t)$ eine monoton wachsende Funktion von t .

Teilt man das Intervall $]0, k/n]$ in k disjunkte Teilintervalle der Länge $1/n$, so sieht man, dass aus (A1) und (A2) $q(k/n) = q(1/n)^k$ folgt. Insbesondere ist $\mathbf{q} = q(1/n)^n$. Also ist $q(k/n) = \mathbf{q}^{k/n}$. Aus Monotoniegründen folgt $q(t) = \mathbf{q}^t$ für alle $t > 0$.

Wäre $\mathbf{q} = 0$, so wäre für jedes $t > 0$ $q(t) = 0$, also $p(t) = 1$. Mit Wahrscheinlichkeit 1 enthielte dann jedes der Intervalle $]0, t/n]$, $]t/n, 2t/n]$, $\dots$, $](n-1)t/n, t]$ mindestens eine Emission. Also gäbe es mindestens n , und da n beliebig war, unendlich viele Emissionen in $]0, t]$. Dies widerspräche (A3). Also ist $\mathbf{q} > 0$.

Aus $p(t) = 1 - \mathbf{q}^t$ folgt

$$\lim_{t \rightarrow 0} \frac{p(t)}{t} = \lim_{t \rightarrow 0} \frac{\mathbf{q}^0 - \mathbf{q}^t}{t} = - \frac{d}{dt} \mathbf{q}^t \Big|_{t=0} = - \mathbf{q}^t \log \mathbf{q} \Big|_{t=0} = - \log \mathbf{q}.$$

Wir untersuchen zunächst den Fall $0 < \mathbf{q} < 1$ und setzen $\lambda = -\log \mathbf{q}$. Dann ist $\lambda > 0$. Wir betrachten nun noch einmal die obige Unterteilung von $]0, t]$ in n gleich lange halboffene disjunkte Intervalle der Länge t/n . Sei $Y_{n,t}$ die Anzahl der Teilintervalle, die mindestens eine Emission enthalten. Dann ist

$$P(Y_{n,t} = k) = \binom{n}{k} p(t/n)^k (1 - p(t/n))^{n-k}. \quad (18.1)$$

Aus $p(t/n)/(t/n) \rightarrow \lambda$ folgt $n \cdot p(t/n) \rightarrow \lambda t$. $P(Y_{n,t} = k)$ strebt also nach der Poisson-Approximation (Folgerung 5.10) für $n \rightarrow \infty$ gegen $e^{-\lambda t} (\lambda t)^k / k!$.

Nun treten aber nach (A4) keine Emissionen gleichzeitig auf. Für großes n ist daher die Zahl der Teilintervalle, die mindestens eine Emission enthalten, gleich der Zahl der Emissionen in $]0, t]$, also $= X_t$. Mit anderen Worten: Es gilt fast sicher $Y_{n,t} \rightarrow X_t$. Es leuchtet ein, dass hieraus

$$P(Y_{n,t} = k) \rightarrow P(X_t = k) \quad (18.2)$$

folgt. Wir stellen den formalen Beweis dieser Aussage kurz zurück. Aus (18.2) und unseren vorangegangenen Überlegungen folgt

$$P(X_t = k) = e^{-\lambda t} \frac{(\lambda t)^k}{k!}. \quad (18.3)$$

Diese Formel bleibt auch im Fall $\mathbf{q} = 1$ richtig. Dann ist $\lambda = 0$.

Sei nun T_1 der Zeitpunkt, zu dem die erste Emission registriert wird. Dann bedeutet $T_1 \in]a, b]$, dass $X_a = 0$ gilt, aber nicht $X_b = 0$. Wegen $\{X_b = 0\} \subset \{X_a = 0\}$ folgt

$$P(T_1 \in]a, b]) = P(X_a = 0) - P(X_b = 0) = e^{-\lambda a} - e^{-\lambda b} = \int_a^b \lambda e^{-\lambda t} dt$$

für $0 \leq a < b < \infty$. X_t hat also eine Poisson-Verteilung mit Parameter λt , und T_1 hat eine Exponentialverteilung mit Parameter λ .

Hätten wir statt des Intervalls $]0, t]$ ein anderes Intervall I der Länge t zerlegt, so hätte die gleiche Überlegung den in (18.3) angegebenen Wert auch für $P(X_I = k)$ ergeben. Wir sehen also, dass aus (A0)–(A4) die folgende stärkere Form von (A2) folgt, die zugleich (A3) und (A4) enthält:

(A2') *Ist I ein beliebiges Intervall der Länge t , so hat X_I eine Poisson-Verteilung mit Parameter λt .*

Auch für (A1) lässt sich eine stärkere Fassung herleiten. Es gilt:

(A1') *Sind $I_1, \dots, I_r$ disjunkte Intervalle, so sind $X_{I_1}, \dots, X_{I_r}$ unabhängig.*

Wir überlassen den Beweis dem Leser als Übung.

Ein Poisson-Prozess lässt sich also auch als Prozess definieren, der die Eigenschaften (A0), (A1') und (A2') hat. λ nennt man den Parameter des Poisson-Prozesses.

[Nebenbei sei vermerkt, dass sich aus (A1') und (A2') die **markowsche Eigenschaft** folgern lässt, die im Fall stetiger Zeit besagt, dass für alle $t_1 < t_2 < \dots < t_{n+1}$ und alle Zustände $i_1, \dots, i_{n+1}$

$$P(X_{t_{n+1}} = i_{n+1} | X_{t_1} = i_1, \dots, X_{t_n} = i_n) = P(X_{t_{n+1}} = i_{n+1} | X_{t_n} = i_n)$$

gelten soll.]

Nun zum Nachweis von (18.2): Sei für $k \geq 1$

$$S_k(\omega) = \inf\{t > 0 : X_t(\omega) \geq k\}$$

der Zeitpunkt der k -ten Emission. (S_k ist wegen $\{S_k > \alpha\} = \cup\{X_t < k, t > \alpha, t \in \mathbb{Q}\}$ messbar.) Sei $B(t, \delta)$ die Menge der ω mit $|S_i(\omega) - S_{i-1}(\omega)| > \delta$ für alle $i \geq 2$ mit $S_i(\omega) \leq t$, also die Menge der ω , für die der Abstand von je zwei Emissionen im Intervall $]0, t]$ größer als δ ist. Die Folge der Mengen $B(t, t/n)$ ist wachsend, und ihre Vereinigung hat Wahrscheinlichkeit 1. Nach Satz 10.5 gilt also $P(B(t, t/n)) \rightarrow 1$ für $n \rightarrow \infty$. Für $\omega \in B(t, t/n)$ ist $Y_{n,t}(\omega) = X_t(\omega)$. Also gilt

$$P(Y_{n,t} \neq X_t) \leq 1 - P(B(t, t/n)) \rightarrow 0, \quad (18.4)$$

und daraus folgt (18.2).

Existenz von Poisson-Prozessen: Wir skizzieren nun noch einen einfachen Beweis der Existenz von Poisson-Prozessen. Wir setzen die Existenz von Folgen von unabhängigen Zufallsvariablen mit gegebenen Verteilungen voraus. (Sie lässt sich aus der Existenz der Gleichverteilung auf $[0, 1]$ herleiten.)

Es genügt einen Poisson-Prozess mit Zeitbereich $]0, 1]$ zu konstruieren, denn aus einer Folge von unabhängigen Prozessen dieser Art lässt sich auf nahe liegende Weise einer mit Zeitbereich $]0, \infty[$ bilden.

Sei S eine $P(\lambda)$ -verteilte Zufallsvariable, seien $Y_1, Y_2, \dots$ gleichverteilt in $]0, 1]$, und seien $S, Y_1, Y_2, \dots$ unabhängig. Bezeichne X_B die Anzahl der $i \leq S$ mit $Y_i \in B$. Wir wollen uns überlegen, dass für disjunkte borelsche Mengen $B_1, \dots, B_r \subset]0, 1]$ die X_{B_j} unabhängig sind. Dazu können wir — notfalls durch Hinzunahme einer weiteren borelschen Menge — annehmen, die Vereinigung der B_j sei $]0, 1]$. Nach Voraussetzung ist $p_j = P(Y_m \in B_j)$ unabhängig von m . Sei $Z_{j,n}$ die Anzahl der $i \leq n$ mit $Y_i \in B_j$. Ist $k_1 + \dots + k_r = n$ und $k_j \in \mathbb{Z}^+$, so ist

$$\begin{aligned} P(X_{B_1} = k_1, \dots, X_{B_r} = k_r) &= P(S = n, Z_{1,n} = k_1, \dots, Z_{r,n} = k_r) \\ &= P(S = n) P(Z_{1,n} = k_1, \dots, Z_{r,n} = k_r) \\ &= e^{-\lambda} \frac{\lambda^n}{n!} \cdot \frac{n!}{k_1! \dots k_r!} p_1^{k_1} \dots p_r^{k_r} \\ &= \prod_{j=1}^r e^{-\lambda p_j} \frac{(\lambda p_j)^{k_j}}{k_j!}. \end{aligned}$$

Daraus folgt (A1') und erst recht (A1), sogar für borelsche Mengen und nicht nur für Intervalle. Die anderen Eigenschaften (A0), (A2), ... sind trivial. Durch $X_t = X_{]0, t]}$ ist also ein Poisson-Prozess definiert.

[Der Poisson-Prozess ist das einfachste Beispiel eines Punktprozesses, d.h. eines Prozesses von zufälligen Punkten auf der Zeitachse oder im Raum. Eine umfassende Darstellung (mit historischen Hinweisen) findet sich in dem Buch von Daley und Vere-Jones (1988).]

18.2 Sprungzeiten beim Poisson-Prozess*

Sei wieder S_i der Zeitpunkt des i -ten Sprungs des Poisson-Prozesses (X_t) . Setzt man $T_1 = S_1$ und $T_i = S_i - S_{i-1}$ für $i \geq 2$, so ist also T_i die Verweildauer des Prozesses (X_t) im Zustand $i - 1$. Es ist $S_i = T_1 + T_2 + \dots + T_i$. Unser Ziel ist jetzt:

Satz 18.1 *Ist $(X_t, t \geq 0)$ ein Poisson-Prozess mit Parameter $\lambda > 0$, so sind die Zufallsvariablen $T_1, T_2, \dots$ unabhängig und exponentialverteilt mit Parameter λ .*

Beweis: Wir benutzen wieder eine diskrete Approximation, müssen aber nun verschiedene t simultan betrachten und verwenden daher nicht Teilintervalle der Länge t/n , sondern der Länge $1/n$.

Bezeichnet $[tn]$ die größte ganze Zahl $\leq tn$, so ist $[tn]$ die Zahl der Intervalle der Form $[(i-1)/n, i/n]$ mit $i \in \mathbb{N}$, die ganz in $]0, t]$ enthalten sind. Sei $X_{n,t}$ die Zahl derjenigen unter diesen Intervallen, die mindestens eine Emission enthalten.

In Bild 18.2 sind auf der Zeitachse die Zeitpunkte der beobachteten Emissionen mit einem kleinen Kreuz markiert. Da X_t die Zahl der Emissionen in $]0, t]$ zählt und niemals zwei Emissionen simultan vorkommen sollen, ist $X_0 = 0$, und X_t macht Sprünge der Höhe 1 nach oben zu den markierten Zeitpunkten.

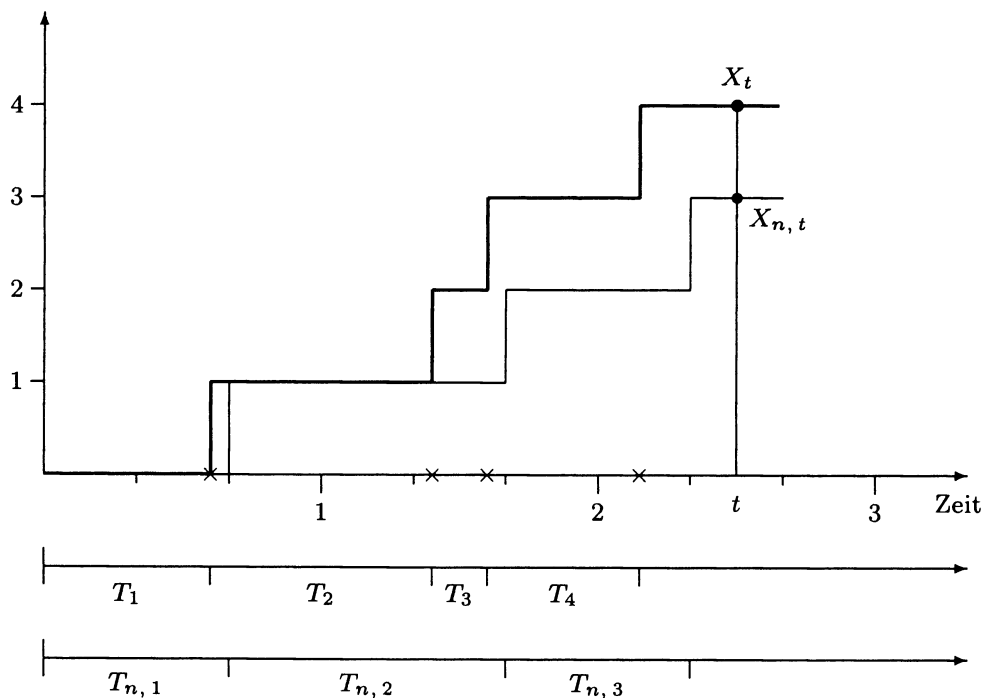


Bild 18.2 Der Poisson-Prozess und seine diskrete Approximation für $n = 3$

Der Prozess $(X_{n,t})$ beginnt ebenfalls bei $X_{n,0} = 0$. Die möglichen Sprungzeiten sind nun die diskreten Zeitpunkte k/n , und zwar erfolgt ein Sprung der Höhe 1, wenn X_t in $[(k-1)/n, k/n]$ mindestens einen Sprung macht, evtl. auch mehr. Seien $T_{n,1}, T_{n,1} + T_{n,2}, \dots$ die Sprungzeiten des Prozesses $(X_{n,t})$.

Setzt man $\lambda_n = np(1/n)$, so gilt $\lambda_n \rightarrow \lambda$. Für $t > 0$ gilt $T_{n,1} > t$ offenbar genau dann, wenn die oben genannten $[tn]$ in $]0, t]$ enthaltenen Teilintervalle alle keine Emissionen enthalten. Daher ist

$$P(T_{n,1} > t) = (1 - p(1/n))^{[tn]} = (1 - \lambda_n/n)^{[nt]}.$$

Aus $[nt] \sim nt$ folgt daher

$$P(T_{n,1} > t) \rightarrow e^{-\lambda t} = P(T_1 > t). \quad (18.5)$$

Offenbar gilt für $j \geq 1$

$$T_1 + \dots + T_j \leq T_{n,1} + \dots + T_{n,j}. \quad (18.6)$$

Ist n so groß, dass im Zeitintervall $]0, t]$ der Abstand zwischen je zwei Emissionen größer ist als $1/n$, so macht der Prozess (X_t) in jedem Intervall $](k-1)/n, k/n]$, das in $]0, t]$ enthalten ist, höchstens einen Sprung. Für alle j mit $T_1 + \dots + T_j \leq t$ gilt dann auch

$$T_{n,1} + \dots + T_{n,j} \leq T_1 + \dots + T_j + 1/n, \quad (18.7)$$

denn die Sprungzeiten des approximierenden Prozesses sind ja höchstens um $1/n$ verzögert.

Gelten (18.6) und (18.7) für alle $j \leq k$, so gilt auch

$$T_j - 1/n \leq T_{n,j} \leq T_j + 1/n, \quad (18.8)$$

denn dann ist

$$\begin{aligned} T_j &= (T_1 + \dots + T_j) - (T_1 + \dots + T_{j-1}) \\ &\leq (T_{n,1} + \dots + T_{n,j}) - (T_{n,1} + \dots + T_{n,j-1} - 1/n) = T_{n,j} + 1/n, \end{aligned}$$

und ebenso prüft man die rechte Ungleichung in (18.8) nach.

Sei nun für einen Moment n fest. Sei $Y_j = 0$, wenn es im Intervall $](j-1)/n, j/n]$ keine Emission gibt, und $= 1$ sonst. Dann ist

$$\begin{aligned} P(T_{n,1} = k/n, T_{n,2} = m/n) &= P(Y_1 = 0, \dots, Y_{k-1} = 0, Y_k = 1, Y_{k+1} = 0, \dots, Y_{k+m-1} = 0, Y_{k+m} = 1) \\ &= P(Y_1 = 0, \dots, Y_k = 1) P(Y_{k+1} = 0, \dots, Y_{k+m} = 1) \\ &= P(T_{n,1} = k/n) P(Y_1 = 0, \dots, Y_{m-1} = 0, Y_m = 1) \\ &= P(T_{n,1} = k/n) P(T_{n,1} = m/n). \end{aligned}$$

Summiert man über alle k , so ergibt sich $P(T_{n,2} = m/n) = P(T_{n,1} = m/n)$. $T_{n,1}$ und $T_{n,2}$ haben also die gleiche Verteilung. Ferner folgt dann $P(T_{n,1} = k/n, T_{n,2} = m/n) = P(T_{n,1} = k/n) P(T_{n,2} = m/n)$, so dass $T_{n,1}$ und $T_{n,2}$ unabhängig sind. Analog zeigt man, dass alle $T_{n,i}$ unabhängig sind und die gleiche Verteilung haben.

Wir sind nun so weit, dass wir beginnen können, die gemeinsame Verteilung von $T_1, \dots, T_k$ zu bestimmen. Seien $t_1, \dots, t_k > 0$. Wir setzen $c = t_1 + t_2 + \dots + t_k + 1$. $B(c, \delta)$ sei wieder die Menge der Ergebnisse ω , für die der Abstand von je zwei Emissionen im Intervall $]0, c]$ größer als δ ist. Auf $B_n = B(c, 1/n)$ gilt (18.7), und also auch (18.8), sofern mindestens eine der in (18.7) auftretenden Summen der T_i oder der $T_{n,i}$ durch $c - 1$ beschränkt ist. Daher folgt nun

$$\begin{aligned}
P(\{T_{n,i} \leq t_i - 1/n \text{ für } i \leq k\} \cap B_n) &\leq P(\{T_i \leq t_i \text{ für } i \leq k\} \cap B_n) \\
&\leq P(\{T_{n,i} \leq t_i + 1/n \text{ für } i \leq k\} \cap B_n).
\end{aligned} \tag{18.9}$$

Wie im Beweis von (18.4) zeigt man $P(B_n) \rightarrow 1$. Für beliebig kleines $\varepsilon > 0$ ist daher die linke Seite in (18.9) für großes n mindestens

$$P(T_{n,i} \leq t_i - \varepsilon \text{ für } i \leq k) - \varepsilon.$$

Da die $T_{n,i}$ unabhängig und identisch verteilt sind, ist nach (18.5) der Limes dieser unteren Schranken

$$\prod_{i=1}^k (1 - \exp(-\lambda(t_i - \varepsilon))) - \varepsilon.$$

Da $\varepsilon > 0$ beliebig klein war, folgt

$$\prod_{i=1}^k (1 - \exp(-\lambda t_i)) \leq P(T_i \leq t_i \text{ für } i \leq k).$$

Analog erkennt man mit Hilfe der rechten Ungleichung in (18.9), dass auch $\geq$ gilt. $\square$

Nun ist es auch leicht zu sehen, dass zu Satz 18.1 eine Umkehrung gilt:

Satz 18.2 Sind $T'_1, T'_2, \dots$ unabhängige Zufallsvariable, die sämtlich exponentiell mit Parameter $\lambda > 0$ verteilt sind, so ist durch

$$X'_t = \sup\{k : T'_1 + \dots + T'_k \leq t\} \quad (t \geq 0)$$

ein Poisson-Prozess mit Parameter λ definiert.

Beweis: Sei $(X_t, 0 \leq t < \infty)$ ein Poisson-Prozess mit Parameter λ und (T_i) wie in Satz 18.1 definiert. Da die T_i die gleiche gemeinsame Verteilung haben wie die T'_i , und da die Beziehung zwischen der Folge (T'_i) und dem Prozess (X'_t) die gleiche ist wie die zwischen (T_i) und (X_t) , folgt die Behauptung. $\square$

[Häufig versucht man die Existenz von Poisson-Prozessen mit Hilfe von Satz 18.2 zu zeigen. Aber ein direkter Beweis, dass (X'_t) die Axiome (A1)–(A4) erfüllt, ist schwierig und wird selten vollständig erbracht.]

Aufgaben

1. Sei $(X_t, t \geq 0)$ ein Poisson-Prozess mit Parameter $\lambda > 0$. Bestimmen Sie die Verteilung der Wartezeit $S_n = \inf\{t : X_t = n\}$ auf den n -ten Sprung. (Hinweis: Anhang zu § 14.) (L)
2. Sei X Poisson-verteilt und Y die Anzahl der Erfolge in X Bernoulli-Versuchen mit Erfolgswahrscheinlichkeit $0 < p < 1$. Zeigen Sie, dass Y und $Z = X - Y$ unabhängig und Poisson-verteilt mit Parametern λp bzw. $\lambda(1 - p)$ sind. (L)
3. Die Anzahl X_t der Autos, die im Zeitintervall $[0, t]$ an einer Tankstelle tanken, sei ein Poisson-Prozess. Mit Wahrscheinlichkeit $0 < p < 1$ tankt ein Auto Dieseldieselkraftstoff. Sei Y_t die Zahl der Autos, die im Intervall $[0, t]$ Dieseldieselkraftstoff tanken, und $Z_t = X_t - Y_t$. Zeigen Sie (unter Verwendung von Aufgabe 2), dass (Y_t) und (Z_t) voneinander unabhängige Poisson-Prozesse mit Parametern λp und $\lambda(1 - p)$ sind. (Dies ist nicht a priori klar. Angenommen es tankten pro Stunde im Mittel etwa 100 Autos und p wäre gleich $1/4$. Dann würden im Mittel pro Stunde ca. 25 Autos Dieseldiesel tanken. In einer bestimmten Stunde werde nun 50-mal Dieseldiesel getankt. Man könnte vermuten, dass in diesem Intervall wohl ca. 150-mal Benzin getankt werden muss, aber diese Vermutung trägt. Die Wahrscheinlichkeit, dass in dieser Stunde 150-mal Benzin getankt wird, ist nicht größer als in jeder anderen. Natürlich sollte man hier realistischerweise eine nichtlineare Zeitskala verwenden. Aber an der Unabhängigkeit der Prozesse (Y_t) und (Z_t) würde dies nichts ändern.)

4. Man zeige: Sind $T_1, T_2, \dots$ unabhängig und geometrisch verteilt, und ist

$$X_k = \sup\{m : T_1 + \dots + T_m \leq k\} \quad (k \in \mathbb{N}),$$

so ist die Folge $X_1, X_2 - X_1, X_3 - X_2, \dots$ Bernoulli-verteilt. (Anleitung: Man imitiere das Argument aus dem Beweis von Satz 18.2.)

5. Zeigen Sie, dass (A1') aus (A0)–(A4) folgt.

Hinweise zum Weiterlesen

Ich hoffe, dass möglichst viele Leser dieses Buches so viel Interesse an Stochastik gefunden haben, dass sie nun mehr davon kennen lernen wollen. Dazu möchte ich ein paar Anregungen geben.

Die Bücher von Engel (1973/76) haben durch viele schöne Beispiele und Aufgaben für den Schulunterricht anregend gewirkt. Systematischer ist der „Leistungskurs Stochastik“ von Barth-Haller (1988), der sich auch durch detaillierte historische Angaben auszeichnet. Chung (1978) didaktisch geschickte „Elementare Wahrscheinlichkeitstheorie“ wendet sich bereits an Studenten. Dass sie bei Adam und Eva (sprich: Mengenalgebra) beginnt und sehr ins Detail geht, wird mancher als Vorteil, mancher als Nachteil empfinden. Eine mathematisch völlig elementar gehaltene Einführung in die Statistik mit zahlreichen Beispielen aus der Praxis, die sehr schön herausarbeitet, worauf man achten muss, haben Freedman, Pisani und Purves (1978) verfasst. Sie dürfte sich auch für die Schule eignen.

Der Einsteigerkurs von Henze (1997) ist zwischen gymnasialen Oberstufenkursen und der Universität angesiedelt.

Fellers klassische zweibändige Einführung in die Wahrscheinlichkeitstheorie (1950/66) enthält einen nach wie vor unübertroffenen Schatz an Ideen und eleganten Argumenten und deckt das Spektrum von elementarer Kombinatorik bis zu fortgeschrittenen Resultaten ab.

Das Buch von Krickeberg-Ziezold (1994) ist eine Einführung in die Stochastik, die schnell zur Statistik kommt. Das eigenwillige und anspruchsvolle Buch von Dinges-Rost (1982) bietet u.a. Einblicke in Fragen der philosophischen Grundlagen und der Anwendungen in der statistischen Physik. Einige attraktive elementar zugängliche Themen, die sich z.B. für Anfängerseminare eignen, findet man bei Jacobs (1969). In Pfanzagl's Buch „Elementare Wahrscheinlichkeitsrechnung“ sind die Anwendungsbeispiele und die Behandlung einiger meist nicht berücksichtigter Themen (Verteilungen auf Kreis und Kugel, Geometrische Wahrscheinlichkeitstheorie) hervorzuheben.

Für das Studium der Grundbegriffe der Maßtheorie möchte ich die Bücher von Bauer (1990) und das knappe Bändchen von Bellach et al. (1978) besonders empfehlen. Weniger auf Stochastiker zugeschnitten, aber vorbildlich in den Motivationen, Anwendungen und historischen Hinweisen ist das Buch von Elstrodt (1996). Gelungene Einführungen in die fortgeschrittene Wahrscheinlichkeitstheorie sind u.a. die Bücher von Gänsler-Stute (1977), Breiman (1968), Bauer (1991) und Chow-Teicher (1978). Das Buch von Schürger (1998) greift einige zusätzliche aktuelle Themen auf. Wer sich nun etwas tiefer in Fragen der Mathematischen Statistik einarbeiten will, könnte mit Bickel-Doksum (1977) beginnen. Das Lehrbuch von Witting (1988) ist eine umfassende mathematisch rigorose Einführung in die Parametrische Statistik, in der Optimalitätsfragen im Zentrum stehen. Der umfangreiche Fortsetzungsband von Witting und Müller-Funk (1995) behandelt asymptotische Methoden.

Für Anwender hat Pfanzagl (1974) die Standardmethoden beschrieben. Wenn diese nicht ausreichen, kann man die Handbücher von Sachs (1968) und Hartung (1982) heranziehen.

Ein wichtiges Nachschlagewerk für alle Fragen der Stochastik ist das von P. H. Müller herausgegebene Lexikon der Stochastik (1991).

Für historisch Interessierte ist die Quellensammlung von I. Schneider (1988) eine Fundgrube. Sie enthält Auszüge aus für die Geschichte der Wahrscheinlichkeitstheorie (bis

1933) bedeutsamen Texten mit informativen Einführungen dazu. Für die Geschichte der Statistik sei auf Stigler (1986) (bis 1900) und Witting (1990) verwiesen.

Einen reichhaltigen Schatz von Übungsaufgaben mit Lösungen haben Wentzel und Owtscharow (1982) und Cacoullos (1989) zusammengestellt.

In den letzten Jahren hat als Anwendungsgebiet der Wahrscheinlichkeitstheorie die Finanzmathematik breites Interesse gefunden. Als deutschsprachige Einführungen sind hier die Bücher von A. Irle (1998) und R. und E. Korn (1999) zu nennen.

Lösungen der mit (L) gekennzeichneten Aufgaben

§ 1

3. Es gibt $(32)_9$ Anordnungen für die ersten neun aufgenommenen Karten in Reihenfolge, 4 Möglichkeiten für das Ass als neunte Karte, dann 3 für das vorangehende Ass, 8 für den Zeitpunkt, an dem das vorangehende Ass kommt, und $(28)_7$ für die anderen Karten in Reihenfolge. Es ergibt sich $4 \cdot 3 \cdot 8 \cdot (28)_7 / (32)_9 = 253/4495$.
11. Wir können annehmen, dass es n Plätze gibt und dass Person A auf Platz 1 sitzt. Jeder der Plätze $2, \dots, n$ für Person B ist gleich wahrscheinlich. Ist $n = 2m + 1$ ungerade und ist $0 \leq k \leq m - 1$, so sitzen k Personen zwischen A und B , wenn B auf Platz $2 + k$ oder auf Platz $n - k$ sitzt. In diesem Fall ist die Wahrscheinlichkeit $2/(n - 1)$. Ist $n = 2m$ gerade, so gilt das Gleiche für $k < (n/2) - 1$. Für $k = (n/2) - 1$ ist die Wahrscheinlichkeit nur $1/(n - 1)$, denn dann muss B auf Platz $(n/2) + 1$ sitzen.
14. Die in (B) angegebene Lösung wäre richtig, wenn die Türme nummeriert wären und gleichartige Stellungen, bei denen nur die Türme untereinander permutiert wären, unterschieden werden sollten. Die richtige Lösung ist $n!$.
15. Durch Fallunterscheidung nach der Ziffer auf der weggelassenen Karte erhält man

$$\frac{5!}{3!} + \frac{5!}{2!2!} + \frac{5!}{2!3!} = 60.$$

§ 2

3. Die mittlere Kinderzahl m einer Familie ist $1 \cdot 0,2 + 2 \cdot 0,2 + 3 \cdot 0,15 + 4 \cdot 0,1 + 5 \cdot 0,05 = 1,70$. Die Wahrscheinlichkeit, dass ein zufällig ausgewählter Junge aus einer Familie mit $1, 2, \dots, 5$ Kindern stammt, ist $0,2/m, 0,4/m, 0,45/m, 0,4/m$ bzw. $0,25/m$. Die Wahrscheinlichkeit, dass ein zufällig ausgewählter Junge, der einer Familie mit i Kindern angehört, mindestens eine Schwester hat, ist $1 - (1/2)^{i-1}$. Nach der Formel von der totalen Wahrscheinlichkeit ergibt sich so als Antwort
- $$\frac{1}{m} \left[0,4 \cdot \frac{1}{2} + 0,45 \cdot \frac{3}{4} + 0,4 \cdot \frac{7}{8} + 0,25 \cdot \frac{15}{16} \right] \approx 0,66.$$
4. Die Wahrscheinlichkeit mindestens eine Sechs zu werfen ist $1 - (5/6)^3 = 91/216$. Die Wahrscheinlichkeit, genau eine Drei und genau eine Sechs zu werfen, ist $(3!) \cdot (1/6)^2 \cdot (4/6) = 4/36$. Die Wahrscheinlichkeit eine Drei und zwei Sechsen oder zwei Dreien und eine Sechs zu werfen ist je $1/72$. Die Wahrscheinlichkeit mindestens eine Drei und mindestens eine Sechs zu werfen ist also $5/36$. Die gesuchte Wahrscheinlichkeit ist also $(5/36)/(91/216) = 30/91$.
 12. Sei A das Ereignis, dass das zuerst gewählte Gemälde ein Original ist, und B das Ereignis, dass der Experte es für ein Original hält. Nach der bayesschen Formel ergibt sich $P(A|B^c) = 5/14$ und $P(A^c|B^c) = 9/14$. Hält also der Experte das zuerst gewählte Gemälde für eine Fälschung, so sind mit Wahrscheinlichkeit $5/14$ noch 9 Originale und 2 Fälschungen wählbar, und mit Wahrscheinlichkeit $9/14$ 10 Originale und eine Fälschung. Die gesuchte Wahrscheinlichkeit ist also $(5/14) \cdot (9/11) + (9/14) \cdot (10/11) = 135/154$.
 13. In der ersten aus zwei Würfeln bestehenden Runde würfelt A zuerst, in jeder folgenden B . Sei $p = 1/6$. Die Wahrscheinlichkeit, dass eine begonnene Runde das Spiel noch nicht beendet, ist $w = (1 - p)^2$. Es ist

$$\begin{aligned} w_A &= p + w(1 - p)p + w^2(1 - p)p + \dots = p - p(1 - p) + p(1 - p) \sum_{i=0}^{\infty} w^i \\ &= p^2 + (1 - p)/(2 - p) = 191/396. \end{aligned}$$

(Für kleines p ist B im Vorteil, für großes p A .)

§ 3

1. Sei $p(k, l) = P(Z = k, X_1 = l)$. Ist $k > l$, so ist $p(k, l) = P(X_2 = k, X_1 = l) = p^2 (1 - p)^{k+l}$. Ist $k < l$, so ist $p(k, l) = 0$. Ist $k = l$, so ist $p(k, k) = P(X_1 = k, X_2 \leq k) = p(1 - p)^k (1 - (1 - p)^{k+1})$. Es ist

$$P(Z = k) = p(k, k) + \sum_{l=0}^{k-1} p(k, l) = p(1 - p)^k [2 - (1 - p)^k (2 - p)].$$

7. Sei X_i = Zahl der Würfe bis die i -te verschiedene Zahl geworfen ist, $Y_1 = 1$, $Y_i = X_i - X_{i-1}$ ($2 \leq i \leq 6$). Ist gerade die $(i - 1)$ -te verschiedene Zahl geworfen worden, so ist in jedem der folgenden Würfe, die Wahrscheinlichkeit p_i , dass die i -te Zahl kommt $= (6 - (i - 1))/6$ (bis sie gekommen ist). Die Y_i sind daher geometrisch verteilt mit Erfolgswahrscheinlichkeit p_i . Es ist $EX_6 = EY_1 + \dots + EY_6 = 1 + 6/5 + 6/4 + \dots + 6/1$. Ferner ist $\text{Var}(Y_3) = (1 - p_3)/p_3^2$. (Die Y_i sind auch unabhängig.)
9. Sei bei einem Wurf p_i die Wahrscheinlichkeit der Augensumme i . Sei $p = p_7$, $r = 1 - p_7 - p_5$. Die gesuchte Wahrscheinlichkeit ist $p + pr + pr^2 + \dots = p_7/(p_7 + p_5) = 3/5$.
10. Es ist $G_a = e \min(N, a) - ca$. Sei $s_a = p_0 + p_1 + \dots + p_a$. Damit ist $EG_a = e(p_1 + 2p_2 + \dots + ap_a + a(1 - s_a)) - ca$. Also ist $EG_a - EG_{a-1} = e(1 - s_{a-1}) - c = (e - c) - es_{a-1}$. Dies ist > 0 (bzw. $= 0$), wenn $s_{a-1} < (bzw. =) (e - c)/e$ gilt. Gibt es ein a mit $s_{a-1} < (e - c)/e < s_a$, so ist EG_a maximal. Ist $s_0 > (e - c)/e$, so ist EG_0 maximal. Ist $s_0 = (e - c)/e = s_1 = \dots = s_{b-1} < s_b$, so sind $EG_0 = \dots = EG_b$ maximal. Andernfalls gibt es ein a mit $s_{a-1} < (e - c)/e = s_a = \dots = s_{b-1} < s_b$. Dann sind $EG_a = \dots = EG_b$ maximal.
11. Sei e_N der gesuchte Erwartungswert. Wir zeigen induktiv $e_N = N$. Natürlich ist $e_1 = 1$ trivial. Sei Y_N die Anzahl der Koinzidenzen in einer zufälligen Permutation von N Elementen und $p_N(n) = P(Y_N = n)$; siehe Satz 3.11. Sei nun $N \geq 2$. Mit Wahrscheinlichkeit $p_N(n)$ sind beim 2-ten Tanz noch $N - n$ Paare dabei. Unter Berücksichtigung des ersten Tanzes folgt

$$e_N = 1 + \sum_{n=0}^N p_N(n) e_{N-n}.$$

Bringen wir das Glied mit e_N auf die linke Seite, so folgt mit der Induktionsannahme

$$(1 - p_N(0)) e_N = 1 + \sum_{n=1}^N p_N(n) e_{N-n} = 1 + \sum_{n=1}^N p_N(n) (N - n) = 1 + (1 - p_N(0)) N - E(Y_N).$$

Da für $k = 1, \dots, N$ das k -te Element von $\{1, \dots, N\}$ bei der Permutation mit Wahrscheinlichkeit $1/N$ eine Koinzidenz liefert, ist $E(Y_N) = 1$. Die Behauptung folgt nun wegen $p_N(0) < 1$.

§ 4

2. Nach Formel (2.10) ist $L_x(p) = \binom{x+r-1}{x} p^r (1-p)^x$. Es folgt $\frac{d}{dp} \log L_x(p) = r/p - x/(1-p)$. Daraus ergibt sich der ML-Schätzer $\hat{p} = r/(x + r)$. (Die zweite Ableitung ist negativ.) Der Schätzer $\hat{p}$ ist nicht erwartungstreu.
6. $\hat{g}(X_1, \dots, X_n) = \max(X_1, \dots, X_n)$.
9. Es ist $g(\lambda) = e^{-3\lambda}$. Ist T erwartungstreu, so gilt für alle λ die Identität

$$e^{-3\lambda} = \sum_{x=0}^{\infty} T(x) e^{-\lambda} \frac{\lambda^x}{x!},$$

also $T(x) = (-2)^x$. Ein Schätzer für eine Wahrscheinlichkeit sollte aber keine negativen Werte annehmen und keine Werte > 1 .

10. Es ist $P_b(X \leq x) = (x/b)^n$. Daraus folgt für $x \leq b$ $P_b(X = x) \geq P_b(X = x - 1)$. Die Mengen $A(b)$ in der Konstruktion von $\tilde{C}$ sind daher Intervalle $\{a(b), a(b) + 1, \dots, b\}$. x gehört zu $A(b)$, wenn $(x/b)^n > \alpha$ ist, d.h. $b < x \sqrt[n]{1/\alpha}$. Bezeichnet $b^*(x)$ das maximale b mit $b < x \sqrt[n]{1/\alpha}$, so ist $C(x) = \{x, x + 1, \dots, b^*(x)\}$.

§ 5

3. Ist S_n die Zahl der wahrgenommenen Reservierungen, so ist S_n $b_{n,p}$ -verteilt mit $p = 0,8$. Es soll $0,025 \approx P(S_n > 200) = P(S_n^* > (200 - np)/s_n)$ sein. Die Approximation mit Korrekturterm und $\Phi^{-1}(0,975) \approx 1,96$ liefert $200,5 - np \approx 1,96 \sigma_n$. Wählt man zur Berechnung von σ_n als erste Näherung $n \approx 250$, so führt dies zum Wert $\sigma_n \approx 6,32$ und zu $n = 235$. Mit $n = 235$ erhält man mit der Normalapproximation eine Überbuchungswahrscheinlichkeit von 0,0208, mit $n = 236$ schon 0,0285. Der Manager darf also 235 Buchungen akzeptieren.
6. Es ist $P(|S_n/n - p| > \varepsilon) = P(|S_n - np| > \varepsilon n) = P(|S_n^*| > \varepsilon n / \sigma_n)$ mit $\sigma_n = \sqrt{npq}$. Wegen $|P(|S_n^*| > \varepsilon n / \sigma_n) - 2\Phi(-\varepsilon n / \sigma_n)| \rightarrow 0$ (Satz 5.5) und $\Phi(-\varepsilon n / \sigma_n) \rightarrow 0$ folgt die Behauptung.
8. Partielle Integration.
9. Nach der Poisson-Approximation mit $p_i = 10^{-4}$ ($i \leq n = 30\,000$) ist die Zahl R der registrierten Teilchen approximativ $P(\lambda)$ -verteilt mit $\lambda = np_i = 3$. Daher ist $P(R = 0) \approx e^{-3}$ und $P(R > 2) \approx 1 - e^{-3}(1 + 3 + 9/2)$. Der Approximationsfehler ist maximal $2 \cdot 30\,000 \cdot 10^{-8} = 6 \cdot 10^{-4}$.

§ 6

2. Unter der Hypothese ist X $b_{6,p}$ -verteilt mit $p = 1/2$. Für große Werte von X wird die Hypothese verworfen. Mit $P(X = 6) + \gamma P(X = 5) = 0,02$ ergibt sich $\gamma = (0,2 - 2^{-6}) / (6 \cdot 2^{-6}) \approx 0,0307$. Ist $X = 6$, so wird die Hypothese verworfen. Ist $X = 5$, so wird sie mit Wahrscheinlichkeit γ verworfen.
4. Die Zahl X der defekten Glühbirnen in der Stichprobe vom Umfang $n = 10$ ist hypergeometrisch verteilt mit $N = 100$ und $S =$ Anzahl der defekten Glühbirnen im Paket. Der Händler testet die Hypothese $S \leq 9$. Er verwirft im Fall $X \geq 1$. Für $S = 9$ ist $P_S(X = 0) = 0,3712$. Also ist für $S = 9$ die Verwerfungswahrscheinlichkeit 0,6288. Dies ist das effektive Niveau des Tests, denn für $S < 9$ ist die Verwerfungswahrscheinlichkeit kleiner.
5. Ist ϑ die Verteilung der Augenzahl eines Würfels, der mit Wahrscheinlichkeit 1 den Wert i würfelt, so ist $\vartheta \in K$. Also muss $E_{\vartheta}(\varphi') \geq E_{\vartheta}(\varphi) = 1$ sein. Damit ist $\varphi'(i, i) = 1$ für $i \in \{1, \dots, 6\}$. Erteilt ϑ jedem der Werte $i \neq j$ die Wahrscheinlichkeit $1/2$, so ist $E_{\vartheta}(\varphi') = \frac{1}{4}(\varphi'(i, i) + \varphi'(j, j) + \varphi'(i, j) + \varphi'(j, i)) \geq \frac{1}{4}E_{\vartheta}(\varphi)$. Wegen $\varphi'(i, j) = \varphi'(j, i)$ und $\varphi(i, j) = \varphi(j, i)$ folgt $\varphi'(i, j) \geq \varphi(i, j)$. Damit ist $\varphi' \geq \varphi$. Wäre für ein $i \neq j$ $\varphi'(i, j) > \varphi(i, j)$, so wäre $E_H(\varphi') > E_H(\varphi)$.
10. Es ist $p_K(k)/p_H(k) = e^{-1/2} \cdot (1/2)^k / (e^{-2} \cdot 2^k) = e^{3/2} 2^{-2k}$. Dies ist fallend in k . Ein nichtrandomisierter NP-Test hat also einen Verwerfungsbereich der Form $R = \{0, 1, \dots, t\}$. Für $t = 0$ ist $P_H(X \in R) = e^{-2} \approx 0,135$, für $t = 1$ ist bereits $P_H(X \in R) > 0,2$. Also kommt nur $R = \{0\}$ in Frage. Der Test mit $R = \{0, 5\}$ hält ebenfalls das Niveau $\alpha = 0,2$ ein und ist schärfer.

§ 7

2. Ist $h = g_{S_T}$, so ist $\text{Var}(S_T) = h''(1-) + h'(1-) - (h'(1-))^2$. Aus $h'(t) = g'(t)g'_T(g(t))$ und $h''(t) = g''(t)g'_T(g(t)) + (g'(t))^2 g''_T(g(t))$ ergibt sich $h'(1-) = (EX)(ET)$ und

$$h''(1-) = (E(X^2) - EX)(ET) + (EX)^2(E(T^2) - ET).$$

Einsetzen liefert die Behauptung.

6. Sei $T_k = \inf\{i \geq 1 : S_i = k\}$. Auf $\{T_1 = t\}$ setze $X'_1 = X_{t+1}$, $X'_2 = X_{t+2}$, ... und $S'_n = X'_1 + \dots + X'_n$. Sei $T'_j = \inf\{i \geq 1 : S'_i = j\}$. Gegeben $\{T_1 = t\}$ hat $X'_1, X'_2, \dots$ die gleiche Verteilung wie $X_1, X_2, \dots$ und also auch (T'_j) die gleiche Verteilung wie (T_j) . Da dies unabhängig vom Wert t richtig ist, ist T_1 unabhängig von jedem T'_j . Aus $T_k = T_1 + T'_{k-1}$ folgt nun $g^{(k)} = g_{T_k} = g_{T_1} \cdot g_{T'_{k-1}} = g \cdot g^{(k-1)}$. Induktiv folgt $g^{(k)} = g^k$. Es ist $P(T_1 = 1) = P(X_1 = 1) = p$. Ist $X_1 = -1$, so muss, von da an gerechnet, die Irrfahrt um 2 Niveaus ansteigen, um zum ersten Mal das Niveau 1 zu erreichen. Also ist für $n > 1$ $w_n^1 = (1 - p)w_{n-1}^2$. Daraus folgt die Behauptung.

§ 8

- Wir können annehmen, dass X und Y Werte in $\mathbb{N}$ haben. Sei $q(k) = P(Y = k)$, $p(i|k) = p(i)/q(k)$, falls $g(i) = k$ ist, und gleich 0 sonst. Setze $H(X|Y = k) = -\sum_i p(i|k) \log_2 p(i|k)$. Dann ist $H(X) - H(Y) = \sum_k q(k) H(X|Y = k) \geq 0$.
- Für alle $P \in \mathcal{P}_\lambda$ ist

$$\begin{aligned} H(P) = -\sum p_k \log_2 p_k &\leq -\sum p_k \log_2 q_k = -\sum p_k \log_2(1/\lambda) - \sum p_k (k-1) \log_2(1-1/\lambda) \\ &= -\log_2(1/\lambda) - \lambda \log_2(1-1/\lambda) + \log_2(1-1/\lambda) \\ &= H(Q). \end{aligned}$$

§ 9

- Die möglichen Werte von $Y = (Y_1, \dots, Y_N)$ sind die Vektoren $(j_1, \dots, j_N)$ mit $0 \leq j_i < i$. Jedem Wert $(j_1, \dots, j_N)$ entspricht genau eine Permutation π , für die $Y(\pi)$ diesen Wert annimmt. Z.B. bedeutet $Y_N(\pi) = j_N = 0$, dass πN der größte der Werte $1, \dots, N$ ist, also $\pi N = N$, $Y_N(\pi) = 1$ bedeutet $\pi N = N-1$, usw. Ist $Y_{N-1}(\pi) = 0$, so ist $\pi(N-1)$ die größte der Zahlen $\{1, \dots, N\} \setminus \{\pi N\}$. Durch $Y_N(\pi)$ und $Y_{N-1}(\pi)$ sind so πN und $\pi(N-1)$ bestimmt, und dann legt $Y_{N-2}(\pi)$ eindeutig $\pi(N-2)$ fest. Induktiv kann man so aus dem Wert von $Y(\pi)$ ablesen, welches π zu diesem Wert gehört. Aus dieser eindeutigen Zuordnung folgt $P(Y = (j_1, \dots, j_N)) = 1/N!$ für alle $(j_1, \dots, j_N)$. Es folgt (a). Die erzeugende Funktion von Y_i ist $g_i(t) = (1+t+\dots+t^{i-1})/i$, die von X ist das Produkt der g_i . Die Verteilung von $\sum_{i=1}^{n+1} Y_i$ ist die Faltung der Verteilung von $\sum_{i=1}^n Y_i$ mit der von Y_{n+1} .

§ 10

- Sei U offen. Ist $U = \mathbb{R}^n$, so ist der Beweis trivial. Sei also $U \neq \mathbb{R}^n$. Für jedes $x \in \mathbb{R}^n$, dessen Koordinaten rational sind, sei $\rho(x)$ der Abstand von x zum Rand von U . Wähle zu x ein halboffenes Intervall $[a, b]$ mit Mitte x , dessen Rand vom Rand von U den Abstand $\rho(x)/3$ hat. Diese Intervalle überdecken U , und es sind nur abzählbar viele.
- Sei für $0 \leq x \leq 1$ und $0 \leq y \leq 1$ $F(x, y) = \max(x, y)$. (F lässt sich stetig auf $\mathbb{R}^2$ fortsetzen.) Wäre F Verteilungsfunktion eines W-Maßes P , so wäre für $A =](0, 0), (1, 1/2)]$ und $B =](0, 0), (1/2, 1)]$ $P(A) = P(B) = 1$ und $P(A \cap B) = 1/2$.
- Ist X der Abstand des zufällig gewählten Punktes vom Mittelpunkt der Kreisscheibe und $x \in [0, 1]$, so ist $P(X \leq x) = F(x) = \pi x^2/\pi$. Nach (10.12) ist also die Dichte in $[0, 1]$ gerade $f(x) = 2x$.
- Sei $F(x) = 1/2$ für $x \in]1/3, 2/3[$, $F(x) = 1/4$ für $x \in]1/9, 2/9[$, $F(x) = 3/4$ für $x \in]7/9, 8/9[$, etc. (Dritteile die Intervalle, auf denen F noch nicht festliegt, und auf dem mittleren Drittel sei $F(x)$ konstant, und zwar soll dort der Wert in der Mitte zwischen dem kleinsten Wert rechts und dem größten links liegen.) Nach abzählbar vielen Schritten ergänze F stetig.

§ 11

- Sei $K = \{(x_1, x_2) : x_1^2 + x_2^2 \leq 1\}$. Für $x \in K$ nimmt $x_1 + x_2$ nur Werte zwischen $-\sqrt{2}$ und $+\sqrt{2}$ an. Für $-\sqrt{2} < t \leq 0$ ist die Fläche des Bereichs $\{(x_1, x_2) \in K : x_1 + x_2 \leq t\}$ — wie man durch Drehung um 45° sieht — gleich

$$\int_{-1}^{t/\sqrt{2}} 2\sqrt{1-x^2} dx = \int_{-\sqrt{2}}^t \sqrt{2-z^2} dz.$$

Da noch durch die Gesamtfläche π von K zu dividieren ist, hat $X_1 + X_2$ die Dichte $\pi^{-1} \sqrt{2-t^2}$. Für $0 \leq t \leq \sqrt{2}$ ergibt sich aus Symmetriegründen die gleiche Dichte.

- Durch Faltung erhält man als Dichte von $U - V$ in $[-1, +1]$ die Funktion $f(x) = 1+x$ (für $x \leq 0$) und $= 1-x$ (für $x \geq 0$). Dann ist die Dichte von $|U - V|$ in $[0, 1]$ gegeben durch $f(x) + f(-x) = 2-2x$. Es ergibt sich $E(|U - V|) = 1/3$ und $\text{Var}(|U - V|) = 1/18$.

9. Die Länge der durch B gehenden zu der Diagonalen durch A und M parallelen Sehne im Abstand $Y = y$ von dieser Diagonalen ist proportional zu $\sqrt{1-y^2}$ ($0 \leq y \leq 1$). Daher ist die Dichte $g(y)$ von Y von der Form $c \cdot \sqrt{1-y^2}$. Wegen $\int g(y) dy = 1$ ist $c = 4/\pi$. Da sich für jede Lage von A die gleiche Dichte ergibt, ist Y von A und damit von X unabhängig. Es ist $F = X \cdot Y/2$. Nach Aufgabe 8 in § 10 hat X die Dichte $f(x) = 2x$ ($0 \leq x \leq 1$). Es ergibt sich $E(F) = (EX)(EY)/2$ mit $EX = 2/3$ und $EY = 4/(3\pi)$, also $EF = 4/(9\pi)$.

§ 12

1. Nach Borel-Cantelli müsste fast jede und also mindestens eine natürliche Zahl durch unendlich viele Primzahlen teilbar sein.
5. Durch Subtraktion von X können wir $X = 0$ annehmen. Sei für $\varepsilon > 0$ $B_n = \{|X_n| \geq \varepsilon\}$. Wegen der stochastischen Konvergenz gilt $P(B_n) \rightarrow 0$. Wegen der Monotonie gilt $B_n \supset B_{n+1}$. Sei $B_\infty = \bigcap B_n$. Dann ist $P(B_\infty) = 0$. Für $\omega \in B_\infty^c$ ist für hinreichend großes n $\omega \in B_n^c$ und also $|X_n(\omega)| < \varepsilon$.
9. Wegen der Stetigkeit der Verteilungsfunktion F von X gibt es zu $\varepsilon > 0$ ein $\delta > 0$ mit

$$\sup\{|F(x) - F(x+h)| : x \in \mathbb{R}, h \in [0, \delta]\} < \varepsilon/2.$$

Für großes n ist $P(|Y_n| \geq \delta) < \varepsilon/2$. Analog zum Argument in Lemma 12.6 ist

$$P(X \leq x - \delta) - P(Y_n > \delta) \leq P(X + Y_n \leq x) \leq P(X \leq x + \delta) + P(Y_n < \delta)$$

und damit $|P(X + Y_n \leq x) - P(X \leq x)| \leq \varepsilon$ gleichmäßig in x .

11. Sei $S_n = X_1 + \dots + X_n$. Es gibt ein $M < \infty$ mit $E(S_n^4) \leq Mn^2$, denn alle $E(X_i X_j X_k X_l)$, in denen ein Index nur einmal vorkommt sind $= 0$. Nach der markowschen Ungleichung ist

$$P(|(\sqrt{n} \log n)^{-1} S_n| \geq \varepsilon) \leq \varepsilon^{-4} E(((\sqrt{n} \log n)^{-1} S_n)^4),$$

und dies strebt gegen 0.

§ 13

2. Zu minimieren ist

$$f(\alpha, \beta, \gamma) = \sum_{i=1}^{10} [(\alpha_i - \alpha)^2 + (\beta_i - \beta)^2 + (\gamma_i - \gamma)^2]$$

unter der Nebenbedingung $\varphi(\alpha, \beta, \gamma) = \alpha + \beta + \gamma - 180^\circ = 0$. Berechnet man die Nullstellen der partiellen Ableitungen der Lagrangefunktion $f(\alpha, \beta, \gamma) + \lambda \varphi(\alpha, \beta, \gamma)$ nach α, β, γ und nach dem Lagrangemultiplikator λ , so erhält man $\lambda = 20(\bar{\alpha} + \bar{\beta} + \bar{\gamma} - 180^\circ)/3$, $\hat{\alpha} = \bar{\alpha} - \lambda/20 = 61,3^\circ + 0,7^\circ/3$. In diesem Fall ist die Summe $\bar{\alpha} + \bar{\beta} + \bar{\gamma} = 179,3^\circ$. Die Differenz $180^\circ - 179,3^\circ$ wird gedrittelt und zu den Schätzern $\bar{\alpha}, \bar{\beta}, \bar{\gamma}$, die man ohne Nebenbedingung erhielte, addiert.

4. $P(X_{(k)} \leq t)$ ist die Wahrscheinlichkeit, dass mindestens k der X_j kleiner oder gleich t sind. Daher ist

$$P(X_{(k)} \leq t) = \sum_{j=k}^n \binom{n}{j} t^j (1-t)^{n-j} = \frac{n!}{(k-1)!(n-k)!} \int_0^t x^{k-1} (1-x)^{n-k} dx.$$

Die letzte Gleichheit prüft man durch partielle Integration. Die Dichte der Verteilung von $X_{(k)}$ ist also

$$g_k(x) = \frac{n!}{(k-1)!(n-k)!} x^{k-1} (1-x)^{n-k} \quad (0 < x < 1).$$

Nun integriere $x g_k(x)$ von 0 bis 1.

7. Die Y_i sind nun $N(\xi_i, w_i \sigma^2)$ -verteilt. Durch Multiplikation der Dichten sieht man, dass die ϑ_k so zu bestimmen sind, dass $\sum_{i=1}^n (y_i - \xi_i)^2 / w_i$ minimiert wird. Wir müssen also nun $\hat{\alpha}$ und $\hat{\beta}$ so bestimmen, dass

$$\sum_{i=1}^n \frac{1}{w_i} (y_i - (\alpha + \beta x_i))^2$$

minimiert wird. Sei $v_i = 1/w_i$ und $\rho_i = v_i / \sum v_i$. Die gleiche Methode wie vorher führt zu den Schätzern

$$\hat{\beta} = \frac{\sum \rho_i x_i y_i - (\sum \rho_i y_i)(\sum \rho_i x_i)}{\sum \rho_i x_i^2 - (\sum \rho_i x_i)^2} \quad \text{und} \quad \hat{\alpha} = \sum \rho_i y_i - \hat{\beta} (\sum \rho_i x_i).$$

§ 14

- Sei $A(\vartheta) = \{x : g(\vartheta) \in C(x)\}$. Da $\{C(x)\}$ Niveau $1 - \alpha$ hat, gilt $P_{\vartheta}(A(\vartheta)) \geq 1 - \alpha$ für alle ϑ . Für $\vartheta \in H$, also für $g(\vartheta) = g_0$, ist $A(\vartheta) = R^c$. Also gilt für diese ϑ $P_{\vartheta}(R) \leq \alpha$.
Sei nun zu jedem möglichen Wert g_0 von g $R(g_0)$ ein Verwerfungsbereich eines nichtrandomisierten Tests der Hypothese $g(\vartheta) = g_0$ zum Niveau α . Setze $C(x) = \{g(\vartheta) : x \in R(g(\vartheta))^c\}$. Wegen $x \in R(g(\vartheta))^c \Leftrightarrow g(\vartheta) \in C(x)$ ist $P_{\vartheta}(\{g(\vartheta) \in C(X)\}) \geq 1 - \alpha$.
- Sei $D_1 = \{(y_1, y_2) : y_1 = y_2\}$ und $D_2 = \{(y_1, y_2) : y_1 = -y_2\}$. Sei $p_i(x)$ die Projektion von $x = (x_1, x_2)$ auf D_i . $|\bar{x}| \sqrt{2}$ ist der Abstand vom Nullpunkt nach $p_1(x)$, und $|s| = |s(x)|$ der von $p_1(x)$ nach x . Also ist $|T(x)|$ das Verhältnis dieser Abstände. $|T(x)|^{-1}$ ist das Verhältnis der entsprechenden Abstände für D_2 . Wegen der Rotationssymmetrie der Verteilung von X für $\mu = 0$ haben $|T(X)|$ und $|T(X)|^{-1}$ die gleiche Verteilung.
- Sei für $x = (x_1, x_2)$ $f_{\mu}(x)$ das Produkt der Dichten der $N(\mu, 1)$ -Verteilung und der $N(\mu, 4)$ -Verteilung. Nach Neyman-Pearson hat der schärfste Test von $\mu = 0$ gegen $\mu = 2$ für geeignetes c die folgende Form: Ist $q(x) = f_2(x)/f_1(x)$, so wird die Hypothese im Fall $q(x) > c$ verworfen und für $q(x) < c$ angenommen. Es ist $q(x) = \exp(2x_1 + x_2/2 - 5/2)$. Für geeignetes c' verwirft der Test im Fall $4x_1 + x_2 > c'$. Aus der Forderung $P_H(4X_1 + X_2 > c') = 0,1$ ergibt sich $c' = 5,73$. Als Wahrscheinlichkeit eines Fehlers 2. Art erhält man $\beta = 0,17$.

§ 15

- Die möglichen Werte von X_n sind 0, 1, 2, 3. Ist X_n bekannt, so sind Y_n und Y_{n+1} bekannt und der Wert von X_{n+1} hängt dann nur noch von Y_{n+2} ab. Da $X_1, \dots, X_{n-1}$ von Y_{n+2} unabhängig sind, ist die markowsche Eigenschaft erfüllt. Es ergibt sich $p_{01} = p_{21} = p$, $p_{00} = p_{20} = 1 - p$, $p_{13} = p_{33} = p$, $p_{12} = p_{32} = 1 - p$. Die übrigen p_{ij} sind gleich 0.
- Die Ruinwahrscheinlichkeit α'_i von Peter bei Startkapital i mit den Einsätzen $1/n$ entspricht dem Problem aus § 15.2 mit $b' = b \cdot n$, $i' = i \cdot n$. Ist $p = 1/2$, so ist $\alpha'_i = 1 - i'/n' = \alpha_i$. Ist $q > p$, so ist $v = p/q < 1$ und $\alpha'_i = (1 - v^{n(b-i)})/(1 - v^{nb}) = 1 + O(v^{n(b-i)})$, die Ruinwahrscheinlichkeit strebt dann für $i < b$ exponentiell schnell mit $n \rightarrow \infty$ gegen 1. Für $p > q$ und $i > 0$ strebt sie exponentiell schnell gegen 0.
- Offenbar ist $P_i(U = i) = 1$ für $i = 0, 3$. Ist der Startzustand 1 oder 2, so betrachten wir eine Kette (Z_n) mit Zustandsraum $\{0, 1, 2, 3, 4, 5\}$. Ist $S \leq n$ und $X_S \in \{0, 3\}$, so sei $Z_n = X_S$; ist $S \leq n$ und $X_S \in \{1, 2\}$, so sei $Z_n = X_S + 3$; ist $S > n$, so sei $Z_n = X_n$. Die Zustände 0, 3, 4, 5 sind absorbierend für (Z_n) . Ist $U = 0$ oder $U = 3$, so erfolgt die Absorption in 0 bzw. in 3. Ist $U = 1$ oder $U = 2$, so erfolgt die Absorption in $U + 3$. Die übrigen Übergangswahrscheinlichkeiten q_{ij} von (Z_n) sind $q_{10} = 1/2$, $q_{12} = (1 - w)/2$, $q_{15} = w/2$, $q_{23} = 1/2$, $q_{21} = (1 - w)/2$, $q_{24} = w/2$. Fragen wir nach $h(i) := P_i(U = 1)$, so betrachten wir $f = 1_{\{4\}}$. Nach (15.17) ist $h(1) = h(2)(1 - w)/2$, $h(2) = w/2 + h(1)(1 - w)/2$. Dadurch sind $h(1)$ und $h(2)$ bestimmt. Fragen wir nach $h(i) = P_i(U = 0)$, so ist $h(1) = 1/2 + h(2)(1 - w)/2$, $h(2) = h(1)(1 - w)/2$. Analog ergeben sich die Wahrscheinlichkeiten $P_i(U = 2)$ und $P_i(U = 3)$.

§ 16

- Es ist $p_{11}^{(n)} = 0$, $p_{33}^{(n)} = 1$ und $p_{55}^{(n)} = 0$ für alle $n \geq 1$. Die Zustände 2, 4, 6, 7 kommunizieren und sind aperiodisch. Ist $\tilde{P}$ die Matrix, die aus P nur die Zeilen und Spalten 2, 4, 6, 7 enthält, so ergibt sich ein invariantes $\tilde{\rho} = (\rho_2, \rho_4, \rho_6, \rho_7)$ aus $\tilde{\rho} = \tilde{\rho} \tilde{P}$ und $\rho_2 + \rho_4 + \rho_6 + \rho_7 = 1$. Es folgt $\rho_2 = 5/76$, $\rho_4 = 34/76$, $\rho_6 = 17/76$, $\rho_7 = 20/76$. Es ist dann $\lim_{n \rightarrow \infty} p_{ii}^{(n)} = \rho_i$ für $i = 2, 4, 6, 7$. Die Verteilung $(0, \rho_2, 0, \rho_4, 0, \rho_6, \rho_7)$ ist invariant. Eine weitere invariante Verteilung ist $(0, 0, 1, 0, 0, 0, 0)$.

5. Sei (X_n) eine markowsche Kette mit Zustandsraum $I = \{0, 1\}$ und $P(X_{n+1} = i | X_n = i) = 1$ für $n = 0$ und $n \geq 2$, aber $P(X_2 = j | X_1 = i) = 1/2$ für alle $i, j \in I$. Sei $T = X_0$. Dann ist $P(Y_2 = 0 | Y_1 = 0, Y_0 = 0) = 1/2$, denn $Y_0 = 0$ impliziert $T = 0$. Andererseits ist, wenn man als Startverteilung $P(X_0 = 0) = P(X_0 = 1) = 1/2$ wählt,

$$P(Y_2 = 0 | Y_1 = 0) = \frac{\frac{1}{4} + \frac{1}{4}}{\frac{1}{2} + \frac{1}{4}} = \frac{2}{3}.$$

9. Sind (p_{ij}) die Übergangswahrscheinlichkeiten von (X_n) und (q_{kl}) die von (Y_n) so hat die aus den Vektoren $Z_n = (X_n, Y_n)$ gebildete Kette die Übergangswahrscheinlichkeiten $p_{(i,k),(j,l)} = p_{ij} q_{kl}$. Die Kette (Z_n) ist aperiodisch und alle Zustände kommunizieren. Daher besucht die Kette (Z_n) für beliebiges $i \in I$ mit Wahrscheinlichkeit 1 irgendwann den Zustand (i, i) . Dies bedeutet $P(\exists n \text{ mit } X_n = Y_n = i) = 1$.

§ 17

4. Nach (17.2) ist $g_f(t) \cdot g_u(t) = \sum_{n=1}^{\infty} u_n t^n = g_u(t) - 1$, also $g_u = 1/(1 - g_f)$. Für festes N gilt, da die $u_n \geq 0$ sind,

$$\sum_{n=0}^N u_n \leq \lim_{t \rightarrow 1-0} g_u(t) \leq \sum_{n=0}^{\infty} u_n = u.$$

Ist $f < 1$, so gilt $g_u(t) \rightarrow (1 - f)^{-1}$ für $t \rightarrow 1 - 0$, also $\sum_{n=0}^N u_n \leq (1 - f)^{-1}$ (für alle N) und $u = (1 - f)^{-1}$. Ist $f = 1$, so gilt $(1 - g_f(t))^{-1} \rightarrow \infty$ für $t \rightarrow 1 - 0$, also $u = \infty$.

5. Trifft zur Zeit 0 kein Teilchen ein, so ist der Zähler zur Zeit 1 unblockiert. Andernfalls wird zur Zeit 0 ein Teilchen registriert. Der Zähler ist dann zur Zeit 2 unblockiert. Also ist $f_1 = 1 - p$, $f_2 = 0$, $f_3 = p$, $f_i = 0$ ($i \geq 4$). Es ist $w_0 = 1$, $w_1 = f_1$, $w_2 = f_1^2$ und $w_n = w_{n-1} f_1 + w_{n-3} f_3$ für $n \geq 3$. Nach dem Erneuerungssatz gilt $w_n \rightarrow 1/\mu$ mit $\mu = f_1 + 3f_3 = 1 + 2p$.

§ 18

1. T_1 hat eine Exponentialverteilung mit Parameter λ . Dies ist eine Γ -Verteilung mit Parametern $\nu = 1$ und $\alpha = \lambda$, also hat nach Satz 14.6 $S_n = T_1 + \dots + T_n$ eine Γ -Verteilung mit Parametern $\nu = n$ und $\alpha = \lambda$.
2. Es ist

$$\begin{aligned} P(Y = k, Z = l) &= P(X = k + l, Y = k) \\ &= e^{-\lambda} \frac{\lambda^{k+l}}{(k+l)!} \binom{k+l}{k} p^k (1-p)^l \\ &= e^{-\lambda p} \frac{(\lambda p)^k}{k!} \cdot e^{-\lambda(1-p)} \frac{(\lambda(1-p))^l}{l!}. \end{aligned}$$

Literaturverzeichnis

- Ahlsvede, R. und I. Wegener (1979): *Suchprobleme*. Teubner, Stuttgart.
- Andrews, D. F., P. J. Bickel, F. R. Hampel, P. J. Huber, W. H. Rogers und J. W. Tukey (1972): *Robust estimates of location*. Princeton Univ. Press, Princeton, New Jersey.
- Ash, R. (1965): *Information Theory*. Interscience Publishers, Wiley, New York, London, Sydney.
- Banach, S. und C. Kuratowski (1929): *Sur une généralisation du problème de la mesure*. Fundamenta Math. **14**, 127–131.
- Barth, F. und R. Haller (1988): *Stochastik, Leistungskurs*. Ehrenwirth, München.
- Bauer, H. (1990): *Maß- und Integrationstheorie*. de Gruyter Lehrbuch, de Gruyter, Berlin, New York.
- Bauer, H. (1991): *Wahrscheinlichkeitstheorie*. 4. Aufl., de Gruyter, Berlin, New York.
- Bellach, J., P. Franken, E. Warmuth, W. Warmuth (1978): *Maß, Integral und bedingter Erwartungswert*. WTB Bd. **226**, Akademie-Verlag, Berlin.
- Bickel, P. J. (1965): *On some robust estimates of location*. Ann. Math. Statist. **43**, 847–858.
- Bickel, P. J. und K. A. Doksum (1977): *Mathematical Statistics*. Holden-Day, Oakland.
- Bickel, P. J. und J. W. O'Connell (1975): *Is there sex bias in graduate admissions*, Science **187**, 398–404.
- Blyth, C. R. und H. A. Still (1983): *Binomial Confidence Intervals*. J. Amer. Statist. Assoc. **78**, 108–116.
- Borel, E. (1909): *Sur les probabilités dénombrables et leurs applications arithmétiques*. Rend. Circ. Mat. Palermo **27**, 247–271.
- Breiman, L. (1968): *Probability*. Addison-Wesley, Reading, Mass.
- Buchner, P. (1951): *Bemerkungen zur Stirlingschen Formel*. Elem. d. Math. **6**, 8–11.
- Cacoullos, T. (1989): *Exercises in probability*. Springer, New York, Berlin, Heidelberg.
- Chow, Y. S. und H. Teicher (1978): *Probability Theory, independence, interchangeability, martingales*. Springer, Berlin, Heidelberg, New York.
- Chung, K. L. (1978): *Elementare Wahrscheinlichkeitstheorie und stochastische Prozesse*. Springer Hochschultext, Springer, Berlin, Heidelberg, New York.
- Courant, R. (1955): *Vorlesungen über Differential- und Integralrechnung I*. Springer, Berlin, Göttingen, Heidelberg.
- Daley, D. J. und D. Vere-Jones (1988): *An introduction to the theory of point processes*. Springer, New York, Berlin, Heidelberg.
- de Moivre, A. (1967): *The doctrine of chances*. Reprint of the third edition. Chelsea Publ. Co., New York. (First edition 1718.)
- Denker, M. (1975): *Elementare Wahrscheinlichkeitstheorie*. Vorlesungsausarbeitung, Göttingen.
- Diaconis, P. (1978): *Statistical Problems in ESP-Research*. Science **201**, 131–136. (Vgl. auch: „Letters“, in Science vom 15.12.1978.)
- Diaconis, P. und B. Efron (Juli 1983): *Statistik per Computer: der Münchhausen-Trick*. Spektrum der Wissenschaft, 56–71.
- Dinges, H. und H. Rost (1982): *Prinzipien der Stochastik*. Teubner Studienbuch Mathematik, Teubner, Stuttgart.
- Dubins, L. E. und L. H. Savage (1976): *Inequalities for stochastic processes — How to gamble if you must*. Dover Publ., New York.
- Dynkin, E. B. und A. A. Juschewitsch (1969): *Sätze und Aufgaben über Markoffsche Prozesse*. Springer, Berlin, Heidelberg, New York.
- Efron, B. (1979): *Bootstrap methods: another look at the jackknife*. Ann. Statist. **7**, 1–26.
- Efron, B. und R. J. Tibshirani (1993): *An introduction to the bootstrap*. Chapman & Hall, New York.
- Elliott, P. D. T. A. (1979): *Probabilistic number theory I, II*. Springer Grundlehren Bd. **239**, **240**, Springer, Berlin, Heidelberg, New York.
- Elstrodt, J. (1996): *Maß- und Integrationstheorie*. Springer, Berlin, Heidelberg, New York.

- Engel, A. (1973/76): *Wahrscheinlichkeitsrechnung und Statistik*. 2 Bände, Klett Studienbücher, Klett, Stuttgart.
- Falk, M. (1987): *Bootstrap: Eine aktuelle statistische Idee*. Mitt. Math. Gesellsch. Hamburg **11**, 475–486.
- Feller, W. (1950/66): *An introduction to probability theory and its applications*. 2 Bände, Wiley, New York, London, Sydney.
- Fisher, R. A. (1935): *The design of experiments*. Oliver and Boyd, Edinburgh.
- Forster, O. (1981): *Analysis 3*. Vieweg Aufbaukurs Mathematik.
- Freedman, D. (1971): *Markov chains*. Holden-Day, San Francisco.
- Freedman, D., R. Pisani und R. Purves (1978): *Statistics*. W. W. Norton & Co., New York, London.
- Gänssler, P. und W. Stute (1977): *Wahrscheinlichkeitstheorie*. Springer Hochschultext, Springer, Berlin, Heidelberg, New York.
- Gallager, R. C. (1968): *Information Theory and Reliable Communication*. Wiley, New York, London, Sydney.
- Gerber, H. U. (1995): *A teacher's remark on exact credibility*. Astin Bulletin **25**, 189–192.
- Greenwood, J. A. und C. E. Stuart (1937): *Mathematical Techniques in ESP-Research*. J. Parapsych. **1**, 206–225.
- Hald, A. (1965): *Statistical Theory with Engineering Applications*. Wiley, New York, London, Sydney.
- Hartung, H. (mit B. Elpelt und K.-H. Klösener) (1984): *Statistik. Lehr- und Handbuch der angewandten Statistik*. 2. Aufl., R. Oldenbourg, München, Wien.
- Hasse, H. (1964): *Vorlesungen über Zahlentheorie*. Springer, Berlin, Heidelberg, New York.
- Henze, N. (1997): *Stochastik für Einsteiger*. Vieweg.
- Hill, T. P. (1995): *A statistical derivation of the significant-digit law*. Statistical Science **10**, 354–363.
- Hoare, C. A. R. (1961): *Algorithm 63, partition; Algorithm 64, quicksort; Algorithm 65, find*. Communications of the ACM **4**, 321–322.
- Hodges, J. L., Jr. und L. Le Cam (1960): *The Poisson approximation to the Poisson binomial distribution*. Ann. Math. Statist. **31**, 737–740.
- Huber, P. J. (1981): *Robust statistics*. Wiley, New York, London, Sydney.
- Irle, A. (1998): *Finanzmathematik*. Teubner, Stuttgart.
- Jacobs, K. (1969): *Selecta Mathematica I*. Springer, Berlin, Heidelberg, New York.
- Jacobs, K. (1983): *Einführung in die Kombinatorik*. de Gruyter Lehrbuch, de Gruyter, Berlin, New York.
- Jaeckel, L. A. (1971): *Robust estimates of location*. Ann. Math. Statist. **42**, 1020–1034.
- Jaglom, A. M. und I. M. Jaglom (1960): *Wahrscheinlichkeit und Information*. VEB Deutscher Verlag der Wissenschaften, Berlin.
- Johnson, N. I. und S. Kotz (1969): *Discrete distributions*. Wiley, New York, London, Sydney.
- Johnson, N. I. (1970): *Continuous univariate distributions*. 2 Bände, Houghton Mifflin Co.
- Kac, M. (1959): *Statistical independence in probability, analysis and number theory*. Carus Math. Monograph **12**, Math. Assoc. of Amer., Distributed by Wiley, New York.
- Kersting, G. D. (1978): *A weak convergence theorem with applications to the Robbins-Monro process*. Ann. Probab. **6**, 1015–1025.
- Knuth, D. E. (1973): *The art of computer programming; fundamental algorithms*. Vol. 1. Addison-Wesley, 2. ed.
- Kolmogorow, A. (1933): *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Ergebn. d. Math., Springer, Berlin.
- Korn, R. und E. (1999): *Optionsbewertung und Portfolio-Optimierung*. Vieweg-Gabler.
- Krengel, U. (1990): *Wahrscheinlichkeitstheorie*. In: Ein Jahrhundert Mathematik 1890–1990. Festschrift zum Jubiläum der DMV. Dokumente zur Geschichte der Mathematik **6**, Vieweg, Braunschweig, Wiesbaden, 457–489.

- Krengel, U. und M. Liese (1978): *Begründung und Kritik der statistischen Auswertung parapsychologischer Experimente*. Zeitschrift für Parapsych. u. Grenzgebiete der Psychologie **20**, 185–212.
- Krickeberg, K. und H. Ziezold (1994): *Stochastische Methoden*. Springer Hochschultext, Springer, Berlin, Heidelberg, New York.
- Laplace, P. S. (1812): *Théorie analytique des probabilités*.
- Linnik, J. W. (1961): *Die Methode der kleinsten Quadrate in moderner Darstellung*. VEB Deutscher Verlag der Wissenschaften, Berlin.
- Ljapounow, A. M. (1900): *Sur une proposition de la théorie des probabilités*. Bull. Acad. Sci. St. Petersburg **5**, 359–386.
- Mardia, K. V. (1972): *Statistics of directional data*. Academic Press, London, New York.
- Mathar, R. (1996): *Informationstheorie*. Teubner, Stuttgart.
- Meyer, P. L. (1971): *Introductory probability and statistical applications*. Addison-Wesley, Reading.
- Morgenstern, D. (1968): *Einführung in die Wahrscheinlichkeitsrechnung und Mathematische Statistik*. Springer Grundlehren **124**, Springer, Berlin, Heidelberg, New York.
- Neyman, J. (1950): *First course in probability and statistics*. Henry Holt & Co., New York.
- Pfanzagl, J. (1974): *Allgemeine Methodenlehre der Statistik II*. Sammlung Götschen, 4. verb. Aufl., de Gruyter, Berlin, New York.
- Pfanzagl, J. (1988): *Elementare Wahrscheinlichkeitsrechnung*. De Gruyter Lehrbuch, Berlin, New York.
- Pflug, G. Ch. (1983): *On Kersting's proof of the central limit theorem*. Statistics and Probability Letters **1**, 323–326.
- Rachev, S. T. und L. Rüschendorf (1995): *Probability metrics and recursive algorithms*. Adv. Appl. Probab. **27**, 770–799.
- Randles, R. H. und D. A. Wolfe (1979): *Introduction to the Theory of Non-parametric Statistics*. Wiley, New York, Chichester, Brisbane, Toronto.
- Rao, C. R. (1973): *Linear Statistical Inference and its Applications*. Wiley, New York, London, Sydney.
- Rényi, A. (1969): *Briefe über die Wahrscheinlichkeit*. Birkhäuser, Basel, Stuttgart.
- Rösler, U. (1987): *An elementary proof of the central limit theorem*. Manuskript.
- Rösler, U. (1991): *A limit theorem for "Quicksort"*. RAIRO. Theoretical Informatics and Applications **25**, 85–100.
- Rösler, U. und L. Rüschendorf (1999): *A contraction method for recursive algorithms*. To appear.
- Rubinstein, R. Y. (1981): *Simulation and the Monte Carlo Method*. Wiley Ser. in Probability and Statistics. Wiley, New York.
- Sachs, L. (1968): *Statistische Auswertungsmethoden*. Springer, Berlin, Heidelberg, New York.
- Schneider, I. (Hrsg.) (1988): *Die Entwicklung der Wahrscheinlichkeitstheorie von den Anfängen bis 1933*. Wiss. Buchgesellschaft, Darmstadt.
- Schürger, K. (1998): *Wahrscheinlichkeitstheorie*. R. Oldenbourg, München.
- Serfling, R. J. (1975): *A general Poisson approximation theorem*. Ann. Probab. **3**, 726–731.
- Stange, K. (1970): *Angewandte Statistik I*. Springer, Berlin, Heidelberg, New York.
- Stange, K. (1977): *Bayes-Verfahren*. Springer Hochschultext, Springer, Berlin, Heidelberg, New York.
- Stigler, S. M. (1977): *Do robust estimators work with real data?* Ann. Statist. **5**, 1055–1098.
- Stigler, S. M. (1981): *Gauß and the invention of least squares*. Ann. Statist. **9**, 465–474.
- Stigler, S. M. (1986): *The history of statistics: the measurement of uncertainty before 1900*. Harvard.
- Stoyan, D., W. S. Kendall und J. Mecke (1987): *Stochastic geometry and its applications*. Chichester.
- Topsøe, F. (1974): *Informationstheorie*. Teubner, Stuttgart.
- Tukey, J. W. (1977): *Exploratory Data Analysis*. Addison-Wesley.
- van Beek, P. (1972): *An application of Fourier methods to the problem of sharpening the Berry-Esseen inequality*. Z. Wahrscheinlichkeitstheorie verw. Geb. **23**, 187–196.

- van der Waerden, B. L. (1971): *Mathematische Statistik*. Springer Grundlehren **87**, 3. Aufl., Springer, Berlin, Heidelberg, New York.
- Wentzel, E. S. und L. A. Owtscharow (1982): *Aufgabensammlung zur Wahrscheinlichkeitsrechnung*. Akademie-Verlag, Berlin.
- Witting, H. (1985): *Mathematische Statistik I*. Teubner, Stuttgart.
- Witting, H. (1990): *Mathematische Statistik*. In: Ein Jahrhundert Mathematik 1890-1990. Festschrift zum Jubiläum der DMV. Dokumente zur Geschichte der Mathematik **6**, Vieweg, Braunschweig, Wiesbaden, 781–815.
- Witting, H. und U. Müller-Funk (1995): *Mathematische Statistik II*. Teubner, Stuttgart.
- Yushkevich, A. P (1986): Nicolas Bernoulli and the publication of James Bernoulli's *Ars Conjectandi*. SIAM, Theory of Probability **31**, 286–303.
- van Zwet, W. R. und J. Osterhoff (1967): *On the combination of independent test statistics*. Ann. Math. Statist. **38**, 659–680.

Tabellen

Die folgenden Tabellen wurden mit Hilfe des mathematischen Programmpakets *scilab*¹ berechnet. Fast identische Tabellen finden sich (z.B.) auch bei Hartung (1984), S. 734–737, und Pfanzagl (1974), S. 286–287.

$n \backslash \gamma$	0,990	0,975	0,950	0,900
1	31,821	12,706	6,314	3,078
2	6,965	4,303	2,920	1,886
3	4,541	3,182	2,353	1,638
4	3,747	2,776	2,132	1,533
5	3,365	2,571	2,015	1,476
6	3,143	2,447	1,943	1,440
7	2,998	2,365	1,895	1,415
8	2,896	2,306	1,860	1,397
9	2,821	2,262	1,833	1,383
10	2,764	2,228	1,812	1,372
11	2,718	2,201	1,796	1,363
12	2,681	2,179	1,782	1,356
13	2,650	2,160	1,771	1,350
14	2,624	2,145	1,761	1,345
15	2,602	2,131	1,753	1,341
16	2,583	2,120	1,746	1,337
17	2,567	2,110	1,740	1,333
18	2,552	2,101	1,734	1,330
19	2,539	2,093	1,729	1,328
20	2,528	2,086	1,725	1,325
21	2,518	2,080	1,721	1,323
22	2,508	2,074	1,717	1,321
23	2,500	2,069	1,714	1,319

$n \backslash \gamma$	0,990	0,975	0,950	0,900
24	2,492	2,064	1,711	1,318
25	2,485	2,060	1,708	1,316
26	2,479	2,056	1,706	1,315
27	2,473	2,052	1,703	1,314
28	2,467	2,048	1,701	1,313
29	2,462	2,045	1,699	1,311
30	2,457	2,042	1,697	1,310
40	2,423	2,021	1,684	1,303
50	2,403	2,009	1,676	1,299
60	2,390	2,000	1,671	1,296
70	2,381	1,994	1,667	1,294
80	2,374	1,990	1,664	1,292
90	2,368	1,987	1,662	1,291
100	2,364	1,984	1,660	1,290
150	2,351	1,976	1,655	1,287
200	2,345	1,972	1,653	1,286
250	2,341	1,969	1,651	1,285
300	2,339	1,968	1,650	1,284
400	2,336	1,966	1,649	1,284
600	2,333	1,964	1,647	1,283
800	2,331	1,963	1,647	1,283
1000	2,330	1,962	1,646	1,282
100000	2,326	1,960	1,645	1,282

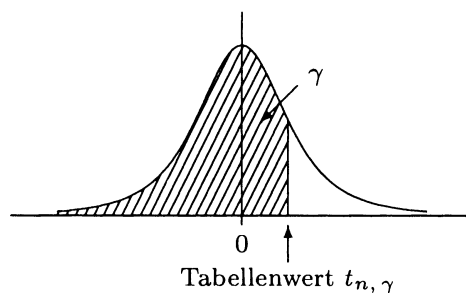


Tabelle I Quantile $t_{n, \gamma}$ der t -Verteilung

¹ Dieses Programm gibt es sowohl für Windows als auch für Unix/Linux. Es darf im nichtkommerziellen privaten Bereich und im Universitätsbereich frei verwendet werden. Es ist (zur Zeit der Drucklegung dieses Buches) im Internet unter <ftp.inria.fr/INRIA/Projects/Meta2/Scilab> erhältlich.

x	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990

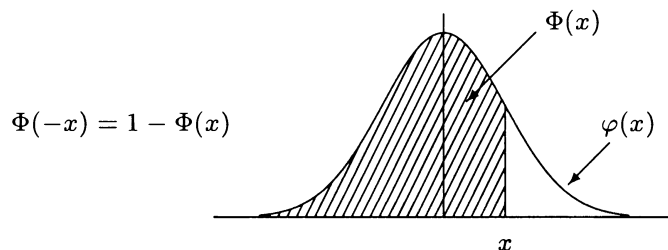
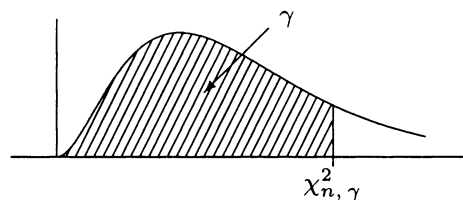


Tabelle II Verteilungsfunktion $\Phi(x)$ der Standard-Normalverteilung $N(0,1)$. Ablesbeispiel:
 $\Phi(1,96) \approx 0,975$.

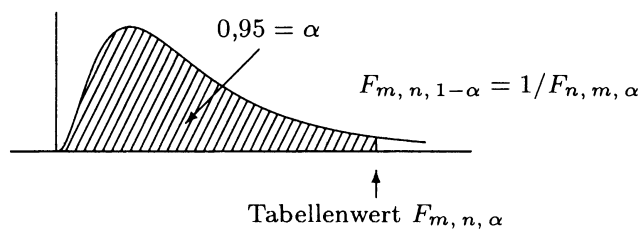
$\gamma \backslash n$	0,995	0,990	0,975	0,950	0,900	0,750	0,500	0,250	0,100	0,050	0,025	0,010	0,005
1	7,879	6,635	5,024	3,841	2,706	1,323	0,455	0,102	0,016	0,004	0,001	0,000	0,000
2	10,60	9,21	7,38	5,99	4,61	2,77	1,39	0,58	0,21	0,10	0,05	0,02	0,01
3	12,84	11,34	9,35	7,81	6,25	4,11	2,37	1,21	0,58	0,35	0,22	0,11	0,07
4	14,86	13,28	11,14	9,49	7,78	5,39	3,36	1,92	1,06	0,71	0,48	0,30	0,21
5	16,75	15,09	12,83	11,07	9,24	6,63	4,35	2,67	1,61	1,15	0,83	0,55	0,41
6	18,55	16,81	14,45	12,59	10,64	7,84	5,35	3,45	2,20	1,64	1,24	0,87	0,68
7	20,28	18,48	16,01	14,07	12,02	9,04	6,35	4,25	2,83	2,17	1,69	1,24	0,99
8	21,95	20,09	17,53	15,51	13,36	10,22	7,34	5,07	3,49	2,73	2,18	1,65	1,34
9	23,59	21,67	19,02	16,92	14,68	11,39	8,34	5,90	4,17	3,33	2,70	2,09	1,73
10	25,19	23,21	20,48	18,31	15,99	12,55	9,34	6,74	4,87	3,94	3,25	2,56	2,16
11	26,76	24,72	21,92	19,68	17,28	13,70	10,34	7,58	5,58	4,57	3,82	3,05	2,60
12	28,30	26,22	23,34	21,03	18,55	14,85	11,34	8,44	6,30	5,23	4,40	3,57	3,07
13	29,82	27,69	24,74	22,36	19,81	15,98	12,34	9,30	7,04	5,89	5,01	4,11	3,57
14	31,32	29,14	26,12	23,68	21,06	17,12	13,34	10,17	7,79	6,57	5,63	4,66	4,07
15	32,80	30,58	27,49	25,00	22,31	18,25	14,34	11,04	8,55	7,26	6,26	5,23	4,60
16	34,27	32,00	28,85	26,30	23,54	19,37	15,34	11,91	9,31	7,96	6,91	5,81	5,14
17	35,72	33,41	30,19	27,59	24,77	20,49	16,34	12,79	10,09	8,67	7,56	6,41	5,70
18	37,16	34,81	31,53	28,87	25,99	21,60	17,34	13,68	10,86	9,39	8,23	7,01	6,26
19	38,58	36,19	32,85	30,14	27,20	22,72	18,34	14,56	11,65	10,12	8,91	7,63	6,84
20	40,00	37,57	34,17	31,41	28,41	23,83	19,34	15,45	12,44	10,85	9,59	8,26	7,43
21	41,40	38,93	35,48	32,67	29,62	24,93	20,34	16,34	13,24	11,59	10,28	8,90	8,03
22	42,80	40,29	36,78	33,92	30,81	26,04	21,34	17,24	14,04	12,34	10,98	9,54	8,64
23	44,18	41,64	38,08	35,17	32,01	27,14	22,34	18,14	14,85	13,09	11,69	10,20	9,26
24	45,56	42,98	39,36	36,42	33,20	28,24	23,34	19,04	15,66	13,85	12,40	10,86	9,89
25	46,93	44,31	40,65	37,65	34,38	29,34	24,34	19,94	16,47	14,61	13,12	11,52	10,52
26	48,29	45,64	41,92	38,89	35,56	30,43	25,34	20,84	17,29	15,38	13,84	12,20	11,16
27	49,64	46,96	43,19	40,11	36,74	31,53	26,34	21,75	18,11	16,15	14,57	12,88	11,81
28	50,99	48,28	44,46	41,34	37,92	32,62	27,34	22,66	18,94	16,93	15,31	13,56	12,46
29	52,34	49,59	45,72	42,56	39,09	33,71	28,34	23,57	19,77	17,71	16,05	14,26	13,12
30	53,67	50,89	46,98	43,77	40,26	34,80	29,34	24,48	20,60	18,49	16,79	14,95	13,79
40	66,77	63,69	59,34	55,76	51,81	45,62	39,34	33,66	29,05	26,51	24,43	22,16	20,71
50	79,49	76,15	71,42	67,50	63,17	56,33	49,33	42,94	37,69	34,76	32,36	29,71	27,99
60	91,95	88,38	83,30	79,08	74,40	66,98	59,33	52,29	46,46	43,19	40,48	37,48	35,53
70	104,2	100,4	95,0	90,5	85,5	77,6	69,3	61,7	55,3	51,7	48,8	45,4	43,3
80	116,3	112,3	106,6	101,9	96,6	88,1	79,3	71,1	64,3	60,4	57,2	53,5	51,2
90	128,3	124,1	118,1	113,1	107,6	98,6	89,3	80,6	73,3	69,1	65,6	61,8	59,2
100	140,2	135,8	129,6	124,3	118,5	109,1	99,3	90,1	82,4	77,9	74,2	70,1	67,3
150	198,4	193,2	185,8	179,6	172,6	161,3	149,3	138,0	128,3	122,7	118,0	112,7	109,1
200	255,3	249,4	241,1	234,0	226,0	213,1	199,3	186,2	174,8	168,3	162,7	156,4	152,2
250	311,3	304,9	295,7	287,9	279,1	264,7	249,3	234,6	221,8	214,4	208,1	200,9	196,2
300	366,8	359,9	349,9	341,4	331,8	316,1	299,3	283,1	269,1	260,9	253,9	246,0	240,7
400	476,6	468,7	457,3	447,6	436,6	418,7	399,3	380,6	364,2	354,6	346,5	337,2	330,9
600	693,0	683,5	669,8	658,1	644,8	623,0	599,3	576,3	556,1	544,2	534,0	522,4	514,5
800	906,8	896,0	880,3	866,9	851,7	826,6	799,3	772,7	749,2	735,4	723,5	709,9	700,7
1000	1119	1107	1090	1075	1058	1030	999	969	943	928	914	899	889

Tabelle III Quantile $\chi_{n, \gamma}^2$ der χ^2 -Verteilung

$n \backslash m$	1	2	3	4	5	6	7	8	9	10	11	12
1	161	200	216	225	230	234	237	239	241	242	243	244
2	18,5	19,0	19,2	19,2	19,3	19,3	19,4	19,4	19,4	19,4	19,4	19,4
3	10,1	9,6	9,3	9,1	9,0	8,9	8,9	8,8	8,8	8,8	8,8	8,7
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,94	5,91
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,70	4,68
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,03	4,00
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,60	3,57
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,31	3,28
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,10	3,07
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,94	2,91
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,82	2,79
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,72	2,69
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,63	2,60
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,57	2,53
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,51	2,48
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,46	2,42
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,41	2,38
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,37	2,34
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,34	2,31
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,31	2,28
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,28	2,25
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,26	2,23
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,24	2,20
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,22	2,18
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,20	2,16
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,18	2,15
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,17	2,13
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,15	2,12
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,14	2,10
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,13	2,09
34	4,13	3,28	2,88	2,65	2,49	2,38	2,29	2,23	2,17	2,12	2,08	2,05
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,04	2,00
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,03	1,99	1,95
70	3,98	3,13	2,74	2,50	2,35	2,23	2,14	2,07	2,02	1,97	1,93	1,89
100	3,94	3,09	2,70	2,46	2,31	2,19	2,10	2,03	1,97	1,93	1,89	1,85
200	3,89	3,04	2,65	2,42	2,26	2,14	2,06	1,98	1,93	1,88	1,84	1,80
300	3,87	3,03	2,63	2,40	2,24	2,13	2,04	1,97	1,91	1,86	1,82	1,78
400	3,86	3,02	2,63	2,39	2,24	2,12	2,03	1,96	1,90	1,85	1,81	1,78
500	3,86	3,01	2,62	2,39	2,23	2,12	2,03	1,96	1,90	1,85	1,81	1,77
100000	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,79	1,75

Tabelle IV Unteres 95%-Quantil $F_{m, n; 0,95}$ der F -Verteilung

$m \backslash n$	14	16	18	20	22	24	30	40	50	60	100	100000
1	245	246	247	248	249	249	250	251	252	252	253	254
2	19,4	19,4	19,4	19,4	19,5	19,5	19,5	19,5	19,5	19,5	19,5	19,5
3	8,71	8,69	8,67	8,66	8,65	8,64	8,62	8,59	8,58	8,57	8,55	8,53
4	5,87	5,84	5,82	5,80	5,79	5,77	5,75	5,72	5,70	5,69	5,66	5,63
5	4,64	4,60	4,58	4,56	4,54	4,53	4,50	4,46	4,44	4,43	4,41	4,37
6	3,96	3,92	3,90	3,87	3,86	3,84	3,81	3,77	3,75	3,74	3,71	3,67
7	3,53	3,49	3,47	3,44	3,43	3,41	3,38	3,34	3,32	3,30	3,27	3,23
8	3,24	3,20	3,17	3,15	3,13	3,12	3,08	3,04	3,02	3,01	2,97	2,93
9	3,03	2,99	2,96	2,94	2,92	2,90	2,86	2,83	2,80	2,79	2,76	2,71
10	2,86	2,83	2,80	2,77	2,75	2,74	2,70	2,66	2,64	2,62	2,59	2,54
11	2,74	2,70	2,67	2,65	2,63	2,61	2,57	2,53	2,51	2,49	2,46	2,40
12	2,64	2,60	2,57	2,54	2,52	2,51	2,47	2,43	2,40	2,38	2,35	2,30
13	2,55	2,51	2,48	2,46	2,44	2,42	2,38	2,34	2,31	2,30	2,26	2,21
14	2,48	2,44	2,41	2,39	2,37	2,35	2,31	2,27	2,24	2,22	2,19	2,13
15	2,42	2,38	2,35	2,33	2,31	2,29	2,25	2,20	2,18	2,16	2,12	2,07
16	2,37	2,33	2,30	2,28	2,25	2,24	2,19	2,15	2,12	2,11	2,07	2,01
17	2,33	2,29	2,26	2,23	2,21	2,19	2,15	2,10	2,08	2,06	2,02	1,96
18	2,29	2,25	2,22	2,19	2,17	2,15	2,11	2,06	2,04	2,02	1,98	1,92
19	2,26	2,21	2,18	2,16	2,13	2,11	2,07	2,03	2,00	1,98	1,94	1,88
20	2,22	2,18	2,15	2,12	2,10	2,08	2,04	1,99	1,97	1,95	1,91	1,84
21	2,20	2,16	2,12	2,10	2,07	2,05	2,01	1,96	1,94	1,92	1,88	1,81
22	2,17	2,13	2,10	2,07	2,05	2,03	1,98	1,94	1,91	1,89	1,85	1,78
23	2,15	2,11	2,08	2,05	2,02	2,01	1,96	1,91	1,88	1,86	1,82	1,76
24	2,13	2,09	2,05	2,03	2,00	1,98	1,94	1,89	1,86	1,84	1,80	1,73
25	2,11	2,07	2,04	2,01	1,98	1,96	1,92	1,87	1,84	1,82	1,78	1,71
26	2,09	2,05	2,02	1,99	1,97	1,95	1,90	1,85	1,82	1,80	1,76	1,69
27	2,08	2,04	2,00	1,97	1,95	1,93	1,88	1,84	1,81	1,79	1,74	1,67
28	2,06	2,02	1,99	1,96	1,93	1,91	1,87	1,82	1,79	1,77	1,73	1,65
29	2,05	2,01	1,97	1,94	1,92	1,90	1,85	1,81	1,77	1,75	1,71	1,64
30	2,04	1,99	1,96	1,93	1,91	1,89	1,84	1,79	1,76	1,74	1,70	1,62
34	1,99	1,95	1,92	1,89	1,86	1,84	1,80	1,75	1,71	1,69	1,65	1,57
40	1,95	1,90	1,87	1,84	1,81	1,79	1,74	1,69	1,66	1,64	1,59	1,51
50	1,89	1,85	1,81	1,78	1,76	1,74	1,69	1,63	1,60	1,58	1,52	1,44
70	1,84	1,79	1,75	1,72	1,70	1,67	1,62	1,57	1,53	1,50	1,45	1,35
100	1,79	1,75	1,71	1,68	1,65	1,63	1,57	1,52	1,48	1,45	1,39	1,28
200	1,74	1,69	1,66	1,62	1,60	1,57	1,52	1,46	1,41	1,39	1,32	1,19
300	1,72	1,68	1,64	1,61	1,58	1,55	1,50	1,43	1,39	1,36	1,30	1,15
400	1,72	1,67	1,63	1,60	1,57	1,54	1,49	1,42	1,38	1,35	1,28	1,13
500	1,71	1,66	1,62	1,59	1,56	1,54	1,48	1,42	1,38	1,35	1,28	1,11
100000	1,69	1,64	1,60	1,57	1,54	1,52	1,46	1,39	1,35	1,32	1,24	1,01



Symbolverzeichnis

Die hinter den Symbolen stehenden Zahlen geben die Nummerierung der Seiten an, auf denen das Symbol aufgeführt ist.

$\mathbb{R}$ reelle Zahlen	P_{ϑ} 61
$\mathbb{N} = \{1, 2, \dots\}$	$L_x(\vartheta), \mathcal{L}_x$ 62, 164
$\mathbb{Z} = \{\dots, -1, 0, 1, 2, \dots\}$	$\bar{X}$ 64, 165
$\emptyset$ leere Menge	s^2 64, 165
A^c Komplement von A	$\varphi(x), \varphi_{\mu, \sigma^2}(x)$ 78, 133
$\mathcal{P}(\Omega)$ Potenzmenge 3	$\Phi(x)$ 81
$\text{card}(A)$ 5	S_n^* 81
$\binom{N}{n}$ 8	$\Phi^{-1}(z)$ 84
$\Omega_I, \Omega_{II}, \dots$ 7–9	$P(\lambda), p(k \lambda)$ 86
$\exp(a) = e^a$ 10	$\beta(\vartheta)$ 95
$h(s; n, N, S)$ 12	$q(x)$ 101
$\binom{n}{k_1, \dots, k_r}$ 14	$g(t), g_X(t)$ 107
$P(A B)$ 21	$H(P), H(X), H(X, Y)$ 116
$\Omega_1 \times \Omega_2 \times \dots \times \Omega_n$ 27	$H(Y X)$ 119
$P_1 \times P_2 \times \dots \times P_n$ 27	$\mathcal{A}(\mathcal{F})$ 129
$\{X \in A\}$ 28	F_X 142
X_I 29	$\int X dP$ 145
$b_{n,p}$ Binomialverteilung 30	$f(x \vartheta)$ 164
$f(k; r, p)$ 31	$g_k(x)$ 178
$P_X, P \circ X^{-1}$ 42, 43, 142	$h_k(x)$ 178
$f \circ X$ 45	$t_{k, \beta}$ 179
$EX, E(X)$ Erwartungswert von X 46, 145	$\Gamma(t)$ 192
$1_A, \mathbf{1}$ 49	p_{ij} 198
$\text{Var}(X)$ 52	$p_{ij}^{(n)}$ 200
$\text{Cov}(X, Y)$ 53	$C(i)$ 209
$\log x$ natürlicher Logarithmus von x 58	$C_r(i)$ 209
	$f_{ij}^{(n)}, f_{ij}^*, p_{ij}^*$ 210
	m_{ii} 221

Namen- und Sachwortverzeichnis

- absolute Häufigkeit, 4
- Absorptionsverteilung, 201
- Absorptionswahrscheinlichkeit, 200
- Abzählprinzip, 7
- additiv, 3
- d'Alembert, Jean le Rond (1717–1783), 17
- Allel, 35
- Alphabet, 114
- Alternative, 93, 94
 - einfache, 101
 - zusammengesetzte, 101
- Anfangszifferverteilung, 137
- A-posteriori-Wahrscheinlichkeit, 39
- A-priori-Wahrscheinlichkeit, 39
- asymptotisch äquivalent, *siehe* asymptotisch gleich
- asymptotisch gleich, 76
- Ausfallrate, 134
- ausgeartet, *siehe* Normalverteilung
- Ausreißer, 170
- Aussterbewahrscheinlichkeit, 111
- austauschbar, 34

- Bacon, Francis (1561–1626), 92
- Banach und Kuratowski
 - Satz von, 136
- Banach, Stefan (1892–1945), 39, 126, 136
- Banachs Fixpunktsatz, 126
- Bayes, Thomas (1702–1761), 39
- bayessche Formel, 23
- bedingt unabhängig, 99
- bedingte Wahrscheinlichkeit, 21
- Behrens-Fisher-Problem, 180
- Benford, Frank, 138
- benfordsches Gesetz,
 - siehe* newcomb-benfordsches Gesetz
- Bernoulli, Jakob (1654–1705), 57
- Bernoulli-Experiment, 29
- Bernoulli-Verteilung, 29
- Berry-Esseen
 - Satz von, 85, 163
- Bertrands Paradox, 137
- Besetzungszahl, 17
- Beta-Funktion, 193
- Bias, 64
- Bienaymé, Irénée Jules (1796–1878), 53, 57, 111
- Bienaymé-Galton-Watson-Prozess, 111
- Bienaymé-Tschebyschew
 - Ungleichung von, 57
- Bindungen, 188
- Binomial
 - koeffizient, 14
 - verteilung, 29, 30
- Binomischer Lehrsatz, 15
- Blyth, Colin R., 74, 85
- Bohlmann, Georg (1869–1928), 39, 136
- Bootstrap-Methode, 173
- Borel, Emile (1871–1956), 157
- Borel-Cantelli-Lemma, 154
- borelsche
 - Menge, 130
 - σ -Algebra, 130
- Bortkiewicz, Ladislaus von (1868–1931), 17, 90
- Bose-Einstein-Statistik, 19
- Broggi, Ugo (1880–1965), 136
- Bruns, Heinrich (1848–1919), 17
- Buchner, P., 90

- Cantelli, Francesco P. (1875–1966), 154
- Cardano, Geronimo (1501–1576), 17
- cauchy-schwarzsche Ungleichung, 54
- Cauchy-Verteilung, 179
- champernownsche Zahl, 157
- Chapman-Kolmogorow-Gleichung, 198–208
- charakteristische Funktionen, 158
- χ^2
 - -Test, 182
 - -Verteilung, 178
- Code, 114
 - Präfixcode, 114
- coupling, 90
- Cramér, Harald (1893–1985), 67
- Cramér-Rao-Ungleichung, *siehe* Informationsungleichung

- Denker, Manfred, 163
- Diaconis, Persi, 100, 173
- Dichte, 128–136
 - der Normalverteilung, 133
 - eindimensionale, 131
 - mehrdimensionale, 135
 - Produktdichte, 136
- diploid, 35
- Dirichlet-Problem, 203
- Doebelin, Wolfgang (1915–1940), 90, 226

- Edgeworth, Francis Ysidro (1845–1926), 163
- Edgeworth-Entwicklung, 163
- Efron, Bradley, 173
- Ehrenfests Modell, 224
- Einschluss-Ausschluss-Formel, 50
- Einstichprobenproblem, 180
- Elementarereignis, 3

empirische
 — Varianz, 165
 — Verteilungsfunktion, 163
 empirischer Regressionskoeffizient, 169
 Entropie, 116
 Erdős, Paul, 226
 Ereignis
 — sicheres, 3
 — unmögliches, 3
 Ereignisse, 3, 130
 — unvereinbare, 3
 Erfolgswahrscheinlichkeit, 29
 Ergebnis, 3
 Ergebnismenge, 3
 Ergodensatz, 162
 Erneuerungsgleichung, 218
 Erneuerungssatz, 218–219
 erwartungstreu, 64
 Erwartungswert, 46
 erzeugende Funktion, 107
 Erzeugendensystem, 140
 exakter Test von Fisher, 187, 194
 explorative Datenanalyse, 173
 Exponentialverteilung, 132, 228–232

F

— -Verteilung, 74, 84, 135
 — — mit m und n Freiheitsgraden, 182
 — — nichtzentrale, 182
 faires Spiel, 58
 Falk, M., 173
 Faltung, 86, 144
 fast
 — alle, 155
 — sicher, 155
 — überall, 155
 Fehler
 — erster Art, 94
 — zweiter Art, 94
 — dritter Art, 97
 Fehlerfortpflanzungsgesetz, 167
 — gaußsches, 168
 Fehlergrenzen, 166, 172
 Feller, William (1906–1970), 109, 163, 226
 Fermat, Pierre de (1601–1665), 6, 17
 Fermi-Dirac-Statistik, 18
 Fisher, Sir Ronald Aylmer (1890–1962), 67, 92, 182, 187
 Fisher-Information, 67
 Fournival, Richard de (1201–1260), 17
 Fréchet, Maurice (1878–1973), 67
 Freiheitsgrad, 74, 84, 178, 182, 187

 Galton, Francis (1822–1911), 111
 Gammaverteilung, 192

Gauß, Carl Friedrich (1777–1855), 80, 168, 172, 173
 Gauß-Markow
 — Satz von, 173
 Geburts- und Todesprozess, 225
 Gen, 35
 Genotyp, 35, 197
 geometrische Verteilung, 30
 Gesamtmittelwert, 181, 182
 Gesamtvarianz, 182
 Gesetz der großen Zahlen, 17, 184, 197
 — kolmogorowsches, 162
 — schwaches, 56–58
 — starkes, 153–162, 171, 215
 getrimmtes Mittel, 171
 Gleichverteilung, 5, 132
 — in einem Gebiet, 135
 Glockenkurve
 — gaußsche, 80
 Gosset, William Sealy (1876–1937), 179
 Gütefunktion, 95, 100

Häufigkeit

— relative, 4
 Hahn-Banach
 — Satz von, 39
 Hardy-Weinberg-Theorem, 37
 harmonisch, 203
 Hausdorff, Felix (1868–1942), 137, 162
 Hilbert, David (1862–1943), 136
 Hill, T. P., 138
 Histogramm, 79
 Hodges, Joseph L., 90, 172
 Huygens, Christiaan (1629–1695), 17
 hypergeometrische Verteilung, 12
 Hypothese, 94
 — Annahme der, 94
 — einfache, 101
 — Verwerfen der, 94
 — zusammengesetzte, 101

Indikatorfunktion, 49

Information, 118
 Informationsungleichung, 67
 invariantes Wahrscheinlichkeitsmaß,
siehe Verteilung, invariante

Irrfahrt

— auf $\mathbb{Z}$, 213
 — auf $\mathbb{Z}^d$, 213
 — mit absorbierendem Rand, 199, 201, 204

Kersting, Götz-Dietrich, 158, 163

Kleinste-Quadrate-Schätzer, 173

Koinzidenz, 52

— verteilung, 52

- Kolmogorow, Andrej Nikolajewitsch
 (1903–1987), 136, 162, 198, 226
 Kombination, 7
 Konfidenz
 — bereich, 70
 — intervall, 69, 70
 — niveau, 70
 — schranke, 70
 Kontingenztafel, 186
 Konvergenz
 — fast sichere, 153
 — stochastische, 153
 Korrelationskoeffizient, 53
 Kovarianz, 53
 Kovarianzmatrix, 149
 kraftsche Ungleichung, 115
 kritischer
 — Bereich, 94
 — Wert, 94
 Kuratowski, K. (1896–1980), 136

 Lagerhaltung, 196
 Laplace
 — -Experiment, 5
 — -Wahrscheinlichkeit, 5
 Laplace, Pierre Simon Marquis de
 (1749–1827), 17, 80, 81, 90
 Le Cam, Lucien, 90
 Legendre, Adrien Marie (1752–1833), 172
 Lehmann-Hodges-Mittel, 172
 Likelihood
 — -Funktion, 62, 164
 — -Quotient, 175
 — -Quotienten-Test, 175
 linearer Schätzer, 173
 lineares Modell, 173
 Ljapounow, Alexandr Michajlowitsch (1857–
 1918), 163
 log-Likelihood-Funktion, 62
 lokaler Grenzwertsatz, 78

 Mantissen- σ -Algebra, 138
 Marginalverteilung, 44
 Markow, Andrej Andrejewitsch (1856–1922),
 57, 136, 173, 195, 197
 markowsch
 — n -Schritt, 204
 markowsche
 — Eigenschaft, 195, 197, 228
 — — starke, 216
 — Kette, 195
 — — aperiodische, 208
 — — homogene, 198
 — — mit stationären Übergangswahrschein-
 lichkeiten, 198
 — — periodische, 208
 — — rekurrente, 212
 — — transiente, 212
 — — umgekehrte, 216
 — Ungleichung, 57
 markowsche Ungleichung, 57
 match, *siehe* Koinzidenz
 Maximum-Likelihood
 — -Ansatz, 60
 — -Gleichung, 62
 — -Schätzer, 60, 62
 — -Schätzung, 62
 Maxwell-Boltzmann-Statistik, 18
 Median, 170
 Méré, Antoine Gombaud Chevalier de (1607–
 1684), 17
 Mergesort, 120
 messbare Abbildung, 140
 messbare Funktion, *siehe* messbare Abbildung
 messbarer Raum, 130
 Methode der kleinsten Quadrate, 57, 168, 172
 Mikrozustand, 17
 Mises, Richard von (1883–1953), 134, 136
 Von-Mises-Verteilung, 134
 Mittelwert, 64
 mittlerer quadratischer Fehler, 65
 Modell
 — statistisches, 61
 — wahrscheinlichkeitstheoretisches, 1
 de Moivre-Laplace
 — Satz von, 80
 Moivre, Abraham de (1667–1754), 80, 81, 90,
 163
 Moment
 — k -tes, 53
 — zweites zentrales, 53
 de morganschen Regel, 49
 Multinomial
 — koeffizient, 14
 — verteilung, 30
 Murmeln
 — unterscheidbare, 9
 — ununterscheidbare, 9

 Nachricht, 114
 negative Binomialverteilung, 31
 Neumann, Johann von (1903–1957), 126
 Newcomb, Simon, 137
 newcomb-benfordsches Gesetz, 138
 Neyman, Jerzy (1894–1981), 92, 93
 Neyman-Pearson
 — -Lemma, 101, 175
 — -Test, 101
 Niveau eines Tests, 95
 — effektives, 105

- Normalapproximation, 76, 183
 - Korrekturterme in der, 83
- normale Zahl, 157
- Normalgleichungen, 169
- Normalverteilung, 133
 - mehrdimensionale, 149
 - — ausgeartete, 149
 - — nichtausgeartete, 149
 - standardisierte, 78, 81
- normierte
 - Mengenfunktion, 3
 - Summe von Zufallsvariablen, 81
- Nullhypothese, 94

- Ordnungsstatistik, 171
- Osterhoff, J., 105

- Paley, Raymond (1907–1933), 136
- Palm, C., 226
- Parameterbereich, 70, 195
- Pascal, Blaise (1623–1662), 6, 17
- Pascal-Verteilung, *siehe* negative Binomialverteilung
- Pascals Dreieck, 16
- Pauli-Verbot, 18
- Pearson, Egon S. (1895–1980), 93
- Pearson, Karl (1857–1936), 184
- Permutation, 7, 8
- Pflug, G. Ch., 163
- Phänotyp, 36
- Phasenraum, 17
- Poisson, Siméon Denis (1781–1840), 90
- Poisson
 - -Approximation, 85
 - -Prozess, 227
 - -Verteilung, 86
- Pollard, H., 226
- Pólya, Georg (1887–1985), 32
- Potenzmenge, 3
- Prinzip des Schäfers, 8
- Produkt von Wahrscheinlichkeitsräumen, 27
- Produktdichte, 136
- Produktexperiment, 27
- Produktformel, 25
- p -Wert, 105

- Quantil, 172, 179
- Quellen-Codierungssatz, 114
- Quicksort, 120

- Rajchman, 155
- Randverteilung, 44
- Rang, 188
 - summentest, 189
- Rao, C. Radhakrishna, 67

- Realisierung, 3
- Regressionsgerade, 169
- Rekurrenz
 - kriterium, 212
 - satz, 212
- resampling, 173
- reversibel, 226
- Rösler, Uwe, 125, 163
- Rückkehrzeit, 208
- Run, 15, 108

- Satz von
 - Banach und Kuratowski, 136
 - Berry-Esseen, 85, 163
 - Hahn-Banach, 39
 - de Moivre-Laplace, 80
- schärfer, 101
- Schätzer, 61
 - konsistenter, 68
 - robuster, 171
- Schätzwert, 62
- Scheinkorrelation, 37
- Shannon, Claude E., 118
- σ
 - -Algebra, 129
 - — erzeugte, 129
 - — Mantissen-, 138
 - -Körper, 129
 - -additiv, 31, 39, 130, 136
- signifikant, 104
- Signifikanzniveau eines Tests, 95
 - beobachtetes, 105
- Simulation, 173
- Skat, 12
- Sprungzeit, 230
- Stabdiagramm, 43
- Standardabweichung, 52
- standardisierte Summe von
 - Zufallsvariablen, 81
- Startverteilung, 199
- Statistische Physik, 17
- Stetigkeit für monotone Folgen von Ereignissen, 32
- Stichprobe, 3
 - geordnete, 7
 - in Reihenfolge, 7
 - mit Wiederholung, 7
 - mit Zurücklegen, 7
 - ohne Rücklegen, 7
 - ohne Reihenfolge, 7
 - ohne Wiederholung, 7
- Stichproben
 - raum, 3
 - umfang, 84
 - unverbundene, 180

- verbundene, 180
- Still, Harold A., 74, 85
- stirlingsche Formel, 76, 89
- stochastische Matrix, 198
- stochastischer Prozess, 195
- Stoppzeit, 211
- Streuung, 52
- studentische Verteilung, *siehe* *t*-Verteilung
- Stufen eines Experiments, 7, 33

- t*
 - -Test
 - — einseitiger, 179
 - — zweiseitiger, 179
 - -Verteilung
 - — mit *k* Freiheitsgraden, 178
 - — nichtzentrale, 180
- Taxiproblem, 63–65
- tea tasting Lady, 92–105
- Test, 92, 94, 100, 175
 - auf Unabhängigkeit, 186
 - einseitiger, 179
 - Macht eines Tests, 95
 - nichtparametrischer, 187
 - randomisierter, 100
 - schärfster, 103
 - zweiseitiger, 179
- Teststatistik, 94
- ties, *siehe* Bindungen
- Tschebyschew, Pafnutij Lwowitsch (1821–1894), 57
- tschebyschewsche Ungleichung, 56

- Übergangswahrscheinlichkeit, 198
 - *m*-Schritt-, 200
- Übergangswahrscheinlichkeiten
 - stationäre, 198
- unabhängig, 25, 26, 45, 143
 - paarweise, 26
- Ungleichung
 - kraftsche, 115
 - markowsche, 57
 - tschebyschewsche, 56
 - von Bienaymé-Tschebyschew, 57
- unkorreliert, 53
- Urnenmodell, 6
 - pólyasches, 32

- Varianz, 52, 146
 - analyse, 180
 - — Klassenbildung in der, 181
 - innerhalb der Klassen, 182
 - zwischen den Klassen, 182
- verbundene Stichproben, 180
- Verteilung, 42, 142
 - gedächtnislose, 132
 - gemeinsame, 43
 - invariante, 206–226
 - mit Mehrfachbesetzung, 10
 - ohne Mehrfachbesetzung, 10
 - stationäre, 162
 - symmetrische, 170
- Verteilungen
 - austauschbare, 34
- Verteilungsfunktion, 81, 131, 135, 142
- Verwerfungsbereich, 94
- Vorzeichenrangtest, 192

- Wärmebewegung, 224
- Wahrscheinlichkeit, 3, 128
 - bedingte, 37
 - totale, 23
- Wahrscheinlichkeitsfunktion, 4
- Wahrscheinlichkeitsmaß, 3, 130
- Wahrscheinlichkeitsraum, 3, 130
 - diskreter, 31
 - laplacescher, 5
- wahrscheinlichkeitstheoretische Zahlentheorie, 39
- Wahrscheinlichkeitstheorie
 - geometrische, 137
- Wahrscheinlichkeitsvektor, 5
- Wahrscheinlichkeitsverteilung, 3, 130
- waldsche Identität, 110
- Warteschlange, 196, 199, 213, 214, 225
- Watson, Henry William, 111
- Weibull-Verteilung, 134
- Wiener, Norbert (1894–1964), 136
- Wilcoxon-Zweistichprobentest, 189
- Wiman, Anders (1865–1959), 136
- Winsorisierte Stichprobenvarianz, 172
- Winsorisiertes Mittel, 172

- Zahlenlotto, 11
- Zeichentest, 187
- Zellen, 9
- Zentraler Grenzwertsatz, 158
- zentrieren, 53
- Zerlegung, 23
- zufällig, 5
- Zufalls
 - experiment, 1
 - größe, 42
 - variable, 42, 140
 - vektor, 42
- Zustand
 - aperiodischer, 208
 - nullrekurrenter, 222
 - Nullzustand, 222
 - Periode eines Zustands, 208

- positiv rekurrenter, 222
- positiver, 222
- rekurrenter, 212
- transienter, 212
- wesentlicher, 208, 212
- Zustände
- kommunizierende, 208
- Zustandsraum, 195
- Zuverlässigkeitstheorie, 49
- Zweistichprobenproblem
- mit unverbundenen Stichproben, 180
- van Zwet, W. R., 105
- Zygmund, Antoni (1900–), 136