

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

LEHRBUCH

Ehrhard Behrends

# Elementare Stochastik

Ein Lernbuch -  
von Studierenden mitentwickelt



Springer Spektrum

---

# Elementare Stochastik

---

Ehrhard Behrends

# Elementare Stochastik

Ein Lernbuch –  
von Studierenden mitentwickelt

 Springer Spektrum

Prof. Dr. Ehrhard Behrends  
Freie Universität Berlin  
Deutschland

ISBN 978-3-8348-1939-0  
DOI 10.1007/978-3-8348-2331-1

ISBN 978-3-8348-2331-1 (eBook)

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie;  
detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de> abrufbar.

Springer Spektrum  
© Spektrum Verlag | Springer Fachmedien Wiesbaden 2013  
Das Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt. Jede Verwertung, die nicht ausdrücklich vom Urheberrechtsgesetz zugelassen ist, bedarf der vorherigen Zustimmung des Verlags. Das gilt insbesondere für Vervielfältigungen, Bearbeitungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Verarbeitung in elektronischen Systemen.

Die Wiedergabe von Gebrauchsnamen, Handelsnamen, Warenbezeichnungen usw. in diesem Werk berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Warenzeichen- und Markenschutz-Gesetzgebung als frei zu betrachten wären und daher von jedermann benutzt werden dürften.

Gedruckt auf säurefreiem und chlorfrei gebleichtem Papier

Springer Spektrum ist eine Marke von Springer DE.  
Springer DE ist Teil der Fachverlagsgruppe Springer Science+Business Media  
[www.springer-spektrum.de](http://www.springer-spektrum.de)

# V o r w o r t

In diesem Buch geht es um die Mathematik des Zufalls. Das scheint auf den ersten Blick ein Widerspruch in sich zu sein, denn wie soll man mit so einer exakten Wissenschaft das Ungewisse beschreiben? Wirklich hat es sehr lange gedauert, bis für die Wahrscheinlichkeitsrechnung eine mathematisch exakte Grundlage geschaffen wurde. Das ist noch nicht einmal hundert Jahre her, und damit ist das Gebiet viel jünger als andere Bereiche wie etwa die Geometrie oder die Algebra, die auf eine Geschichte von mehreren tausend Jahren zurückblicken können.

Die Wahrscheinlichkeitstheorie ist inzwischen sehr weit entwickelt, und immer wieder werden neue Anwendungsbereiche erschlossen. Sie, liebe Leserinnen und Leser, können dieses faszinierendes Teilgebiet der Mathematik hier kennen lernen. Es gibt – anders als bei vielen anderen mathematischen Theorien – eine Fülle von Beziehungen zur Alltagserfahrung, und die meisten der wichtigen Konzepte sind nichts weiter als eine Präzisierung von Mechanismen, mit denen wir täglich mit dem Zufall umgehen.

Sie können sich davon überzeugen, dass uns die Evolution für das intuitive Erfassen von zufälligen Phänomenen sehr unterschiedlich vorbereitet hat. Manche – wie etwa die in Kapitel 4 zu besprechenden bedingten Wahrscheinlichkeiten – können wir in Bruchteilen von Sekunden richtig einschätzen, bei anderen liegen wir mit unserer Beurteilung völlig falsch. Deswegen gibt es in diesem Gebiet auch so viele Paradoxien: Sachverhalte, die beweisbar richtig sind, bei denen aber die meisten etwas ganz anderes erwartet hätten.

Zwei Besonderheiten dieses Buches sollen noch hervorgehoben werden. Erstens wird hier ein Kompromiss versucht: Alle, die das Gebiet so faszinierend finden, dass sie darin weiterarbeiten wollen, werden dafür nach dem Durcharbeiten der zwölf Kapitel eine belastbare Grundlage haben. Und für diejenigen, die nur die wesentlichen Ideen kennen lernen wollen, finden sich es am Ende der Einleitung Hinweise darauf, auf welche Aspekte sie sich konzentrieren sollten.

Und zweitens: Ich habe mich sehr darüber gefreut, dass eine Reihe von Studierenden, die meine Veranstaltungen in den letzten Jahren besucht haben, dazu bereit waren, mich während des Entstehungsprozesses des vorliegenden Buches zu beraten. So konnte der Nutzen für zukünftigen Leser ganz bestimmt gesteigert werden, und dafür möchte ich mich an dieser Stelle ganz herzlich bedanken.

Berlin, im Juni 2012

Ehrhard Behrends



Das Team: Maximilian Kollock, Adam Schienle, Ehrhard Behrends, Michael Brückner,  
Thomas Stollin (hinten); Kristine Kaiser, Ekkehard Schnoor, Sophie Knell (vorn).

## Vorwort des studentischen Teams

Wenn man ausgerechnet während einer Klausur vom prüfenden Dozenten angesprochen wird, lässt es einen schwer etwas Gutes vermuten. So ging es einigen von uns, auf die Professor Behrends während einer Prüfung zur Elementaren Stochastik zukam. Die Befürchtung, eines vermeintlichen Schummelversuchs bezichtigt zu werden, war jedoch unbegründet. Ganz im Gegenteil erfolgte eine Einladung zur Teilnahme an diesem Projekt, zu dem sich insgesamt sieben Studenten gefunden haben, um mit Vorschlägen, Ideen, aber auch Kritik und Änderungswünschen bei der Entstehung des vorliegenden Buches mitzuhelfen.

Wir hoffen, dass die Arbeit daran auch anderen Studenten und den Lesern dieses Buches insgesamt zu Gute kommen wird. Für uns alle war sie eine große Freude und Bereicherung. Ziel dieses Buches ist es, nicht nur Mathematikstudenten anzusprechen, die sich im Rahmen ihres Studiums mit Elementarer Stochastik befassen und denen dieses Buch als Begleitmaterial zur Vorlesung dient, sondern es soll auch jenen Lesern einen Einblick in die Materie gewähren, die sich als Laien mit diesem Gebiet beschäftigen wollen. Um für alle Leser eine größtmögliche Verständlichkeit des Textes zu erreichen, hat Herr Behrends uns als seine Studenten einbezogen, um durch unsere Mitarbeit auch solche Schwierigkeiten des Stoffes nicht aus den Augen zu verlieren, derer sich Hochschulprofessoren aufgrund ihrer jahrelangen Berufserfahrung mitunter gar nicht mehr bewusst sind. Unsere Probleme und Anmerkungen wurden regelmäßig und angeregt mit Herrn Behrends diskutiert und die Stellen, die uns bei unserer Lektüre als schwer verständlich erschienen, sind so mit zusätzlichen Erläuterungen versehen worden, von denen wir hoffen, dass sie zur Lesbarkeit dieses Buches für alle beitragen.

Für einige der Übungsaufgaben, die sich am Ende jedes Kapitels befinden, haben wir Musterlösungen erstellt, die auf der Homepage zum Buch zu finden sind. (Die Adresse ist in der Einleitung zu finden.)

Im Bewusstsein, aus der Mitarbeit an diesem Buch selbst Vieles gelernt zu haben, bedanken wir uns hiermit bei Herrn Behrends für viele Stunden guter, intensiver und produktiver Diskussion bei Tee und Keksen und freuen uns sehr, dieses spannende und - wie wir hoffen - für alle Leser hilfreiche Buch mitentwickelt zu haben.

Berlin, im Juni 2012

Michael Brückner, Kristine Kaiser, Sophie Knell, Maximilian Kollock,  
Adam Schienle, Ekkehard Schnoor, Thomas Stollin

## E i n l e i t u n g

Es wurde schon im Vorwort bemerkt, dass viele Begriffe, mit denen man sich im Studium beschäftigt, mit unserer Erfahrungswelt auf den ersten Blick recht wenig zu tun haben. Wo findet man denn dort – zum Beispiel – endliche Körper oder konvergente Folgen? Ganz anders ist es mit dem Zufall. Jeder weiß, dass er bei Glücksspielen wie etwa beim Lotto eine Rolle spielt und dass das Thema eine fundamentale Bedeutung hat, wenn es um Versicherungen oder Zuverlässigkeitberechnungen geht. Auch ist seit fast hundert Jahren bekannt, dass die Natur nur noch in Begriffen der Wahrscheinlichkeitstheorie beschrieben werden kann, wenn es um die Mikrowelt geht. Die Vorhersagen der Quantentheorie stimmen hervorragend mit den Messungen überein. Unsere Illusion, in einer von deterministischen Naturgesetzen regierten Welt zu leben, röhrt daher, dass wir immer nur das Zusammenwirken von gigantisch vielen Elementarteilchen wahrnehmen, und dabei verschwindet das Zufällige. (Mehr dazu in Teil 3 dieses Buches: „Der Zufall verschwindet im Unendlichen“.)

Überraschenderweise hat sich die Wahrscheinlichkeitsrechnung erst vergleichsweise spät gleichberechtigt neben anderen mathematischen Gebieten etabliert. Glücksspiele sind schon bei den ältesten Kulturen anzutreffen, und sicher hatten die Spieler auch ein Gefühl für die dabei relevanten Wahrscheinlichkeiten, aber Mathematiker haben sich damit nicht beschäftigt.

Die Aufnahme des Gebiets in die Mathematik geschah – wenn man es stark vereinfacht beschreibt – in zwei Schritten. Der erste ist auf das 17. Jahrhundert zu datieren. In dieser Zeit gab es die ersten Versuche, Gesetzmäßigkeiten bei Zufallsphänomenen zu beschreiben. Als historischer Ausgangspunkt gilt die Frage des Adligen Antoine Gombaud de Méré an den Mathematiker und Philosophen Blaise Pascal, wie man bei einem Spiel die Gewinnsumme aufteilen sollte, wenn aus irgendwelchen Gründen ein vorzeitiger Abbruch erforderlich ist. Pascal korrespondierte darüber mit dem Mathematiker Pierre de Fermat, und recht bald gab es die ersten allgemeinen Erkenntnisse zum Phänomen „Zufall“.

Es ging dann rasant aufwärts, und viele wichtige Ergebnisse wurden schon im 17. und 18. Jahrhundert gefunden. Die klügsten Köpfe der Mathematikgeschichte haben Ideen beigesteuert, aber eine exakte Grundlegung des Gebietes ließ lange auf sich warten. Damit war die Wahrscheinlichkeitsrechnung in einer ähnlichen Situation wie die Analysis, in der ja auch große Fortschritte unter Verwendung des vagen Konzepts der „unendlich kleinen Größen“ erzielt wurden, bis im 19. Jahrhundert die Grundlagen geklärt wurden.

In der Theorie des Zufalls dauerte diese Klärung wesentlich länger. 1933 gilt als das Geburtsjahr der modernen Wahrscheinlichkeitstheorie. In diesem Jahr stellte der russischen Mathematiker Kolmogoroff ein Axiomensystem vor, das bis heute Bestand hat. Seit dieser Zeit liegt die Grundlage in der Maßtheorie, einem Gebiet, in dem es ursprünglich um das Messen von Flächen und Volumina ging.

Die Wissenschaft vom Zufall wird heute als „Stochastik“ bezeichnet. Der Name ist vom griechischen „στοχαστικὴ τέχνη“ (*stochastike techne*) abgeleitet,

das bedeutet so viel wie „die Kunst des Vermutens“. (Im Lateinischen heißt das Gebiet genauso: „*ars conjectandi*“.) Stochastik ist der Sammelbegriff für Wahrscheinlichkeitsrechnung und Statistik, beide Teilgebiete kann man in diesem Buch kennen lernen.

Die Stochastik hat eine große Bedeutung erlangt, die immer noch zunimmt. Es gibt praktisch kein Teilgebiet der Mathematik mehr, in dem nicht auch stochastische Methoden eingesetzt werden, und es werden immer neue Anwendungsbereiche erschlossen. Ein vergleichsweise junges Beispiel ist die Finanzmathematik, in der jetzt tiefliegende wahrscheinlichkeitstheoretische Ergebnisse zur Risikoberechnung und zur Bewertung von Optionen eingesetzt werden.

Was erwartet die Leser in diesem Buch? In *Kapitel 1* werden die wichtigsten Aspekte des Phänomens „Zufall“ herausgearbeitet, Ausgangspunkt ist dabei die Alltagserfahrung. Dadurch soll die Definition von Wahrscheinlichkeitsräumen motiviert werden. Gegen Ende des Kapitels wird es dann etwas technischer, weil wir da schon einige allgemeine Ergebnisse zusammenstellen, die in den folgenden Kapiteln eine wichtige Rolle spielen werden. (Wer möchte, kann das systematische Durcharbeiten dieser Abschnitte erst einmal vertagen.) *Kapitel 2* ist dann der Vorstellung der wichtigsten Beispielklassen von Wahrscheinlichkeitsräumen gewidmet. Sie haben alle eine besondere Bedeutung, einigen wird sogar ein eigenes Kapitel gewidmet sein.

Danach geht es in *Kapitel 3* um Informationskompression: Wie lassen sich Wahrscheinlichkeitsräume vereinfachen, wenn man es gar nicht so genau wissen möchte. (Zum Beispiel ist es bei einem Lottoschein nicht so wichtig, welche Zahlen darauf angekreuzt sind, sondern nur, wie viele Richtige dabei sind.) Hier werden auch kombinatorische Tatsachen wichtig, um konkrete Rechnungen durchführen zu können. *Kapitel 4* steht dann unter dem Motto „Wie verändern Informationen Wahrscheinlichkeiten?“. Das klingt abstrakt, es handelt sich aber um einen Vorgang, auf den uns die Evolution gut vorbereitet hat; er ist im Gehirn quasi fest verdrahtet. Der Spezialfall, dass Informationen *keine* Auswirkungen haben, ist dabei besonders wichtig, er führt zum Begriff der Unabhängigkeit.

In den anschließenden Kapiteln, in *Kapitel 5* und in *Kapitel 6*, beschäftigen wir uns mit zwei besonders wichtigen Klassen von Wahrscheinlichkeitsräumen: Mit der *Binomialverteilung* und mit der *Exponentialverteilung*. Es geht dabei zum ersten Mal um die „Überlagerung“ von vielen Zufallseinflüssen und um die Frage, wie man beim Warten den Begriff „gedächtnislos“ präzisieren kann und welche Schlussfolgerungen dann daraus gezogen werden können.

Danach beginnen die Untersuchungen zum Verhalten des Zufalls im Unendlichen. *Kapitel 7* hat dabei vorbereitenden Charakter, dort wird präzisiert, welche Konvergenzbegriffe im Folgenden eine Rolle spielen werden. In dem recht umfangreichen *Kapitel 8* stehen dann die wichtigsten Sätze, mit denen das Konvergenzverhalten beschrieben wird. Alle sagen aus, dass der Zufallseinfluss bei der Überlagerung von vielen Einzelergebnissen mehr und mehr verschwindet, doch je nach Situation wird das unterschiedlich präzisiert. Eine ganz besondere Bedeutung hat dabei der *zentrale Grenzwertsatz*, in dem die universelle

Bedeutung der Gaußschen Glockenkurve bewiesen wird.

Damit ist der wahrscheinlichkeitstheoretische Teil des Buches abgeschlossen, die verbleibenden vier Kapitel beschäftigen sich mit *mathematischer Statistik*. Vereinfacht kann man sagen:

- In der Wahrscheinlichkeitstheorie sind die Wahrscheinlichkeitsräume gegeben, und daraus leitet man Prognosen ab.
- In der Statistik ist der gerade relevante Wahrscheinlichkeitsraum unbekannt, man weiß nur, dass er zu einer speziellen Klasse gehört. Nun wird eine Stichprobe genommen, und daraus sollen Rückschlüsse auf diesen Raum gezogen und vielleicht auch Empfehlungen für damit zusammenhängende Entscheidungen gegeben werden.

*Kapitel 9* hat den Charakter einer Einführung in die neuen Fragestellungen: Welche Daten sind interessant, wie kann man sie visualisieren, welche Größen gestatten einen ersten Überblick? Etwas formaler wird es in *Kapitel 10*, in dem *statistische Modelle* als genügend allgemeine Definition zur Behandlung von statistischen Fragestellungen eingeführt werden. Es wird untersucht, was „gute“ Lösungen auszeichnet und wie man sie in vielen wichtigen Fällen konkret angeben kann. Danach, in *Kapitel 11*, geht es darum, statistische Methoden zur Entscheidungshilfe heranzuziehen: „Sollte man dieses Medikament freigeben?“, „Ist es gerechtfertigt zu glauben, dass diese Münze fair ist?“, usw. Das wird präzisiert, und in vielen Fällen kann man ein bestmögliches Verfahren vorschlagen. Den Schluss bildet *Kapitel 12*, in dem so genannte parameter-unabhängige Verfahren studiert werden. Damit werden Probleme behandelt, bei denen es nicht primär um Zahlen geht. (Wie stellt man zum Beispiel fest, ob die vorliegenden Ergebnisse mit einem ganz bestimmten Wahrscheinlichkeitsraum erzeugt wurden?)

Damit sind wir noch nicht am Ende dieses Buches angelangt, es gibt noch einen *Anhang*. Er besteht aus mehreren Teilen, in denen an Sachverhalte aus anderen Bereichen der Mathematik erinnert wird, die in diesem Buch verwendet werden. Auch gibt es Tabellen, mit denen man – ohne erst mit Google-Hilfe im Internet suchen zu müssen – die hier entwickelten Methoden gleich an konkreten Beispielen behandeln kann.

Manche werden in diesem Buch eine philosophische Auseinandersetzung mit dem Thema „Zufall“ vermissen. Die ist bewusst ausgespart worden. Erstens gibt es bis heute keine allgemein akzeptierte Antwort auf die Frage, was denn „Zufall“ oder „Wahrscheinlichkeit“ eigentlich sind, wenn man es ganz genau nimmt. Und zweitens ist eine solche Antwort für das erfolgreiche Modellieren stochastischer Aspekte der Welt gar nicht erforderlich. Wichtig ist, dass die Voraussagen erfolgreich genutzt werden können, und das wird von allen – von der Lotteriesellschaft über die Spielbankbetreiber bis zu den Versicherungsunternehmen und vielen anderen Anwendern – bestätigt werden.

**Was sollte man schon wissen, wenn man dieses Buch liest?** Den meisten Untersuchungen wird man folgen können, wenn man mit der in allen mathematischen Theorien verwendeten Sprache vertraut ist: Was sind Mengen, Teilmengen, Mengensysteme und so weiter? Elementare analytische Kenntnisse sind jedoch schon ab Kapitel 2 erforderlich: Wie differenziert man? Wie ist das Integral  $\int_a^b f(x) dx$  für stetige Funktionen  $f$  definiert? ... Das wird üblicher Weise in den letzten Jahren der Schule und den ersten Semestern der Universität behandelt, es wird hier vorausgesetzt. An wenigen Stellen werden auch weitergehende Begriffe und Resultate aus der Analysis wichtig: Supremum und Infimum, Feinheiten der Reihenrechnung, der Transformationssatz für Gebietsintegrale. Das Wichtigste dazu ist im Anhang zusammengestellt.

Eine entsprechende Bemerkung ist zu weiteren Voraussetzungen zu machen: Es ist eher von Vorteil, wenn man sich in Maß- und Integrationstheorie und bei euklidischen Räumen auskennt. Zwingend notwendig ist es nicht, und die wichtigsten Tatsachen sind im Anhang zu finden.

**Die Internetseite zum Buch.** Unter der Adresse

<http://www.springer-spektrum.de/Buch/978-3-8348-1939-0/Elementare-Stochastik.html> wird eine Internetseite zu diesem Buch verfügbar sein. Dort findet man ergänzende Materialien:

- Die Lösungen ausgewählter Übungsaufgaben.
- Ein Computerprogramm, mit dem sich einige der hier behandelten Räume und Tatsachen visualisieren lassen (s.u.).
- Einige weitere Ergänzungen, zum Beispiel Filme zur Illustration der winzigen Wahrscheinlichkeiten, sechs Richtige im Lotto zu haben.

→ **Homepage!** Im Buch sind diejenigen Stellen, zu denen es ergänzendes Material gibt, am Rand wie nebenstehend gekennzeichnet.

**Das Computerprogramm zum Buch.** Ich bin der festen Überzeugung, dass man abstrakt definierte Wahrscheinlichkeitsräume und theoretische Ergebnisse durch Simulationen illustrieren sollte, um sie besser zu verstehen. Deswegen wird auf der oben angegebenen Internetseite ein Computerprogramm zum Buch zur Verfügung gestellt. (Mehr dazu im Anhang auf Seite 369).

An den entsprechenden Stellen in den zwölf Kapiteln findet man Hinweise, ob es zum gerade behandelten Thema eine Ergänzung durch Simulation und Visualisierung gibt. Am Rand findet man dann jeweils (so wie hier) eine entsprechende Notiz.

### Besonderheiten

Wie bei anderen Büchern, die in ein mathematisches Teilgebiet einführen, werden Kenner auch hier kaum etwas für sie Neues finden. Wichtig waren mir eine ausführliche Motivation der grundlegenden Ideen sowie eine mathematisch belastbare und gleichzeitig verständliche Herleitung der wichtigsten Resultate. Die Diskussionen mit meinem studentischen Team waren bei der Umsetzung dieser Idee eine große Hilfe.

Ich möchte zwei Ergebnisse hervorheben, die man in anderen Lehrbüchern zur elementaren Stochastik nicht findet. Erstens enthält das Buch einen neuen Beweis zu einem überraschenden Ergebnis über  $\sigma$ -Algebren (siehe Anhang, Seite 354), und zweitens zeigen wir, dass das zweite Lemma von Borel-Cantelli schon unter der Voraussetzung der paarweisen Unabhängigkeit der betrachteten Ereignisse richtig ist (Abschnitt 8.6): Das ist überraschend, denn im „Standardbeweis“ spielt die Unabhängigkeit der gesamten Familie eine wichtige Rolle.

### **Wie sollten Sie dieses Buch lesen?**

Es richtet sich an verschiedene Zielgruppen:

- An interessierente Nicht-Mathematiker (Oberschüler; Wissenschaftler aus Bereichen, für die Mathematik ein wichtiges Hilfsmittel darstellt; Mathematik-affine Laien.)
- Studierende, die die wichtigsten Ideen der Stochastik kennen lernen wollen.
- Studierende, die es ganz genau wissen wollen. Zum Beispiel deswegen, weil sie das Gebiet faszinierend finden und sich weiter darin vertiefen wollen. Oder deswegen, weil sie die stochastischen Aspekte ihres Lieblingsgebiets besser verstehen möchten.

Alles gleichzeitig kann ein einziger Text sicher nicht leisten. Ich habe versucht, den Ansprüchen der verschiedenen Zielgruppen dadurch gerecht zu werden, dass an verschiedenen Stellen Hinweise dazu gegeben werden, welche Teile man später lesen oder ganz weglassen kann. Die dritte Gruppe wird sicher alles systematisch durcharbeiten wollen, wobei es der persönlichen Vorliebe überlassen bleibt, ob man sich die eher technischen Abschnitte (zum Beispiel 1.4 bis 1.6) gleich zumutet oder erst dann, wenn die entsprechenden Ergebnisse gebraucht werden.

Alle anderen sollten sich die einzelnen Abschnitte bis zu ihrer persönlichen Belastbarkeitsgrenze vornehmen, rechtzeitig vor Frustrationsbeginn zum nächsten Abschnitt übergehen, bei Bedarf das Fehlende nachlesen und möglichst viele Übungsaufgaben selbstständig lösen.

Ich wünsche allen viel Erfolg beim Einarbeiten in die Geheimnisse der Stochastik!

Ehrhard Behrends

Berlin, im Juni 2012

# Inhaltsverzeichnis

|   |           |
|---|-----------|
| <b>I Wahrscheinlichkeitsräume</b>                         | <b>1</b>  |
| <b>1 Wie wird der Zufall modelliert?</b>                  | <b>3</b>  |
| 1.1 Ein sehr naiver Ansatz: Zufallsautomaten . . . . .    | 4         |
| 1.2 Die Präzision: $\sigma$ -Algebren . . . . .           | 6         |
| 1.3 Wahrscheinlichkeitsräume: Eigenschaften . . . . .     | 11        |
| 1.4 Erzeugte $\sigma$ -Algebren . . . . .                 | 14        |
| 1.5 Borelmengen . . . . .                                 | 17        |
| 1.6 Zwei wichtige Beweistechniken . . . . .               | 21        |
| 1.7 Ergänzungen . . . . .                                 | 26        |
| 1.8 Verständnisfragen . . . . .                           | 30        |
| 1.9 Übungsaufgaben . . . . .                              | 33        |
| <b>2 Erste Beispiele</b>                                  | <b>37</b> |
| 2.1 Diskrete Wahrscheinlichkeitsräume . . . . .           | 37        |
| 2.2 Wahrscheinlichkeitsdichten . . . . .                  | 42        |
| 2.3 Simulation diskreter Räume . . . . .                  | 52        |
| 2.4 Simulation: Räume mit Dichtefunktionen . . . . .      | 56        |
| 2.5 Ergänzungen . . . . .                                 | 62        |
| 2.6 Verständnisfragen . . . . .                           | 64        |
| 2.7 Übungsaufgaben . . . . .                              | 65        |
| <b>II Wichtige Konzepte</b>                               | <b>69</b> |
| <b>3 Zufallsvariable</b>                                  | <b>71</b> |
| 3.1 Was ist eine Zufallsvariable? . . . . .               | 71        |
| 3.2 Induzierte Wahrscheinlichkeitsräume . . . . .         | 75        |
| 3.3 Erwartungswert, Varianz und Streuung . . . . .        | 79        |
| 3.4 Elementare Kombinatorik . . . . .                     | 91        |
| 3.5 Berechnung induzierter Wahrscheinlichkeiten . . . . . | 97        |
| 3.6 Ergänzungen . . . . .                                 | 106       |
| 3.7 Verständnisfragen . . . . .                           | 109       |
| 3.8 Übungsaufgaben . . . . .                              | 111       |

|   |            |
|---|------------|
| <b>4 Bedingte Wahrscheinlichkeiten</b>                    | <b>115</b> |
| 4.1 Bedingte Wahrscheinlichkeiten: die Idee . . . . .     | 116        |
| 4.2 Der Satz von Bayes . . . . .                          | 122        |
| 4.3 Unabhängigkeit für mehr als zwei Ereignisse . . . . . | 128        |
| 4.4 Unabhängigkeit für Zufallsvariable . . . . .          | 134        |
| 4.5 Der „Klonsatz“ . . . . .                              | 141        |
| 4.6 Folgerungen aus der Unabhängigkeit . . . . .          | 146        |
| 4.7 Verständnisfragen . . . . .                           | 156        |
| 4.8 Übungsaufgaben . . . . .                              | 157        |
| <b>III Binomial- und Exponentialverteilung</b>            | <b>163</b> |
| <b>5 Die Binomialverteilung</b>                           | <b>165</b> |
| 5.1 Binomialverteilung: Definition . . . . .              | 166        |
| 5.2 Hypergeometrische Verteilung: Approximation . . . . . | 169        |
| 5.3 Approximation durch die Poissonverteilung . . . . .   | 171        |
| 5.4 Der Satz von de Moivre-Laplace . . . . .              | 175        |
| 5.5 Verständnisfragen . . . . .                           | 185        |
| 5.6 Übungsaufgaben . . . . .                              | 186        |
| <b>6 Die Exponentialverteilung</b>                        | <b>189</b> |
| 6.1 Gedächtnislose Wartezeiten . . . . .                  | 189        |
| 6.2 Kombinationen gedächtnisloser Wartezeiten . . . . .   | 194        |
| 6.3 Diskrete gedächtnislose Wartezeiten . . . . .         | 200        |
| 6.4 Verständnisfragen . . . . .                           | 203        |
| 6.5 Übungsaufgaben . . . . .                              | 204        |
| <b>IV Der Zufall verschwindet im Unendlichen</b>          | <b>207</b> |
| <b>7 Konvergenz von Zufallsvariablen</b>                  | <b>209</b> |
| 7.1 Konvergenz in Wahrscheinlichkeit . . . . .            | 210        |
| 7.2 Fast sicher punktweise Konvergenz . . . . .           | 211        |
| 7.3 Konvergenz in Verteilung . . . . .                    | 213        |
| 7.4 Verständnisfragen . . . . .                           | 219        |
| 7.5 Übungsaufgaben . . . . .                              | 220        |
| <b>8 Die Gesetze der großen Zahlen</b>                    | <b>223</b> |
| 8.1 Die Lemmata von Borel-Cantelli . . . . .              | 224        |
| 8.2 Das schwache Gesetz der großen Zahlen . . . . .       | 230        |
| 8.3 Das starke Gesetz der großen Zahlen . . . . .         | 237        |
| 8.4 Der zentrale Grenzwertsatz . . . . .                  | 244        |
| 8.5 Der Satz vom iterierten Logarithmus . . . . .         | 255        |
| 8.6 Ergänzungen . . . . .                                 | 260        |
| 8.7 Verständnisfragen . . . . .                           | 263        |
| 8.8 Übungsaufgaben . . . . .                              | 265        |

|  |            |
|--|------------|
| <b>V Grundlagen der Statistik</b>                        | <b>267</b> |
| <b>9 Beschreibende Statistik</b>                         | <b>271</b> |
| 9.1 Statistische Daten . . . . .                         | 271        |
| 9.2 Visualisierung von statistischen Daten . . . . .     | 272        |
| 9.3 Stichprobenmittel und Stichprobenvarianz . . . . .   | 275        |
| 9.4 Korrelation und Regression . . . . .                 | 279        |
| 9.5 Verständnisfragen . . . . .                          | 284        |
| 9.6 Übungsaufgaben . . . . .                             | 285        |
| <b>10 Schätzen</b>                                       | <b>289</b> |
| 10.1 Das statistische Modell, Schätzfunktionen . . . . . | 290        |
| 10.2 Güteeigenschaften für Schätzer . . . . .            | 293        |
| 10.3 Beispiele für Punktschätzer . . . . .               | 300        |
| 10.4 Konfidenzbereiche . . . . .                         | 304        |
| 10.5 Konfidenzintervalle: Normalverteilung . . . . .     | 307        |
| 10.6 Verständnisfragen . . . . .                         | 314        |
| 10.7 Übungsaufgaben . . . . .                            | 315        |
| <b>11 Entscheiden</b>                                    | <b>317</b> |
| 11.1 Hypothesen . . . . .                                | 317        |
| 11.2 Testfunktionen . . . . .                            | 320        |
| 11.3 Neyman-Pearson-Theorie . . . . .                    | 326        |
| 11.4 Verständnisfragen . . . . .                         | 333        |
| 11.5 Übungsaufgaben . . . . .                            | 334        |
| <b>12 Nichtparametrische Statistik</b>                   | <b>337</b> |
| 12.1 Der $\chi^2$ -Anpassungstest . . . . .              | 338        |
| 12.2 Der $\chi^2$ -Test auf Unabhängigkeit . . . . .     | 341        |
| 12.3 Rangtests . . . . .                                 | 342        |
| 12.4 Der Kolmogoroff-Smirnoff-Test . . . . .             | 346        |
| 12.5 Verständnisfragen . . . . .                         | 349        |
| 12.6 Übungsaufgaben . . . . .                            | 350        |
| <b>Anhänge</b>   | <b>353</b> |
| Mengenlehre . . . . .                                    | 353        |
| Vereinigungen von $\sigma$ -Algebren . . . . .           | 354        |
| Maßtheorie . . . . .                                     | 356        |
| Das Skalarprodukt auf dem $\mathbb{R}^n$ . . . . .       | 359        |
| Analysis . . . . .                                       | 359        |
| Tabellen . . . . .                                       | 362        |
| Die Computerprogramme zum Buch . . . . .                 | 369        |
| Literatur . . . . .                                      | 370        |
| <b>Register</b>  | <b>371</b> |

## Teil I

# Wahrscheinlichkeitsräume

# Kapitel 1

## Wie wird der Zufall modelliert?

Eine wichtige Aufgabe der Mathematik besteht darin, Modelle zur Beschreibung gewisser Aspekte unserer Erfahrungswelt bereitzustellen. Schon in der Antike wurden mit Hilfe elementarer Geometrie Felder vermessen, seit dem 17. Jahrhundert kann man auch „Bewegung“ beschreiben usw.

Das angemessene mathematische Modell, um über den Zufall präzise reden zu können, sind *Wahrscheinlichkeitsräume*, die werden wir in diesem ersten Kapitel einführen. Bemerkenswerterweise kann man dabei die Frage ausklammern, was denn „Zufall“ eigentlich ganz genau ist. Das ist bei den anderen Modellen aber ähnlich: Was ein Kreis ist, kann in der Mathematik exakt definiert werden. In der Natur gibt es aber keine exakten Kreise, doch trotzdem kann man die Formeln für Umfang und Flächeninhalt oft erfolgreich zur Lösung konkreter Probleme anwenden. Schlimmer noch, es ist sogar so, dass viele der verwendeten Begriffe gar nicht erklärt sind, wenn man es genau nimmt. Anwender haben zum Beispiel kein Problem damit, für die Länge und Breite eines rechteckigen Tisches eine konkrete reelle Zahl einzusetzen. Doch was genau ist „Länge“? Nicht einmal die fünfte Ziffer nach dem Komma dürfte wohldefiniert sein, wenn wir in Metern messen, und dass alles nur bis auf einen gewissen Fehler bekannt ist, wird sofort klar, wenn man über den Satz „die Länge dieses Tisches ist irrational“ nachdenkt. Im mathematischen Modell mag das eine sinnvolle Aussage sein, in der realen Welt ist sie völlig simileer.

Glücklicherweise macht das aber nichts, denn die am idealisierten Modell durchgeführten Rechnungen lassen sich trotzdem praktisch nutzen. Es ist nämlich wirklich bedeutungslos, dass ich die Fläche des Tisches nur bis auf einen Quadratmillimeter genau kenne, wenn ich wissen möchte, wie viel Farbe ich zum Streichen brauchen werde.

Diese Überlegungen gelten sinngemäß auch für Wahrscheinlichkeitsräume: Es sind idealisierte Modelle zur Beschreibung zufälliger Phänomene, doch solche Modelle sind in der realen Welt nicht zu finden, wenn man es ganz genau

nimmt. Sie eignen sich aber trotzdem hervorragend für Anwendungen in vielen verschiedenen Bereichen: Wahlprognosen, Risikoabschätzungen bei Banken und Versicherungen, Gewinnaussichten beim Lotto oder in der Spielbank usw.

Es folgt eine Übersicht über dieses Kapitel. In Abschnitt 1.1 werden einige charakteristische Aspekte des Zufalls zusammengestellt, um die folgenden Definitionen zu motivieren. Was Mathematiker unter einem Wahrscheinlichkeitsraum verstehen, lernen Sie dann in Abschnitt 1.2. Dort werden auch einige erste Beispiele vorgestellt. Danach geht es weiter in Abschnitt 1.3 mit einigen allgemeinen Eigenschaften von solchen Räumen. In den Abschnitten 1.4, 1.5 und 1.6 wird es dann etwas technischer. Wir beweisen Ergebnisse, die unsere spätere Arbeit mit Wahrscheinlichkeitsräumen wesentlich erleichtern werden. Das Kapitel schließt in den Abschnitten 1.7, 1.8 und 1.9 mit Ergänzungen, Verständnisfragen und Übungsaufgaben.

## 1.1 Ein sehr naiver Ansatz: Zufallsautomaten

Wir beginnen mit einer Zusammenstellung von Situationen aus unserem Alltag, die nach allgemeiner Einschätzung als „zufällig“ bezeichnet werden:



Bild 1.1.1: Der Zufall: Glücksspiele.

- Sie werfen einen Würfel. Welche Zahl wird gewürfelt werden?
- Sie schlagen ein Buch auf, das auf Französisch geschrieben ist. Wird es auf dieser Seite eine Vokabel geben, die Sie nicht kennen?
- Jetzt nehmen Sie an einer Skatrunde teil. Wie viele Buben werden Sie haben?
- Sie rufen Ihren besten Freund / Ihre beste Freundin an, es ist besetzt. Wie lange werden Sie warten müssen, bis die Leitung wieder frei ist? Werden es mehr als zehn Minuten sein?
- Ein Geigerzähler wird an ein schwach radioaktives Präparat gehalten. Wie oft wird es in der nächsten Minute einen Klick geben?
- Wird Sie Ihr Lottoschein bei der nächsten Ausspielung reich machen? Wie viele Richtige wird es für Sie geben, werden es mehr als vier sein?

Es ist klar, dass sich diese Liste beliebig verlängern ließe. Es gibt *eine Reihe von Gemeinsamkeiten*:

- Man macht so etwas wie ein „Zufallsexperiment“: Es gibt eine genau umschriebene „Versuchsanordnung“, und dann wird das „Experiment“ durchgeführt.
- Es ist nicht klar, welches Ergebnis herauskommen wird. Man weiß jedoch, dass es in einer ganz bestimmten Menge liegen wird: In den Beispielen waren es die Mengen  $\{1, 2, 3, 4, 5, 6\}$  (Würfel) bzw.  $\{\text{ja,nein}\}$  (Vokabel) bzw.  $\{0, 1, 2, 3, 4\}$  (Skat) bzw.  $[0, +\infty[$  (Telefon) bzw.  $\{0, 1, 2, \dots\}$  (Geigerzähler) bzw.  $\{0, 1, 2, 3, 4, 5, 6\}$  (Lotto).
- Manchmal interessiert einen das genaue Ergebnis, in anderen Fällen möchte man nur wissen, ob es in einer bestimmten Teilmenge der möglichen Ergebnisse liegt. Zum Beispiel in  $[10, \infty[$  im Telefonbeispiel oder in  $\{5, 6\}$  im Lottobeispiel.
- Wenn man das „Experiment“ sehr, sehr oft durchgeführt hat, zeichnen sich gewisse „Tendenzen“ ab: In etwa einem Sechstel aller Würfelwürfe erscheint eine 1, in etwa 5 Prozent aller Fälle müssen Sie eine Vokabel nachschlagen, usw.
- Viel mehr weiß man eigentlich nicht. Wenn Sie das „Experiment“ noch einmal machen, können Sie immer noch nicht mit Sicherheit voraussagen, wie es ausgehen wird. Auch wenn Sie es schon sehr oft durchgeführt haben.

Damit wir für so eine Situation ein Bild vor Augen haben, soll der Begriff des *Zufallsautomaten* eingeführt werden. Das ist ein Kasten mit einem Knopf oben drauf. Wenn wir auf den drücken, wird eine „Zufallsausgabe“ produziert und sofort ausgegeben.

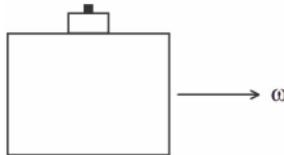


Bild 1.1.2: Der „Zufallsautomat“.

Aufgrund der vorstehenden Zusammenstellung der für den Zufall charakteristischen Aspekte ist ein Zufallsautomat wie folgt beschreibbar:

- Die Ausgaben des Automaten liegen in einer bekannten Menge  $\Omega$ . (Der griechische Buchstabe  $\Omega$  ist in der Wahrscheinlichkeitsrechnung das für die „möglichen Ergebnisse“ übliche Symbol. Es wird „Omega“ ausgesprochen.) Die Elemente aus  $\Omega$  heißen *Elementarereignisse*, und deswegen nennt man – logisch, aber schwerfällig –  $\Omega$  die Menge der Elementarereignisse.

- Wenn man den Automaten abfragt, wird er also ein Ergebnis  $\omega \in \Omega$  liefern<sup>1)</sup>. Häufig will man gar nicht wissen, welches  $\omega$  genau erzeugt wurde, sondern nur, ob es in einer gewissen Teilmenge  $E$  von  $\Omega$  liegt. So ein  $E$  heißt dann ein *Ereignis*.

(Es ist ein bisschen verwirrend, aber der Unterschied ist wesentlich: Elementarereignisse sind *Elemente* aus  $\Omega$ , Ereignisse sind *Teilmengen*. In den meisten Fällen – aber nicht immer – ist  $\{\omega\}$  ein Ereignis für Elementarereignisse  $\omega$ .)

- Wie oft kommt es vor, dass  $\omega$  in  $E$  liegt? Fragt man den Zufallsautomaten „sehr oft“ ab, so gibt es „Tendenzen“: Man kann erwarten, dass sich der Anteil der Experimente, in denen die Frage „ $\omega$  in  $E$ ?“ positiv beantwortet wird, mit zunehmender Anzahl der Experimente einem festen Wert nähern wird. (Zum Beispiel sollte bei „vielen“ Würfelwürfen etwa die Hälfte der Ergebnisse eine gerade Zahl sein.) Dieser Wert heißt die *Wahrscheinlichkeit von E*, sie wird mit  $\mathbb{P}(E)$  (gesprochen „P von E“) bezeichnet.
- Ansonsten ist das Verhalten des Automaten völlig unvorhersehbar. Auch wenn Sie ihn noch so oft abgefragt haben.

Damit ist so ein Zufallsautomat durch drei Dinge charakterisiert. Erstens durch eine Menge  $\Omega$ , zweitens durch die Gesamtheit der Ereignisse  $E$ , die uns interessieren werden, und drittens durch die Zahlen  $\mathbb{P}(E)$ , die offensichtlich im Intervall  $[0, 1]$  liegen müssen<sup>2)</sup>.

## 1.2 Die Präzision: $\sigma$ -Algebren, Wahrscheinlichkeitsmaße und Wahrscheinlichkeitsräume

Im vorigen Abschnitt war alles noch ein bisschen vage. Es gab „Tendenzen“, es wurde von „vielen“ Experimenten gesprochen usw. So etwas eignet sich natürlich nicht als Ausgangspunkt einer mathematischen Theorie. Deswegen soll nun alles weggelassen werden, was unklar ist. Am Ende steht die Definition des Wahrscheinlichkeitsraums, da gibt es dann wirklich nur noch  $\Omega$ , die Ereignisse  $E$  und die Zahlen  $\mathbb{P}(E)$ .

Schritt 1:  $\Omega$ , die Menge der Elementarereignisse

Das ist unproblematisch. Als  $\Omega$  wird *jede* nichtleere Menge zugelassen. In diesem Buch werden das zwar so gut wie immer „einfache“ Mengen sein (endliche Mengen,  $\mathbb{N}$ , Intervalle usw.), doch kommen in Fortgeschrittenenveranstaltungen auch Situationen vor, bei denen  $\Omega$  viel komplizierter ist, zum Beispiel aus allen stetigen Funktionen auf  $\mathbb{R}$  besteht.

---

<sup>1)</sup>Auch  $\omega$  wird als „Omega“ ausgesprochen. Wenn Verwechslungen möglich sind, nennt man  $\omega$  „Klein-Omega“ und  $\Omega$  „Groß-Omega“.

<sup>2)</sup>Beachte: In der Mathematik sind Wahrscheinlichkeiten Zahlen  $p$  in  $[0, 1]$ , im täglichen Leben gibt man sie oft in Prozent an. Zum Umrechnen ist  $p$  einfach mit 100 zu multiplizieren.

Schritt 2: Was sind Mengensysteme?

Dieser Unterabschnitt ist eingeschoben, weil Mengensysteme in diesem Buch sehr häufig auftreten werden, sie aber für Anfänger etwas gewöhnungsbedürftig sind. Im Grunde ist es aber gar nicht so schwer. Wir starten mit irgendeiner Menge  $M$ . Die *Potenzmenge* von  $M$  ist diejenige Menge, die aus allen Teilmengen von  $M$  besteht. Sie wird mit  $\mathcal{P}(M)$  („P von M“) bezeichnet:

$$\mathcal{P}(M) := \{E \mid E \subset M\}^3).$$

Wenn  $M$   $k$  Elemente hat, so gibt es in der Potenzmenge schon  $2^k$  Elemente. Die Potenzmenge von  $\{0, 1\}$  zum Beispiel sieht so aus:

$$\mathcal{P}(\{0, 1\}) = \{\emptyset, \{0\}, \{1\}, \{0, 1\}\}.$$

Ein *Mengensystem auf  $M$*  ist dann nichts weiter als eine Teilmenge von  $\mathcal{P}(M)$ : Gewisse Teilmengen sind zu einer neuen Gesamtheit zusammengefasst. Wie üblich bei Mengendefinitionen gibt es mehrere Möglichkeiten, so eine Gesamtheit exakt zu beschreiben. Erstens kann man ein Mengensystem durch explizite Angabe der Elemente, die dazugehören sollen, erklären. So wäre etwa

$$\mathcal{M} := \{\{1, 2\}, \{2, 3\}, \{3, 4, 5\}\}$$

ein Mengensystem auf  $\{1, 2, 3, 4, 5\}$ . Eine zweite Möglichkeit besteht in der Angabe einer definierenden Eigenschaft. Zum Beispiel wäre

$$\mathcal{M} := \{E \mid E \subset \mathbb{R}, [0, 1] \subset E\}$$

ein Mengensystem auf  $\mathbb{R}$ . Wichtig ist dabei nur, dass sich für alle  $E \subset M$  eindeutig entscheiden lässt, ob  $E \in \mathcal{M}$  gilt oder nicht. Im vorstehenden Beispiel etwa ist klar, dass  $[-1, 3]$  zu  $\mathcal{M}$  gehört, die Menge  $\mathbb{Q}$  der rationalen Zahlen aber nicht.

Schritt 3: Das Mengensystem der Ereignisse

Zurück zum Thema Wahrscheinlichkeitstheorie. Wenn  $E \subset \Omega$  ein uns interessierendes Ereignis ist, wollen wir doch über  $\mathbb{P}(E)$ , die Wahrscheinlichkeit von  $E$ , sprechen können. In einer idealen Mathematikwelt könnte man sich wünschen, *alle* Teilmengen als Ereignisse zuzulassen. Leider geht das in vielen Fällen nicht. Die Gründe sind etwas verwicket, Sie können sie bei Gelegenheit auf Seite 28 nachlesen. Deswegen muss man an dieser Stelle etwas sorgfältiger sein.

Wir beginnen mit einer Beobachtung. Teilmengen von  $\Omega$  entsprechen doch Fragen über die  $\omega \in \Omega$ . Zum Beispiel entspricht die Frage „Zeigt der Würfel eine gerade Zahl?“ der Teilmenge  $\{2, 4, 6\}$  von  $\{1, 2, 3, 4, 5, 6\}$ . Es ist sicher ein legitimer Wunsch, aus Fragen neue Fragen abzuleiten, also zum Beispiel von der Frage „ $F$ ?“ zu „nicht  $F$ ?“ überzugehen oder zwei mögliche Fragen „ $F_1$ ?“ und

---

<sup>3)</sup>Achtung: Das Symbol „ $\subset$ “ bezeichnet in diesem Buch nicht nur echte Teilmengen. Auch  $M \subset M$  ist eine richtige Aussage.

„ $F_2$ ?“ zu „ $F_1$  oder  $F_2$ ?“ oder zu „ $F_1$  und  $F_2$ ?“ zu kombinieren. In der Sprache der Mengenlehre entspricht das dem Übergang zur Komplementärmenge, zur Vereinigung bzw. zum Durchschnitt.

Es wird allerdings auch manchmal notwendig sein, *unendlich viele Fragen zu kombinieren*. Ist zum Beispiel  $\Omega = \mathbb{R}$  und besagt Frage Nummer  $n$ : „Ist  $\omega \in [2n, 2n+1]$ ?“ (für  $n = 1, 2, \dots$ ), so könnte man daraus die Frage ableiten: „Ist  $\omega$  in  $[2, 3]$  oder in  $[4, 5]$  oder in  $[6, 7]$  oder in ...?“ In die Sprache der Mengenlehre übersetzt, heißt das: Sind  $E_1, E_2, \dots$  Ereignisse, so soll auch die Vereinigung  $\bigcup_n E_n$  ein Ereignis sein<sup>4)</sup>.

Will man diese Wünsche in unserer Situation umsetzen, so muss man verlangen, dass das System der Ereignisse gewisse Eigenschaften hat. Hier die für das Folgende relevante Definition:

**Definition 1.2.1.** Ein Mengensystem  $\mathcal{E}$  auf der Menge  $\Omega$  heißt eine  $\sigma$ -Algebra (gesprochen „sigma-Algebra“), wenn die folgenden drei Eigenschaften erfüllt sind:

- (i)  $\Omega$  und die leere Menge  $\emptyset$  gehören zu  $\mathcal{E}$ .
- (ii) Für jedes  $E \in \mathcal{E}$  gilt auch  $\Omega \setminus E$ , die Komplementärmenge von  $E$ , in  $\mathcal{E}$ .
- (iii) Sind  $E_1, E_2, \dots$  Elemente aus  $\mathcal{E}$ , so gehört auch  $\bigcup_n E_n$  zu  $\mathcal{E}$ .

Anders ausgedrückt: Man fordert, dass  $\emptyset$  und  $\Omega$  zu  $\mathcal{E}$  gehören und dass  $\mathcal{E}$  unter der Bildung von Komplementen und abzählbaren Vereinigungen abgeschlossen ist.

Wenn Missverständnisse möglich sind, sagt man ausführlicher „ $\sigma$ -Algebra auf  $\Omega$ “ statt  $\sigma$ -Algebra. Der Name ist etwas verwirrend. Algebraische Strukturen sind auf den ersten Blick nicht zu sehen, und was soll das „ $\sigma$ “ hier bedeuten? Die Erklärung: Auf einer  $\sigma$ -Algebra kann man wirklich so etwas wie eine algebraische Struktur erklären, vgl. dazu Übungsaufgabe Ü1.2.11 auf Seite 34. Und „ $\sigma$ “ wird in verschiedenen Bereichen der Mathematik dann verwendet, wenn in der Definition Abzählbarkeit eine Rolle spielt:  $\sigma$ -kompakte Räume z.B. sind solche, die abzählbare Vereinigung von kompakten Teilmengen sind<sup>5)</sup>.

Man sollte sich mit dieser Definition gut vertraut machen, sie wird im Folgenden eine unverzichtbare Rolle spielen. Hier einige erste *Beispiele* (die für uns wichtigste  $\sigma$ -Algebra, die  $\sigma$ -Algebra der Borelmengen, werden wir im Abschnitt 1.4 ausführlich besprechen).

1. Für jede Menge  $\Omega$  ist  $\mathcal{P}(\Omega)$ , die Potenzmenge von  $\Omega$ , eine  $\sigma$ -Algebra. Das ist die *größte*  $\sigma$ -Algebra auf  $\Omega$ : Jede andere  $\sigma$ -Algebra auf  $\Omega$  ist darin enthalten.
2.  $\{\emptyset, \Omega\}$  ist die *kleinste*  $\sigma$ -Algebra auf  $\Omega$ : Jede andere  $\sigma$ -Algebra auf  $\Omega$  enthält  $\{\emptyset, \Omega\}$ .

---

<sup>4)</sup>Zur Erinnerung: Sind  $E_1, E_2, E_3, \dots$  Teilmengen einer Menge  $\Omega$ , so versteht man unter der Vereinigung  $\bigcup_n E_n$  die Menge aller  $\omega \in \Omega$ , die in mindestens einer der Mengen  $E_n$  enthalten sind.

<sup>5)</sup>Zur Erinnerung: Eine Menge  $M$  heißt *abzählbar*, wenn es eine bijektive Abbildung von  $\mathbb{N}$  nach  $M$  gibt, wenn man also die Elemente von  $M$  mit den Zahlen  $1, 2, 3, \dots$  durchnummernieren kann. Und „ $M$  ist höchstens abzählbar“ soll bedeuten, dass  $M$  endlich oder abzählbar ist.

**3.** Es sei  $\mathcal{D}$  eine *Partition* der Menge  $\Omega$ . Das bedeutet, dass  $\mathcal{D}$  ein Mengensystem auf  $\Omega$  ist, so dass je zwei verschiedene Elemente aus  $\mathcal{D}$  einen leeren Durchschnitt haben und die Vereinigung über alle  $D \in \mathcal{D}$  gleich  $\Omega$  ist. (So ist zum Beispiel  $\{\{1\}, \{2, 3\}, \{4, 5, 6\}\}$  eine Partition von  $\{1, 2, 3, 4, 5, 6\}$ .) Betrachte dann für beliebige Teilmengen  $\mathcal{D}'$  von  $\mathcal{D}$  die Vereinigung  $\bigcup_{D \in \mathcal{D}'} D$ . Die Gesamtheit der so entstehenden Mengen bildet dann eine  $\sigma$ -Algebra, und im Fall endlicher  $\Omega$  entstehen alle  $\sigma$ -Algebren auf  $\Omega$  auf diese Weise<sup>6)</sup>.

Schritt 4: Wahrscheinlichkeitsmaße

Das bisherige Vorgehen lässt sich so zusammenfassen:

Um den Zufall in einer speziellen Situation zu beschreiben, brauchen wir zunächst eine Menge  $\Omega$ . Das sind die möglichen Ergebnisse des „Zufallsexperiments“. Und dann muss eine  $\sigma$ -Algebra  $\mathcal{E}$  auf  $\Omega$  geben sein. Die Interpretation: Für die  $E \in \mathcal{E}$  ist die Frage „Lieg das Ergebnis  $\omega$  des Experiments in  $E$ ?“ zulässig.

Es fehlt nur noch, den  $E \in \mathcal{E}$  eine Zahl  $\mathbb{P}(E)$  in  $[0, 1]$  zuzuordnen. Das soll als *Wahrscheinlichkeit von E* interpretiert werden.

Damit so eine Interpretation möglich ist, muss man zwei nahe liegende Forderungen an die Zuordnung  $E \mapsto \mathbb{P}(E)$  stellen. Sie stehen in der

**Definition 1.2.2.** Es sei  $\mathcal{E}$  eine  $\sigma$ -Algebra auf der Menge  $\Omega$ . Eine Abbildung  $\mathbb{P} : \mathcal{E} \rightarrow [0, 1]$  heißt ein Wahrscheinlichkeitsmaß, wenn gilt:

- (i) Es ist  $\mathbb{P}(\Omega) = 1$ .
- (ii) Für jede disjunkte Folge  $E_1, E_2, \dots$  in  $\mathcal{E}$  ist

$$\mathbb{P}(E_1 \cup E_2 \cup \dots) = \mathbb{P}(E_1) + \mathbb{P}(E_2) + \dots$$

(Der Begriff „disjunkte Folge“ soll bedeuten, dass die Mengen  $E_i, E_j$  für  $i \neq j$  disjunkt sind, dass also  $E_i \cap E_j = \emptyset$  gilt.)

Man sagt, dass  $\mathbb{P}$   $\sigma$ -additiv ist.

**Bemerkungen:** 1. Die  $\sigma$ -Additivität kann man etwas kürzer als  $\mathbb{P}(\bigcup_k E_k) = \sum_{k=1}^{\infty} \mathbb{P}(E_k)$  (für disjunkte Folgen) schreiben. Die linke Seite ist wohldefiniert, da  $\mathcal{E}$  eine  $\sigma$ -Algebra ist: Deswegen gehören beliebige abzählbare Vereinigungen wieder zu  $\mathcal{E}$ . Die rechte Seite ist eine unendliche Reihe, die Reihensumme ist also der Grenzwert der Partialsummen  $\sum_{k=1}^n \mathbb{P}(E_k)$  für  $n \rightarrow \infty$ .

2. Da  $\Omega$  als disjunkte Vereinigung  $\Omega \cup \emptyset \cup \emptyset \cup \emptyset \cup \dots$  geschrieben werden kann, muss

$$1 = \mathbb{P}(\Omega) = 1 + \mathbb{P}(\emptyset) + \mathbb{P}(\emptyset) + \dots$$

gelten. Notwendig ist damit  $\mathbb{P}(\emptyset) = 0$ . Und wenn man für disjunkte  $E, F$  die Vereinigung  $E \cup F$  als abzählbare Vereinigung  $E \cup F \cup \emptyset \cup \emptyset \cup \dots$  schreibt, folgt  $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F)$ .

---

<sup>6)</sup>Vgl. Übungsaufgabe Ü1.2.5.

Kurz: Wahrscheinlichkeitsmaße sind auch endlich additiv, d.h. auch bei endlich vielen disjunkten Ereignissen addieren sich die Wahrscheinlichkeiten.

All das ist in Hinblick auf unsere Interpretation sehr plausibel: Garantiert passiert irgendetwas ( $\mathbb{P}(\Omega) = 1$ ), es ist auszuschließen, dass nichts passiert ( $\mathbb{P}(\emptyset) = 0$ ), und für disjunkte  $E, F$  ist doch wirklich der prozentuale Anteil der  $\omega \in E \cup F$  gleich der Summe aus den Anteilen der  $\omega \in E$  und der  $\omega \in F$ .

Es folgen noch zwei einfache *Beispiele*:

**1.** Dieses Beispiel modelliert das Werfen eines fairen Würfels. Wir definieren  $\Omega$  als  $\{1, 2, 3, 4, 5, 6\}$ ,  $\mathcal{E}$  besteht aus allen Teilmengen von  $\Omega$ , und für  $E \in \mathcal{E}$  definieren wir  $\mathbb{P}(E)$  durch  $\#E/6$ ; dabei bedeutet  $\#E$  die Anzahl der Elemente in  $E$ . Dass dadurch wirklich ein Wahrscheinlichkeitsmaß definiert wird, liegt daran, dass die Anzahl der Elemente einer disjunkten Vereinigung  $\bigcup E_n$  gleich der Summe der  $\#E_n$  ist<sup>7)</sup>.

**2.** Wir betrachten irgendeine nichtleere Menge, in ihr zeichnen wir ein Element  $\omega_0$  aus.  $\mathcal{E}$  soll die Potenzmenge von  $\Omega$  sein, und  $\mathbb{P}$  wird so definiert: Gehört  $\omega_0$  zu  $E$ , so setzen wir  $\mathbb{P}(E) := 1$ . Andernfalls soll  $\mathbb{P}(E) := 0$  sein. Machen Sie sich zur Übung klar, dass das wirklich ein Wahrscheinlichkeitsmaß ist.



Kolmogoroff

Finale: Die Definition „Wahrscheinlichkeitsraum“

Es fehlt eigentlich nur noch, alles zusammenzufassen. Die folgende wichtige Definition geht auf den russischen Mathematiker Kolmogoroff<sup>8)</sup> zurück:

**Definition 1.2.3.** Ein Wahrscheinlichkeitsraum  $(\Omega, \mathcal{E}, \mathbb{P})$  besteht aus drei Dimensionen:

- (i) Einer nichtleeren Menge  $\Omega$ , der Menge der Elementarereignisse.
- (ii) Einer  $\sigma$ -Algebra  $\mathcal{E}$  auf  $\Omega$ , der  $\sigma$ -Algebra der Ereignisse.
- (iii) Einem Wahrscheinlichkeitsmaß  $\mathbb{P} : \mathcal{E} \rightarrow [0, 1]$ .

Man beachte, dass diese Definition nur noch präzise definierte mathematische Begriffe enthält. Alles, was bei der Behandlung von Zufallsphänomenen nur vage ausgedrückt werden kann, wurde weggelassen.

Zwei einfache Beispiele haben wir eben schon kennen gelernt, die für uns wichtigsten Wahrscheinlichkeitsräume werden in Kapitel 2 vorgestellt.

Wir werden in den folgenden Kapiteln sehen, dass die Definition geeignet ist, eine Vielfalt von Situationen der realen Welt angemessen zu beschreiben, in denen es um den Zufall geht.

---

<sup>7)</sup>Im vorliegenden Fall kann es bei disjunkten Vereinigungen  $\bigcup_n E_n$  natürlich nur höchstens sechs Mengen  $E_n$  geben, die nicht leer sind.

<sup>8)</sup>Andrey Kolmogoroff, 1903 bis 1987. Er schrieb schon als Student wichtige mathematische Arbeiten. Mit 28 Jahren wurde er Professor in Moskau. Sein Buch „Grundbegriffe der Wahrscheinlichkeitsrechnung“ (1933) gilt als Beginn der heute allgemein akzeptierten Wahrscheinlichkeitstheorie auf einer präzisen axiomatischen Grundlage.

### 1.3 Allgemeine Eigenschaften von Wahrscheinlichkeitsräumen

Wir wissen nun, was Wahrscheinlichkeitsräume sind. Aus der Definition ergeben sich viele Folgerungen, die wir immer und immer wieder verwenden werden. Deswegen ist es sinnvoll, die wichtigsten übersichtlich zusammenzustellen.

#### $\sigma$ -Algebren

Die Betrachtung von  $\sigma$ -Algebren war dadurch motiviert worden, dass es legitim sein soll, Fragen über Ereignisse zu kombinieren: mit „ $F$ “ auch „nicht  $F$ “ usw. Bei der Definition der  $\sigma$ -Algebren treten aber nur noch Komplemente und Vereinigungen, also „nicht“ und „oder“ auf. Ist der Rest vergessen worden?

Der folgende Satz zeigt, dass das nicht der Fall ist:

**Satz 1.3.1.** *Es sei  $\mathcal{E}$  eine  $\sigma$ -Algebra auf der Menge  $\Omega$ . Dann gilt:*

- (i) *Für  $E, F \in \mathcal{E}$  gilt auch  $E \cup F \in \mathcal{E}$ . Nicht nur abzählbare, sondern auch endliche Vereinigungen führen also aus  $\mathcal{E}$  nicht hinaus.*
- (ii) *Sind  $E$  und  $F$  Ereignisse, so auch  $E \cap F$ , und mit einer Folge  $E_1, E_2, \dots$  von Ereignissen gehört auch der Durchschnitt<sup>9)</sup>  $\bigcap_n E_n$  zu  $\mathcal{E}$ .*
- (iii) *Es können auch relative Komplemente gebildet werden: Für  $E, F \in \mathcal{E}$  ist auch  $E \setminus F$  ( $= \{\omega \mid \omega \in E, \omega \notin F\}$ ) in  $\mathcal{E}$ .*

**Beweis:** (i) Das wird den meisten spitzfindig vorkommen, doch in der Definition ist wirklich nur von abzählbaren Vereinigungen die Rede. Wir mussten schon in Bemerkung 2 nach Definition 1.2.2 auf diese Tatsache hinweisen, um auf die endliche Additivität von Wahrscheinlichkeitsmaßen hinweisen zu können. Der Beweis ist nicht schwer, man muss  $E \cup F$  nur etwas gekünstelt als abzählbare Vereinigung schreiben, z.B. als  $E \cup F \cup \emptyset \cup \emptyset \cup \dots$ .

(ii) Man wendet hier die elementare Gleichung

$$E \cap F = \Omega \setminus [(\Omega \setminus E) \cup (\Omega \setminus F)]$$

aus der Mengenlehre an<sup>10)</sup>. Damit sieht man, dass man  $E \cap F$  erhalten kann, indem man nur Operationen ausführt, die in einer  $\sigma$ -Algebra zulässig sind.

Für abzählbare Durchschnitte muss die Mengengleichheit

$$\bigcap_n E_n = \Omega \setminus \left[ \bigcup_n (\Omega \setminus E_n) \right]$$

verwendet werden.

---

<sup>9)</sup>Der Durchschnitt  $\bigcap_n E_n$  der  $E_n$  ist die Teilmenge aller  $\omega$ , die zu allen  $E_n$  gehören.

<sup>10)</sup>Von der Gültigkeit dieser Beziehung sollte man sich durch eine Skizze überzeugen.

(iii) Auch das ist mit einer geschickten Darstellung leicht zu begründen: Beachte, dass  $E \setminus F = E \cap (\Omega \setminus F)$  gilt.  $\square$

Wenn man die einzelnen Ergebnisse kombiniert, kann man aus Ereignissen schon ziemlich komplizierte neue Ereignisse zusammensetzen. Mit  $E, F, G, H \in \mathcal{E}$  ist zum Beispiel auch

$$(E \cup F) \setminus (G \cap H)$$

ein Ereignis<sup>11)</sup>, und abzählbare Operationen können auch verwendet werden. In Kapitel 8 wird es zum Beispiel wichtig sein zu wissen, dass mit Ereignissen  $E_1, E_2, \dots$  auch  $\bigcap_{m \in \mathbb{N}} \bigcup_{n \in \mathbb{N}, n \geq m} E_n$  Ereignis ist. Das ist durch die bereits bewiesenen Ergebnisse garantiert.

Wenn man den vorstehenden Beweis verstanden hat, sollte es klar sein, dass man für  $\sigma$ -Algebren die folgende Faustregel<sup>12)</sup> aufstellen kann:

#### Faustregel für das Arbeiten mit $\sigma$ -Algebren

Es sei  $\mathcal{E}$  eine  $\sigma$ -Algebra und  $E \subset \Omega$ . Wenn in der Definition von  $E$  nur höchstens abzählbar viele Mengen auftreten, die alle in  $\mathcal{E}$  liegen und wenn nur die Standard-Mengenoperationen (also  $\cap, \bigcap, \cup, \bigcup, \setminus$ ) verwendet wurden, dann gehört auch  $E$  zu  $\mathcal{E}$ .

#### Wahrscheinlichkeitsmaße

Auch an Wahrscheinlichkeitsmaße haben wir nur recht bescheidene Forderungen gestellt. Es lassen sich daraus aber wichtige Folgerungen herleiten. Teilweise handelt es sich um Eigenschaften, die für das Konzept „Wahrscheinlichkeit“ plausibel sind („Größere Ereignisse haben eine nicht kleinere Wahrscheinlichkeit“, Teil (iii) des folgenden Satzes), teilweise geht es um eher technische Tatsachen, die für spätere Beweise wichtig werden.

**Satz 1.3.2.** *Es sei  $(\Omega, \mathcal{E}, \mathbb{P})$  ein Wahrscheinlichkeitsraum. Dann gilt:*

- (i) *Es ist  $\mathbb{P}(\emptyset) = 0$ .*
- (ii) *Sind  $E$  und  $F$  disjunkte Ereignisse, so ist  $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F)$ .*
- (iii) *Für Ereignisse  $E, F$  mit  $E \subset F$  ist  $\mathbb{P}(E) \leq \mathbb{P}(F)$ .*
- (iv) *Stets gilt  $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F)$ . Insbesondere ist  $\mathbb{P}(E \cup F) \leq \mathbb{P}(E) + \mathbb{P}(F)$ . (Diese Tatsache wird Subadditivität genannt.)*

*Eine entsprechende Aussage ist auch für abzählbar viele Ereignisse  $E_1, E_2, \dots$  richtig:*

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} E_n\right) \leq \sum_{n=1}^{\infty} \mathbb{P}(E_n).$$

---

<sup>11)</sup>Übersetzt als Frage heißt das: Gilt  $E$  oder  $F$ , aber nicht gleichzeitig  $G$  und  $H$ ?

<sup>12)</sup>Das wird ausdrücklich als Faustregel und nicht als Satz formuliert, da eine präzise Formulierung recht schwerfällig wäre.

(v) Es seien  $E$  und  $E_1, E_2, \dots$  Ereignisse mit  $E_1 \subset E_2 \subset \dots$  und  $\bigcup_n E_n = E$ . Es ist also  $E$  die „aufsteigende Vereinigung“ der  $E_n$ . Dann gilt  $\mathbb{P}(E) = \lim_{n \rightarrow \infty} \mathbb{P}(E_n)$ ; man nennt diese Tatsache die „Stetigkeit nach oben“.

(vi) Es seien  $E$  und  $E_1, E_2, \dots$  Ereignisse mit  $E_1 \supset E_2 \supset \dots$  und  $\bigcap_n E_n = E$ . Es ist also  $E$  der „absteigende Durchschnitt“ der  $E_n$ . Auch dann gilt  $\mathbb{P}(E) = \lim_{n \rightarrow \infty} \mathbb{P}(E_n)$ ; man spricht von der „Stetigkeit nach unten“.

**Beweis:** Dass (i) und (ii) richtig sind, wurde schon auf Seite 9 begründet.

(iii) Wegen  $E \subset F$  ist  $F$  die disjunkte Vereinigung von  $E$  und  $F \setminus E$ . Folglich gilt  $\mathbb{P}(F) = \mathbb{P}(E) + \mathbb{P}(F \setminus E)$ , und da  $\mathbb{P}(F \setminus E)$  nichtnegativ ist, ergibt sich die Behauptung.

(iv) Stets ist  $E \cup F$  die disjunkte Vereinigung aus  $E$  und  $F \setminus E$ , auch ist  $F$  die disjunkte Vereinigung von  $E \cap F$  und  $F \setminus E$ . Wir können also so rechnen:

$$\begin{aligned}\mathbb{P}(E \cup F) &= \mathbb{P}(E) + \mathbb{P}(F \setminus E) \\ &= \mathbb{P}(E) + \mathbb{P}(F \setminus E) + \mathbb{P}(E \cap F) - \mathbb{P}(E \cap F) \\ &= \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F).\end{aligned}$$

Der Beweis der Subadditivität im abzählbaren Fall ist ähnlich. Wir definieren  $F_1 := E_1$ ,  $F_2 := E_2 \setminus E_1$ ,  $F_3 := E_3 \setminus (E_1 \cup E_2)$ , usw. (allgemein, für  $n \in \mathbb{N}$ :  $F_n := E_n \setminus (\bigcup_{k=1}^{k=n-1} E_k)$ ). Dann ist  $F_n \subset E_n$ , und  $\bigcup_n E_n$  ist die *disjunkte* Vereinigung der  $F_n$ . Mit dem vorstehenden Teil (iii) folgt

$$\mathbb{P}\left(\bigcup_n E_n\right) = \mathbb{P}\left(\bigcup_n F_n\right) = \sum_n \mathbb{P}(F_n) \leq \sum_n \mathbb{P}(E_n).$$

(v)  $E$  kann als disjunkte Vereinigung geschrieben werden:

$$E = E_1 \cup (E_2 \setminus E_1) \cup (E_3 \setminus E_2) \cup (E_4 \setminus E_3) \cup \dots,$$

es ist also

$$\mathbb{P}(E) = \mathbb{P}(E_1) + \mathbb{P}(E_2 \setminus E_1) + \mathbb{P}(E_3 \setminus E_2) + \mathbb{P}(E_4 \setminus E_3) + \dots.$$

Die rechts stehende Reihe ist der Limes der Partialsummen, und in denen hebt sich fast alles weg: Wegen  $E_{k-1} \subset E_k$  ist nämlich

$$\mathbb{P}(E_k \setminus E_{k-1}) = \mathbb{P}(E_k) - \mathbb{P}(E_{k-1}),$$

und es folgt

$$\begin{aligned}\mathbb{P}(E) &= \lim_n \mathbb{P}(E_1) + \sum_{k=2}^n (\mathbb{P}(E_k) - \mathbb{P}(E_{k-1})) \\ &= \lim_n \mathbb{P}(E_n).\end{aligned}$$

(vi) Wir gehen zu Komplementen über:  $\Omega \setminus E$  ist die aufsteigende Vereinigung der  $\Omega \setminus E_n$ . Folglich ist  $\mathbb{P}(\Omega \setminus E) = \lim_n \mathbb{P}(\Omega \setminus E_n)$ . Die linke Seite ist aber gleich  $1 - \mathbb{P}(E)$ , die rechte gleich  $\lim_n (1 - \mathbb{P}(E_n)) = 1 - \lim_n \mathbb{P}(E_n)$ . Das beweist die Behauptung.  $\square$

## 1.4 Erzeugte $\sigma$ -Algebren

In Abschnitt 1.2 wurde definiert, was eine  $\sigma$ -Algebra ist. In den späteren Anwendungen wird es allerdings oft so sein, dass wir an Wahrscheinlichkeiten für Elemente eines Mengensystems interessiert sind, das nicht alle Forderungen an eine  $\sigma$ -Algebra erfüllt. Wenn zum Beispiel  $\Omega$  gleich  $\mathbb{R}$  ist, könnten die Wahrscheinlichkeiten für Intervalle wichtig sein, doch das System der Intervalle bildet keine  $\sigma$ -Algebra. (Warum eigentlich nicht?)

In diesem Abschnitt beschreiben wir, was zu tun ist und welche Techniken zur Verfügung stehen, vorgelegte Mengensysteme zu  $\sigma$ -Algebren zu ergänzen. Das wird an vielen Stellen in diesem Buch wichtig werden.

Die Überlegungen sind allerdings in diesem und in den beiden folgenden Abschnitten ein bisschen technisch. Daher mein Rat für alle, die das genaue Durcharbeiten auf später verschieben wollen: Lesen Sie die Zusammenfassung der Abschnitte 1.4, 1.5 und 1.6 auf Seite 26 und arbeiten Sie die Einzelheiten nach und nach durch, wenn die Ergebnisse in späteren Kapiteln gebraucht werden.

Es sei  $\Omega$  irgendeine Menge und  $\mathcal{M}$  irgendein Mengensystem auf  $\Omega$ , d.h. irgendeine Teilmenge der Potenzmenge von  $\Omega$ . Wir suchen eine  $\sigma$ -Algebra  $\mathcal{A}$  auf  $\Omega$ , die alle Elemente aus  $\mathcal{M}$  enthält und „möglichst klein“ ist. Das sieht nicht besonders schwierig aus: Man muss einfach die leere Menge und  $\Omega$  zu  $\mathcal{M}$  hinzufügen, wenn sie noch nicht dazugehören und dann Komplemente und abzählbare Vereinigungen von Elementen aus  $\mathcal{M}$  dazunehmen.

In ganz einfachen Fällen klappt das wirklich so. Wenn etwa  $\mathcal{M}$  nur aus einer einzigen Menge  $A$  besteht, erhält man auf diese Weise das Mengensystem

$$\{\emptyset, \Omega, A, \Omega \setminus A\},$$

und das ist wirklich die kleinstmögliche  $\sigma$ -Algebra  $\mathcal{A}$ , für die  $\mathcal{M} \subset \mathcal{A}$  gilt.

Im Allgemeinen ist dieses „Konstruktionsverfahren“ aber nicht ausreichend. Etwas mehr wird in den Ergänzungen auf Seite 29 dazu gesagt, hier soll gleich die richtige Lösung verraten werden:

**Satz 1.4.1.** *Sei  $\mathcal{M} \subset \mathcal{P}(\Omega)$ . Dann gibt es eine  $\sigma$ -Algebra  $\mathcal{A} \subset \mathcal{P}(\Omega)$  mit den folgenden Eigenschaften:*

- (i)  $\mathcal{M}$  ist eine Teilmenge von  $\mathcal{A}$ .
- (ii)  $\mathcal{A}$  ist kleinstmöglich im folgenden Sinn: Ist  $\mathcal{A}'$  eine weitere  $\sigma$ -Algebra auf  $\Omega$  mit  $\mathcal{M} \subset \mathcal{A}'$ , so gilt  $\mathcal{A} \subset \mathcal{A}'$ .

$\mathcal{A}$  ist durch diese Eigenschaften eindeutig bestimmt. Wir sprechen von der von  $\mathcal{M}$  erzeugten  $\sigma$ -Algebra und bezeichnen sie mit  $\sigma(\mathcal{M})$  (gesprochen „sigma von  $M$ “).

**Beweis:** Hier ist die Definition:  $\mathcal{A}$  soll der Durchschnitt aller  $\sigma$ -Algebren sein, die  $\mathcal{M}$  enthalten.

Das ist, mindestens beim ersten Kennenlernen, nicht ganz leicht zu verstehen, denn erstens geht es um einen Durchschnitt von möglicherweise unendlich vielen Objekten und zweitens spielt sich alles in der Potenzmenge  $\mathcal{P}(\Omega)$  von  $\Omega$  ab.

Daher folgt hier die Definition noch einmal etwas ausführlicher. Für eine beliebige Teilmenge  $E$  von  $\Omega$  gibt es doch nur eine von zwei Möglichkeiten:

1. Wann immer jemand eine  $\sigma$ -Algebra  $\mathcal{B}$  auf  $\Omega$  mit  $\mathcal{M} \subset \mathcal{B}$  gefunden hat, so gehört  $E$  zu  $\mathcal{B}$ .
2. Das Gegenteil von „1.“ ist der Fall: Man kann eine  $\sigma$ -Algebra  $\mathcal{B}$  auf  $\Omega$  mit  $\mathcal{M} \subset \mathcal{B}$  finden, so dass  $E$  nicht zu  $\mathcal{B}$  gehört.

Im ersten Fall wird  $E$  ein Element von  $\mathcal{A}$ , im zweiten Fall nicht.

#### 1. Schritt: $\mathcal{A}$ ist eine $\sigma$ -Algebra.

Wir zeigen zunächst, dass  $\Omega$  in  $\mathcal{A}$  liegt. Dazu ist zu zeigen: Ist  $\mathcal{B}$  eine beliebige  $\sigma$ -Algebra, die  $\mathcal{M}$  enthält, so ist  $\Omega \in \mathcal{B}$ . Das ist aber klar, denn alle  $\sigma$ -Algebren enthalten  $\Omega$ . Ganz genau so ergibt sich, dass  $\emptyset \in \mathcal{A}$ .

Nun sei  $E \in \mathcal{A}$ , wir wollen zeigen, dass auch  $\Omega \setminus E$  zu  $\mathcal{A}$  gehört. Dazu müssen wir eine beliebige  $\sigma$ -Algebra  $\mathcal{B}$  mit  $\mathcal{M} \subset \mathcal{B}$  betrachten und zeigen, dass  $\Omega \setminus E \in \mathcal{B}$  gilt. Nach Voraussetzung ist aber  $E \in \mathcal{B}$ , und da  $\mathcal{B}$  eine  $\sigma$ -Algebra ist, bedeutet das  $\Omega \setminus E \in \mathcal{B}$ . Das ist genau das, was wir zeigen mussten.

Und der Beweis, dass  $\mathcal{A}$  abzählbare Vereinigungen enthält, verläuft nach dem gleichen Muster:  $E_1, E_2, \dots \in \mathcal{A}$  gegeben; eine  $\sigma$ -Algebra  $\mathcal{B}$  mit  $\mathcal{M} \subset \mathcal{B}$  wählen; die  $E_1, E_2, \dots$  liegen nach Definition von  $\mathcal{A}$  in  $\mathcal{B}$ , und da das eine  $\sigma$ -Algebra ist, gilt  $\bigcup_n E_n \in \mathcal{B}$ ; fertig.

#### 2. Schritt: Es gilt die Aussage (i) der Behauptung.

Das ist einfach: Ist  $E \in \mathcal{M}$ , so erfüllt  $E$  ganz offensichtlich die Bedingung, zu allen  $\sigma$ -Algebren  $\mathcal{B}$  zu gehören, die  $\mathcal{M}$  enthalten.

#### 3. Schritt: Es gilt die Aussage (ii) der Behauptung.

Auch das ist nicht schwer: Ist  $E \in \mathcal{A}$  und  $\mathcal{A}'$  eine  $\sigma$ -Algebra, die  $\mathcal{M}$  enthält, so ist  $E$  nach Definition ein Element von  $\mathcal{A}'$ . Das zeigt  $\mathcal{A} \subset \mathcal{A}'$ .

#### 4. Schritt: $\mathcal{A}$ ist durch diese Eigenschaften eindeutig bestimmt.

$\mathcal{A}_1$  und  $\mathcal{A}_2$  seien  $\sigma$ -Algebren auf  $\Omega$ , für die jeweils (i) und (ii) des Satzes erfüllt ist. Da insbesondere  $\mathcal{M} \subset \mathcal{A}_1$  gilt, liefert die Eigenschaft (ii) von  $\mathcal{A}_2$ , dass  $\mathcal{A}_2 \subset \mathcal{A}_1$  gelten muss. Und vertauscht man die Rollen von  $\mathcal{A}_1$  und  $\mathcal{A}_2$ , so ergibt sich  $\mathcal{A}_1 \subset \mathcal{A}_2$ . Zusammen heißt das  $\mathcal{A}_1 = \mathcal{A}_2$ .  $\square$

Hier noch ein *allgemeiner Kommentar*. Der vorstehende Beweis war recht technisch, es handelt sich aber um ein in vielen Bereichen der Mathematik übliches Verfahren. Wann immer man eine Menge  $M$  hat, die man durch Hinzunahme möglichst weniger Elemente so vergrößern möchte, dass gewisse Eigenschaften erfüllt sind, so kann man den Durchschnitt über alle Obermengen von  $M$  bilden, die diese Eigenschaften haben. Das klappt immer dann, wenn die fraglichen Eigenschaften bei Durchschnittsbildung erhalten bleiben. Bei uns ging es um die erzeugte  $\sigma$ -Algebra, der „erzeugte Unterraum einer Teilmenge eines Vektorraumes“, die „erzeugte Untergruppe einer Teilmenge einer Gruppe“, die „konvexe Hülle einer Teilmenge eines  $\mathbb{R}$ -Vektorraums“, die „abgeschlossene Hülle einer Teilmenge eines metrischen Raumes“, ... entstehen genauso.

Wie kann man denn einer Menge  $E \subset \Omega$  aber nun ansehen, ob sie zu  $\sigma(\mathcal{M})$  gehört? Das ist so allgemein nicht zu beantworten, in der Regel muss man in jedem Einzelfall auf besondere Weise argumentieren:

- Es seien etwa  $E, F \in \mathcal{M}$ . Dann ist  $E \cup F \in \sigma(\mathcal{M})$ . Beweis:  $E, F$  gehören zu  $\sigma(\mathcal{M})$ , und  $\sigma(\mathcal{M})$  ist ein  $\sigma$ -Algebra.
- Angenommen, man weiß schon, dass  $E_1, E_2, \dots$  zu  $\sigma(\mathcal{M})$  gehören. Dann liegt auch  $\bigcup_n E_n$  in  $\sigma(\mathcal{M})$ . Beweis: Das folgt wieder daraus, dass  $\sigma(\mathcal{M})$  eine  $\sigma$ -Algebra ist.
- Klar ist auch, dass dann  $\Omega \setminus (\bigcup_n E_n)$  in  $\sigma(\mathcal{M})$  liegt.
- Und so weiter ...

Das ist auch schon im Wesentlichen eine vollständige Beschreibung der Möglichkeiten: Alle diejenigen Mengen gehören zu  $\sigma(\mathcal{M})$ , die sich so aus Elementen aus  $\mathcal{M}$  aufbauen lassen, dass nur Mengenkonstruktionen (Schnitt, Vereinigung, Komplemente) vorkommen und höchstens abzählbar viele Elemente aus  $\mathcal{M}$  verwendet werden. Mehr dazu finden Sie im nächsten Abschnitt.

### *Einige Beispiele*

1.  $\mathcal{M}$  ist endlich. In diesem Fall kann  $\sigma(\mathcal{M})$  explizit angegeben werden.  $\mathcal{M}$  bestehe aus  $n$  Mengen:

$$\mathcal{M} = \{A_1, A_2, \dots, A_n\}.$$

Wir verfahren in zwei Schritten. Zunächst konstruieren wir alle Mengen der Form  $C_1 \cap \dots \cap C_n$ , wobei  $C_i$  eine der Mengen  $A_i$  oder  $\Omega \setminus A_i$  ist. Das sind  $2^n$  Möglichkeiten. Dann gilt:

- Je zwei verschiedene dieser Mengen haben einen leeren Durchschnitt (denn im Durchschnitt tritt für irgendein  $i$  der Schnitt von  $A_i$  und  $\Omega \setminus A_i$  auf, und das ist die leere Menge).
- Die Vereinigung aller dieser Mengen ist  $\Omega$ : Ist  $\omega$  beliebig, definiere für jedes  $i$  die Menge  $C_i$  als  $A_i$  bzw. als  $\Omega \setminus A_i$ , je nachdem, ob  $\omega$  in  $A_i$  liegt oder nicht; dann ist auf jeden Fall  $\omega \in C_1 \cap \dots \cap C_n$ ,  $\omega$  liegt also in der Vereinigung der so konstruierten Mengen.

Anders ausgedrückt: Bezeichnet man die nichtleeren, auf diese Weise entstehenden Mengen mit  $D_1, \dots, D_r$  (wobei  $r$  höchstens gleich  $2^n$  ist), so sind die  $D_i$  paarweise disjunkt und es gilt  $\bigcup_{\rho=1}^r D_\rho = \Omega$ .

Im zweiten Schritt bilden wir alle möglichen Vereinigungen der  $D_i$ , wir betrachten also die Mengen  $\bigcup_{i \in \Delta} D_i$ , wobei  $\Delta$  alle Teilmengen von  $\{1, \dots, r\}$  durchläuft<sup>13)</sup>. Wir behaupten, dass die Gesamtheit dieser Vereinigungen – nennen wir sie für die nächsten Zeilen  $\mathcal{V}$  – gleich  $\sigma(\{A_1, \dots, A_n\})$  ist.

Klar ist, dass  $\mathcal{V}$  in  $\sigma(\{A_1, \dots, A_n\})$  enthalten ist, denn alle Elemente entstehen durch – sogar endlich viele – Mengenoperationen aus den  $A_i$ . Und für die

---

<sup>13)</sup>Im Fall  $\Delta = \emptyset$  wollen wir die Vereinigung als die leere Menge interpretieren.

andere Inklusion argumentiert man so:  $\mathcal{V}$  ist eine  $\sigma$ -Algebra, die alle  $A_i$  enthält (das ist leicht zu sehen), und deswegen muss  $\sigma(\mathcal{M}) \subset \mathcal{V}$  gelten.

Es lohnt sich, die Strategie, die zur konkreten Beschreibung von  $\sigma(\mathcal{M})$  geführt hat, herauszuarbeiten: Wir haben so lange aus  $\mathcal{M}$  – unter Verwendung von jeweils höchstens abzählbar vielen Mengen – neue Mengen gebildet, bis eine  $\sigma$ -Algebra entstand. Das muss dann  $\sigma(\mathcal{M})$  sein. In einfachen Fällen führt dieser Weg wirklich zum Ziel.

**2.**  $\Omega$  ist beliebig, und  $\mathcal{M}$  besteht aus allen einpunktigen Teilmengen von  $\Omega$ . Die erzeugte  $\sigma$ -Algebra muss dann sicher alle Teilmengen  $A$  von  $\Omega$  enthalten, für die  $A$  oder  $\Omega \setminus A$  höchstens abzählbar ist, denn solche Mengen entstehen durch Anwendung von Mengenoperationen auf höchstens abzählbar viele ein-elementige Mengen. Die Gesamtheit dieser  $A$  ist aber schon eine  $\sigma$ -Algebra, die  $\sigma$ -Algebra der höchstens abzählbaren oder höchstens co-abzählbaren Teilmengen. (Zum Nachweis, dass das wirklich eine  $\sigma$ -Algebra ist, muss man sich nur an das folgende Ergebnis über abzählbare Mengen erinnern: Sind  $A_1, A_2, \dots$  höchstens abzählbare Mengen, so ist auch  $\bigcup_n A_n$  höchstens abzählbar.) Damit sind wir auch schon fertig, aufgrund der Bemerkung vor dem diesem Beispiel haben wir damit schon  $\sigma(\mathcal{M})$  identifiziert.

## 1.5 Borelmengen

Die  $\sigma$ -Algebra der Borelmengen, die wir gleich definieren werden, ist das mit Abstand wichtigste Beispiel für erzeugte  $\sigma$ -Algebren. Ihre Bedeutung ergibt sich aus den folgenden Tatsachen:

- Eine „vernünftige“ Wahrscheinlichkeitstheorie, in der alle Teilmengen von  $\mathbb{R}$  (oder gar des  $\mathbb{R}^n$ ) zugelassen sind, kann es nicht geben. Die Präzisierung dieser Aussage finden Sie in den Ergänzungen auf Seite 28.
- Die Borelmengen bilden eine Klasse von Teilmengen, die einerseits geeignet für die Wahrscheinlichkeitstheorie ist und die andererseits alle Mengen enthält, die in möglichen Anwendungen interessant sein könnten.

### Die Borelmengen in $\mathbb{R}$

Vorbereitend sollte man sich daran erinnern, dass eine Teilmenge  $O$  von  $\mathbb{R}$  offen heißt, wenn für jedes  $x \in O$  ein  $\varepsilon > 0$  existiert, so dass das ganze Intervall  $]x - \varepsilon, x + \varepsilon[$  zu  $O$  gehört. Einfache Beispiele sind alle offenen Intervalle  $]a, b[$ , aber auch Vereinigungen beliebig vieler offener Intervalle.

Wir bezeichnen mit  $\mathcal{O}$  die Gesamtheit der offenen Teilmengen. Die  $\sigma$ -Algebra der Borelmengen auf  $\mathbb{R}$  ist dann  $\sigma(\mathcal{O})$ , die von  $\mathcal{O}$  erzeugte  $\sigma$ -Algebra. Unter einer Borelmenge versteht man ein Element dieser  $\sigma$ -Algebra<sup>14)</sup>.

<sup>14)</sup> Auf den ersten Blick sieht die Definition „eine Borelmenge ist ein Element der  $\sigma$ -Algebra der Borelmengen“ zirkular aus. Sie ist es aber nicht, man könnte ja auch „Element von  $\sigma(\mathcal{O})$ “ sagen.

Wir kümmern uns zunächst um einige *Beispiele*. Dazu verwenden wir die Strategie, die am Ende des vorigen Abschnitts beschrieben wurde: Baue die interessierenden Mengen „auf abzählbare Weise“ unter Verwendung offener Mengen oder schon konstruierter Borelmengen auf.

**1.** Eine Teilmenge  $A$  von  $\mathbb{R}$  heißt bekanntlich *abgeschlossen*, wenn die Menge  $\mathbb{R} \setminus A$  offen ist. Es ist leicht einzusehen, dass solche  $A$  Borelmengen sind:  $\mathbb{R} \setminus A$  ist offen und folglich Borelmenge, und  $A$  kann als  $\mathbb{R} \setminus (\mathbb{R} \setminus A)$  geschrieben werden. Insbesondere sind einpunktige Mengen Borelmengen.

**2.** Halboffene Intervalle, z.B. Intervalle der Form  $]a, b]$  sind Borelmengen. Das kann man auf verschiedene Weise begründen. Es ist richtig, weil  $]a, b] = ]a, b[ \cup \{b\}$  ist, eine Vereinigung von zwei Borelmengen.

Es folgt aber auch aus der Gleichung  $]a, b] = \bigcap_n ]a, b + 1/n[$  (abzählbarer Schnitt von Borelmengen) oder der Gleichung  $]a, b] = \bigcup_n [a + 1/n, b]$  (abzählbare Vereinigung von Borelmengen).

**3.**  $\mathbb{Q}$ , die Menge der rationalen Zahlen ist eine Borelmenge, denn  $\mathbb{Q}$  kann als abzählbare Vereinigung einpunktiger Mengen geschrieben werden. Folglich ist auch  $\mathbb{R} \setminus \mathbb{Q}$ , die Menge der irrationalen Zahlen, eine Borelmenge.

**4.** Sei  $f : \mathbb{R} \rightarrow \mathbb{R}$  eine stetige Funktion. Für beliebige Zahlen  $a, b$  mit  $a < b$  ist dann

$$B := \{x \mid x \in \mathbb{R}, a \leq f(x) < b\}$$

eine Borelmenge. Um das einzusehen, muss man sich erstens daran erinnern, dass Urbilder offener Mengen unter stetigen Funktionen offen sind<sup>15)</sup>, und zweitens kann man  $B$  „geschickt“ als

$$B = f^{-1}(]-\infty, b]) \setminus f^{-1}(]-\infty, a[)$$

schreiben: Das folgt daraus, dass „ $a \leq f(x) < b$ “ gleichwertig zu „ $f(x) < b$  und nicht  $f(x) < a$ “ ist.  $B$  ist also Komplementärmenge offener Mengen und damit eine Borelmenge.

Wenn man so fortfährt, entpuppen sich immer mehr Teilmengen von  $\mathbb{R}$  als Borelmengen. Man gelangt zu der folgenden

#### Faustregel für Borelmengen in $\mathbb{R}$ :

Alle Teilmengen von  $\mathbb{R}$ , in deren Definition nur höchstens abzählbar viele stetige Funktionen und die Symbole  $<, \leq, >, \geq, =$  vorkommen, sind Borelmengen.

Das muss im Einzelfall wie in den vorstehenden Beispielen eigentlich immer nachgeprüft werden. Wenn man das aber einige Male gemacht hat, spart man sich, die immer gleichen Argumente zu wiederholen. Diese Faustregel ist gemeint, wenn man sagt, dass „alle für die Anwendungen wichtigen Mengen“ Borelmengen sind. (Wie man eine Teilmenge von  $\mathbb{R}$  findet, die *keine* Borelmenge ist, wird in Satz 1.7.2 gezeigt.)

---

<sup>15)</sup>Genauer: Ist  $O$  offen, so ist  $f^{-1}(O) := \{x \mid f(x) \in O\}$  offen.

Die Borelmengen des  $\mathbb{R}^n$

Im  $\mathbb{R}^n$  ist es im Wesentlichen genauso. Wir erinnern uns, dass eine Teilmenge  $O$  offen genannt wird, wenn es zu jedem  $x \in O$  ein  $\varepsilon > 0$  so gibt, dass die (euklidische) Kugel um  $x$  mit dem Radius  $\varepsilon$  ganz in  $O$  liegt. Und ist dann  $\mathcal{O}$  die Gesamtheit aller offenen Mengen, so nennt man  $\sigma(\mathcal{O})$  die  $\sigma$ -Algebra der Borelmengen des  $\mathbb{R}^n$ .

Wie im Fall der Borelmengen in  $\mathbb{R}$  kann man sich nun nach und nach davon überzeugen, dass „alles, was man hinschreiben kann“, zu einer Borelmenge im  $\mathbb{R}^n$  führt. Wie im Eindimensionalen sind abgeschlossene und höchstens abzählbare Mengen Borelmengen, und man kann sich auch hier zunutze machen, dass Urbilder offener Mengen unter stetigen Funktionen wieder offen sind. So ist zum Beispiel

$$K := \{(x, y, z) \mid (x, y, z) \in \mathbb{R}^3, 1 < \sqrt{x^2 + y^2 + z^2} \leq 2\}$$

(aus der Kugel mit dem Radius 2 wurde die Einheitskugel entfernt) deswegen eine Borelmenge, weil man  $K$  als

$$K = f^{-1}([1, +\infty)) \setminus f^{-1}([2, +\infty))$$

schreiben kann, wobei  $f : \mathbb{R}^3 \rightarrow \mathbb{R}$  die stetige Funktion  $f(x, y, z) := \sqrt{x^2 + y^2 + z^2}$  ist. ( $f(x, y, z)$  ist die euklidische Norm  $\|(x, y, z)\|$  des Vektors  $(x, y, z)$ .)

Kurz: Auch im  $\mathbb{R}^n$  gilt, dass alles, was jemals für die Wahrscheinlichkeitsrechnung gebraucht werden wird, aus dem Bereich der Borelmengen nicht hinausführt.

Die Borelmengen in beliebigen metrischen oder topologischen Räumen

Nur der Vollständigkeit halber sei erwähnt, dass Borelmengen auch in viel allgemeineren Situationen auftreten können: Wo immer man mit offenen Mengen arbeitet, also in metrischen Räumen oder allgemeiner in topologischen Räumen, ist die  $\sigma$ -Algebra der Borelmengen die von den offenen Mengen erzeugte  $\sigma$ -Algebra.

Das wird in diesem Buch keine Rolle spielen. Es sollte aber erwähnt werden, dass es eine fundamentale Schwierigkeit gibt, die bei den „kleinen“ Räumen wie dem  $\mathbb{R}^n$  nicht auftritt: Mitunter kann das, was einen interessiert, nicht durch abzählbar viele Bedingungen definiert werden. Zum Beispiel ist die Frage „Ist  $f$  stetig?“ für eine Funktion  $f : \mathbb{R} \rightarrow \mathbb{R}$  nicht dadurch zu entscheiden, dass man abzählbar viele Funktionswerte  $f(x)$  untersucht.

Andere Erzeuger der Borelmengen

Später wird es wichtig sein, gewisse Eigenschaften für alle Borelmengen nachweisen zu können. Dazu sollen allgemeine Ergebnisse des folgenden Typs angewendet werden: Wenn man etwas über  $\mathcal{M}$  weiß, dann weiß man manchmal auch gleich etwas über die erzeugte  $\sigma$ -Algebra  $\sigma(\mathcal{M})$ . Es wäre daher wünschenswert zu wissen, ob man die  $\sigma$ -Algebra der Borelmengen auch als  $\sigma(\mathcal{M})$  schreiben

kann, wobei  $\mathcal{M}$  ein anderes (und hoffentlich viel einfacheres) Mengensystem ist als das System der offenen Mengen.

Um so etwas zu untersuchen, ist es nützlich, mit einer elementaren Beobachtung zu beginnen:

**Lemma 1.5.1.** *Es seien  $\mathcal{M}_1$  und  $\mathcal{M}_2$  Mengensysteme auf  $\Omega$ . Ist dann  $\mathcal{M}_1 \subset \sigma(\mathcal{M}_2)$  und  $\mathcal{M}_2 \subset \sigma(\mathcal{M}_1)$ , so gilt  $\sigma(\mathcal{M}_1) = \sigma(\mathcal{M}_2)$ .*

**Beweis:** Die erste Bedingung impliziert  $\sigma(\mathcal{M}_1) \subset \sigma(\mathcal{M}_2)$ , die zweite  $\sigma(\mathcal{M}_2) \subset \sigma(\mathcal{M}_1)$ . Damit ist alles gezeigt.  $\square$

Es folgen die wichtigsten Beispiele für Erzeuger der Borelmengen auf  $\mathbb{R}$ :

**Satz 1.5.2.** *Für jedes der folgenden Mengensysteme  $\mathcal{M}$  gilt:  $\sigma(\mathcal{M})$  ist die  $\sigma$ -Algebra der Borelmengen von  $\mathbb{R}$*

- (i)  $\mathcal{M}$  besteht aus allen offenen Intervallen.
- (ii)  $\mathcal{M}$  besteht aus allen abgeschlossenen Intervallen.
- (iii)  $\mathcal{M}$  besteht aus allen offenen Intervallen mit rationalen Endpunkten.
- (iv)  $\mathcal{M}$  besteht aus allen abgeschlossenen Intervallen mit rationalen Endpunkten.
- (v)  $\mathcal{M}$  besteht aus allen offenen Intervallen des Typs  $]a, +\infty[$ .
- (vi)  $\mathcal{M}$  besteht aus allen abgeschlossenen Intervallen des Typs  $[a, +\infty[$ .
- (vii)  $\mathcal{M}$  besteht aus allen offenen Intervallen des Typs  $]a, +\infty[$  mit  $a \in \mathbb{Q}$ .
- (viii)  $\mathcal{M}$  besteht aus allen abgeschlossenen Intervallen des Typs  $[a, +\infty[$  mit  $a \in \mathbb{Q}$ .

Man beachte, dass die Mengensysteme in (iii), (iv), (vii) und (viii) abzählbar sind<sup>16)</sup>.

**Beweis:** Die Beweise sind alle sehr ähnlich. Mit  $\mathcal{O}$  bezeichnen wir wieder das System der offenen Teilmengen von  $\mathbb{R}$ , als typisches Beispiel führen wir den Beweis von (i).  $\mathcal{M}$  besteht also aus allen offenen Intervallen.

Klar ist, dass  $\mathcal{M}$  eine Teilmenge von  $\mathcal{O}$  ist, es fehlt noch der Nachweis, dass jede offene Menge in  $\sigma(\mathcal{M})$  liegt. Sei dazu  $O \subset \mathbb{R}$  offen. Wir nummerieren die rationalen Zahlen in  $O$  durch:  $r_1, r_2, \dots$  und legen um jedes  $r_n$  ein möglichst großes offenes Intervall mit Mittelpunkt  $r_n$ , das ganz in  $O$  liegt. Genauer: Setze

$$\varepsilon_n := \sup\{\varepsilon \mid ]r_n - \varepsilon, r_n + \varepsilon[ \subset O\}.$$

Dann lässt sich zeigen, dass  $O$  gleich  $\bigcup_n ]r_n - \varepsilon_n, r_n + \varepsilon_n[$  ist<sup>17)</sup>, die Menge  $O$  ist also abzählbare Vereinigung von Elementen aus  $\mathcal{M}$  und liegt deswegen in  $\sigma(\mathcal{M})$ . Wegen des vorstehenden Lemmas ist (i) damit bewiesen.

<sup>16)</sup>Eine  $\sigma$ -Algebra  $\mathcal{E}$ , zu der man eine abzählbare Teilmenge  $\mathcal{M}$  mit der Eigenschaft  $\mathcal{E} = \sigma(\mathcal{M})$  finden kann, heißt *abzählbar erzeugt*. Die  $\sigma$ -Algebra der Borelmengen in  $\mathbb{R}$  (und auch im  $\mathbb{R}^n$ ) hat diese Eigenschaft.

<sup>17)</sup>Zunächst überlegt man sich, dass stets  $]r_n - \varepsilon_n, r_n + \varepsilon_n[ \subset O$  gilt. Deswegen ist  $\bigcup_n ]r_n - \varepsilon_n, r_n + \varepsilon_n[ \subset O$ , und „ $\subset$ “ sieht man so ein: Ist  $x \in O$ , wähle  $\varepsilon > 0$  mit  $]x - \varepsilon, x + \varepsilon[ \subset O$ . Ist dann  $r_n$  eine rationale Zahl, mit  $|x - r_n| < \varepsilon/2$ , so ist  $x \in ]r_n - \varepsilon_n, r_n + \varepsilon_n[$ . So ein  $r_n$  existiert, da die rationalen Zahlen dicht in den reellen Zahlen liegen.

Der Beweis für die anderen Aussagen verläuft ganz analog. Es ist zwischen-durch mehrfach wichtig zu wissen, dass die rationalen Zahlen dicht liegen.  $\square$

Im  $\mathbb{R}^n$  ist es ganz ähnlich, auch da gibt es die Möglichkeit, die Borelmengen durch andere Systeme zu erzeugen als durch das System der offenen Teilmengen. Nachstehend finden sich einige Beispiele. Auf Beweise verzichten wir hier, sie sind ähnlich einfach wie im eindimensionalen Fall.

Die Borelmengen des  $\mathbb{R}^n$  werden erzeugt durch

- das Mengensystem der offenen Kugeln (das sind die Mengen der Form  $\{x \mid \|x - x_0\| < r\}$ );
- das Mengensystem der abgeschlossenen Kugeln (das sind die Mengen der Form  $\{x \mid \|x - x_0\| \leq r\}$ );
- das Mengensystem der Mengen des Typs

$$\{(x_1, \dots, x_n) \mid x_1 \geq a_1, \dots, x_n \geq a_n\},$$

wobei  $(a_1, \dots, a_n) \in \mathbb{R}^n$ ;

- das Mengensystem der Hyperquader, das sind Mengen des Typs

$$\{(x_1, \dots, x_n) \mid a_1 \leq x_1 \leq b_1, \dots, a_n \leq x_n \leq b_n\},$$

wobei  $(a_1, \dots, a_n), (b_1, \dots, b_n) \in \mathbb{R}^n$  mit  $a_1 < b_1, \dots, a_n < b_n$ .

Und da  $\mathbb{Q}^n$  im  $\mathbb{R}^n$  dicht liegt, kann man sich sogar auf Kugeln mit Mittelpunkt im  $\mathbb{Q}^n$  und rationalem Radius beschränken. Das impliziert, dass auch im  $\mathbb{R}^n$  die Borelmengen durch ein abzählbares Mengensystem erzeugt werden.

## 1.6 Zwei wichtige Beweistechniken

In vielen Beweisen in späteren Kapiteln wird es darum gehen, für alle Elemente einer  $\sigma$ -Algebra etwas nachzuweisen. Dabei wird man oft sehr effektiv und elegant die Beweistechniken einsetzen können, die in diesem Abschnitt vorgestellt werden.

Von  $\mathcal{M}$  zu  $\sigma(\mathcal{M})$

Wieder sei  $\mathcal{M}$  ein Mengensystem auf  $\Omega$ . Wir haben  $\mathcal{M}$  auf sparsamste Weise zur  $\sigma$ -Algebra  $\sigma(\mathcal{M})$  ergänzt. Genau genommen wissen wir über diese  $\sigma$ -Algebra nur zwei Dinge. *Erstens* ist es eine  $\sigma$ -Algebra, und *zweitens* ist sie in jeder  $\sigma$ -Algebra enthalten, die  $\mathcal{M}$  enthält. Und das muss reichen, um alle Aussagen über erzeugte  $\sigma$ -Algebren zu beweisen!

Zum späteren Zitieren formulieren wir diese Beobachtung noch einmal als

**Satz 1.6.1.** Es sei  $X$  eine Eigenschaft, die Teilmengen von  $\Omega$  haben können. Es soll möglich sein, die folgenden beiden Tatsachen nachzuweisen:

(i) Alle  $E \in \mathcal{M}$  haben  $X$ .

(ii) Die Gesamtheit der  $E \subset \Omega$ , die  $X$  haben, bildet eine  $\sigma$ -Algebra.

Dann haben alle Elemente von  $\sigma(\mathcal{M})$  die Eigenschaft  $X$ .

Wir wollen diese Beweisstrategie<sup>18)</sup> im nächsten Satz illustrieren. Es geht dabei um Borelmengen, unserem wichtigsten Beispiel für erzeugte  $\sigma$ -Algebren.

**Satz 1.6.2.** (i) Sei  $B \subset \mathbb{R}^n$  eine Borelmenge und  $x_0 \in \mathbb{R}^n$ . Dann ist auch  $x_0 + B$ , also die Menge  $\{x_0 + x \mid x \in B\}$ , eine Borelmenge.

Kurz: Translationen von Borelmengen sind Borelmengen.

(ii) Sei  $B \subset \mathbb{R}^n$  eine Borelmenge und  $a \in \mathbb{R}$ . Dann ist auch  $a \cdot B$ , also die Menge  $\{a \cdot x \mid x \in B\}$ , eine Borelmenge.

Kurz: Vielfache von Borelmengen sind Borelmengen.

(iii) Sei  $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$  eine stetige Funktion und  $B$  eine Borelmenge im  $\mathbb{R}^m$ . Dann ist die Urbildmenge  $f^{-1}(B)$ , also die Menge  $\{x \in \mathbb{R}^n \mid f(x) \in B\}$ , eine Borelmenge im  $\mathbb{R}^n$ .

**Beweis:** (i) Wir definieren

$$\mathcal{B}_{x_0} := \{B \mid B \text{ und } x_0 + B \text{ sind Borelmengen}\}.$$

Das ist eine  $\sigma$ -Algebra: Für diesen Beweis muss man wirklich nur die Definition kennen und einige offensichtliche Tatsachen im Zusammenhang mit Translationen beachten. (Z.B., dass disjunkte Teilmengen nach dem Verschieben immer noch disjunkt sind.) Da die Borelmengen von den offenen Teilmengen erzeugt sind und Translationen offener Mengen wieder offen sind, ist der Beweis schon geführt.

(ii) Im Fall  $a \neq 0$  geht das ganz analog zum vorigen Beweis, diesmal sind die Borelmengen  $B$  zu betrachten, für die  $a \cdot B$  Borelmenge ist. Und im Fall  $a = 0$  ist  $a \cdot B$  die leere Menge oder  $\{0\}$ , das sind beides Borelmengen.

(iii) Hier beginnt man mit

$$\mathcal{B}_f := \{B \mid B \text{ ist Borelmenge im } \mathbb{R}^m, \text{ und } f^{-1}(B) \text{ ist Borelmenge im } \mathbb{R}^n\}.$$

Das ist eine  $\sigma$ -Algebra, dazu sind nur einfache mengentheoretische Tatsachen im Zusammenhang mit  $B \mapsto f^{-1}(B)$  auszunutzen (etwa:  $f^{-1}(\mathbb{R}^m) = \mathbb{R}^n$ ). Und da Urbilder offener Mengen unter stetigen Abbildungen wieder offen sind, enthält  $\mathcal{B}_f$  alle offenen Teilmengen des  $\mathbb{R}^m$ . Fertig<sup>19)</sup>.  $\square$

---

<sup>18)</sup>Sie wird manchmal die „Strategie der guten Mengen“ genannt.

<sup>19)</sup>Man kann übrigens (ii) im Fall  $a \neq 0$  als Spezialfall von (iii) auffassen, wenn man (iii) für die stetige Abbildung  $f : x \mapsto x/a$  anwendet. Dann ist nämlich  $f^{-1}(B) = a \cdot B$ .

**Das Dynkin-Argument**

Das zweite Beweisverfahren ist etwas komplizierter zu beschreiben. Es wird in diesem Buch mehrfach eine wichtige Rolle spielen. Es erwartet Sie eine Definition, ein Satz und eine Anwendung zur Illustration.

Zunächst die

**Definition 1.6.3.** Sei  $\Omega$  eine Menge und  $\mathcal{D}$  ein Mengensystem auf  $\Omega$ . Das System  $\mathcal{D}$  heißt ein Dynkinsystem<sup>20)</sup>, wenn die folgenden drei Bedingungen erfüllt sind:

- (i) Die leere Menge und  $\Omega$  gehören zu  $\mathcal{D}$ .
- (ii) Mit  $D$  gehört auch  $\Omega \setminus D$  zu  $\mathcal{D}$ .
- (iii) Ist  $D_1, D_2, \dots$  eine disjunkte Folge in  $\mathcal{D}$  (die  $D_n$  liegen also in  $\mathcal{D}$ , und für  $n \neq m$  gilt  $D_n \cap D_m = \emptyset$ ), so gehört auch  $\bigcup_n D_n$  zu  $\mathcal{D}$ .

Anders ausgedrückt: Es sind fast die gleichen Bedingungen wie bei den Forderungen an eine  $\sigma$ -Algebra. Der einzige Unterschied besteht darin, dass man nur verlangt, dass Vereinigungen *disjunkter* Folgen wieder dazugehören. Offensichtlich sind damit alle  $\sigma$ -Algebren Dynkinsysteme. Umgekehrt muss das nicht gelten. Zum Beispiel bilden die Teilmengen von  $\{1, \dots, 100\}$ , deren Anzahl durch 10 teilbar ist, ein Dynkinsystem, doch das ist keine  $\sigma$ -Algebra. (Warum?) Dynkinsysteme sind immer dann gut zu beherrschen, wenn man etwas über *disjunkte* Vereinigungen weiß; das wird zum Beispiel bei Wahrscheinlichkeiten der Fall sein.

Es ist klar, dass der Durchschnitt beliebig vieler Dynkinsysteme wieder ein Dynkinsystem ist<sup>21)</sup>. Folglich kann man wieder für ein beliebiges Mengensystem  $\mathcal{M}$  das kleinste Dynkinsystem betrachten, das  $\mathcal{M}$  enthält: Schneide einfach alle Dynkinsysteme, die  $\mathcal{M}$  enthalten. Dieses *von  $\mathcal{M}$  erzeugte* Dynkinsystem wollen wir  $\mathcal{D}(\mathcal{M})$  nennen. Aufgrund der Definition ist dann klar:

Ist  $\mathcal{D}$  ein Dynkinsystem auf  $\Omega$  und gilt  $\mathcal{M} \subset \mathcal{D}$ , so ist  $\mathcal{D}(\mathcal{M}) \subset \mathcal{D}$ .

Diese Beobachtung wird gleich wichtig werden.

Hier nun das für uns wichtigste Ergebnis über Dynkinsysteme. Man sieht ihm seine Bedeutung für die zukünftigen Untersuchungen wirklich nicht an:

**Satz 1.6.4.** Es sei  $\mathcal{M}$  ein schnitt-stabiles Mengensystem auf der Menge  $\Omega$ : Für  $E, F \in \mathcal{M}$  soll stets auch  $E \cap F \in \mathcal{M}$  gelten. Dann ist  $\mathcal{D}(\mathcal{M}) = \sigma(\mathcal{M})$ . Kurz: Das erzeugte Dynkinsystem ist bereits die von  $\mathcal{M}$  erzeugte  $\sigma$ -Algebra.

Speziell für Borelmengen heißt das: Ist  $X$  eine Eigenschaft, die Borelmengen haben können und gilt

<sup>20)</sup>Evgenii Dynkin, geboren 1924. Dynkin wurde in Russland wegen seiner jüdischen Abstammung verfolgt. Erst nach Stalins Tod wurde er Professor in Moskau, 1968 emigrierte er in die USA. Dort ist er Professor an der Cornell-Universität in Ithaka, New York. Von ihm stammen wichtige Beiträge zur Wahrscheinlichkeitstheorie, insbesondere zur Theorie stochastischer Prozesse.

<sup>21)</sup>Man kopiere einfach die Ideen aus dem Beweis auf Seite 14, dass der Schnitt von  $\sigma$ -Algebren eine  $\sigma$ -Algebra ist.



Dynkin

- (i) alle  $B$  in einem schnitt-stabilen Erzeuger der Borelmengen haben  $X$ , sowie
- (ii) das System der Borelmengen, die  $X$  haben, bildet ein Dynkinsystem, so haben alle Borelmengen die Eigenschaft  $X$ .

**Beweis:** Wir beweisen die Behauptung in drei Schritten:

1. Schritt:  $\mathcal{D}(\mathcal{M})$  ist schnitt-stabil. Fixiere zunächst ein  $E \in \mathcal{M}$ . Wir behaupten, dass  $E \cap F \in \mathcal{D}(\mathcal{M})$  für alle  $F \in \mathcal{D}(\mathcal{M})$  gilt. Zum Beweis definieren wir

$$\mathcal{D}_E := \{F \mid F \in \mathcal{D}(\mathcal{M}), E \cap F \in \mathcal{D}(\mathcal{M})\}.$$

Das ist ein Dynkinsystem, dazu muss man sich nur an einige elementare Rechenregeln aus der Mengenlehre erinnern:

- $\Omega \cap E = E$ : Wegen  $E \in \mathcal{M} \subset \mathcal{D}(\mathcal{M})$  beweist das  $\Omega \in \mathcal{D}_E$ .
- $\emptyset \cap E = \emptyset$ : Da die leere Menge zu  $\mathcal{D}(\mathcal{M})$  gehört, folgt  $\emptyset \in \mathcal{D}_E$ .
- Um zu zeigen, dass Komplemente wieder dazugehören, benötigen wir eine Vorbereitung.

Nach Voraussetzung enthalten Dynkinsysteme  $\mathcal{D}$  mit jeder Menge auch die Komplementärmenge. Wir behaupten, dass auch *relative Komplemente* enthalten sind, d.h. für  $A, B \in \mathcal{D}$  mit  $A \subset B$  ist  $B \setminus A \in \mathcal{D}$ . Zur Begründung ist nur zu bemerken, dass man  $B \setminus A$  als

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

schreiben kann, wobei die Vereinigung aufgrund der Voraussetzung  $A \subset B$  eine disjunkte Vereinigung ist, also wieder zu  $\mathcal{D}$  gehört.

Sei nun  $F \in \mathcal{D}_E$ , die Menge  $E \cap F$  liegt also in  $\mathcal{D}(\mathcal{M})$ . Da das ein Dynkin-system ist, gehört aufgrund der Vorbemerkung auch  $E \setminus (E \cap F)$  zu  $\mathcal{D}(\mathcal{M})$ , und  $E \setminus (E \cap F)$  stimmt mit  $E \cap (\Omega \setminus F)$  überein. Damit gilt  $\Omega \setminus F \in \mathcal{D}_E$ .

- Sei  $(F_n)$  eine disjunkte Folge in  $\mathcal{D}_E$ . Aus der Formel

$$E \cap \left( \bigcup_n F_n \right) = \bigcup_n E \cap F_n$$

schließen wir, dass  $E \cap (\bigcup_n F_n)$  zu  $\mathcal{D}(\mathcal{M})$  gehört, denn alle  $E \cap F_n$  gehören nach Voraussetzung zu  $\mathcal{D}(\mathcal{M})$ , und  $(E \cap F_n)_n$  ist eine disjunkte Folge.

Die Tatsache, dass  $\mathcal{M}$  schnitt-stabil ist, impliziert sofort, dass  $\mathcal{M} \subset \mathcal{D}_E$ . Und daraus können wir mit dem vor dem Satz beschriebenen Argument schließen, dass  $\mathcal{D}(\mathcal{M}) \subset \mathcal{D}_E$  gilt: Für alle  $F \in \mathcal{D}(\mathcal{M})$  ist also  $E \cap F \in \mathcal{D}(\mathcal{M})$ .

Nun fixieren wir ein  $F' \in \mathcal{D}(\mathcal{M})$  und betrachten  $\mathcal{D}_{F'}$ : Dieses Mengensystem ist genauso definiert wie eben, nämlich als das System der  $F \in \mathcal{D}(\mathcal{M})$ , für die auch  $F' \cap F \in \mathcal{D}(\mathcal{M})$  gilt. Das ist wieder ein Dynkinsystem, das aufgrund des ersten Beweisteils alle  $E \in \mathcal{M}$  enthält. Es folgt:  $\mathcal{D}(\mathcal{M}) \subset \mathcal{D}_{F'}$ . Das ist aber

gerade die Aussage, dass für beliebige  $F \in \mathcal{D}(\mathcal{M})$  die Menge  $F \cap F'$  zu  $\mathcal{D}(\mathcal{M})$  gehört, und folglich ist  $\mathcal{D}(\mathcal{M})$  schnitt-stabil.

*2. Schritt:*  $\mathcal{D}(\mathcal{M})$  ist eine  $\sigma$ -Algebra. Allgemein gilt, dass ein schnitt-stabiles Dynkinsystem  $\mathcal{D}$  eine  $\sigma$ -Algebra ist. Es ist doch nur zu zeigen, dass mit  $D_1, D_2, \dots$  auch  $\bigcup_n D_n$  zu  $\mathcal{D}$  gehört. Auch wenn die  $D_n$  möglicherweise nicht paarweise disjunkt sind.

Das kann man dadurch einsehen, dass man  $\bigcup_n D_n$  geschickt aus Durchschnitten, relativen Komplementen und disjunkten Vereinigungen zusammensetzt:  $D_1 \cup D_2$  zum Beispiel ist doch die disjunkte Vereinigung von  $D_1$  und  $D_2 \setminus (D_1 \cap D_2)$ , gehört also wieder zu  $\mathcal{D}$ . Für die hier auftretende abzählbare Vereinigung  $\bigcup_n D_n$  bauen wir diese Idee ein bisschen aus: Wir definieren  $E_1 := D_1$  und dann rekursiv  $E_{n+1} := D_{n+1} \setminus (D_{n+1} \cap E_n)$ . Dann gehören alle  $E_n$  zu  $\mathcal{D}$ , und  $\bigcup_n D_n$  stimmt mit der (disjunkten!) Vereinigung  $\bigcup_n E_n$  überein und liegt deswegen in  $\mathcal{D}$ .

*3. Schritt:*  $\mathcal{D}(\mathcal{M}) = \sigma(\mathcal{M})$ .

Das ist überraschend einfach: Wir wissen schon, dass  $\mathcal{D}(\mathcal{M})$  eine  $\sigma$ -Algebra ist, die  $\mathcal{M}$  enthält. Also muss  $\sigma(\mathcal{M}) \subset \mathcal{D}(\mathcal{M})$  gelten, denn  $\sigma(\mathcal{M})$  ist ja die kleinste  $\sigma$ -Algebra, die  $\mathcal{M}$  enthält. Umgekehrt:  $\sigma(\mathcal{M})$  ist ja insbesondere auch ein Dynkinsystem, in dem  $\mathcal{M}$  enthalten ist, und daher wissen wir, dass  $\mathcal{D}(\mathcal{M}) \subset \sigma(\mathcal{M})$  gelten muss (denn  $\mathcal{D}(\mathcal{M})$  ist das kleinste Dynkinsystem, in dem  $\mathcal{M}$  liegt). Zusammen zeigt das  $\mathcal{D}(\mathcal{M}) = \sigma(\mathcal{M})$ .  $\square$

Im nächsten Satz stellen wir eine wichtige Anwendung vor. Sie zeigt, dass Dynkinsysteme für Situationen maßgeschneidert sind, in denen man eventuell keine Informationen über *beliebige*, wohl aber über *disjunkte* Vereinigungen hat.

**Satz 1.6.5.** *Es seien  $\mathbb{P}_1$  und  $\mathbb{P}_2$  Wahrscheinlichkeitsmaße, die auf den Borelmengen von  $\mathbb{R}$  definiert sind. Stimmen dann  $\mathbb{P}_1$  und  $\mathbb{P}_2$  auf allen abgeschlossenen Intervallen überein, gilt also stets  $\mathbb{P}_1([a, b]) = \mathbb{P}_2([a, b])$ , so ist  $\mathbb{P}_1 = \mathbb{P}_2$ , d.h. es ist sogar  $\mathbb{P}_1(B) = \mathbb{P}_2(B)$  für alle Borelmengen.*

*Allgemeiner gilt: Sind zwei Wahrscheinlichkeitsmaße auf  $(\Omega, \mathcal{E})$  definiert und stimmen sie auf einem schnitt-stabilen Erzeuger von  $\mathcal{E}$  überein, so sind sie identisch.*

**Beweis:** Wir wenden Satz 1.6.3 an, indem wir zwei Tatsachen kombinieren:

- Das System der Borelmengen, für das  $\mathbb{P}_1(B) = \mathbb{P}_2(B)$  gilt, ist ein Dynkinsystem. Hier wird es wichtig, dass wir nur *disjunkte* Vereinigungen behandeln müssen: Gilt  $\mathbb{P}_1(E_i) = \mathbb{P}_2(E_i)$  für  $i = 1, 2, \dots$ , so gilt auch  $\mathbb{P}_1(\bigcup E_i) = \mathbb{P}_2(\bigcup E_i)$ , falls die  $E_i$  paarweise disjunkt sind. Das folgt daraus, dass beide Maße  $\sigma$ -additiv sind. (Ohne die Voraussetzung der Disjunkttheit könnte man nicht so schließen.)
- Die abgeschlossenen Intervalle bilden einen schnitt-stabilen Erzeuger der Borelmengen von  $\mathbb{R}$  (Satz 1.5.2).

(Die „Eigenschaft  $X$ “ aus dem Satz 1.6.4 ist also „ $\mathbb{P}_1(B) = \mathbb{P}_2(B)$ “.)

Die allgemeinere Variante wird genauso bewiesen.  $\square$

Die Bedeutung dieses Ergebnisses kann kaum überbetont werden. Es besagt doch, dass es ausreicht, sich um Wahrscheinlichkeiten für Intervalle zu kümmern, also um sehr einfache Teilmengen von  $\mathbb{R}$ .

### Zusammenfassung der Abschnitte 1.4 bis 1.6

Die *Borelmengen* in  $\mathbb{R}$  (bzw. im  $\mathbb{R}^n$ ) bilden eine  $\sigma$ -Algebra von Teilmengen von  $\mathbb{R}$  (bzw. des  $\mathbb{R}^n$ ). Die Gesamtheit der Borelmengen ist die kleinste  $\sigma$ -Algebra, die alle offenen Teilmengen enthält. Die in den Anwendungen vorkommenden Teilmengen von  $\mathbb{R}$  (bzw. des  $\mathbb{R}^n$ ) sind immer Borelmengen.

Es gibt verschiedene Techniken, mit denen man zeigen kann, dass eine gewisse Eigenschaft, die man für Mengen sinnvoll formulieren kann, für alle Borelmengen erfüllt ist. Die Einzelheiten können Sie später durcharbeiten.

## 1.7 Ergänzungen

Wir beginnen diesen Abschnitt mit einem

Gegenbeispiel: Eine Teilmenge von  $\mathbb{R}$ , die keine Borelmenge ist

Aus der Maßtheorie übernehmen wir ohne Beweis die folgende Tatsache:

**Satz 1.7.1.** Bezeichne mit  $\mathcal{B}$  die  $\sigma$ -Algebra der Borelmengen auf  $\mathbb{R}$ . Dann gibt es eine eindeutig bestimmte Abbildung  $\lambda : \mathcal{B} \rightarrow [0, +\infty]$  mit den folgenden Eigenschaften:

(i) Für disjunkte  $B_1, B_2, \dots$  in  $\mathcal{B}$  ist  $\lambda(\bigcup_n B_n) = \sum_n \lambda(B_n)$ . Dabei ist die rechte Seite als Reihe in  $[0, +\infty]$  zu verstehen. Die Reihensumme ist dann das Supremum der Partialsummen<sup>22)</sup>.

(ii)  $\lambda([a, b]) = b - a$  für alle Intervalle  $[a, b]$ .

Dieses Maß heißt das Borel-Lebesgue-Maß auf  $\mathbb{R}$ .

Ebenfalls ohne Beweis werden wir verwenden:

1.  $\lambda$  ist translationsinvariant: Es ist  $\lambda(x + B) = \lambda(B)$  für alle Borelmengen  $B$ . (Dabei ist  $x + B := \{x + y \mid y \in B\}$  die Translation um  $x$  von  $B$ .)
2.  $\lambda$  ist monoton: Für  $B \subset B'$  ist  $\lambda(B) \leq \lambda(B')$ . (Der Beweis ist völlig analog zum Beweis von Satz 1.3.2 (iii).)

Hier nun der Satz, der dafür verantwortlich ist, dass Wahrscheinlichkeitstheorie für Anfänger so kompliziert sein muss:

---

<sup>22)</sup>Hier muss man sich daran erinnern, dass  $\infty + x = x + \infty = \infty$  für alle  $x \in [0, +\infty]$  als  $\infty$  definiert ist.

**Satz 1.7.2.** Es gibt eine Teilmenge  $B_0$  von  $\mathbb{R}$ , die keine Borelmenge ist.

**Beweis:** Wir werden Mengen  $B_1, B_2, \dots$  mit den folgenden Eigenschaften „konstruieren“:

- Alle  $B_n$  entstehen als geeignete Translationen  $x_n + B_0$  einer Menge  $B_0$ .
- Die  $B_n$  sind paarweise disjunkt.
- $[0, 1] \subset \bigcup_n B_n \subset [-1, 2]$ .

Dann kann die Menge  $B_0$  keine Borelmenge sein, wir beweisen das indirekt. Wäre sie doch eine, so wären auch alle  $B_n$  als Translationen Borelmengen.  $\lambda(\bigcup_n B_n)$  ist wegen der Disjunktheit gleich  $\sum_n \lambda(B_n)$ , wobei aufgrund der Translationsinvarianz von  $\lambda$  jedes  $\lambda(B_n)$  gleich  $\lambda(B_0)$  ist:

$$\lambda\left(\bigcup_n B_n\right) = \lambda(B_0) + \lambda(B_0) + \dots$$

Andererseits liegt  $\bigcup_n B_n$  zwischen  $[0, 1]$  und  $[-1, 2]$ , die Monotonie impliziert dann

$$\begin{aligned} 1 &= \lambda([0, 1]) \\ &\leq \lambda\left(\bigcup_n B_n\right) \\ &= \lambda(B_0) + \lambda(B_0) + \dots \\ &\leq \lambda([-1, 2]) \\ &= 3. \end{aligned}$$

Das ist ein Widerspruch, denn es gibt keine reelle Zahl  $a \geq 0$ , für welche die Reihensumme  $a + a + \dots$  zwischen 1 und 3 liegt. (Für  $a = 0$  kommt Null heraus, und für  $a > 0$  ergibt sich Unendlich.)

Es fehlt noch der Nachweis, dass es so eine Folge  $B_1, B_2, \dots$  gibt. Dazu betrachten wir auf  $[0, 1]$  eine spezielle Relation<sup>23)</sup>  $\pi$ . Sie ist dadurch definiert, dass  $x \pi y$  genau dann gelten soll, wenn  $x - y$  eine rationale Zahl ist. Es ist dann nicht schwer zu sehen, dass  $\pi$  eine Äquivalenzrelation ist.

Für jedes  $x \in [0, 1]$  bezeichnen wir dann die zugehörige Äquivalenzklasse mit  $D_x$ : Das ist die Gesamtheit aller  $y \in [0, 1]$  mit  $x \pi y$ . Zwei Äquivalenzklassen  $D_{x_1}, D_{x_2}$  sind dann entweder identisch (wenn  $x_1 \pi x_2$ ) oder disjunkt. Offensichtlich ist auch  $x \in D_x$  für  $x \in [0, 1]$ , und deswegen ist  $[0, 1]$  die disjunkte Vereinigung dieser Äquivalenzklassen.

Nun wird  $B_0$  definiert: Wir wählen aus jeder der Äquivalenzklassen genau ein Element aus. Dazu muss man das Auswahlaxiom anwenden (mehr dazu finden Sie im Anhang auf Seite 353). Wir schreiben noch die (abzählbar vielen) rationalen Zahlen in  $[-1, 1]$  als Folge  $q_1, q_2, \dots$  und definieren dann  $B_n := q_n + B_0$ . Dann gilt wirklich:

<sup>23)</sup>Siehe Anhang, Seite 353.

- Jedes  $B_n$  ist eine Translation von  $B_0$ .
- Jedes  $B_n$  (und damit die Vereinigung) liegt in  $[-1, 2]$ .
- Ist  $y \in [0, 1]$ , so kann man ein  $x \in B_0$  wählen, das in  $D_y$  liegt: Aus jeder Äquivalenzklasse gibt es ja einen Vertreter in  $B_0$ . Die Zahl  $y - x$  ist also rational, und sie liegt in  $[-1, 1]$ , denn sowohl  $x$  als auch  $y$  liegen in  $[0, 1]$ . Folglich gibt es ein  $n$  mit  $y - x = q_n$ , und das bedeutet  $y = x + q_n \in B_n$ . Damit ist  $[0, 1] \subset \bigcup_n B_n$  gezeigt.
- Wir beweisen noch, dass die  $B_n$  paarweise disjunkt sind. Sei dazu  $n \neq m$ , also auch  $q_n \neq q_m$ . Gäbe es ein  $y \in B_n \cap B_m$ , könnte man  $y$  sowohl als  $q_n + x$  als auch als  $q_m + x'$  für geeignete  $x, x' \in B_0$  schreiben. Es wäre also  $q_n - q_m = x' - x$ . Das würde  $x\pi x'$  implizieren: ein Widerspruch, denn wegen  $q_n \neq q_m$  ist  $x \neq x'$ , und deswegen müssen  $x$  und  $x'$  in verschiedenen Äquivalenzklassen liegen. (Beachte, dass  $B_0$  aus jeder Äquivalenzklasse nur ein Element enthält.)

Damit ist die Existenz einer Menge bewiesen, die keine Borelmenge ist.  $\square$

#### Die unerfreulichen Konsequenzen des vorstehenden Satzes

Wahrscheinlichkeitstheorie könnte doch so einfach sein! Die Lehrbücher könnten so beginnen: „ $\Omega$  ist die Menge der Elementarereignisse, und für jede Teilmenge  $E \subset \Omega$  ist die Wahrscheinlichkeit  $\mathbb{P}(E)$  definiert. Von der die Wahrscheinlichkeiten beschreibenden Abbildung  $\mathbb{P} : \mathcal{P}(\Omega) \rightarrow [0, 1]$  verlangen wir nur, dass  $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F)$  für disjunkte  $E, F$  gilt, dass also  $\mathbb{P}$  endlich additiv ist.“ Niemand bräuchte also zu lernen, was eine  $\sigma$ -Algebra oder ein Dynkinsystem ist, wie erzeugte  $\sigma$ -Algebren definiert sind, was man unter Borelmengen versteht und was  $\sigma$ -Additivität bedeutet.

Wenn man bereit ist, sich auf ganz einfache Situationen – auf endliche  $\Omega$  – zu beschränken, kann man es wirklich so machen. Das würde aber nur einen winzigen Bruchteil der interessanten Anwendungsgebiete abdecken. Sehr oft nämlich ist  $\Omega$  unendlich, zum Beispiel ein Intervall. Und dann kann es nicht so einfach gehen, wie es im vorigen Absatz versucht wurde: Auch für zukünftige Generationen wird es solche Lehrbücher leider nicht geben.

Es gibt zwei Gründe:

1. *Warum reicht es nicht, dass  $\mathbb{P}$  endlich additiv ist?* Das liegt daran, dass es für endlich additive Maße nur eine sehr schwerfällige Maß- und Integrationstheorie gibt. Wir brauchen aber später dringend eine reichhaltige Theorie der Integrale, um über Erwartungswerte und Streuungen reden zu können und in der Lage zu sein, dazu auch nichttriviale Ergebnisse beweisen zu können.
2. *Warum kann man nicht alle Teilmengen als Ereignisse zulassen?* Das liegt am vorstehenden Gegenbeispiel. In vielen Fällen wird es nämlich so sein, dass Wahrscheinlichkeiten translationsinvariant sein sollen, wenn  $\Omega$  eine Teilmenge von  $\mathbb{R}$  ist: Zum Beispiel soll bei der Gleichverteilung auf  $[0, 1]$  die Wahrscheinlichkeit von einem Teilintervall  $[a, b]$  gleich  $b - a$  sein, also nur von der Länge,

nicht aber von der Position abhängen. Anders ausgedrückt heißt das, dass wir auf das Borel-Lebesguemaß  $\lambda$  nicht verzichten können, das nur auf den Borelmengen definiert ist. Und der vorstehende Satz zeigt, dass es Teilmengen von  $\mathbb{R}$  gibt, die nicht zu diesem Mengensystem gehören.

Man könnte ja hoffen, dass man die Definition von  $\lambda$  auf alle Teilmengen von  $\mathbb{R}$  ausdehnen kann. Im Beweis von Satz 1.7.2 haben wir aber nur ausgenutzt, dass wir es mit einem translationsinvarianten Maß zu tun haben, das auf den Intervallen die richtigen Ergebnisse liefert: In solchen Fällen kann dieses Maß für unsere Menge  $B_0$  nicht definiert sein. Das gilt insbesondere für das *Lebesguemaß*, da nimmt man zu den Borelmengen noch alle Nullmengen hinzu (das sind Teilmengen von Borelmengen  $B$  mit  $\lambda(B) = 0$ ), betrachtet die erzeugte  $\sigma$ -Algebra (die *Lebesgumengen*) und setzt die Definition von  $\lambda$  auf diese größere  $\sigma$ -Algebra fort, indem man allen Nullmengen das Maß Null zuordnet.

Anders ausgedrückt: Es wird nie ein Maß geben, das translationsinvariant ist, alle Intervalle richtig misst und auf allen Teilmengen von  $\mathbb{R}$  definiert ist.

*Und wo bleibt das Positive?* Positiv ist, dass es einen Ansatz gibt, mit dem man sehr gut arbeiten kann: Den werden Sie in diesem Buch kennen lernen. Da alle in den Anwendungen vorkommenden Mengen Borelmengen sind, ist die Einschränkung, dass nicht immer alle Teilmengen von  $\Omega$  Ereignisse sind, praktisch ohne Bedeutung. Und mit erzeugten  $\sigma$ -Algebren, Borelmengen, Dynkinsystemen usw. kann man alles, was wir brauchen werden, so exakt entwickeln, wie es sich für eine mathematische Theorie gehört.

#### Die „Konstruktion“ der erzeugten $\sigma$ -Algebra

Es wurde schon auf Seite 15 erwähnt, dass die Definition „kleinste  $\sigma$ -Algebra, die  $\mathcal{M}$  enthält“, analog zu Definitionen in anderen Bereichen ist: lineare Hülle, konvexe Hülle usw.

In vielen Fällen ist es möglich, die Elemente dieser „Hülle“ explizit anzugeben. Ist zum Beispiel  $M$  Teilmenge eines  $K$ -Vektorraums, so ist die lineare Hülle von  $M$  die Menge aller Linearkombinationen aus Elementen aus  $M$ , d.h. die Menge der  $\sum_{i=1}^n a_i x_i$  mit  $n \in \mathbb{N}$ ,  $a_i \in K$ ,  $x_i \in M$ . Ähnliche konkrete Beschreibungen gibt es bei konvexen Hüllen, erzeugten Untergruppen, dem Abschluss einer Menge usw.

Warum sollte es bei der erzeugten  $\sigma$ -Algebra anders sein? Auf den ersten Blick sieht es doch ganz einfach aus: Starte mit  $\mathcal{M}$  und nimm  $\emptyset$ ,  $\Omega$  sowie alle Komplemente und alle abzählbaren Vereinigungen von Elementen aus  $\mathcal{M}$  dazu. Das ist im Allgemeinen noch keine  $\sigma$ -Algebra, denn die abzählbaren Vereinigungen der Komplemente und die Komplemente der Vereinigungen sind noch nicht berücksichtigt.

Daher versuchen wir es noch einmal, diesmal etwas sorgfältiger: Ist  $\mathcal{M}$  ein Mengensystem auf  $\Omega$ , so bezeichnen wir mit  $H(\mathcal{M})$  das System  $\mathcal{M}$  zusammen mit  $\emptyset$ ,  $\Omega$  und allen Komplementen und allen abzählbaren Vereinigungen von Elementen aus  $\mathcal{M}$ . Und das iterieren wir: Setze  $H_1(\mathcal{M}) := H(\mathcal{M})$ ,  $H_2(\mathcal{M}) := H(H_1(\mathcal{M}))$ , allgemein  $H_{n+1}(\mathcal{M}) := H(H_n(\mathcal{M}))$ .

Es ist sicher  $\mathcal{M} \subset H_1(\mathcal{M}) \subset H_2(\mathcal{M}) \subset \dots$ , und  $H_\infty := \bigcup_n H_n(\mathcal{M})$  sollte doch eine  $\sigma$ -Algebra sein. Das muss aber leider nicht stimmen, denn wählt man

$M_n \in H_n(\mathcal{M})$ , so kann man nicht garantieren, dass  $\bigcup_n M_n$  zu  $H_\infty(\mathcal{M})$  gehört. Wir müssen also weiter machen:  $H(H_\infty(\mathcal{M}))$  bilden usw.

Unter Verwendung der Ordinalzahltheorie kommt man so wirklich ans Ziel. Wenn Sie die Grundbegriffe schon kennen, lässt sich dieser Weg so beschreiben: Definiere  $H_\alpha(\mathcal{M})$  für alle Ordinalzahlen  $\alpha$  durch transfinite Rekursion. Wir setzen  $H_1(\mathcal{M}) := H(\mathcal{M})$ , und für  $\alpha > 1$  ist

$$H_\alpha(\mathcal{M}) := H\left(\bigcup_{\beta < \alpha} H_\beta(\mathcal{M})\right).$$

Dann ist wirklich  $\sigma(\mathcal{M}) = H_\omega(\mathcal{M})$ , wobei  $\omega$  die kleinste überabzählbare Ordinalzahl bezeichnet.

Das ist doch recht verwickelt, für uns wird die Definition „ $\sigma(\mathcal{M})$ “ ist die kleinste  $\sigma$ -Algebra auf  $\Omega$ , die  $\mathcal{M}$  enthält“ genügen. Das reicht auch völlig aus, da wir – zum Beispiel – nie eine exakte Beschreibung der allgemeinsten Borelmenge benötigen. Es wird immer nur wichtig sein zu wissen, dass eine konkret gegebene Menge eine Borelmenge ist. Und alles Weitere wird mit den in den Abschnitten 1.4 bis 1.6 beschriebenen Techniken erledigt.

## 1.8 Verständnisfragen

Zum Ende jedes Kapitels in diesem Buch gibt es einen Abschnitt „Verständnisfragen“. Dort findet man eine Zusammenstellung von *Sachfragen* (was sollte man unbedingt *kennen*) und *Methodenfragen* (was sollte man *können*). Beide Aspekte des Wissens sind in der Mathematik unverzichtbar. Weiteres Material zum Einarbeiten in die unter „*Methodenfragen*“ genannten Verfahren finden Sie im folgenden Abschnitt.

### Zu Abschnitt 1.2

#### *Sachfragen*

**S1:** Fragen über Elementarereignisse kann man durch „und“, „oder“ und „nicht“ kombinieren bzw. modifizieren. Was entspricht diesen Operationen in der Sprache der Mengenlehre?

**S2:** Was ist eine  $\sigma$ -Algebra?

**S3:** Was ist die kleinste, was die größte  $\sigma$ -Algebra auf einer Menge?

**S4:** Was ist ein Wahrscheinlichkeitsmaß?

**S5:** Was ist ein Elementarereignis, was ist ein Ereignis?

**S6:** Was ist ein Wahrscheinlichkeitsraum?

**S7:** Was ist eine „disjunkte Folge“ von Teilmengen einer Menge?

#### *Methodenfragen*

**M1:** Nachweisen können, dass ein vorgelegtes System von Teilmengen eine  $\sigma$ -Algebra ist.

**M2:** Nachprüfen können, ob eine auf einer  $\sigma$ -Algebra definierte Funktion ein Wahrscheinlichkeitsmaß ist.

### Zu Abschnitt 1.3

#### *Sachfragen*

**S1:** Die Faustregel zum Umgang mit  $\sigma$ -Algebren lautet: Ist  $\mathcal{E}$  eine  $\sigma$ -Algebra und sind  $X$  viele Elemente aus  $\mathcal{E}$  gegeben, so gehört auch jede Menge zu  $\mathcal{E}$ , die aus diesen Elementen durch die üblichen Mengenoperationen (Schnitt, Vereinigung, Komplement) gebildet wird. Was ist für „ $X$ “ einzusetzen? Endlich? Abzählbar? Höchstens abzählbar?

**S2:** Was versteht man unter der „Subadditivität“ eines Wahrscheinlichkeitsmaßes?

**S3:** Was bedeutet die „Stetigkeit nach oben“ bzw. die „Stetigkeit nach unten“ für ein Wahrscheinlichkeitsmaß?

#### *Methodenfragen*

**M1:** Einfache Sachverhalte für  $\sigma$ -Algebren beweisen können. Zum Beispiel:

a) Der Durchschnitt einer Familie von  $\sigma$ -Algebren auf einer Menge ist wieder eine  $\sigma$ -Algebra.

b) Richtig oder falsch: Das System der Teilmengen von  $\{1, \dots, 100\}$ , deren Elementanzahl durch 20 teilbar ist, ist eine  $\sigma$ -Algebra.

**M2:** Einfache Sachverhalte für Wahrscheinlichkeitsmaße beweisen können. Beispiele:

a) Es seien  $a_1, \dots, a_n \in \mathbb{R}$ . Wir definieren  $\mathbb{P}(E) := \sum_{i \in E} a_i$  für  $E \subset \{1, \dots, n\}$ . Unter welchen Bedingungen an die  $a_i$  ist das ein Wahrscheinlichkeitsmaß auf  $\{1, \dots, n\}$ ?

b) Sind  $\mathbb{P}_1$  und  $\mathbb{P}_2$  Wahrscheinlichkeitsmaße auf  $(\Omega, \mathcal{E})$ , so wird auch durch  $E \mapsto (\mathbb{P}_1(E) + \mathbb{P}_2(E))/2$  ein Wahrscheinlichkeitsmaß definiert.

**M3:** Den Satz von der Stetigkeit von Wahrscheinlichkeitsmaßen anwenden können.

Beispiel: Sei  $\mathbb{P}$  ein Wahrscheinlichkeitsmaß auf  $\mathbb{N}$ . Für jedes  $\varepsilon > 0$  gibt es dann eine Primzahl  $p$ , so dass  $\mathbb{P}(\{p, p+1, \dots\}) \leq \varepsilon$ .

### Zu Abschnitt 1.4

#### *Sachfragen*

**S1:** Was ist die von einem Mengensystem erzeugte  $\sigma$ -Algebra?

**S2:** Wie kann man sie „explizit“ definieren?

#### *Methodenfragen*

**M1:** In einfachen Fällen die erzeugte  $\sigma$ -Algebra bestimmen können. Zum Beispiel: Welche Teilmengen von  $\{1, \dots, 6\}$  gehören zu der von  $\{\{1, 2, 3, 4\}, \{4, 5, 6\}\}$  erzeugten  $\sigma$ -Algebra?

**M2:** Typische Schlussweisen im Zusammenhang mit erzeugten  $\sigma$ -Algebren beherrschen. Zum Beispiel: Ist  $\mathcal{M}$  ein Mengensystem auf  $\mathbb{R}$  mit der Eigenschaft,

dass für jedes  $n \in \mathbb{N}$  und jedes  $E \in \mathcal{M}$  die Menge  $n + E := \{n + y \mid y \in E\}$  auch zu  $\mathcal{M}$  gehört, so ist auch  $n + E \in \sigma(\mathcal{M})$  für jedes  $E \in \sigma(\mathcal{M})$  und alle  $n$ .

### Zu Abschnitt 1.5

#### *Sachfragen*

**S1:** Was ist eine Borelmenge im  $\mathbb{R}^n$ ?

**S2:** Das System der Borelmengen besteht nach Definition aus den Mengen, die in einer ganz bestimmten erzeugten  $\sigma$ -Algebra liegen. Welche der folgenden Mengensysteme erzeugen ebenfalls die Borelmengen von  $\mathbb{R}$ ?

- a) Die Intervalle der Form  $[a, \infty[$ .
- b) Die dreielementigen Teilmengen von  $\mathbb{R}$ .

**S3:** Ist jede Teilmenge von  $\mathbb{R}$  eine Borelmenge?

#### *Methodenfragen*

**M1:** Nachprüfen können, ob eine konkret gegebene Menge Borelmenge ist. Warum zum Beispiel ist  $\mathbb{Q} \setminus [3, 12[$  eine Borelmenge?

**M2:** Allgemeine Ergebnisse über Borelmengen beweisen können. Beispiele:

- a) Man zeige, dass  $B \times \mathbb{Q}$  für jede Borelmenge  $B$  in  $\mathbb{R}$  eine Borelmenge im  $\mathbb{R}^2$  ist.
- b) Ist  $C$  keine Borelmenge, so ist auch  $3 + C (= \{3 + x \mid x \in C\})$  keine Borelmenge.

**M3:** Nachprüfen können, ob ein Mengensystem alle Borelmengen erzeugt. Warum zum Beispiel werden die Borelmengen des  $\mathbb{R}^2$  vom System der offenen Kreisscheiben erzeugt?

### Zu Abschnitt 1.6

#### *Sachfragen*

**S1:** Was ist ein Dynkinsystem?

**S2:** Wie ist das von einem Mengensystem erzeugte Dynkinsystem definiert?

**S3:** Der wichtigste Satz lautet: Ist  $\mathcal{M}$  ein Mengensystem mit der Eigenschaft X, so stimmt das von  $\mathcal{M}$  erzeugte Dynkinsystem mit der von  $\mathcal{M}$  erzeugten  $\sigma$ -Algebra überein. Was ist die Eigenschaft X?

#### *Methodenfragen*

**M1:** Nachprüfen können, ob ein vorgelegtes Mengensystem ein Dynkinsystem ist. Beispiel: Ist das Mengensystem der offenen Teilintervalle von  $\mathbb{R}$  ein Dynkin-System?

**M2:** Die Dynkintechnik beherrschen. Beispiel: Ist  $\mathcal{D}$  ein Dynkinsystem, das alle Intervalle  $] -\infty, a ]$  ( $a \in \mathbb{R}$ ) enthält, so gehören auch alle Borelmengen zu  $\mathcal{D}$ .

## 1.9 Übungsaufgaben

### Zu Abschnitt 1.2

**Ü1.2.1** Sei  $\Omega = \{1, 2, 3, 4\}$ . Geben Sie drei verschiedene  $\sigma$ -Algebren auf  $\Omega$  an.

**Ü1.2.2** Sei  $\Omega$  beliebig und  $\mathcal{E}$  eine  $\sigma$ -Algebra auf  $\Omega$ . Für eine Teilmenge  $C$  von  $\Omega$  definieren wir die *Spur* von  $\mathcal{E}$  in  $C$  durch  $\mathcal{E}_C := \{E \cap C \mid E \in \mathcal{E}\}$ . Zeigen Sie, daß  $\mathcal{E}_C$  eine  $\sigma$ -Algebra ist und dass im Fall  $C \in \mathcal{E}$  gilt:

$$\mathcal{E}_C = \{E \in \mathcal{E} \mid E \subset C\}.$$

**Ü1.2.3** Es sei  $\Omega$  eine Menge, und  $\mathcal{E}$  wird als die Potenzmenge von  $\Omega$  definiert. Wir fixieren ein  $\omega_0 \in \Omega$  und definieren dann  $\mathbb{P} : \mathcal{E} \rightarrow [0, 1]$  wie folgt: Es ist  $\mathbb{P}(E) := 1$ , wenn  $\omega_0$  zu  $E$  gehört, für alle anderen  $E$  ist  $\mathbb{P}(E) := 0$ .

Zeigen Sie, dass  $\mathbb{P}$  ein Wahrscheinlichkeitsmaß ist. (Es wird das zu  $\omega_0$  gehörige *Punktmaß* genannt.)

**Ü1.2.4** Es gibt keine abzählbar unendliche  $\sigma$ -Algebra.

Tipp: Wenn  $\mathcal{E}$  eine unendliche  $\sigma$ -Algebra ist, kann man eine Folge  $E_1, E_2, \dots$  disjunkter nichleerer Elemente in  $\mathcal{E}$  finden. Es folgt, dass  $\mathcal{E}$  mindestens so viele Elemente enthalten muss, wie es Teilmengen von  $\mathbb{N}$  gibt.

**Ü1.2.5** Sei  $\Omega$  eine überabzählbare Menge. Geben Sie die kleinste  $\sigma$ -Algebra an, die alle eingleitigen Teilmengen von  $\Omega$  enthält. Zeigen Sie, dass auf dieser  $\sigma$ -Algebra durch

$$\mathbb{P}(E) = \begin{cases} 0 & \text{falls } E \text{ ist abzählbar,} \\ 1 & \text{falls } E \text{ ist überabzählbar} \end{cases}$$

ein Wahrscheinlichkeitsmaß definiert wird.

**Ü1.2.6** Sei  $\Omega$  eine Menge und  $\mathcal{E}$  eine  $\sigma$ -Algebra auf  $\Omega$ . Ein Ereignis  $A \in \mathcal{E}$  heißt *Atom*, falls es kein  $B \in \mathcal{E}$  gibt mit  $B \subset A$  und  $B \neq A$ . Zeigen Sie:

a) Zwei verschiedene Atome sind disjunkt.

b) Ist  $\Omega$  höchstens abzählbar, so existiert zu jedem  $\omega \in \Omega$  genau ein Atom  $A(\omega)$  mit  $\omega \in A(\omega)$ .  $\mathcal{E}$  ist die Menge aller Vereinigungen seiner Atome.

**Ü1.2.7** Sei  $\Omega = \{-100.000, \dots, 100.000\}$ . Welches der folgenden Mengensysteme ist eine  $\sigma$ -Algebra auf  $\Omega$ ?

a) Alle Teilmengen, für die die Summe der Elemente Null ergibt (die leere Summe ist als Null definiert). Zum Beispiel gehört  $\{-2, -1, 3\}$  zu diesem Mengensystem,  $\{-1, 3\}$  aber nicht.

b) Sei  $A := \{1, 2, \dots, 12\}$ . Man betrachte das System aller Teilmengen von  $\Omega$ , für die der Schnitt mit  $A$  eine gerade Anzahl von Elementen hat. (Insbesondere gehören alle zu  $A$  disjunkten Mengen dazu, denn 0 ist gerade.)

c) Das System, das aus der leeren Menge und allen Teilmengen der Form  $\{-100.000, \dots, 0\} \cup E$  mit  $E \subset \{1, \dots, 100.000\}$  besteht.

**Ü1.2.8**  $\Omega$  sei eine  $k$ -elementige Menge. Für welche Zahlen  $n$  gibt es eine  $\sigma$ -Algebra auf  $\Omega$  mit  $n$  Elementen?

**Ü1.2.9** Es seien  $\mathcal{E}_1$  und  $\mathcal{E}_2$   $\sigma$ -Algebren auf  $\Omega$ . Beweisen Sie: Wenn  $\mathcal{E}_1 \cup \mathcal{E}_2$  eine  $\sigma$ -Algebra ist, so gilt  $\mathcal{E}_1 \subset \mathcal{E}_2$  oder  $\mathcal{E}_2 \subset \mathcal{E}_1$ .

**Ü1.2.10** Es sei  $\Omega$  eine Menge, und  $\mathcal{E}$  sowie  $\mathcal{E}_1, \mathcal{E}_2, \dots$  seien  $\sigma$ -Algebren auf  $\Omega$ . Es gelte  $\mathcal{E}_1 \subset \mathcal{E}_2 \subset \dots \subset \mathcal{E}$ .

Dann muss  $\bigcup_{n=1}^{\infty} \mathcal{E}_n$  keine  $\sigma$ -Algebra sein.

Die Situation ist sogar viel dramatischer:  $\bigcup_{n=1}^{\infty} \mathcal{E}_n$  ist *nie* eine  $\sigma$ -Algebra. (Abgesehen von der trivialen Situation, dass die  $\mathcal{E}_n$  von einer Stelle ab alle übereinstimmen.)

Einen Beweis dieser Aussage findet man im Anhang auf Seite 354.

**Ü1.2.11** Eine  $\sigma$ -Algebra  $\mathcal{E}$  ist mit den Verknüpfungen  $\Delta$  und  $\cap$  ein kommutativer Ring. (Dabei steht „ $E \Delta F$ “ für die symmetrische Differenz:  $E \Delta F := (E \setminus F) \cup (F \setminus E)$ .)

Man kann also wirklich eine algebraische Struktur in einer  $\sigma$ -Algebra finden.

### Zu Abschnitt 1.3

**Ü1.3.1**  $\mathbb{P}_1$  und  $\mathbb{P}_2$  seien Wahrscheinlichkeitsmaße auf  $(\Omega, \mathcal{E})$ . Wir definieren  $\mathbb{P} : \mathcal{P}(\mathbb{N}) \rightarrow \mathbb{R}$  durch

$$\mathbb{P}(E) := \min\{\mathbb{P}_1(E), \mathbb{P}_2(E)\}.$$

Zeigen Sie, dass  $\mathbb{P}$  genau dann ein Wahrscheinlichkeitsmaß ist, wenn  $\mathbb{P}_1 = \mathbb{P}_2$  gilt.

**Ü1.3.2**  $\mathbb{P}_1$  und  $\mathbb{P}_2$  seien Wahrscheinlichkeitsmaße auf  $(\Omega, \mathcal{E})$ . Gilt dann  $\mathbb{P}_1(E) \leq \mathbb{P}_2(E)$  für alle  $E \in \mathcal{E}$ , so ist  $\mathbb{P}_1 = \mathbb{P}_2$ .

**Ü1.3.3** Es sei  $(\Omega, \mathcal{E}, \mathbb{P})$  ein Wahrscheinlichkeitsraum. Wir nennen eine Teilmenge  $N$  von  $\Omega$  eine *Nullmenge*, wenn es ein  $E \in \mathcal{E}$  mit  $N \subset E$  und  $\mathbb{P}(E) = 0$  gibt. Beweisen Sie:

- a) Abzählbare Vereinigungen von Nullmengen sind wieder eine Nullmenge.
- b) Jede Teilmenge einer Nullmenge ist eine Nullmenge.
- c) Definiere  $\mathcal{E}' := \{E \cup N \mid E \in \mathcal{E}, N \text{ ist Nullmenge}\}$ . Dann ist  $\mathcal{E}'$  eine  $\sigma$ -Algebra, die  $\mathcal{E}$  enthält.
- d) Eine Funktion  $\mathbb{P}' := \mathcal{E}' \rightarrow [0, 1]$  sei durch  $\mathbb{P}'(E \cup N) := \mathbb{P}(E)$  erklärt. Diese Abbildung ist wohldefiniert, es ist ein Wahrscheinlichkeitsmaß, und für  $E \in \mathcal{E}$  ist  $\mathbb{P}'(E) = \mathbb{P}(E)$ . (Man spricht von der *Vervollständigung* von  $(\Omega, \mathcal{E}, \mathbb{P})$ .)

**Ü1.3.4** Sei  $\Omega = \mathbb{N}$ , versehen mit irgendeiner  $\sigma$ -Algebra  $\mathcal{E}$ . Zeigen Sie, dass die Menge  $\mathcal{P}$  der Wahrscheinlichkeitsmaße auf  $(\Omega, \mathcal{E})$  eine konvexe Menge ist. Wie sehen die Extrempunkte von  $\mathcal{P}$  aus?

(Ein Element  $x$  einer konvexen Menge  $K$  heißt extremal, wenn es sich nicht als echte Konvexitätskombination zweier verschiedener Punkte  $x_1, x_2 \in K, x_1, x_2 \neq x$  darstellen lässt, wenn also für  $\lambda \in ]0, 1[$  und  $x = \lambda x_1 + (1 - \lambda)x_2$  stets folgt, dass  $x = x_1 = x_2$  gilt. So sind zum Beispiel die Extrempunkte eines Quadrats gerade die Ecken.)

### Zu Abschnitt 1.4

**Ü1.4.1** Es sei  $\mathcal{E}$  eine  $\sigma$ -Algebra auf  $\mathbb{R}^2$ , die alle offenen Kreisscheiben enthält. Dann enthält sie auch alle offenen Rechtecke.

**Ü1.4.2** Sei  $\Omega := \{0, 1\}^{\mathbb{N}}$ , also die Menge aller Folgen mit Werten in  $\{0, 1\}$ . Darauf sei eine  $\sigma$ -Algebra  $\mathcal{E}$  gegeben, die alle Mengen der Form

$$\{(x_n) \mid x_{n_0} = 0\}, n_0 = 1, 2, \dots$$

enthält.

Zeigen Sie, dass dann auch die folgenden Mengen in  $\mathcal{E}$  enthalten sind:

- alle einpunktigen Teilmengen;
- die Teilmenge der konvergenten Folgen in  $\Omega$ ;
- die Teilmenge derjenigen Folgen, die an genau 122 Stellen den Wert 0 annehmen.

**Ü1.4.3** Vorgelegt sei eine  $n$ -elementige Menge, und es sei  $k \leq n$  eine natürliche Zahl. Betrachten Sie das System derjenigen Teilmengen, die genau  $k$  Elemente enthalten und berechnen Sie die erzeugte  $\sigma$ -Algebra.

### Zu Abschnitt 1.5

**Ü1.5.1** Beweisen Sie, dass die folgenden Teilmengen von  $\mathbb{R}$  Borelmengen sind:

- Die Menge  $\mathcal{A}$  der algebraischen Zahlen. (Eine Zahl heißt algebraisch, wenn sie Nullstelle eines Polynoms mit rationalen Koeffizienten ist.)
- Die Menge  $\{x \mid 0 < f(x) \leq (f(x))^2\}$ , wobei  $f : \mathbb{R} \rightarrow \mathbb{R}$  eine stetige Funktion ist.

**Ü1.5.2** Man zeige: Ist  $B \subset \mathbb{R}$  keine Borelmenge, so ist auch  $a \cdot B := \{ax \mid x \in B\}$  keine Borelmenge für  $a \neq 0$ .

**Ü1.5.3** Sei  $M$  eine Menge, wir verstehen sie mit der diskreten Metrik  $d$ . (Es ist also  $d(x, y) := 1$  für  $x \neq y$  und  $d(x, y) := 0$  sonst.) Bestimmen Sie die Borelmengen in diesem metrischen Raum.

**Ü1.5.4** Es seien  $f, g : \mathbb{R}^n \rightarrow \mathbb{R}$  stetige Funktionen. Zeigen Sie, dass

$$\{(x_1, \dots, x_{n+1}) \mid f(x_1, \dots, x_n) \leq x_{n+1} < g(x_1, \dots, x_n)\}$$

eine Borelmenge im  $\mathbb{R}^{n+1}$  ist.

### Zu Abschnitt 1.6

**Ü1.6.1** Sei  $\Omega$  eine  $n$ -elementige Menge, und es sei  $k$  ein Teiler von  $n$ . Man zeige: Das Mengensystem aller  $E \subset \Omega$ , für die die Anzahl der Elemente ein Vielfaches von  $k$  ist, ist ein Dynkinsystem auf  $\Omega$ . Handelt es sich auch um eine  $\sigma$ -Algebra?

**Ü1.6.2** Es seien  $\mathbb{P}_1, \mathbb{P}_2, \mathbb{P}_3$  Wahrscheinlichkeitsmaße auf  $(\Omega, \mathcal{E})$ . Zeigen Sie, dass

$$\{E \in \mathcal{E} \mid \mathbb{P}_1(E) = \mathbb{P}_2(E) = \mathbb{P}_3(E)\}$$

ein Dynkinsystem ist.

**Ü1.6.3** Finden Sie eine endliche Menge  $\Omega$  und ein Mengensystem  $\mathcal{M}$ , so dass gilt:

- Die von  $\mathcal{M}$  erzeugte  $\sigma$ -Algebra ist die Potenzmenge von  $\Omega$ .

b) Es gibt zwei verschiedene Wahrscheinlichkeitsmaße  $\mathbb{P}_1, \mathbb{P}_2$  auf  $(\Omega, \mathcal{P}(\Omega))$ , die auf  $\mathcal{M}$  übereinstimmen. (Das zeigt, dass die Forderung der Durchschnittsstabilität in Satz 1.6.5 wesentlich ist.)

**Ü1.6.4** Welches der Mengensysteme aus Übungsaufgabe Ü1.2.7 ist ein Dynkin-system?

# Kapitel 2

## Erste Beispiele für Wahrscheinlichkeitsräume

In diesem Kapitel geht es um das erste Kennenlernen einiger wichtiger Wahrscheinlichkeitsräume. Sie werden im Folgenden eine wichtige Rolle bei der Illustration allgemeiner Ergebnisse und neuer Konzepte spielen, und für manche wird es später ein eigenes Kapitel geben, in dem ihre Bedeutung für ein spezielles wahrscheinlichkeitstheoretisches Problem dargestellt wird.

Es wird sich um besonders leicht zugängliche Beispiele zur Modellierung des Zufalls handeln. In Abschnitt 2.1 geht es um *diskrete Räume*, da ist die Welt besonders einfach: Alle Teilmengen von  $\Omega$  sind Ereignisse, und das Wahrscheinlichkeitsmaß ist übersichtlich zu beschreiben. Die zweite große Beispielklasse sind die *durch Dichtefunktionen definierten Räume*, die wir in Abschnitt 2.2 einführen. Da ist  $\Omega$  eine „einfache“ Menge, z.B. ein Intervall in  $\mathbb{R}$ , und Wahrscheinlichkeiten werden durch Integrale über eine feste Funktion, die Dichtefunktion, definiert.

Dass man die zugehörigen „Zufallsautomaten“ schnell und unproblematisch auf jedem Computer simulieren kann, wird dann in den Abschnitten 2.3 und 2.4 gezeigt. Am Ende des Kapitels finden Sie dann Ergänzungen (Abschnitt 2.5), Verständnisfragen (Abschnitt 2.6) und Übungsaufgaben (Abschnitt 2.7).

### 2.1 Diskrete Wahrscheinlichkeitsräume

Das Wort „diskret“ verwendet man in der Mathematik immer dann, wenn die gerade interessierenden Objekte als voneinander isoliert betrachtet werden können. „Kontinuierlich“ dagegen ist für Situationen reserviert, bei denen die Objekte – wie bei einer durchgezogenen Linie – sozusagen ineinander übergehen<sup>1)</sup>.

<sup>1)</sup>Das ist als Versuch zu verstehen, die Begriffe zu unterscheiden. Er erhebt nicht den Anspruch einer präzisen Definition.

In der Wahrscheinlichkeitstheorie spricht man dann von „*diskreten Räumen*“ ( $\Omega, \mathcal{E}, \mathbb{P}$ ), wenn  $\Omega$  höchstens abzählbar (also endlich oder abzählbar) ist. In diesem Fall wird auch immer angenommen, dass  $\mathcal{E}$  gleich der Potenzmenge von  $\Omega$  ist: *Jede Teilmenge von  $\Omega$  ist Ereignis.*

Wie kann man denn in solchen Fällen ein Wahrscheinlichkeitsmaß definieren? Es muss doch eine Abbildung  $\mathbb{P} : \mathcal{P}(\Omega) \rightarrow [0, 1]$  sein. Nun kann  $\mathcal{P}(\Omega)$  schon für endliche und „nicht zu große“  $\Omega$  eine riesige Menge sein: Wenn  $\Omega$   $n$  Elemente hat, hat die Potenzmenge  $2^n$  Elemente. (Für  $n = 50$  ist schon  $2^n = 1.125.899.906.842.624$ , das ist etwas mehr als eine Billiarde!)

Eine Überlegung hilft weiter. Wenn  $\mathbb{P}$  ein Wahrscheinlichkeitsmaß auf der höchstens abzählbaren Menge  $\Omega$  ist und *wenn* ich die Zahlen  $\mathbb{P}(\{\omega\})$ , also die Wahrscheinlichkeiten der einelementigen Ereignisse, schon kenne, dann kenne ich wegen der  $\sigma$ -Additivität schon alle Wahrscheinlichkeiten. Ist nämlich  $E \subset \Omega$  eine beliebige Teilmenge, so kann man doch  $E$  als disjunkte Vereinigung der Mengen  $\{\omega\}$  schreiben, wobei  $\omega$  alle Elemente aus  $E$  durchläuft. (Zum Beispiel wäre  $\{1, 4, 5\} = \{1\} \cup \{4\} \cup \{5\}$ , eine offensichtlich disjunkte Vereinigung.) Wegen der  $\sigma$ -Additivität von  $\mathbb{P}$  muss deswegen  $\mathbb{P}(E)$  die Summe der  $\mathbb{P}(\{\omega\})$  sein, wobei alle  $\omega \in E$  zu berücksichtigen sind. In Kurzfassung:  $\mathbb{P}(E) = \sum_{\omega \in E} \mathbb{P}(\{\omega\})$ .

Die Moral: Man muss gar nicht alle  $\mathbb{P}(E)$  kennen, es reicht zu wissen, wie groß die  $\mathbb{P}(\{\omega\})$  für alle  $\omega \in \Omega$  sind. Das ist eine gewaltige Einsparung, denn im Fall  $n$ -elementiger  $\Omega$  brauchen wir nur noch  $n$  Informationen und nicht  $2^n$ . (Also z.B. nur 50 statt 1.125.899.906.842.624.)

Diese Beobachtung lässt sich zu einem Definitionsverfahren erweitern. In der Formulierung wird die „ungeordnete Summation“ eine Rolle spielen, also Ausdrücke der Form  $\sum_{\omega \in E} p_\omega$ . Die naive Interpretation sollte klar sein, etwas genauer ist der Begriff im Anhang auf Seite 360 erläutert.

**Satz 2.1.1.**  $\Omega$  sei eine höchstens abzählbare Menge. Wir versehen  $\Omega$  mit der Potenzmenge als  $\sigma$ -Algebra der Ereignisse:  $\mathcal{E} := \mathcal{P}(\Omega)$ .

- (i) Ein Wahrscheinlichkeitsmaß  $\mathbb{P} : \mathcal{E} \rightarrow [0, 1]$  ist durch die Werte auf den einelementigen Mengen  $\{\omega\}$  eindeutig bestimmt. Es gilt  $\sum_{\omega \in \Omega} \mathbb{P}(\{\omega\}) = 1$ .
- (ii) Umgekehrt: Für  $\omega \in \Omega$  seien Zahlen  $p_\omega \in [0, 1]$  vorgegeben. Wir setzen voraus, dass  $\sum_{\omega \in \Omega} p_\omega = 1$  gilt. Dann wird durch  $\mathbb{P}(E) := \sum_{\omega \in E} p_\omega$  (für  $E \subset \Omega$ ) ein Wahrscheinlichkeitsmaß definiert.

**Beweis:** Die Begründung für (i) wurde vor dem Satz schon gegeben. (Dass die Summe der  $\mathbb{P}(\{\omega\})$  gleich Eins ist, folgt aus  $\mathbb{P}(\Omega) = 1$ .)

Für den Beweis von (ii) ist zunächst nachzuprüfen, ob  $\mathbb{P}(\Omega) = 1$  gilt: Das ist wegen der Voraussetzung  $\sum_{\omega} p_\omega = 1$  klar. Dass sich die Wahrscheinlichkeiten für (abzählbar viele) disjunkte Ereignisse addieren, ist eine Folgerung aus der Kommutativität und der Assoziativität der Addition reeller Zahlen. Ist zum Beispiel  $\Omega = \{1, 2, 3, 4, 5, 6, 7\}$ ,  $E_1 = \{2, 3\}$  und  $E_2 = \{1, 6, 7\}$ , so ist die Gleichung  $\mathbb{P}(E_1 \cup E_2) = \mathbb{P}(E_1) + \mathbb{P}(E_2)$  gerade die Aussage  $p_1 + p_2 + p_3 + p_6 + p_7 = (p_2 + p_3) + (p_1 + p_6 + p_7)$ . Falls es um abzählbar viele  $E_n$  geht und/oder manche

$E_n$  unendlich viele Elemente enthalten, muss man etwas sorgfältiger argumentieren. Es ist ein eher analytisches als wahrscheinlichkeitstheoretisches Problem, einiges dazu findet man im Anhang auf Seite 359.  $\square$

Nach diesen Vorbereitungen ist es leicht, Beispiele für Wahrscheinlichkeitsräume anzugeben, wenn  $\Omega$  höchstens abzählbar ist.

Der allgemeinste Wahrscheinlichkeitsraum auf  $\Omega = \{1, \dots, n\}$

Hier sind  $n$  Zahlen  $p_1, \dots, p_n$  vorgegeben, für die zwei Bedingungen erfüllt sind: Erstens müssen alle  $p_i$  in  $[0, 1]$  liegen, und zweitens muss  $\sum_{i=1}^n p_i = 1$  gelten. Als „Konstruktionsanleitung“ bedeutet das: Man nehme eine Strecke der Länge 1 und zerlege sie auf irgendeine Weise in  $n$  Teilstrecken.

Man kann das dadurch veranschaulichen, dass man die Elementarereignisse  $1, \dots, n$  durch Punkte auf der  $x$ -Achse markiert und darüber einen Balken der Länge  $p_i$  abträgt<sup>2)</sup>. Im folgenden Bild zum Beispiel ist  $n = 4$ , und die  $p_i$  wurden so gewählt:  $p_1 = p_2 = p_3 = 1/5, p_4 = 2/5$ .



Bild 2.1.1: Veranschaulichung eines Wahrscheinlichkeitsraums.

Und wenn die  $p_i$  bekannt sind, kann man alle Wahrscheinlichkeiten leicht ermitteln:  $\mathbb{P}(E)$  ist – für beliebige Teilmengen  $E$  von  $\{1, \dots, n\}$  – die Summe der  $p_i$ , wobei  $i$  alle Elemente aus  $E$  durchläuft. Im vorstehenden Beispiel etwa ist  $\mathbb{P}(\{2, 4\}) = p_2 + p_4 = 3/5$ .

Es stellen sich im Zusammenhang mit diesen Räumen *zwei naheliegende Fragen*:

*Frage 1: Ist der Fall  $n = 1$  zugelassen?* Antwort: Ja! Das ist ein eigentlich uninteressanter Grenzfall. Der zugehörige „Zufallsautomat“ würde mit Sicherheit bei jeder Abfrage die 1 ausgeben.

*Frage 2: Dürfen einige  $p_i$  gleich Null sein?* Antwort: Ja! Doch solche  $i$  sind für Wahrscheinlichkeitsabfragen praktisch unsichtbar und könnten gleich weggelassen werden. Zum Beispiel sollte man von  $\Omega = \{1, 2, 3, 4\}$  und  $p_1 = p_2 = 0, p_3 = 0.2, p_4 = 0.8$  zu  $\Omega = \{1, 2\}$  und  $p_1 = 0.2, p_2 = 0.8$  übergehen.

Der allgemeinste Wahrscheinlichkeitsraum auf  $\mathbb{N}$  oder auf  $\mathbb{N}_0$

Das geht ganz ähnlich. Diesmal ist eine Folge  $(p_n)_{n=1,2,\dots}$  (bzw.  $(p_n)_{n=0,1,\dots}$ ) in  $[0, 1]$  vorzugeben, und die Reihensumme  $\sum_{n=1}^{\infty} p_n$  (bzw.  $\sum_{n=0}^{\infty} p_n$ ) muss

<sup>2)</sup>Der Längenmaßstab wird dabei von Fall zu Fall unterschiedlich gewählt. Andernfalls könnte man in manchen Fällen nur winzige Balken sehen.

gleich Eins sein. Auch diesmal ist eine Veranschaulichung durch eine „Balkenfolge“ möglich (Beispiele folgen gleich), und die  $i$  mit  $p_i = 0$  spielen keine Rolle.

### Laplaceräume

Das ist unser erster „konkreter“ Wahrscheinlichkeitsraum. Hier die Definition: Sei  $(\Omega, \mathcal{E}, \mathbb{P})$  ein Wahrscheinlichkeitsraum, wobei  $\Omega$  endlich ist. Wenn dann alle  $\mathbb{P}(\{\omega\})$  gleich sind, spricht man von einem *Laplaceräum*.

Wenn  $\Omega$   $n$  Elemente hat (z.B. im Fall  $\Omega = \{1, \dots, n\}$ ), muss – weil die Summe Eins ist – jedes  $\mathbb{P}(\{\omega\})$  gleich  $1/n$  sein.

Für Laplaceräume sind Wahrscheinlichkeiten von Teilmengen besonders leicht zu berechnen: Es ist doch

$$\mathbb{P}(E) = \#E \cdot \frac{1}{n} = \frac{\#E}{\#\Omega},$$

was man manchmal als „Wahrscheinlichkeit gleich günstige Fälle durch mögliche Fälle“ ausdrückt. (Zur Erinnerung:  $\#E$  ist das Symbol für die Anzahl der Elemente in  $E$ .) Doch *Achtung*: So einfach ist es wirklich nur in Laplaceräumen.

Die Tatsache, dass das Zählen – wie viele Elemente hat  $E$ ? – in der Wahrscheinlichkeitsrechnung eine wichtige Rolle spielt, ist auch der Grund dafür, dass wir später in Abschnitt 3.4 einige Ergebnisse aus der Kombinatorik herleiten werden.

Als Spezialfälle für Laplaceräume könnte man die faire Münze ( $n = 2$ ) und den fairen Würfel ( $n = 6$ ) erwähnen.

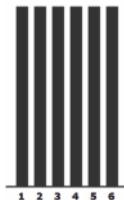


Bild 2.1.2: Der faire Würfel.

*Ein Nachtrag:* Haben Sie die Definition für Laplaceräume auf abzählbaren Mengen vermisst? Warum kann man z.B. nicht allen natürlichen Zahlen die gleiche Wahrscheinlichkeit zuordnen? Der Grund: So einen Wahrscheinlichkeitsraum gibt es nicht: Wäre nämlich  $a \in [0, 1]$  die Wahrscheinlichkeit aller  $\{n\}$ , so müsste  $a + a + a + \dots = 1$  gelten. Keine Zahl  $a$  erfüllt diese Bedingung.

### Der Bernoulliraum

Der Bernoulliraum ist der einfachste (nichttriviale) Wahrscheinlichkeitsraum: Es gibt genau zwei Elementarereignisse, üblicherweise wählt man  $\Omega = \{0, 1\}$ . Und dann ist alles durch eine einzige Zahl  $p = \mathbb{P}(\{1\})$  festgelegt: Wegen  $\mathbb{P}(\Omega) = 1$  muss nämlich  $\mathbb{P}(\{0\}) = 1 - p$  gelten.

In der üblichen Terminologie wird 1 als „Erfolg“ bezeichnet. Es gibt jedoch sehr unterschiedliche „Erfolge“:

- „Erfolg“ kann bedeuten, eine 6 zu würfeln (oder das Kreuz Ass aus einem Skatspiel zu ziehen oder 6 Richtige im Lotto zu haben). Dann ist  $p = 1/6$  bzw.  $p = 1/32$  bzw.  $1/13.983.816$ .
- In anderen Situationen könnte es ein „Erfolg“ sein, dass ein Medikament wirkt oder dass bei einem Crashtest der Airbag nicht funktioniert oder dass ein Unkrautvernichter wirklich den Rasen vom Löwenzahn befreit.

Zu Bernoulliräumen selbst ist nicht viel zu sagen. Damit zusammenhängende Fragen werden aber sehr ausführlich in Kapitel 5 untersucht werden.

### Die Poissonverteilung

Dieses Beispiel kommt irgendwie „aus heiterem Himmel“, es sieht auf den ersten Blick ziemlich beliebig aus. Hier die Definition:

Es ist  $\Omega = \mathbb{N}_0 = \{0, 1, \dots\}$ , und es ist eine Zahl  $\lambda \geq 0$  vorgegeben. Die *Poissonverteilung*<sup>3)</sup> zum Parameter  $\lambda$  ist dann durch

$$\mathbb{P}(\{n\}) := p(n; \lambda) := \frac{\lambda^n}{n!} e^{-\lambda}$$

für  $n \in \mathbb{N}_0$  erklärt. (Damit das in allen Fällen definiert ist, sollte man vorher  $0!$  als 1 und  $0^0$  ebenfalls als 1 erklärt haben.)

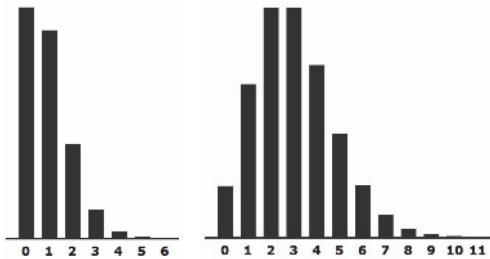


Bild 2.1.3: Poissonverteilungen zu  $\lambda = 0.9$  (links) und  $\lambda = 3$  (rechts).

Wird dadurch wirklich ein Wahrscheinlichkeitsmaß erklärt, ist wirklich die Reihensumme  $\sum_{n=0}^{\infty} p(n; \lambda)$  gleich Eins? An dieser Stelle sollte man sich an die Reihe für die Exponentialfunktion erinnern:  $e^x = 1 + x + x^2/2! + \dots = \sum_{n=0}^{\infty} x^n/n!$ . Aufgrund dieser Formel ist

$$\sum_{n \in \mathbb{N}_0} p(n; \lambda) = \left( \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} \right) e^{-\lambda} = e^{\lambda} e^{-\lambda} = 1.$$



Poisson

→  
Programm!

<sup>3)</sup>Siméon Poisson, 1781 bis 1840. Seine Jugend fiel in die Zeit der französischen Revolution. Er studierte bei Laplace und Lagrange, 1806 wurde er Professor an der École Polytechnique. Seine Interessen galten zunächst überwiegend der Analysis und der mathematischen Physik. Die (heute so genannte) Poissonverteilung wurde erstmals 1837 von ihm untersucht.

Die Bedeutung der Poissonverteilung zur Modellierung der „Überlagerung seltener Ereignisse“ wird erst in Abschnitt 5.3 klar werden. Da werden wir sehen, dass – zum Beispiel – die Anzahl der Anrufe in einem festen Zeitintervall mit diesem Wahrscheinlichkeitsmaß beschrieben werden kann.

### Die geometrische Verteilung

In diesem Beispiel ist  $\Omega = \mathbb{N}$ . Es ist ein Parameter  $q \in [0, 1[$  vorgegeben, und die Wahrscheinlichkeiten für die einpunktigen Teilmengen von  $\Omega$  werden so definiert:

$$\mathbb{P}(\{n\}) := q^{n-1}(1-q),$$

wobei wir im Fall  $q = 0$  und  $n = 1$  den Ausdruck  $0^0$  wieder als Eins definieren. Dann spricht man von der *geometrischen Verteilung zum Parameter  $q$* .

Um sich davon zu überzeugen, dass dadurch wirklich ein Wahrscheinlichkeitsmaß definiert wird, muss man nur die Formel für die geometrische Reihe anwenden:  $1 + q + q^2 + \dots = \sum_{n=0}^{\infty} q^n = 1/(1 - q)$ . Hier die Veranschaulichung der Wahrscheinlichkeiten für  $q = 0.8$ :

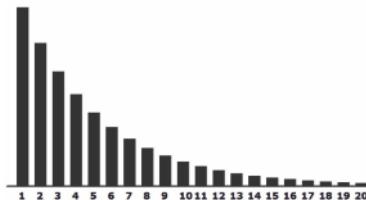


Bild 2.1.4: Die geometrische Verteilung für  $q = 0.8$ .

→  
Programm!

Die geometrische Verteilung spielt in der Wahrscheinlichkeitstheorie des Wartens eine wichtige Rolle: Wie lange muss man wohl zum Beispiel warten, um die erste Sechs zu würfeln? Präzisionen folgen in Kapitel 6.

## 2.2 Wahrscheinlichkeitsdichten

In diesem Abschnitt lernen wir eine weitere große Klasse von Wahrscheinlichkeitsräumen kennen: Räume, bei denen die Wahrscheinlichkeiten durch eine *Dichtefunktion* definiert sind. Allen diesen Beispielen ist gemeinsam, dass  $\Omega$  eine „einfache“ Teilmenge von  $\mathbb{R}$  oder des  $\mathbb{R}^n$  ist. Insbesondere wird  $\Omega$  eine Borelmenge sein<sup>4)</sup>. Als  $\sigma$ -Algebra  $\mathcal{E}$  wird dann immer die  $\sigma$ -Algebra der Borelmengen betrachtet, die in  $\Omega$  liegen<sup>5)</sup>.

<sup>4)</sup>Falls Sie es noch nicht getan haben, wäre jetzt eine gute Gelegenheit, sich die Definition „Borelmenge“ in Abschnitt 1.5 anzusehen. Falls Sie lieber erst einmal weiterlesen wollen, ersetzen Sie überall das Wort „Borelmenge“ durch „Teilmenge von  $\mathbb{R}$  (bzw. des  $\mathbb{R}^n$ )“.

<sup>5)</sup>Es ist nicht schwer zu sehen, dass das wirklich eine  $\sigma$ -Algebra ist. Vgl. Übungsaufgabe Ü1.2.2.

Sei etwa  $\Omega$  ein Intervall  $[a, b]$ . Um für diese Situation ein Wahrscheinlichkeitsmaß zu definieren, müsste man doch festsetzen, was  $\mathbb{P}(E)$  für beliebige  $E \in \mathcal{E}$  bedeuten soll. Das sieht hoffnungslos aus, denn  $\mathcal{E}$  enthält unendlich viele (sogar überabzählbar viele) Elemente.

In vielen Fällen ist es aber so, dass die  $\mathbb{P}(E)$  durch eine Integration definiert werden können: Es ist  $\mathbb{P}(E) = \int_E f(x) dx$ . Das sieht ein bisschen furchterregend aus, denn was soll denn  $\int_E f(x) dx$  für beliebige Borelmengen bedeuten? In der Analysis hat man doch nur gelernt, was man unter  $\int_c^d f(x) dx$  versteht? Das wird aber im Folgenden für eine Wahrscheinlichkeitstheorie völlig ausreichen. Präzisiert wird das in

**Satz 2.2.1.** *Es sei  $f : [a, b] \rightarrow [0, +\infty]$  eine stetige Funktion. Wir setzen voraus, dass  $\int_a^b f(x) dx = 1$  gilt.*

*Dann gibt es ein eindeutig bestimmtes Wahrscheinlichkeitsmaß  $\mathbb{P}$ , das auf den in  $[a, b]$  enthaltenen Borelmengen definiert ist, für das gilt:*

$$\mathbb{P}([c, d]) = \int_c^d f(x) dx$$

*für alle in  $[a, b]$  enthaltenen Intervalle  $[c, d]$ . Dabei ist das Integral das aus der Analysis bekannte Riemann-Integral.*

*Die Funktion  $f$  heißt die zu diesem Wahrscheinlichkeitsraum gehörige Dichtefunktion (oder auch die Wahrscheinlichkeitsdichte).*

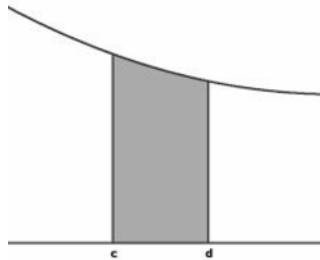


Bild 2.2.1: Die Wahrscheinlichkeit von  $[c, d]$  ist gleich der grau eingezeichneten Fläche.

**Beweis:** Zunächst zur Existenz. Dazu müssen wir aus der Maß- und Integrationstheorie einige Ergebnisse übernehmen<sup>6)</sup>:

- Man kann  $\int_E f(x) dx$  für alle Borelmengen in  $[a, b]$  sinnvoll definieren. Ist  $E = [c, d]$  ein Intervall, so stimmt dieses Integral mit dem Riemann-Integral überein.
- Sind  $E_1, E_2, \dots$  disjunkte Borelmengen, so ist

$$\int_{E_1 \cup E_2 \cup \dots} f(x) dx = \int_{E_1} f(x) dx + \int_{E_2} f(x) dx + \dots$$

<sup>6)</sup>Mehr dazu findet man im Anhang auf Seite 357.

Dann ist es klar, dass durch  $\mathbb{P}(E) := \int_E f(x) dx$  ein Wahrscheinlichkeitsmaß definiert ist.

Zum Nachweis der *Eindeutigkeit* ist nur an Satz 1.6.5 zu erinnern. Danach sind Wahrscheinlichkeitsmaße durch ihre Werte auf den Teilintervallen eindeutig bestimmt.  $\square$

Für alle, die schon wissen, wie Integrale in komplizierteren Situationen definiert sind, ist der Satz wesentlich verallgemeinerbar:

- $f$  muss nicht stetig sein. Es reicht die stückweise Stetigkeit. Es genügt sogar vorauszusetzen, dass  $f$  integrierbar und nichtnegativ ist und dass das Integral über  $\Omega$  Eins ist. Der Satz gilt nicht nur für  $\Omega = [a, b]$ , sondern auch für  $\Omega = \mathbb{R}$  und  $\Omega = [a, +\infty[$  (mit  $a \in \mathbb{R}$ ).
- Im  $\mathbb{R}^n$  ist die Situation ähnlich. Wenn  $\Omega \subset \mathbb{R}^n$  und eine Funktion  $f : \Omega \rightarrow [0, +\infty[$  mit  $\int_{\Omega} f(x) dx = 1$  vorgegeben sind, induziert das ein Wahrscheinlichkeitsmaß  $\mathbb{P}$  durch die Definition

$$\mathbb{P}(E) := \int_E f(x) dx$$

für Borelmengen  $E$  in  $\Omega$ . Auch hier ist  $\mathbb{P}$  durch die Werte auf den „einfachen“  $E$  eindeutig bestimmt. (Zum Beispiel durch die Werte auf den „Hyperquadern“, falls  $\Omega = \mathbb{R}^n$ . Die Definition finden Sie auf Seite 21, und die Aussage kann wieder dadurch begründet werden, dass die Hyperquader einen schnitt-stabilen Erzeuger der Borelmengen des  $\mathbb{R}^n$  bilden.)

- Und falls Sie schon das Integral für beliebige Maßräume beherrschen, sollte klar sein, dass durch jede nichtnegative integrierbare Funktion mit Integral Eins auf jedem beliebigen Maßraum ein Wahrscheinlichkeitsraum definiert wird.

Wir betrachten zum Kennenlernen dieses neuen Definitionsverfahrens für Wahrscheinlichkeitsräume ein einfaches Beispiel: Auf  $[0, 1]$  ist die Dichtefunktion  $f(x) := 2x$  definiert. Sie ist zulässig, denn es gilt  $f(x) \geq 0$  für alle  $x$ , und  $\int_0^1 f(x) dx = 1$ .

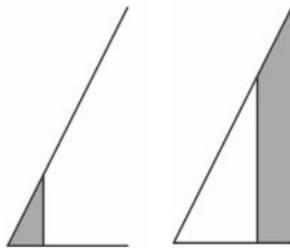


Bild 2.2.2: Die Wahrscheinlichkeiten für  $[0, 0.3]$  und  $[0.7, 1]$ .

Wahrscheinlichkeiten sind in diesem Fall sehr leicht zu berechnen: Es ist zum Beispiel  $\mathbb{P}([0, 0.3]) = \int_0^{0.3} 2x dx = 0.09$ , oder  $\mathbb{P}([0.7, 1]) = \int_{0.7}^1 2x dx = 0.51$  usw. Obwohl also die Intervalle gleich lang sind – beide haben die Länge 0.3

– sind die Wahrscheinlichkeiten sehr unterschiedlich. Das liegt natürlich daran, dass die Dichtefunktion auf  $[0.7, 1]$  größere Werte hat als auf  $[0, 0.3]$ <sup>7)</sup>.

Und hier noch *eine letzte Bemerkung zur allgemeinen Situation*. Wir hatten vorausgesetzt, dass  $f$  nichtnegativ ist und dass das Integral über  $\Omega$  gleich Eins ist. Oft kann leicht durch einen kleinen Trick erreicht werden, dass die zweite Bedingung erfüllt ist: Ist das Integral von  $f$  über  $\Omega$  endlich und positiv, ist etwa  $\int_{\Omega} f(x) dx = I$ , so ist offensichtlich die Funktion  $f/I$  eine Dichtefunktion. (Zur Illustration betrachten wir  $f(x) := x^4$  auf  $\Omega = [-1, 2]$ . Das Integral  $\int_{\Omega} f(x) dx$  ist gleich  $33/5$ , und deswegen ist  $5x^4/33$  eine Dichtefunktion auf  $\Omega$ .)

Es folgen einige *spezielle Beispiele* aus der unübersehbaren Vielfalt von Dichtefunktionen. Sie alle haben für die Wahrscheinlichkeitstheorie eine besondere Bedeutung, und deswegen werden wir den meisten auch später noch einmal begegnen.

### Die Gleichverteilung

Sei zunächst  $\Omega = [a, b]$  ein beschränktes Intervall. Die Gleichverteilung ist dadurch definiert, dass die Dichtefunktion gleich einer Konstante  $c$  ist. Da das Integral über  $\Omega$  gleich Eins sein soll, muss  $c = 1/(b - a)$  gelten.

Die Wahrscheinlichkeiten sind dann leicht zu berechnen:  $\mathbb{P}(E)$  ist gleich der „relativen Länge“ von  $E$  in  $\Omega$ , also gleich  $\lambda(E)/(b - a)$  ( $\lambda$  bezeichnet wieder das Borel-Lebesguemaß, vgl. Satz 1.7.1). Insbesondere für Intervalle  $[c, d] \subset \Omega$  gilt:

$$\mathbb{P}([c, d]) = \frac{d - c}{b - a}.$$

Die Gleichverteilung ist das kontinuierliche Analogon zu den Laplaceräumen: Kein Teilbereich von  $\Omega$  ist bevorzugt, alle sind gleichberechtigt. Der wichtigste Spezialfall betrifft das Einheitsintervall  $[0, 1]$ . Im Abschnitt 2.4 werden wir zeigen, wie man sich die Tatsache zunutze macht, dass der zugehörige Wahrscheinlichkeitsraum in jedem Computer simuliert werden kann.

### Bemerkungen:

1. Auch auf offenen und halboffenen beschränkten Intervallen kann man die Gleichverteilung durch eine konstante Dichtefunktion definieren. Wir werden uns aber auf abgeschlossene Intervalle beschränken, da Ausdrücke der Form  $\int_{[a,b]} f(x) dx$  in der Analysis nicht behandelt werden.
2. Auf unbeschränkten Intervallen gibt es keine Gleichverteilung<sup>8)</sup>. Der Grund: Ist  $\Omega$  unbeschränkt und  $c$  eine Konstante, so ist das Integral  $\int_{\Omega} c dx$  gleich  $c \cdot \infty$ , also entweder Null (im Fall  $c = 0$ ) oder Unendlich (im Fall  $c > 0$ ). Eins kommt jedenfalls nie heraus.

<sup>7)</sup>Das gilt offensichtlich allgemein: Vergleicht man die Wahrscheinlichkeiten von Teilintervallen gleicher Länge, so sind die besonders groß (bzw. besonders klein), falls die Dichtefunktion  $f$  auf diesen Teilintervallen besonders große (bzw. kleine) Werte annimmt.

<sup>8)</sup>Auf unendlichen diskreten Mengen geht das ja auch nicht, vgl. Seite 40.

**3.** Alles geht ganz analog im  $\mathbb{R}^n$ . Zur Illustration betrachten wir als  $\Omega$  ein Rechteck mit den Seitenlängen 3 und 5 im  $\mathbb{R}^2$ . Damit eine konstante Funktion beim Integrieren über  $\Omega$  den Wert Eins liefert, muss die Konstante gleich  $1/15$  sein. Das ist in diesem Fall die Dichtefunktion der Gleichverteilung. Ist  $E \subset \Omega$  eine Teilmenge, für die Sie die Fläche bestimmen können, so ist

$$\mathbb{P}(E) = (\text{Flächeninhalt von } E) / (\text{Flächeninhalt von } \Omega).$$

Ist zum Beispiel  $E$  ein Teilquadrat von  $\Omega$  mit Kantenlänge 2, so ist  $\mathbb{P}(E) = 4/15$ .

Als weiteres Beispiel betrachten wir eine kreisrunde Stadt, der Radius betrage 10 Kilometer. Wir nehmen an, dass sie gleichmäßig besiedelt ist: keine Wälder, keine Seen usw.

Genau im Zentrum befindet sich eine Diskothek, die in allen Stadtteilen gleichermaßen beliebt ist. Wie wahrscheinlich ist es, dass eine zufällig ausgewählte Person mehr als 5 Kilometer vom Zentrum entfernt wohnt?

Die Lösung: Bezeichnet man mit  $E$  die Fläche „Abstand zum Zentrum zwischen 5 und 10 Kilometer“ und betrachtet man die Gleichverteilung auf  $\Omega$  (eine Kreisscheibe mit 10 Kilometer Radius), so ist die gesuchte Wahrscheinlichkeit gleich „Flächeninhalt von  $E$  durch Flächeninhalt von  $\Omega$ “, also gleich  $(100 - 25)\pi/100\pi = 0.75$ .

Entsprechend wird man auf Volumenverhältnisse im  $\mathbb{R}^3$  oder allgemeiner auf das Verhältnis  $n$ -dimensionaler Volumina im  $\mathbb{R}^n$  geführt.

Als konkretes Beispiel betrachten wir die Gleichverteilung auf einem *Hyperwürfel* im  $\mathbb{R}^n$ :  $\Omega$  soll die Menge aller derjenigen  $n$ -Tupel  $(x_1, \dots, x_n)$  sein, für die  $x_1, \dots, x_n \in [-1, 1]$  gilt<sup>9)</sup>. Das Volumen ist offensichtlich  $2^n$ , die Dichtefunktion ist deswegen die Konstante  $1/2^n$ . Wie wahrscheinlich ist es, dass ein zufällig gewählter Punkt „nahe am Rand“ liegt? Genauer: Für  $\varepsilon \in ]0, 1[$  betrachten wir als  $E$  die Menge der  $(x_1, \dots, x_n)$ , für die der Abstand zum Rand höchstens  $\varepsilon$  ist.  $E$  ist dann gerade die Menge

$$\Omega \setminus \{(x_1, \dots, x_n) \mid x_1, \dots, x_n \in [-1 + \varepsilon, 1 - \varepsilon]\},$$

für die Wahrscheinlichkeit von  $E$  ergibt sich damit der Wert

$$\frac{2^n - (2 - 2\varepsilon)^n}{2^n} = 1 - (1 - \varepsilon)^n.$$

Und die rechte Seite geht für  $n \rightarrow \infty$  gegen Eins, egal wie klein  $\varepsilon$  ist. Anders ausgedrückt: Für „große“  $n$  ist es sehr wahrscheinlich, dass ein zufällig ausgewählter Punkt nahe am Rand liegt.

Zum Abschluss der Überlegungen im Zusammenhang mit der Gleichverteilung soll noch ein Problem angesprochen werden, das bei allen durch Dichtefunktionen definierten Wahrscheinlichkeitsräumen auftritt und das Anfängern manchmal erhebliche Verständnisprobleme macht:

---

<sup>9)</sup>Das ist ein Intervall im  $\mathbb{R}^1$  bzw. ein Quadrat im  $\mathbb{R}^2$  bzw. ein Würfel im  $\mathbb{R}^3$ .

### Ein quasi „philosophisches“ Problem

Wir betrachten  $\Omega = [0, 1]$  mit der Gleichverteilung. Für Teilintervalle  $[c, d]$  von  $\Omega$  ist doch dann die Wahrscheinlichkeit gleich  $d - c$ . Das gilt für alle Teileintervalle, also auch für solche, bei denen  $c = d$  ist. So ist zum Beispiel  $\mathbb{P}([0.25, 0.25]) = 0$ , es ist also nur mit Wahrscheinlichkeit Null zu erwarten, dass bei einer Abfrage exakt 0.25 ausgegeben wird. Heißt das nicht, dass es *unmöglich* ist, dass der Zufall die Zahl 0.25 produziert?

Die Lösung des Problems: Auch Ereignisse  $E$  mit Wahrscheinlichkeit Null müssen mitberücksichtigt werden. Man ändert den Raum allerdings nicht wesentlich ab, wenn man von  $\Omega$  zu  $\Omega \setminus E$  übergeht. Es ist jedoch nicht möglich, alle Ereignisse mit Wahrscheinlichkeit Null zu entfernen, dazu sind es zu viele: Wenn man aus  $[0, 1]$  alle Teileintervalle  $[c, c]$  herausnimmt, würde ja gar nichts übrig bleiben! Im diskreten Fall kann man das noch machen, denn da sind ja höchstens abzählbar viele  $E$  zu entfernen, im kontinuierlichen Fall geht das nicht.

Das *Fazit* lautet also: Ist  $\mathbb{P}(E) = 0$ , so heißt das nicht, dass es *unmöglich* ist, dass der Zufall ein  $c \in E$  erzeugt. Im Gegenteil wird bei jeder Abfrage ein  $c$  erzeugt, und das liegt dann in  $[c, c]$ , einem Ereignis mit Wahrscheinlichkeit Null.

#### Eine Anwendung der Gleichverteilung: das Buffonsche Nadelexperiment

Es geht um ein berühmtes Experiment aus der Wahrscheinlichkeitstheorie, mit dem man überraschenderweise die Kreiszahl  $\pi$  approximativ bestimmen kann. Es handelt sich um das erste Monte-Carlo-Verfahren der Mathematikgeschichte<sup>10)</sup>.



Bild 2.2.3: Man werfe Nadeln ganz zufällig auf liniertes Papier ...

Die Idee geht auf den Comte de Buffon<sup>11)</sup> zurück, der im 18. Jahrhundert

<sup>10)</sup>Darunter versteht man Verfahren, bei denen der Zufall zu Berechnungen eingesetzt wird.

<sup>11)</sup>George Comte de Buffon, 1707 bis 1788. Er war ein wohlhabender französischer Adliger, der sich für viele Bereiche der Naturwissenschaften interessierte. Große Teile des Wissens der damaligen Zeit wurden in seiner Enzyklopädie zusammengefasst. Unter Mathematikern wurde er durch das in diesem Buch beschriebene Nadelexperiment bekannt.



Buffon

lebte. Er war Privatgelehrter mit vielen Interessen, berühmt ist er durch das folgende *Nadelexperiment* geworden:

Man nehme liniertes Papier (Linienabstand  $d$ ) und eine Nadel der Länge  $l$ ; es soll  $l < d$  sein. Nun wird die Nadel „zufällig“ auf das Papier geworfen, und uns interessiert, ob sie eine der Linien schneidet. Wie groß ist diese Wahrscheinlichkeit?

Um das zu analysieren, überlegen wir zunächst, was „zufälliges Werfen“ bedeuten soll. Der Mittelpunkt der Nadel wird doch irgendeinen Abstand  $y$  zur nächstgelegenen Linie haben: Es ist dann  $y \in [0, d/2]$ , und es ist plausibel anzunehmen, dass die  $y$  entsprechend der Gleichverteilung erzeugt werden.

Dann gibt es noch eine Neigung  $\alpha$  der Nadel gegen die Linien, wobei  $\alpha$  zwischen 0 (parallel) und  $\pi/2$  (senkrecht) variiert. Auch hier ist es sinnvoll davon auszugehen, dass  $\alpha$  gleichverteilt ist.

Zusammen heißt das: „Nadel zufällig werfen“ entspricht der gleichverteilten Auswahl eines Tupels  $(\alpha, y)$  aus  $\Omega = [0, \pi/2] \times [0, d/2]$ .

Welche  $(\alpha, y)$  führen denn dazu, dass die Nadel die nächstgelegene Linie schneidet? Dazu muss doch der Abstand  $y$  des Mittelpunktes der Nadel (das ist im nachstehenden Bild der Punkt  $M$ ) zur oberen Linie durch die Steigung der Nadel überbrückt werden. Der überbrückte Abstand – die Länge der Strecke von  $A$  nach  $B$  – ist aufgrund elementarer Trigonometrie gleich  $(l/2) \sin \alpha$ :

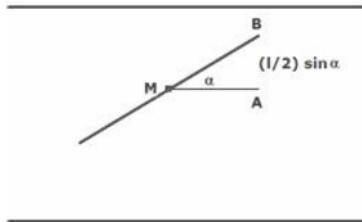


Bild 2.2.4: Wann schneidet die Nadel eine Linie?

Die Nadel schneidet also die Linie genau dann, wenn  $y \leq (l/2) \sin \alpha$  gilt. Die Gesamtheit der  $(y, \alpha)$  ist in Bild 2.2.5 grau eingezzeichnet:

*Das* ist das uns interessierende Ereignis, wir bezeichnen es mit  $E$ . Aufgrund der bisherigen Überlegungen gilt:

- $E$  ist die Fläche unter dem Graphen der Funktion  $\alpha \mapsto (l/2) \sin \alpha$ , diese Funktion ist auf  $[0, \pi/2]$  definiert. Folglich ist der Flächeninhalt gleich  $\int_0^{\pi/2} (l/2) \sin \alpha d\alpha = l/2$ .
- Die Wahrscheinlichkeit von  $E$  ist „Flächeninhalt von  $E$  durch Flächeninhalt von  $\Omega$ “, also gleich

$$\mathbb{P}(E) = \frac{l/2}{(\pi/2)(d/2)} = \frac{2l}{\pi d}.$$

Das Ergebnis ist auch plausibel: Die Wahrscheinlichkeit wird mit größerem  $l$  (bzw. größerem  $d$ ) zunehmen (bzw. abnehmen).

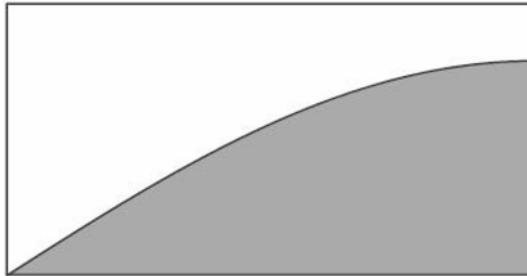


Bild 2.2.5: Die  $(\alpha, y)$ , die zu einem Schnittpunkt führen (grau).

Man könnte damit zufrieden sein, denn  $\mathbb{P}(E)$  ist ja nun ermittelt. Man kann die Formel aber auch nach  $\pi$  auflösen, d.h.

$$\pi = \frac{2 \cdot l}{d \cdot \mathbb{P}(E)}$$

berechnen und sich dann daran erinnern, dass  $\mathbb{P}(E)$  doch die relative Anzahl der „Experimente“ ist, bei denen ein  $\omega \in E$  erzeugt wurde.

Und das führt zu der folgenden *Anleitung zur Bestimmung von  $\pi$* :

→ **Programm!**

- Man werfe die Nadel „sehr oft“ und protokolliere, ob sie eine der Linien trifft.
- Wenn das in  $m$  von  $n$  Fällen so war, approximiere  $\mathbb{P}(E)$  durch  $m/n$ , und daraus wird die folgende Näherung für  $\pi$  hergeleitet:

$$\pi \approx \frac{2l}{(m/n)d}.$$

Es ist fast überflüssig zu bemerken, dass das kein besonders effektiver und zuverlässiger Weg ist, die Kreiszahl zu bestimmen. Die Handlungsanweisung ist aber sehr einfach ...

Statt einer Nadel und liniertem Papier kann man natürlich auch einen Dielenfußboden und ein Stöckchen nehmen. Der Comte soll das Experiment angeblich in dieser Variante durchgeführt haben.

#### Flächenbestimmung mit Monte-Carlo-Verfahren

Mit der gleichen Idee kann man auch Flächeninhalte näherungsweise bestimmen. Mal angenommen, wir wollen die Fläche  $F$  zwischen dem Graphen einer Funktion  $f$  und der  $x$ -Achse, also  $\int_a^b f(x) dx$  ermitteln. Dazu schließen wir  $F$  in

→ **Programm!**

ein Rechteck  $R$  ein, dann ist der Flächeninhalt von  $F$  gleich dem Produkt aus dem Flächeninhalt von  $R$  und der Wahrscheinlichkeit von  $F$ , wenn  $R$  mit der Gleichverteilung versehen ist. Man muss also nur „sehr viele“ Zufallspunkte in  $R$  erzeugen<sup>12)</sup> und zählen, wie viele davon in  $F$  liegen. Wenn das zum Beispiel in der Hälfte der Fälle so ist, wird  $F$  etwa den halben Flächeninhalt von  $R$  haben.

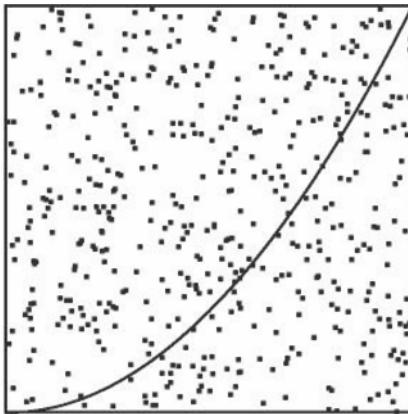


Bild 2.2.6: Monte-Carlo-Verfahren: Bestimmung des Flächeninhalts.

Dieses Verfahren ist einfach zu programmieren, es ist auf die kompliziertesten Funktionen anwendbar und man kann es auch in mehr als zwei Dimensionen einsetzen.

#### Die Exponentialverteilung

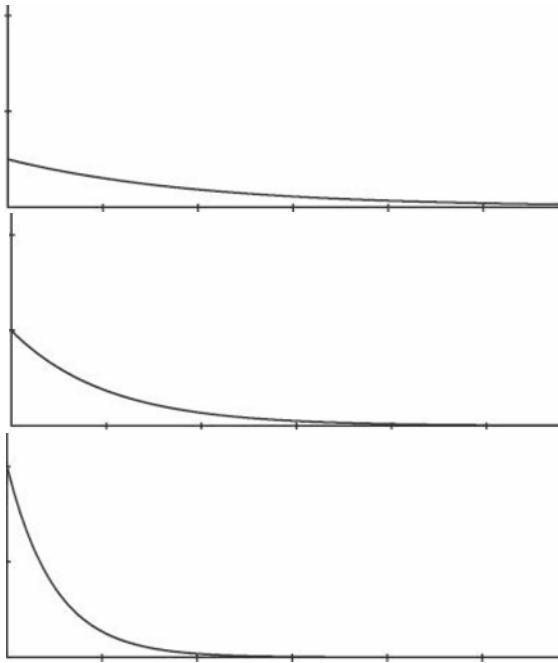
Es handelt sich um eine ganze Klasse von Verteilungen. Immer ist  $\Omega = [0, \infty[$ , und es ist ein Parameter  $\lambda > 0$  vorgegeben. Die zugehörige Dichtefunktion der Exponentialverteilung ist durch  $f(x) := \lambda e^{-\lambda x}$  gegeben. Es ist schnell einzusehen, dass das wirklich eine Dichtefunktion ist: Offensichtlich ist stets  $f(x) \geq 0$ , und

$$\int_0^\infty \lambda e^{-\lambda x} dx = \lim_{r \rightarrow \infty} \int_0^r \lambda e^{-\lambda x} dx = \lim_{r \rightarrow \infty} 1 - e^{-\lambda r} = 1.$$

Ist  $\lambda$  groß, hat die Dichtefunktion zunächst einen großen Wert bei 0 und fällt dann sehr schnell ab: Teilintervalle  $[c, d]$  werden also nur dann eine nicht vernachlässigbare Wahrscheinlichkeit haben, wenn  $[c, d]$  nahe bei der Null liegt. Ist dagegen  $\lambda$  „klein“, so fällt die Dichtefunktion nur „sehr langsam“. Dadurch bekommen möglicherweise auch Intervalle  $[c, d]$  mit „großen“  $c, d$  eine hohe Wahrscheinlichkeit.

---

<sup>12)</sup>Wie das geht, steht in Abschnitt 2.4.

Bild 2.2.7: Exponentialverteilungen zu  $\lambda = 0.5$ ,  $\lambda = 1$  und  $\lambda = 2$ .

Wir werden uns in Kapitel 6 ausführlich um die Exponentialverteilung kümmern. Sie ist unter bestimmten Bedingungen gut dazu geeignet, Wartezeiten zu modellieren: Die Wahrscheinlichkeit, dass die Wartezeit zwischen  $c$  und  $d$  liegt, ist  $\int_c^d \lambda e^{-\lambda x} dx$  für ein geeignetes  $\lambda$ . Alle diese Zahlen sind übrigens leicht auszurechnen: Es ist  $\int_c^d \lambda e^{-\lambda x} dx = e^{-\lambda c} - e^{-\lambda d}$ .

→  
Programm!

### Die Normalverteilungen

Das ist sicher das interessanteste und wichtigste Beispiel eines durch Dichtefunktionen definierten Wahrscheinlichkeitsraumes. Die Begründung dieser Aussage muss allerdings auf Abschnitt 9.6 vertagt werden.

Es ist in diesem Fall  $\Omega = \mathbb{R}$ , und es gibt zwei Parameter: eine Zahl  $a \in \mathbb{R}$  und eine Zahl  $\sigma > 0$ . Die Dichtefunktion ist durch

$$f_{a,\sigma^2}(x) := \frac{1}{\sqrt{2\pi} \cdot \sigma} \exp\left(-\frac{(x-a)^2}{2\sigma^2}\right)$$

erklärt; dabei haben wir im Interesse der besseren Lesbarkeit die Exponentialfunktion  $e^x$  als  $\exp(x)$  geschrieben.

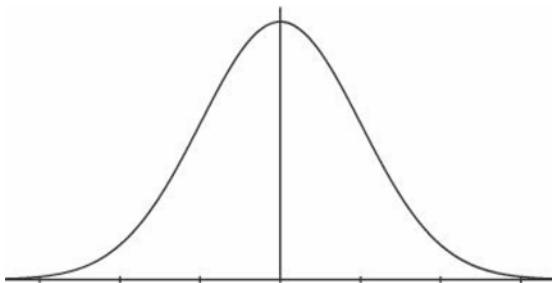
Hat man  $\mathbb{R}$  mit dieser Dichtefunktion zu einem Wahrscheinlichkeitsraum gemacht, so spricht man von der *Normalverteilung*  $N(a, \sigma^2)$ .

Zum ersten Kennenlernen gibt es hier eine Skizze:

→  
Programm!



Gauß

Bild 2.2.8: Die Standardnormalverteilung  $N(0, 1)$ .

Die Dichtefunktion ist also die (manchmal so genannte) *Gaußsche Glockenkurve*<sup>13)</sup>. Für verschiedene  $a, \sigma$  gehen die  $f_{a,\sigma^2}$  durch einfache Transformationen ineinander über, es ist nur der Maßstab auf der  $x$ - und auf der  $y$ -Achse zu verändern.

Im Zusammenhang mit der Normalverteilung gibt es *zwei Schwierigkeiten*. Erstens haben wir noch gar nicht begründet, dass es wirklich eine Dichtefunktion ist: Warum ist  $\int_{-\infty}^{+\infty} f_{a,\sigma^2}(x) dx = 1$ ? Man kann das zwar durch eine Substitution auf die Frage zurückführen, ob  $\int_{-\infty}^{+\infty} e^{-x^2/2} dx = \sqrt{2\pi}$  gilt, doch das lässt sich mit elementaren Mitteln wirklich nicht entscheiden. Wir werden es hier einfach als Tatsache verwenden, eine Begründung finden Interessierte im Anhang auf Seite 362.

Die zweite Schwierigkeit betrifft *die konkrete Auswertung von Wahrscheinlichkeiten bei der Normalverteilung*. Die Wahrscheinlichkeit für ein Intervall  $[c, d]$  ist doch  $\int_c^d f_{a,\sigma^2}(x) dx$ , es ist aber nicht zu sehen, wie dieses Integral ausgerechnet werden kann. Die ganze Wahrheit ist sogar noch ernüchternder: Man kann beweisen, dass es keine geschlossenen darstellbare Stammfunktion zu diesen Dichtefunktionen gibt<sup>14)</sup>.

Glücklicherweise hat dieses Problem keine Konsequenzen, denn man kann die Integrale ja mit numerischen Verfahren in jeder beliebigen Genauigkeit bestimmen. Für alle praktisch wichtigen Probleme reicht es sogar aus, die in Tabellen zusammengestellten Werte abzulesen. Das soll hier kurz erläutert werden:

- Auf Seite 364 finden Sie eine *Tabelle der Normalverteilung*. Dort sind für verschiedene  $x$  die Wahrscheinlichkeiten der Intervalle  $] -\infty, x ]$  für den Fall der  $N(0, 1)$ -Verteilung (der so genannten *Standard-Normalverteilung*) auf vier Stellen genau aufgeführt.

Zum Beispiel ist  $\mathbb{P}(] -\infty, -0.13 ]) = 0.4483$  und  $\mathbb{P}(] -\infty, 1.66 ]) = 0.9515$ .

<sup>13)</sup>Carl Friedrich Gauß, 1777 bis 1855. Für viele gilt Gauß als der bedeutendste Mathematiker, den es je gegeben hat. Zu fast allen mathematischen Teilgebieten hat er wesentliche Beiträge geleistet.

<sup>14)</sup>Das ist ein Ergebnis von Liouville aus dem 19. Jahrhundert, einen Beweis dieser Tatsache findet man in meinem Buch zur Analysis II (Kapitel 6, Abschnitt 6.6).

- Möchte man  $\mathbb{P}([c, d])$  bestimmen, so muss man beachten, dass

$$]c, d] = ]-\infty, d] \setminus ]-\infty, c]$$

gilt und dass  $\mathbb{P}([c, d]) = \mathbb{P}(]c, d])$ , denn einpunktige Mengen haben Wahrscheinlichkeit Null. Deswegen ist

$$\mathbb{P}([c, d]) = \mathbb{P}(]-\infty, d]) - \mathbb{P}(]-\infty, c]).$$

So ist zum Beispiel  $\mathbb{P}[-0.13, 1.66] = 0.9515 - 0.4483 = 0.5032$ .

- Sind die Werte für allgemeine  $N(a, \sigma^2)$ -Verteilungen interessant, so muss man auf die Standard-Normalverteilung transformieren: Wir werden in Kapitel 7 beweisen, dass für ein standard-normalverteiltes  $x$  die Zahl  $a + \sigma x$   $N(a, \sigma^2)$ -verteilt ist. Es folgt sofort: Die Wahrscheinlichkeit des Intervalls  $[c, d]$  unter der Dichte  $f_{a, \sigma^2}$  ist gleich der Wahrscheinlichkeit von  $[(c - a)/\sigma, (d - a)/\sigma]$  bezüglich der Standard-Normalverteilung.

Geht es zum Beispiel um die Wahrscheinlichkeit von  $[3, 7]$  unter  $N(3, 100)$  (es ist also  $a = 3$  und  $\sigma = 10$ ), so muss man die Wahrscheinlichkeit von  $[0, 0.4]$  unter  $N(0, 1)$  berechnen. Mit Tafelhilfe ergibt sich der Wert  $0.6554 - 0.5000 = 0.1554$ .

## 2.3 Simulation diskreter Räume

In Kapitel 1 hatten wir den Zufall durch „Zufallsautomaten“ illustriert: Das sind „schwarze Kästen“, die auf Knopfdruck Ausgaben  $\omega$  produzieren, die erstens in der vorgegebenen Menge  $\Omega$  liegen und für die zweitens die Antwort auf die Frage „ $\omega \in E$ ?“ bei  $n$  Versuchen in etwa  $n\mathbb{P}(E)$  Fällen „ja“ sein wird, wenn nur  $n$  groß genug ist. Dabei sind die Zahlen  $\mathbb{P}(E)$ , die „Wahrscheinlichkeiten von  $E$ “, vorgegeben.

Gibt es solche „Zufallsautomaten“ wirklich? Die erfreuliche Antwort besteht aus zwei Teilen: Erstens sind ganz spezielle Zufallsautomaten in jedem Computer verfügbar, und zweitens kann man damit ohne großen Aufwand maßgeschneiderte Zufallsautomaten für jeden erdenklichen Zweck simulieren.

Vorgefertigt findet man:

- *Laplaceräume*. Es ist in so gut wie allen Programmiersprachen ein Unterprogramm verfügbar, das – bei Vorgabe einer natürlichen Zahl  $n$  – eine Zahl in  $\{0, 1, \dots, n-1\}$  ausgibt. Alle Ausgaben haben die gleiche Wahrscheinlichkeit, das Programm verhält sich wie ein (fast) perfekter Zufallsautomat, der einen Laplace Raum simuliert.

Oft heißt dieses Unterprogramm `random(n)`. Will man damit zum Beispiel Würfelausgaben simulieren, so muss man nur die Programmschleife

```
...; k:=1+random(6); Ausgabe k; ...
```

einbauen. Das Ergebnis könnte dann so ausssehen:

→  
Programm!

```
3, 3, 4, 2, 5, 5, 6, 3, 3, 3, 1, 4, 4, 2, 1, 3, 3, 1, 3, 2, 2, 1,
6, 3, 1, 6, 2, 2, 5, 3, 4, 6, 2, 2, 6, 1, 5, 6, 3, 1, 1, 1, 6, 4,
1, 2, 6, 3, 4, 3, 6, 3, 6, 6, 2, 6, 5, 4, 2, 6, 1, 5, 2, 4, ...
```

Zur Simulation einer fairen Münze, die auf ihren beiden Seiten die Zahlen 0 und 1 trägt, müsste man einfach `...; k:=random(2); Ausgabe k; ...` so oft wie gewünscht durchlaufen lassen. Hier ein Simulationsbeispiel<sup>15)</sup>:

```
1, 1, 0, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 0, 1, 0,
0, 0, 0, 0, 1, 0, 1, 1, 1, 0, 1, 1, 0, 0, 1, 0, 1, 0, 0, 1, 1,
1, 1, 0, 1, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 1,
0, 1, 1, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 1, ...
```

Wie verlässlich sind diese Simulationen? Sind unsere Anforderungen an einen Zufallsautomaten wirklich erfüllt? Die Antwort: Streng genommen gilt das nicht, und es ist auch nicht zu erwarten, da ein Computer Ergebnisse nur auf deterministische Weise erzeugen kann. Allerdings sind die Abweichungen vom „perfekten Zufall“ so winzig, dass Computersimulationen für praktisch alle Fälle gut zu verwenden sind.

- *Die Gleichverteilung auf  $[0, 1]$ .* Bei diesem Unterprogramm werden Ausgaben in  $[0, 1]$  erzeugt, und zwar so, dass bei „vielen“ Abfragen der Anteil der Ausgaben in einem Teilintervall  $[c, d]$  von  $[0, 1]$  nahe bei  $d - c$  ist. (Zum Beispiel sollten etwa 20 Prozent der ausgegebenen Ergebnisse in  $[0.3, 0.5]$  liegen. Das entspricht also den Abfragen zu dem Wahrscheinlichkeitsraum  $\Omega = [0, 1]$  mit der Dichtefunktion  $f(x) = 1$ .)

Es ist klar, dass man die Gleichverteilung auf  $[0, 1]$  niemals ganz exakt erzeugen kann, denn es werden ja nur Zahlen mit endlich vielen Dezimalziffern erzeugt. Für alle praktischen Zwecke reicht das aber völlig aus.

Oft heißt dieses Unterprogramm `random`. Die zugehörige Schleife sieht so aus:

```
...; x:=random; Ausgabe x; ...
```

Hier ein Beispiel, wir haben Zahlen mit drei Dezimalziffern erzeugt:

```
0.430, 0.871, 0.778, 0.228, 0.303, 0.337, 0.764, 0.180, 0.963, 0.070,
0.192, 0.177, 0.336, 0.419, 0.173, 0.303, 0.258, 0.468, 0.365, 0.779,
0.253, 0.591, 0.179, 0.178, 0.485, 0.921, 0.598, 0.692, 0.076, ...
```

Dieses Erzeugen von Zufallsausgaben im Computer ist bemerkenswert effektiv, pro Sekunde lassen sich mehrere Millionen Abfragen durchführen. In den

---

<sup>15)</sup>Man beachte, dass das weniger zufällig aussieht, als man erwartet hätte. Jemand, der sich so eine Folge hätte ausdenken sollen, würde sicher nicht so viele Einsen oder Nullen hintereinander auftreten lassen. Das ist eines der vielen Beispiele dafür, dass unsere Intuition in Bezug auf das Thema „Zufall“ nicht besonders gut ausgeprägt ist.

Ergänzungen zu diesem Kapitel findet man einige Informationen darüber, wie so etwas realisiert werden kann (Seite 64).

#### Simulation: ein beliebiger endlicher Wahrscheinlichkeitsraum

Ist  $\Omega = \{1, \dots, n\}$  durch Vorgabe von Zahlen  $p_1, \dots, p_n \in [0, 1]$  mit  $\sum p_i = 1$  zu einem Wahrscheinlichkeitsraum gemacht worden, so könnte die Simulation so aussehen:

*Erster Schritt:* Unterteile das Intervall  $[0, 1]$  in  $n$  Intervalle  $I_1, \dots, I_n$ , so dass  $I_i$  die Länge  $p_i$  hat ( $i = 1, \dots, n$ ). Zum Beispiel

$$I_1 := [0, p_1[, I_2 := [p_1, p_1 + p_2[, \dots, I_n := [p_1 + \dots + p_{n-1}, 1].$$

*Zweiter Schritt:* Erzeuge eine gleichverteilte Zufallsabfrage  $x \in [0, 1]$  mit `random`. Liegt  $x$  in  $I_i$ , so wird  $i$  ausgegeben.

Nachstehend findet man eine Illustration dieses Verfahrens für das erste Beispiel aus Abschnitt 2.1 ( $\Omega = \{1, 2, 3, 4\}$ ,  $p_1 = p_2 = p_3 = 1/5$ ,  $p_4 = 2/5$ ):

```
1, 3, 4, 4, 4, 1, 4, 3, 1, 4, 1, 4, 4, 1, 4, 1, 3, 1, 2, 3, 3, 2, 4, 2,
3, 4, 4, 4, 2, 4, 2, 2, 4, 2, 4, 1, 4, 3, 1, 2, 1, 2, 4, 3, 1, 4, 4, 2,
1, 2, 4, 3, 4, 2, 3, 2, 1, 4, 3, 2, ...
```

Ganz analog geht man vor, wenn  $\Omega$  endlich, aber nicht von der Form  $\{1, \dots, n\}$  ist. Da unterteilt man  $[0, 1]$  in Intervalle  $I_\omega$  mit der Länge  $p_\omega$ .

#### Simulation: abzählbare Wahrscheinlichkeitsräume

Die vorstehende Idee lässt sich nicht direkt auf den Fall abzählbarer  $\Omega$  übertragen, da in einem Computer immer nur endlich viele Operationen möglich sind. Sie kann aber so modifiziert werden, dass es für alle praktisch wichtigen Fälle ausreicht.

Gegeben seien Zahlen  $p_1, p_2, \dots$  in  $[0, 1]$  mit  $\sum_n p_n = 1$ . Fixiere nun ein „kleines“  $\varepsilon > 0$  (etwa  $\varepsilon = 1/10.000$ ) und wähle  $n_0$  so groß, dass  $\sum_{n=1}^{n_0} p_n \geq 1 - \varepsilon$ . Gehe dann über zu  $\Omega' := \{1, \dots, n_0\}$  mit den folgenden Wahrscheinlichkeiten: Für  $n < n_0$  ist  $p'_n := p_n$ , und  $p'_{n_0} := \sum_{n=n_0}^{\infty} p_n$ . (Die Zahlen  $n_0, n_0+1, \dots$  wurden also zu  $n_0$  zusammengefasst.)  $\Omega'$  ist dann fast gar nicht von  $\Omega$  zu unterscheiden. Nur ganz selten, mit Wahrscheinlichkeit höchstens  $\varepsilon$ , wird ein falsches Ergebnis erzeugt (nämlich  $n_0$  statt evtl. eine größere Zahl).

Und  $\Omega'$  kann, wie vorstehend beschrieben, simuliert werden.

#### Simulation: Laplaceräume

Hier gibt es mehrere Möglichkeiten. Man kann einen Laplaceraum auf der Menge  $\{1, \dots, n\}$  wie folgt simulieren:

- Am einfachsten ist es, durch `1+random(n)` direkt eine Zufallszahl mit den richtigen Eigenschaften auszugeben.

- Natürlich kann man auch das allgemein für endliche Wahrscheinlichkeitsräume gültige Verfahren anpassen: Zufallszahl  $x$  in  $[0, 1]$  erzeugen; nachprüfen, in welchem der Intervalle  $[(i-1)/n, i/n]$  ( $i = 1, \dots, n$ ) die Zahl  $x$  liegt;  $i$  ausgeben<sup>16)</sup>.
- Wenn das Programm „abschneiden“, also die Nachkommastellen unterdrücken kann, geht es auch eleganter:  $x \in [0, 1]$  erzeugen;  $n \cdot x$  berechnen; Nachkommastellen abschneiden und das Ergebnis ausgeben. Das sind dann – wie bei `random(n)` – in  $\{0, \dots, n-1\}$  gleichverteilte Zahlen.

Wenn es Zahlen in  $1, \dots, n$  sein sollen, muss man noch Eins addieren.

#### Simulation: Bernoulliraum

Das allgemeine Verfahren kann bei gegebener Wahrscheinlichkeit  $p$  für „Erfolg“ sehr effektiv eingesetzt werden: Erzeuge eine gleichverteilte Zufallszahl  $x$  in  $[0, 1]$ ; gilt  $x \leq p$ , wird „1“ ausgegeben, andernfalls Null. Im Fall  $p = 0.3$  ergaben sich die folgenden Werte:

```

0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 1,
0, 0, 0, 1, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 1,
0, 1, 1, 0, 1, 0, 0, ...

```

#### Simulation: Poissonverteilung

Natürlich könnte man das allgemeine Verfahren für abzählbare  $\Omega$  hier anpassen. Es gibt jedoch eine weit effektivere Möglichkeit, die wir allerdings erst viel später begründen können (vgl. Korollar 6.2.2). Bis dahin ist es so etwas wie ein Kochrezept:

*Anleitung zur Simulation der Poissonverteilung zum Parameter  $\lambda > 0$ :*

1. Setze  $y = 0$ .
2. Erzeuge eine gleichverteilte Zufallszahl  $x \in [0, 1]$ .
3. Addiere  $-\log x$  zu  $y$ .
4. Wiederhole – falls erforderlich – die Schritte „2.“ und „3.“ so oft, bis erstmals  $y > \lambda$  gilt. Wenn das im  $k$ -ten Durchlauf passiert ist, wird  $k-1$  ausgegeben<sup>17)</sup>.

Zur Illustration findet man nachstehend ein Beispiel für den Fall  $\lambda = 4$ :

```

6, 5, 3, 0, 6, 6, 4, 2, 6, 4, 4, 1, 3, 1, 4, 4, 4, 3, 5, 5, 2, 2, 5, 3,
7, 8, 3, 1, 3, 4, 4, 5, 3, 5, 2, 5, 3, 4, 4, 3, 5, 3, 5, 3, 4, 8, 4, 1,
7, 3, 2, 3, 6, 4, 5, 1, 2, 6, 6, 5, 2, 2, 5, 3, 2, 5, 7, 2, 4, 3, 6, 6,
7, 8, 6, 2, 4, 4, 6, 4, ...

```

<sup>16)</sup>Nur mit Wahrscheinlichkeit Null wird  $x$  gleichzeitig in zwei Intervallen liegen.

<sup>17)</sup>Die Ausgabe ist damit 0, falls schon für das erste  $x$  die Bedingung  $-\log x > \lambda$  erfüllt war.

Simulation: geometrische Verteilung

Wie schon erwähnt, beschreibt die geometrische Verteilung das „Warten auf den ersten Erfolg“. Das wird zwar erst in Abschnitt 6.4 auf Seite 167 begründet, wir wollen diese Tatsache jetzt schon zu einer Simulationsanleitung umschreiben.

$q$  sei vorgegeben. Setze  $p := 1 - q$ . Führe dann so lange Bernoulli-Experimente mit Erfolgswahrscheinlichkeit  $p$  durch, bis zum ersten Mal eine 1 erscheint. Wenn das im  $k$ -ten Versuch passiert, wird  $k$  ausgegeben.

Das Beispiel  $q = 5/6$  (also  $p = 1/6$ ) kann durch das Warten auf die erste Sechs beim Würfeln illustriert werden. In der folgenden Simulation hat das manchmal ziemlich lange gedauert:

```
2, 8, 4, 8, 3, 8, 6, 21, 2, 1, 1, 7, 13, 1, 9, 8, 6, 13, 2, 11, 4, 20,
16, 10, 19, 10, 9, 5, 12, 7, 1, 1, 1, 3, 9, 1, 1, 1, 1, 3, 2, 24, 9, 11,
9, 6, 9, 14, 15, 6, 25, 10, 5, 1, 1, 4, 2, 8, 11, 3, 6, 2, 4, 6, 1, 15,
17, 8, 4, 3, 1, 1, 19, 2, 4, 9, 11, 12, 2, 6, ...
```

→  
Programm!

## 2.4 Simulation: Räume mit Dichtefunktionen

Im vorigen Abschnitt wurde schon bemerkt, dass man  $[0, 1]$  mit der Gleichverteilung sehr einfach simulieren kann, denn dieser „Zufallsautomat“ ist in jedem Computer schon werkseitig eingebaut.

Das sieht auf den ersten Blick recht spartanisch aus, wir werden aber gleich sehen, dass man damit alle Räume simulieren kann, bei denen die Wahrscheinlichkeiten durch eine Dichtefunktion auf einem Intervall definiert sind.

Wir beginnen mit einigen Vorbereitungen. Eine fundamentale Rolle wird bei den Überlegungen dieses Abschnitts der *Hauptsatz der Differential- und Integralrechnung* spielen:

Ist  $F$  eine stetig differenzierbare Funktion und  $f = F'$ , so gilt  
 $\int_a^c f(x) dx = F(c) - F(a)$ .

(Das lässt sich auch auf uneigentliche Integrale verallgemeinern. So ist zum Beispiel  $\int_a^\infty f(x) dx = \lim_{r \rightarrow \infty} F(r) - F(a)$ , falls das Integral existiert.)

Ist  $f : [a, b] \rightarrow \mathbb{R}$  eine Dichtefunktion und erfüllt  $F$  die Bedingung  $F(a) = 0$ , so ist folglich die Zahl  $F(c)$  für jedes  $c \in [a, b]$  die Wahrscheinlichkeit von  $[a, c]$ . Diese Wahrscheinlichkeiten spielen auch für beliebige Wahrscheinlichkeitsmaße auf  $\mathbb{R}$  – also auch für solche, die nicht durch eine Dichtefunktion definiert sind, eine wichtige Rolle:

**Definition 2.4.1.** Es sei  $\mathbb{P}$  ein Wahrscheinlichkeitsmaß auf der  $\sigma$ -Algebra der Borelmengen in  $\mathbb{R}$ . Wir definieren dann die zu  $\mathbb{P}$  geförmige Verteilungsfunktion  $F_{\mathbb{P}}$  durch

$$F_{\mathbb{P}} : \mathbb{R} \rightarrow [0, +\infty[, x \mapsto \mathbb{P}(-\infty, x]).$$

Ist  $\Omega$  eine Borelmenge in  $\mathbb{R}$ , so kann jedes Wahrscheinlichkeitsmaß  $\mathbb{P}$  auf  $\Omega$  auch als Wahrscheinlichkeitsmaß auf  $\mathbb{R}$  aufgefasst werden: Man muss für Borelmengen  $B \subset \mathbb{R}$  die Wahrscheinlichkeit von  $B$  nur durch  $\mathbb{P}(B) := \mathbb{P}(\Omega \cap B)$  erklären. Folglich lässt sich auch für solche  $\mathbb{P}$  die Funktion  $F_{\mathbb{P}}$  definieren.

Hier sieht man Ausschnitte aus den Graphen der Verteilungsfunktionen, die zum fairen Würfel und zur Gleichverteilung auf  $[0, 1]$  gehören:

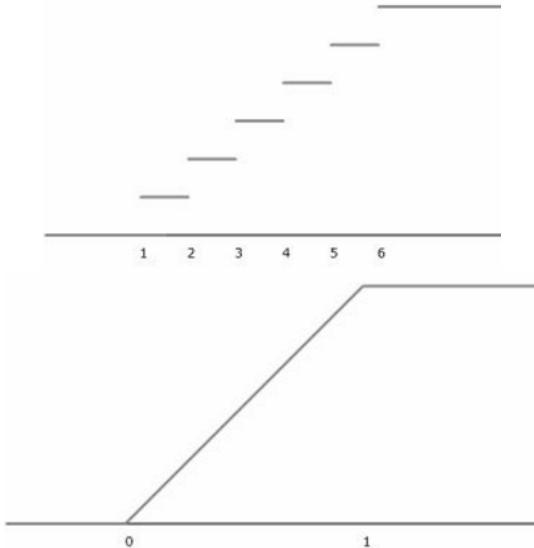


Bild 2.4.1: Verteilungsfunktionen: Würfel (oben) und Gleichverteilung auf  $[0, 1]$  (unten).

Bemerkenswerter Weise ist die Zuordnung  $\mathbb{P} \mapsto F_{\mathbb{P}}$  injektiv, d.h.  $\mathbb{P}$  kann auf eindeutige Weise aus  $F_{\mathbb{P}}$  rekonstruiert werden. Aus  $F_{\mathbb{P}}$  ergeben sich nämlich die Wahrscheinlichkeiten für Intervalle  $]a, b]$ , denn  $\mathbb{P}(]a, b]) = F_{\mathbb{P}}(b) - F_{\mathbb{P}}(a)$ . Und da die Gesamtheit dieser Intervalle einen durchschnitts-stabilen Erzeuger der Borelmengen bildet, ist wegen der in Abschnitt 1.6 hergeleiteten Ergebnisse das Maß  $\mathbb{P}$  eindeutig bestimmt.

Man kann auch genau beschreiben, welche Funktionen  $F$  als  $F_{\mathbb{P}}$  auftreten können. Wir werden das allerdings nur für einen Spezialfall benötigen:

**Lemma 2.4.2.** *Es sei  $F : \mathbb{R} \rightarrow [0, \infty[$  eine monoton steigende stetig differenzierbare Funktion, für die  $\lim_{x \rightarrow -\infty} F(x) = 0$  und  $\lim_{x \rightarrow +\infty} F(x) = 1$  gilt.*

*Dann ist  $f := F'$  eine Dichtefunktion auf  $\mathbb{R}$ , und für das zugehörige Wahrscheinlichkeitsmaß  $\mathbb{P}$  gilt  $F_{\mathbb{P}} = F$ .*

**Beweis:** Das folgt fast unmittelbar aus dem Hauptsatz der Differential- und

Integralrechnung. Danach ist nämlich für alle  $x$

$$\begin{aligned} F_{\mathbb{P}}(x) &= \mathbb{P}([0, x]) \\ &= \int_0^x f(t) dt \\ &= F(x) - F(0) \\ &= F(x). \end{aligned}$$

Der Hauptsatz wurde beim dritten Gleichheitszeichen verwendet.  $\square$

Wir kommen zurück zum Problem, Räume mit Dichten zu simulieren. Hier ist unser Hauptergebnis:

**Satz 2.4.3.** *Es sei  $\Omega = [a, b]$ ,  $\Omega = [a, +\infty[$  oder  $\Omega = \mathbb{R}$ . Durch eine stetige Dichtefunktion  $f$  sei  $\Omega$  zu einem Wahrscheinlichkeitsraum gemacht worden. Diesen Raum kann man wie folgt simulieren:*

- Definiere  $F : \Omega \rightarrow \mathbb{R}$  als Verteilungsfunktion, also in den ersten beiden Fällen durch

$$F(x) := \mathbb{P}([a, x]) = \int_a^x f(t) dt$$

und im Fall  $\Omega = \mathbb{R}$  durch

$$F(x) := \mathbb{P}([-∞, x]) = \int_{-∞}^x f(t) dt.$$

Man weiß aus der Analysis, dass  $F$  eine Stammfunktion zu  $f$  ist, d.h., dass  $F' = f$  gilt<sup>18)</sup>.

- Zur Simulation verfahre dann wie folgt: Erzeuge ein gleichverteiltes  $y$  in  $[0, 1]$ ; suche ein  $x \in \Omega$  mit  $F(x) = y$  (so ein  $x$  existiert nach dem Zwischenwertsatz, denn  $F$  ist eine stetige Funktion, die monoton von 0 nach Eins wächst<sup>19)</sup>; dann wird  $x$  ausgegeben. Falls es mehrere  $x$  mit  $F(x) = y$  gibt, soll die Ausgabe das kleinste dieser  $x$  sein.

Wir behaupten, dass damit die richtigen Wahrscheinlichkeiten simuliert werden: Der Anteil der Ausgaben in  $[c, d] \subset \Omega$  ist bei „vielen“ Ausgaben gleich  $\int_c^d f(x) dx$ .

**Beweis:** Sei  $[c, d]$  ein Teilintervall von  $\Omega$ . Setze  $c' := F(c)$  und  $d' := F(d)$ . Da  $F$  eine monoton steigende Funktion ist, gilt  $c' \leq d'$ .

Wie wahrscheinlich ist es, dass unser Verfahren ein  $x \in [c, d]$  erzeugt? Das passiert doch genau dann, wenn das zuerst produzierte  $y$  in  $[c', d']$  liegt, und

<sup>18)</sup>  $F$  ist diejenige Stammfunktion, für die  $F(a) = 0$  ist bzw. – im Fall  $\Omega = \mathbb{R}$  – für die  $\lim_{a \rightarrow -\infty} F(a) = 0$  gilt.

<sup>19)</sup> Genau genommen gibt es im Fall  $\Omega = \mathbb{R}$  eventuell keine Urbilder, falls  $y = 0$  oder  $y = 1$  erzeugt wurde. Doch das ist nur mit Wahrscheinlichkeit 0 zu erwarten.

die Wahrscheinlichkeit, dass das passiert, ist gleich  $d' - c'$  (denn  $y$  wurde gemäß der Gleichverteilung erzeugt).

Anders ausgedrückt: Die Wahrscheinlichkeit für  $x \in [c, d]$  ist gleich  $d' - c'$ , also gleich  $F(d) - F(c)$ . Nach dem Hauptsatz der Differential- und Integralrechnung stimmt diese Zahl aber mit  $\int_c^d f(x) dx$  überein, und damit ist alles bewiesen.  $\square$

Wir fassen zusammen: Wenn man in der Lage ist, eine Stammfunktion  $F$  zu  $f$  zu bestimmen und die inverse Abbildung zu  $F$  beherrscht, ist es ganz leicht,  $\Omega$  mit der Dichtefunktion  $f$  zu simulieren:  $y \in [0, 1]$  gleichverteilt erzeugen, dann  $x := F^{-1}(y)$  ausgeben.

Hier einige **Beispiele** dazu:

**1.** Wir betrachten  $\Omega = [2, 11]$  mit der Gleichverteilung. Hier ist  $f(x)$  die Konstante  $1/9$ , und damit ist  $F$  durch  $F(x) := (x - 2)/9$  für  $x \in \Omega$  definiert. Die Simulation sieht also so aus:  $y \in [0, 1]$  gleichverteilt erzeugen, dann die Gleichung  $y = (x - 2)/9$  nach  $x$  auflösen und  $x$  ausgeben:  $x = 9y + 2$ . (Als Computerprogramm: `... y:=random; x:=9y+2; Ausgabe x ...`)

Allgemein: Um die Gleichverteilung auf  $[a, b]$  zu simulieren, erzeuge gleichverteilte  $y$  in  $[0, 1]$ , dann ist  $x = a + (b - a)y$  auszugeben.

**2.**  $\Omega = [0, 1]$ , versehen mit der Dichtefunktion  $f(x) = (1 + x)/1.5$ . (Ausgaben in der Nähe von 1 sollten also wahrscheinlicher sein als Ausgaben in der Nähe von Null.) Die passende Stammfunktion zu  $f$  ist schnell gefunden:  $F(x) = (x + x^2/2)/1.5$ . Und damit kann die Simulationsvorschrift angegeben werden:  $y \in [0, 1]$  gleichverteilt erzeugen, dann  $x \in [0, 1]$  so bestimmen, dass  $(x + x^2/2)/1.5 = y$ , dann  $x$  ausgeben. Die zugehörige quadratische Gleichung ist leicht zu lösen, wir erhalten  $x = -1 + \sqrt{1 + 3y}$ .

**3.** Es sei  $n \in \mathbb{N}$ . Die Abbildung  $x \mapsto x^n$  ist noch keine Dichtefunktion auf  $[0, 1]$ , denn das Integral ist  $1/(n+1)$ . Deswegen betrachten wir  $f(x) := (n+1)x^n$ . Es ist zu erwarten, dass bevorzugt Werte in der Nähe von 1 bei Abfragen erzeugt werden, und zwar umso extremer, je größer  $n$  ist.

Um diesen Raum zu simulieren, bestimmen wir zunächst die Funktion  $F(x) = \int_0^x (n+1)t^n dt = x^{n+1}$ . So erhalten wir die Simulationsvorschrift:  $y \in [0, 1]$  gleichverteilt wählen, dann  $x := \sqrt[n+1]{y}$  ausgeben. Dadurch werden aus den gleichverteilten  $y$  wirklich Werte, die überwiegend nahe bei 1 liegen.

**4.** Diesmal betrachten wir die *Exponentialverteilung* zum Parameter  $\lambda > 0$ . Es ist also  $\Omega = [0, \infty[$  und  $f(x) = \lambda e^{-\lambda x}$ . Die Funktion  $F$  ist durch  $F(x) = 1 - e^{-\lambda x}$  definiert, und deswegen lautet die Simulationsvorschrift: Erzeuge  $y$  gleichverteilt in  $[0, 1]$  und löse dann  $y = 1 - e^{-\lambda x}$  nach  $x$  auf; dann ist  $x$  auszugeben. Es folgt:  $x = -(\log(1 - y))/\lambda$ . Hier ist noch eine kleine *Vereinfachung* möglich: Die Wahrscheinlichkeit, dass  $y$  in einem Intervall  $[c, d]$  liegt, ist genauso groß wie die Wahrscheinlichkeit, dass  $y$  in  $[1 - d, 1 - c]$  liegt, denn beide Intervalle haben die gleiche Länge. Anders ausgedrückt: Die Wahrscheinlichkeiten für die Ausgabe  $y$  sind identisch mit den Wahrscheinlichkeiten für die Ausgabe  $1 - y$ . Deswegen

darf in unserem Simulationsverfahren  $y$  durch  $1 - y$  ersetzt werden, und wir erhalten die folgende *Vorschrift für die Simulation der Exponentialverteilung*:  $y$  gleichverteilt in  $[0, 1]$  erzeugen, dann  $x := -(\log y)/\lambda$  ausgeben<sup>20)</sup>.

Hier sind einige (auf drei Stellen gerundete) Simulationen im Fall  $\lambda = 1$ :

→  
Programm!

1.330, 0.992, 2.130, 0.854, 0.815, 0.783, 3.904, 0.309, 1.305, 0.817,  
 1.826, 0.583, 1.791, 0.451, 0.006, 0.996, 1.337, 0.975, 1.241, 1.354, 0.025,  
 1.221, 0.351, 0.319, 0.451, 0.088, 0.062, 0.891, 0.224, 0.484, 0.324, 0.097,  
 2.672, 1.417, 2.452, 1.963, 0.532, 0.064, 0.367, 0.096, 0.677, 0.025, ...

Die *Simulation der Normalverteilung* ist etwas schwieriger, denn explizit angebbare Stammfunktionen zu diesen Dichtefunktionen gibt es nicht. Man könnte zwar mit Approximationen arbeiten, doch wäre das recht schwerfällig. Es gibt aber andere Methoden.

Als Vorüberlegung erinnern wir daran, dass die verschiedenen Normalverteilungen durch Skalierungen auseinander hervorgehen<sup>21)</sup>: Erzeugt man ein  $x$  gemäß der Standard-Normalverteilung  $N(0, 1)$ , so verhält sich  $a + \sigma x$ , als wenn es gemäß  $N(a, \sigma^2)$  erzeugt worden wäre.

Das hat die erfreuliche Konsequenz, dass wir uns nur um das Simulieren der Standardnormalverteilung kümmern müssen. Hier sind die zwei am häufigsten verwendeten Verfahren, beide benutzen im Vorgriff Ergebnisse, die wir erst später beweisen werden:

#### *Simulation der Standardnormalverteilung, Version 1:*

Wenn man  $N(0, 1)$  zweimal abfragt, entsteht doch ein Tupel  $(x, y) \in \mathbb{R}^2$ . Mit welchen Wahrscheinlichkeiten entstehen diese Punkte, wie wahrscheinlich ist es zum Beispiel, dass  $(x, y)$  in einem bestimmten Rechteck  $R = [a, b] \times [c, d]$  liegt? Die (später zu begründende) Lösung: Es gibt wieder eine Dichtefunktion  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ , nämlich

$$f(x, y) = \frac{1}{2\pi} \exp(-(x^2 + y^2)/2),$$

und die gesuchte Wahrscheinlichkeit ist das Integral von  $f$  über  $R$ .

Trägt man über jedem  $(x, y)$  die Höhe  $f(x, y)$  ab, betrachtet man also den Graphen von  $f$ , so ergibt sich so etwas wie ein Sombrero. Er hat eine bemerkenswerte Eigenschaft, er ist *rotationssymmetrisch*. Deswegen ist die Darstellung in Polarkoordinaten  $(r, \phi)$  besonders einfach: Ist  $(x, y)$  durch  $(r, \phi)$  dargestellt, also  $(x, y) = (r \cos \phi, r \sin \phi)$ , so ist der Wert von  $f$  dort gleich  $(1/2\pi) \exp(-r^2/2)$ .

Wie wahrscheinlich ist es, dass  $(x, y)$  in einem „sehr dünnen“ Kreisring liegt (der Abstand zum Nullpunkt soll zwischen  $r$  und  $r + dr$  liegen)? Da die Dichtefunktion dort fast konstant ist, nämlich gleich  $(1/2\pi) \exp(-r^2/2)$ , ist die Wahrscheinlichkeit gleich „Produkt aus Flächeninhalt des Kreisrings und  $f$ -Wert“,

<sup>20)</sup>Lassen Sie sich nicht von dem Minuszeichen verwirren: Da  $y$  kleiner als Eins ist, ist der Logarithmus negativ. Die  $x$ -Ausgabe ist also positiv.

<sup>21)</sup>Vgl. Seite 53, der Beweis wird allerdings erst in Kapitel 8 auf Seite 246 gegeben.

also  $2\pi r \cdot dr \cdot (1/2\pi) \exp(-r^2/2) = r \exp(-r^2/2) \cdot dr =: g(r) \cdot dr$ . Anders ausgedrückt heißt das, dass der Abstand  $r$  zur Null von  $(x, y)$  die Dichtefunktion  $g$  hat. Damit ist  $r$  leicht zu simulieren, denn  $g$  hat die explizit angebbare Stammfunktion  $\exp(-r^2/2)$ , und die ist auch leicht zu invertieren. Und hat man erst einmal  $r$ , muss man nur noch auf dem Kreis mit dem Radius  $r$  einen Punkt gleichverteilt aussuchen: *Das* ist dann unsere  $(x, y)$ -Simulation, sowohl  $x$  als auch  $y$  können verwendet werden. Hier das Ganze noch einmal als „Rezept“:

- Erzeuge  $a$  gleichverteilt in  $[0, 1]$  ( $\dots a := \text{random}; \dots$ ).
- Finde  $r$  mit  $a = \exp(-r^2/2)$ , also  $r := \sqrt{-2 \log a}$ .
- Erzeuge ein gleichverteiltes  $b$  in  $[0, 2\pi]$  ( $\dots b := 2\pi \cdot \text{random}; \dots$ ).
- Definiere  $x := r \cos b$  und  $y := r \sin b$ . Ausgabe:  $x$  und  $y$ .

Die bei diesem Verfahren erzeugten  $x, y$  sind exakt standard-normalverteilt.

*Simulation der Standardnormalverteilung, Version 2:* Bei diesem Verfahren muss man nur Zufallszahlen addieren. Es ist nicht hundertprozentig exakt, die Genauigkeit ist allerdings für alle praktischen Zwecke ausreichend. Begründung für die Gültigkeit ist der *zentrale Grenzwertsatz*, den wir in Abschnitt 9.6 beweisen werden (siehe insbesondere Seite 253). Danach haben „richtig skalierte Überlagerungen“ von „vielen“ Zufallsergebnissen immer (fast) die gleichen Wahrscheinlichkeiten wie die Standardnormalverteilung. So gelangt man zu der folgenden Vorschrift:

- Erzeuge zwölf in  $[0, 1]$  gleichverteilte Zahlen, berechne die Summe und subtrahiere 6:  $\dots x := 0; \text{for } i := 1 \text{ to } 12 \text{ } x := x + \text{random}; x := x - 6; \dots$
- Dieses  $x$  ist unsere Ausgabe.

Das kann natürlich nicht exakt sein, da nur Zahlen in  $[-6, 6]$  produziert werden, doch sind die Wahrscheinlichkeiten für Ausgaben  $x$  mit  $|x| > 6$  so gering, dass dieser Nachteil praktisch keine Rolle spielt.

Simulationen mit dem zweiten Verfahren führten zu dem folgenden (auf drei Stellen gerundeten) Ergebnis:

|        |        |        |        |        |        |        |        |        |        |
|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 0.966  | -0.506 | 0.492  | 0.810  | -0.854 | -0.563 | 0.247  | -1.527 | -1.319 | 1.380  |
| -0.441 | 0.775  | -0.863 | -0.532 | -0.021 | -1.247 | -0.633 | -1.056 | -1.030 | -0.832 |
| -0.244 | 0.399  | 1.156  | -0.248 | 1.477  | 0.020  | 0.569  | -0.537 | 0.736  | ...    |

Wir haben in den letzten beiden Abschnitten eine Reihe von Simulationsmöglichkeiten kennengelernt. Damit lassen sich zu vorgegebenen Wahrscheinlichkeitsräumen  $(\Omega, \mathcal{E}, \mathbb{P})$  „Zufallsautomaten“ konstruieren, die Elementarereignisse so erzeugen, dass der Anteil der  $\omega$  in einem Ereignis  $E$  bei großen Versuchszahlengegen  $\mathbb{P}(E)$  geht.

Das macht es möglich, unbekannte  $\mathbb{P}(E)$  experimentell (wenigstens approximativ) zu bestimmen. Im Fall des Buffonschen Nadelexperiments und der stochastischen Flächenbestimmung wurde das schon in Abschnitt 2.2 bemerkt. Das

Verfahren ist aber universell einsetzbar. Wenn man zum Beispiel wissen möchte, wie wahrscheinlich es ist, dass bei einer Lotterziehung nur Zahlen zwischen 11 und 33 gezogen werden und man keine Lust hat zu rechnen, so kann man das „experimentell“ angehen: Erzeuge eine Million Lottotipps und prüfe nach, wie oft alle Zahlen zwischen 11 und 33 lagen. Dieser Anteil ist eine Approximation an die gesuchte Wahrscheinlichkeit<sup>22)</sup>.

Wir illustrieren das Verfahren noch durch ein weiteres Beispiel: Es soll der *Flächeninhalt des Einheitskreises* bestimmt werden. So könnte man – als Alternative zum Buffonschen Nadelexperiment – die Kreiszahl  $\pi$  „experimentell“ ermitteln. Dazu erzeugen wir  $n$  Punkte im Quadrat  $Q = \{(x, y) \mid x, y \in [-1, 1]\}$  durch  $x := 2 * \text{random} - 1$ ,  $y := 2 * \text{random} - 1$  und zählen, für wie viele die Bedingung  $x^2 + y^2 \leq 1$  erfüllt ist. Wenn das in  $m$  Fällen eintritt, sollte  $4m/n$  eine Approximation der Kreisfläche (also von  $\pi$ ) sein, denn  $Q$  hat die Fläche 4.

→  
Programm!

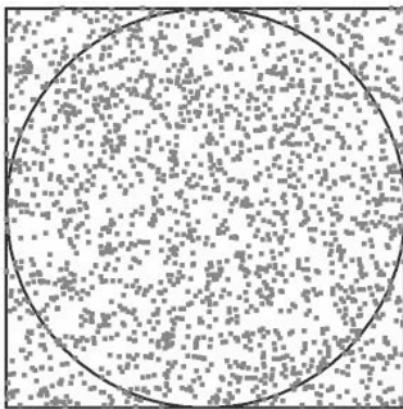


Bild 2.4.2: Stochastische  $\pi$ -Bestimmung.

Im Fall  $n = 10.000.000$  ergab sich der Wert 3.141384. Das ist noch nicht besonders gut, doch die Rechenzeit war nur der Bruchteil einer Sekunde. Und genau so schnell hätte man ein approximatives Ergebnis in Situationen bekommen, in denen man keine andere Möglichkeit hat, den wirklichen Wert zu ermitteln.

## 2.5 Ergänzungen

### Weitere Wahrscheinlichkeitsräume

In den vorigen Abschnitten haben Sie einige wichtige Beispiele von Wahrscheinlichkeitsräumen kennen gelernt. Der hier gewählte Zugang (diskrete Räume und die Definition mit Hilfe von Dichtefunktionen) hat den Vorteil, dass keine

<sup>22)</sup>Die Präzisierung und die theoretische Absicherung dieser Aussage werden in Abschnitt 8.2 nachgetragen.

besonderen mathematischen Vorkenntnisse erforderlich sind und dass man damit trotzdem in der Lage ist, viele wichtige Ideen der Wahrscheinlichkeitstheorie gut behandeln zu können.

Ein einfaches Beispiel für einen Wahrscheinlichkeitsraum, der *nicht diskret* ist und bei dem das Wahrscheinlichkeitsmaß auch *nicht durch eine Dichtefunktion definiert* ist, könnte so aussehen:

- Es ist  $\Omega = [0, 1]$ , und  $\mathcal{E}$  ist die  $\sigma$ -Algebra der Borelmengen.
- $\mathbb{P}$  wird so definiert: Für Intervalle<sup>23)</sup>  $E$ , die 0 enthalten, ist  $\mathbb{P}(E)$  gleich  $0.5 + 0.5 \int_E dx$ , für alle anderen ist  $\mathbb{P}(E) = 0.5 \int_E dx$ .

Dann ist  $(\Omega, \mathcal{E}, \mathbb{P})$  offensichtlich nicht diskret, denn  $\Omega$  ist nicht abzählbar. Es kann auch keine Dichtefunktion  $f$  geben, denn es ist einerseits  $\mathbb{P}(\{0\}) = 0.5$  und andererseits gilt  $\int_0^0 f(x) dx = 0$  für jedes  $f$ . Formal gesehen ist  $\mathbb{P}$  eine Konvexitätskombination einfacherer Wahrscheinlichkeitsmaße: Bezeichnet  $\mathbb{P}_1$  das zu 0 gehörige Punktmaß (vgl. Übungsaufgabe Ü1.2.3) und  $\mathbb{P}_2$  die Gleichverteilung auf  $[0, 1]$ , so ist  $\mathbb{P} = (\mathbb{P}_1 + \mathbb{P}_2)/2$ .

Dieser Raum kann auch bei naheliegenden Fragestellungen auftreten, wir werden ihn schon im nächsten Kapitel (auf Seite 77) wieder antreffen. Trotzdem wird es reichen, sich mit den in den vorstehenden Abschnitten behandelten Beispielen beschäftigt zu haben.

### Pseudozufallszahlen

Nun soll noch kurz beschrieben werden, *wie es Computer fertigbringen, den Zufall zu simulieren*. Bevor Sie weiterlesen, sollten Sie ein Gedankenexperiment machen: Wie würden *Sie* denn etwas Zufälliges erzeugen? Würfel? Münze? Das Kennzeichen des nächsten Autos, das vorbeikommt?

Sie werden schnell feststellen, dass das ein langwieriges Verfahren ist. Wünschenswert wären aber gewaltige Mengen von Zufallsergebnissen in kürzester Zeit. In der Frühzeit der Computer ist viel experimentiert worden, so wurden etwa Signale aus einem verrauschten Radiokanal zur Erzeugung des Zufalls verwendet. Heute arbeiten praktisch alle Rechner mit dem gleichen Verfahren. Dabei wählt man, ein für allemal, drei große natürliche Zahlen  $a_0, a_1$  und  $m$ . Man startet mit einer Zahl  $x_0$  und definiert dann eine Folge  $x_1, x_2, \dots$  von „Zufallszahlen“ induktiv durch die Vorschrift

$$x_{n+1} := a_0 x_n + a_1 \bmod m.$$

Dabei bezeichnet  $a \bmod b$  den Rest, der beim Teilen von  $a$  durch  $b$  übrig bleibt. (Zum Beispiel gilt  $22 \bmod 7 = 1$ .)

Bei geschickter Wahl von  $a_0, a_1, m$  und  $x_0$  entstehen auf diese Weise Zahlen, die von einer Zufallsfolge praktisch nicht zu unterscheiden sind. Diese „Pseudozufallszahlen“ lassen sich sehr schnell erzeugen, und man kann mit ihrer Hilfe so

---

<sup>23)</sup>Die gleiche Definition gilt für beliebige Borelmengen  $E$ , doch dazu müsste man sich vorher klar gemacht haben, was  $\int_E dx$  in diesem Fall bedeutet. Vgl. den Anhang zur Maß- und Integrationstheorie auf Seite 356.

gut wie alle Wahrscheinlichkeitsräume effektiv simulieren. Wir haben ja schon in den Abschnitten 2.3 und 2.4 gesehen, wie man den Zufall sozusagen „transformieren“ kann, dass man zum Beispiel eine Simulation der Gleichverteilung auf  $[0, 1]$  dazu verwenden kann, eine beliebige durch eine Dichte definierte Wahrscheinlichkeit zu simulieren.

## 2.6 Verständnisfragen

### Zu Abschnitt 2.1

#### *Sachfragen*

**S1:** Was ist ein diskreter Wahrscheinlichkeitsraum?

**S2:** Was ist ein Laplaceraum?

**S3:** Was versteht man unter einem Bernoulliraum?

**S4:** Wie ist die Poissonverteilung definiert?

**S5:** Wie ist die geometrische Verteilung definiert?

#### *Methodenfragen*

**M1:** Wahrscheinlichkeiten im Zusammenhang mit diskreten Wahrscheinlichkeitsräumen bestimmen können.

### Zu Abschnitt 2.2

#### *Sachfragen*

**S1:** Was ist eine Dichtefunktion?

**S2:** Wie kann mit Hilfe einer Dichtefunktion ein Wahrscheinlichkeitsraum definiert werden?

**S3:** Wie ist die Gleichverteilung auf einem beschränkten Intervall definiert?

**S4:** Wie kann man die Kreiszahl  $\pi$  approximativ à la Buffon bestimmen?

**S5:** Was ist ein Monte-Carlo-Verfahren?

**S7:** Was versteht man unter der Exponentialverteilung?

**S7:** Wie ist die Normalverteilung definiert?

#### *Methodenfragen*

**M1:** Nachprüfen können, ob eine vorgelegte Funktion eine Dichtefunktion ist.

**M2:** Wahrscheinlichkeiten im Zusammenhang mit Wahrscheinlichkeitsräumen, die durch eine Dichtefunktion definiert sind, bestimmen können.

### Zu Abschnitt 2.3

#### *Sachfragen*

**S1:** Welche Simulationsmöglichkeiten werden von jedem Computer zur Verfügung gestellt?

**S2:** Wie simuliert man einen beliebigen endlichen Wahrscheinlichkeitsraum?

**S3:** Wie simuliert man die Poissonverteilung?

**S4:** Wie simuliert man die geometrische Verteilung?

*Methodenfragen*

**M1:** Programme für die Simulation beliebiger diskreter Wahrscheinlichkeitsräume entwerfen können.

**M2:** Konkrete Wahrscheinlichkeiten im Zusammenhang mit diskreten Wahrscheinlichkeitsräumen durch Simulation approximativ bestimmen können.

### Zu Abschnitt 2.4

*Sachfragen*

**S1:** Was ist die zu einem Wahrscheinlichkeitsmaß gehörige Verteilungsfunktion?

**S2:** Welche Dichte hat ein Maß  $\mathbb{P}$ , für das die Verteilungsfunktion  $F_{\mathbb{P}}$  stetig differenzierbar ist?

**S3:** Angenommen, man kann die Gleichverteilung auf  $[0, 1]$  simulieren. Wie kann man dann einen Raum mit Dichtefunktion simulieren?

**S4:** Wie simuliert man die Exponentialverteilung?

**S5:** Wie simuliert man schnell und in guter Näherung die Normalverteilung?

*Methodenfragen*

**M1:** Programme für die Simulation von Wahrscheinlichkeitsräumen entwerfen können, die durch eine Dichtefunktion definiert sind und bei denen die Voraussetzungen „geeignet“ sind (was muss man für die Simulation ausrechnen können?).

**M2:** Konkrete Wahrscheinlichkeiten im Zusammenhang mit durch Dichtefunktionen definierten Wahrscheinlichkeitsräumen durch Simulation approximativ bestimmen können.

## 2.7 Übungsaufgaben

### Zu Abschnitt 2.1

**Ü2.1.1** Man werfe zwei (sechsseitige) Würfel  $W_1$  und  $W_2$ , wobei  $\mathbb{P}(W_1 = 1) = \mathbb{P}(W_1 = 2) = \mathbb{P}(W_1 = 3) = 1/9$ ,  $\mathbb{P}(W_1 = 4) = \mathbb{P}(W_1 = 5) = \mathbb{P}(W_1 = 6) = 2/9$  und  $\mathbb{P}(W_2 = 1) = \mathbb{P}(W_2 = 2) = \mathbb{P}(W_2 = 3) = 2/9$ ,  $\mathbb{P}(W_2 = 4) = \mathbb{P}(W_2 = 5) = \mathbb{P}(W_2 = 6) = 1/9$  sei. Mit welcher Wahrscheinlichkeit ist die Augensumme 4 bzw. 7?

**Ü2.1.2** Betrachte auf  $\mathbb{N}_0$  die Poissonverteilung zum Parameter  $\lambda$ . Hat dann die Menge der geraden oder die der ungeraden Zahlen die größere Wahrscheinlichkeit?

Tipp: Wie kann man die Differenz dieser Wahrscheinlichkeiten übersichtlicher mit Hilfe der e-Funktion schreiben?

**Ü2.1.3** Wie vorstehend, aber für die geometrische Verteilung zum Parameter  $q$  auf  $\mathbb{N}$ .

**Ü2.1.4** Jedes Wahrscheinlichkeitsmaß auf  $\{1, \dots, n\}$  kann mit einem Vektor  $p \in \mathbb{R}^n$  identifiziert werden: Dem Maß wird der Vektor  $p := (\mathbb{P}(\{1\}), \dots, \mathbb{P}(\{n\}))$  zugeordnet. Sei  $\mathcal{P}$  die Gesamtheit der so entstehenden Vektoren. Zeigen Sie, dass  $\mathcal{P}$  eine kompakte und konvexe Teilmenge des  $\mathbb{R}^n$  ist.

**Ü2.1.5** Es sei  $\mathcal{P}$  wie vorstehend. Beschreiben Sie die Extrempunkte von  $\mathcal{P}$  (vgl. Übungsaufgabe Ü1.3.4.)

**Ü2.1.6** Auf  $\mathbb{N}_0$  sei eine Poissonverteilung zum Parameter  $\lambda$  gegeben. Man weiß, dass  $\mathbb{P}(\{3\}) = \mathbb{P}(\{5\})$  gilt. Kann man daraus darauf schließen, wie groß  $\lambda$  ist? Wenn ja: Wie groß ist  $\lambda$ ?

## Zu Abschnitt 2.2

**Ü2.2.1** Es sei  $\Omega = [0, 1]$ , versehen mit der Dichtefunktion  $(\alpha + 1)x^\alpha$  mit einer Zahl  $\alpha > 0$ .

- a) Ist das wirklich eine Dichtefunktion?
- b) Man bestimme  $\alpha$  so, dass  $\mathbb{P}([0, 0.5]) = 0.001$  ist.

**Ü2.2.2** In einem Kreis mit Radius 1 wird eine Sehne „zufällig“ gewählt. Wie groß ist die Wahrscheinlichkeit, dass die Sehnenlänge  $S$  größer als 1 ist?

Man löse die Aufgabe, indem man den Begriff „zufällig“ auf folgende Arten präzisiert:

- a) Der Sehnenmittelpunkt ist gleichverteilt im Inneren des Kreises.
- b) Die Polarkoordinaten des Sehnenmittelpunktes sind gleichverteilt auf der Menge  $[0, 1] \times [0, 2\pi]$ .
- c) Aus Symmetriegründen denke man sich einen Sehnenendpunkt festgelegt; der andere soll dann gleichverteilt auf dem Kreisrand sein.

**Ü2.2.3** Sei  $p > 0$  eine vorgegebene Zahl. Unter welchen Bedingungen an  $p$  gibt es eine Zahl  $\lambda$ , so dass bezüglich der zugehörigen Exponentialverteilung das Intervall  $[1, 2]$  die Wahrscheinlichkeit  $p$  hat? Wie ist  $\lambda$  in diesem Fall zu wählen?

**Ü2.2.4** Sei  $\mathcal{P}$  die Menge der Wahrscheinlichkeitsmaße auf  $[0, 1]$ , die eine stetige Dichte haben. Man zeige, dass  $\mathcal{P}$  eine konvexe Menge ist (die Elemente von  $\mathcal{P}$  werden als Abbildungen von den Borelmengen in  $[0, 1]$  nach  $[0, 1]$  aufgefasst). Es soll weiter gezeigt werden, dass  $\mathcal{P}$  keine Extrempunkte hat.

**Ü2.2.5** Eine fiktive Geschichte: Ein Zeitgenosse von Buffon hat keine Stöckchen zur Hand, er wirft runde Bierdeckel auf den Dielenfußboden. (Dielenbreite  $d$ , Bierdeckeldurchmesser  $b$ .) Prüfen Sie, ob man aus der Wahrscheinlichkeit, dass der Bierdeckel eine Kante trifft, die Zahl  $\pi$  approximativ bestimmen kann.

Eine etwas anspruchsvollere Version der fiktiven Geschichte lautet: Ein Zeitgenosse von Buffon wirft quadratische Bierdeckel auf den Dielenfußboden. (Dielenbreite  $d$ , Bierdeckelseitenkante  $b$ .) Prüfen Sie, ob man aus der Wahrscheinlichkeit, dass der Bierdeckel eine Kante trifft, die Zahl  $\pi$  approximativ bestimmen kann.

**Ü2.2.6** Sei  $f : [a, b] \rightarrow [0, +\infty[$  eine stetige Dichtefunktion und  $r$  sei kleiner als  $b - a$ . Man zeige, dass es ein Teilintervall von  $[a, b]$  der Länge  $r$  mit maximaler Wahrscheinlichkeit gibt. Finden Sie auch ein Beispiel, wo es genau zwei derartige Intervalle gibt.

**Zu Abschnitt 2.3**

**Ü2.3.1** Schreiben Sie ein Programm, das Zufallspermutationen von  $\{1, \dots, n\}$  erzeugt. Bestimmen Sie dann mit einem Monte-Carlo-Verfahren die Wahrscheinlichkeit, dass in einer Zufallspermutation kein Element fixiert bleibt.

**Ü2.3.2** Wir haben gezeigt, wie man mit Hilfe der Gleichverteilung auf  $[0, 1]$  einen Laplace Raum auf  $\{1, \dots, n\}$  erzeugen kann. Geht das auch umgekehrt? Wenigstens in guter Näherung?

**Zu Abschnitt 2.4**

**Ü2.4.1** Finden Sie mit einem Monte-Carlo-Verfahren die Wahrscheinlichkeit von  $[1, 2]$  unter  $N(0, 1)$  und vergleichen Sie mit dem aus der Tabelle abgelesenen wirklichen Wert.

**Ü2.4.2** Beweisen Sie, dass die Verteilungsfunktion  $F_{\mathbb{P}}$  monoton steigend und von rechts stetig ist. (D.h. es gilt stets  $\lim_{n \rightarrow \infty} F_{\mathbb{P}}(x_n) = F_{\mathbb{P}}(x)$  für Folgen  $(x_n)$  mit  $x_n \geq x$  und  $x_n \rightarrow x$ .)

**Ü2.4.3**  $x \mapsto \sin x$  ist eine Dichtefunktion auf  $[0, \pi/2]$ . Schreiben Sie ein Programm, das Punkte aus dem entsprechenden Wahrscheinlichkeitsraum simuliert.

## **Teil II**

# **Wichtige Konzepte**

# Kapitel 3

## Zufallsvariable

Im zweiten Teil dieses Buches geht es um fundamentale Begriffe der Wahrscheinlichkeitstheorie: *Zufallsvariable* und *bedingte Wahrscheinlichkeiten*. Kapitel 3 könnte man unter das Leitmotiv „Kompression von Informationen“ stellen. Manchmal möchte man es nämlich gar nicht so genau wissen: Es ist zum Beispiel ziemlich unerheblich, welche sechs Lottozahlen am Sonnabend gezogen wurden, interessant ist doch eher, wie viele Richtige ich angekreuzt habe.

Um das zu präzisieren, werden in Abschnitt 3.1 *Zufallsvariable* eingeführt. Jede Zufallsvariable gibt Anlass zu einem neuen Wahrscheinlichkeitsraum, durch den die Wahrscheinlichkeiten der „komprimierten Informationen“ beschrieben werden. Das ist der Gegenstand von Abschnitt 3.2.

Wenn es um Zahlen geht, möchte man wissen, was denn „im Mittel“ zu erwarten ist und ob die Ergebnisse mehr oder weniger stark streuen werden: Die zugehörigen Definitionen – *Erwartungswert* und *Streuung* – finden Sie in Abschnitt 3.3. Dann gibt es in Abschnitt 3.4 ein *Zwischenspiel zur elementaren Kombinatorik*: Wie viele Möglichkeiten gibt es, dieses oder jenes zu tun? Das ist eine notwendige Vorbereitung, um in Situationen, die mit Laplaceräumen zusammenhängen, Wahrscheinlichkeiten ausrechnen zu können. Diese Rechnungen findet man in Abschnitt 3.5. Sie werden Überraschendes (*Geburtstagsparadoxon!*) finden, die deprimierend kleinen Wahrscheinlichkeiten für lohnende Gewinne im *Lotto* bestimmen können und *erste statistische Methoden* kennen lernen.

Das Kapitel endet in den Abschnitten 3.6, 3.7 und 3.8 mit Ergänzungen, Verständnisfragen und Übungsaufgaben.

### 3.1 Was ist eine Zufallsvariable?

Abbildungen lernt man schon in der ersten Woche des Studiums kennen. Manchmal „vergessen“ Abbildungen Informationen: Ist  $g : M \rightarrow N$  eine Abbildung und kennt man  $y = g(x)$ , so kann man in der Regel nicht mit Sicherheit auf  $x$  schließen. Meist ist die Information über  $x$  unwiederbringlich verloren, nur im

Fall injektiver Abbildungen (falls also aus  $x \neq y$  stets  $g(x) \neq g(y)$  folgt) ist  $x$  rekonstruierbar. Diesen Nachteil nimmt man aber oft bewusst in Kauf, um Wesentliches herauszuarbeiten.

Im Fall von Wahrscheinlichkeitsräumen  $(\Omega, \mathcal{E}, \mathbb{P})$  spielt das auch eine wichtige Rolle: Ein Elementarereignis  $\omega$  wird vom Zufall erzeugt, aber statt  $\omega$  gilt das Interesse  $X(\omega)$ , wobei  $X$  eine Abbildung von  $\Omega$  in irgendeine Menge  $C$  ist. Hier einige Beispiele:

- $\Omega$  besteht aus allen Lottotipps, und  $X(\omega)$  ist die Anzahl der Richtigen. In diesem Fall ist  $X : \Omega \rightarrow \{0, 1, 2, 3, 4, 5, 6\}$ .
- $\Omega$  beschreibe alle Möglichkeiten, mit zwei fairen Würfeln zu würfeln. Es ist also  $\Omega = \{1, 2, 3, 4, 5, 6\} \times \{1, 2, 3, 4, 5, 6\}$ . Und  $X : \Omega \rightarrow \{2, \dots, 12\}$  soll einem  $\omega = (i, j)$  die Augensumme zuordnen:  $X(i, j) := i + j$ .
- $\Omega$  ist die Einheitskreisscheibe im  $\mathbb{R}^2$  mit der Gleichverteilung, und die Abbildung  $X : \Omega \rightarrow [0, 1]$  ordnet jedem  $\omega$  den Abstand zum Nullpunkt zu:  $X(\omega) := ||\omega||$ .

Solche auf  $\Omega$  definierten Abbildungen  $X$  werden *Zufallsvariable* genannt (die präzise Definition folgt gleich). Zwar variiert – in Abhängigkeit davon, welches  $\omega$  erzeugt wurde –  $X(\omega)$  mit dem Zufall, an  $X$  selber ist aber nichts Zufälliges und es variiert auch nichts, es ist einfach eine Abbildung.

Es ist noch ein technischer Aspekt zu berücksichtigen, denn im Zusammenhang mit Zufallsvariablen interessieren wieder Wahrscheinlichkeiten: Wie wahrscheinlich etwa ist es in den Beispielen, dass ich mindestens fünf Richtige habe, oder dass mit zwei Würfeln eine ungerade Zahl gewürfelt wird, oder dass der Abstand zum Nullpunkt zwischen 0.5 und 0.7 liegt?

Allgemein: Ist  $X : \Omega \rightarrow C$  und ist  $F \subset C$ , so könnte es wichtig sein zu wissen, mit welcher Wahrscheinlichkeit  $X(\omega)$  in  $F$  liegen wird. Das ist sicher gleich  $\mathbb{P}(\{\omega \mid X(\omega) \in F\})$ , doch damit das definiert ist, muss  $\{\omega \mid X(\omega) \in F\}$  Ereignis sein, also zu  $\mathcal{E}$  gehören. Es wird meist nicht notwendig sein, das für alle  $F \subset C$  zu fordern. Mindestens sollten diejenigen  $F$  zulässig sein, die auch in den Beispielen in Kapitel 2 auftraten. So gelangt man zu der folgenden

**Definition 3.1.1.** Es sei  $(\Omega, \mathcal{E}, \mathbb{P})$  ein Wahrscheinlichkeitsraum und  $C$  eine Menge, und auf  $C$  sei eine  $\sigma$ -Algebra  $\mathcal{F}$  vorgegeben. Wir vereinbaren, dass  $\mathcal{F}$  die Potenzmenge von  $C$  ist, wenn  $C$  endlich oder abzählbar ist, und dass  $\mathcal{F}$  aus den Borelmengen in  $C$  besteht, wenn  $C = \mathbb{R}$  oder  $C = \mathbb{R}^n$  gilt.

Eine  $C$ -wertige Zufallsvariable ist dann eine Abbildung  $X : \Omega \rightarrow C$ , so dass  $\{\omega \mid X(\omega) \in F\}$  für alle  $F \in \mathcal{F}$  zu  $\mathcal{E}$  gehört.

*Bemerkungen und Beispiele:*

**1.** Bald werden wir, nach dem Beweis einiger Ergebnisse, eine *Faustregel* formulieren: „Alles, was man sinnvoll definieren kann, ist eine Zufallsvariable“. Sie

machen also nicht allzu viel falsch, wenn Sie vorläufig *jede* Abbildung als Zufallsvariable auffassen und sich erst nach und nach an die technischen Aspekte im Zusammenhang mit dieser Definition gewöhnen.

**2.** Der Ausdruck  $\{\omega \mid X(\omega) \in F\}$  wird oft vorkommen. In der Mengenlehre ist dafür das Symbol  $X^{-1}(F)$  üblich, in der Wahrscheinlichkeitstheorie schreibt man oft auch kurz  $\{X \in F\}$ .

**3.** Drei Beispiele für Zufallsvariable haben wir schon vor der Definition kennen gelernt. Dass das dritte Beispiel auch im strengen Sinn eine Zufallsvariable ist, folgt aus der Stetigkeit von  $X$ : Solche Funktionen sind immer Zufallsvariable (s.u., Korollar 3.1.3).

**4.** Sei  $E$  ein Ereignis in  $\Omega$ . Wir definieren  $X : \Omega \rightarrow \mathbb{R}$  als *Indikatorfunktion* der Menge  $E$ : Es ist  $X(\omega) := 1$ , falls  $\omega \in E$  und  $X(\omega) := 0$  sonst. Für diese Indikatorfunktion werden wir im Folgenden auch  $\chi_E$  (gesprochen „chi  $E$ “) schreiben.

Ist  $B \subset \mathbb{R}$  eine Borelmenge, so können als  $\{\chi_E \in B\}$  nur vier Fälle auftreten: Diese Menge kann die leere Menge,  $E$ ,  $\Omega \setminus E$  oder  $\Omega$  sein, je nachdem, welche der Zahlen 0, 1 zu  $B$  gehören. Da alle diese Mengen in  $\mathcal{E}$  liegen, ist  $\chi_E$  eine Zufallsvariable.

Die Überlegung kann auch umgekehrt werden: Betrachtet man  $\chi_E$  für ein  $E$ , das *kein* Ereignis ist, so wird  $\chi_E$  auch keine Zufallsvariable sein. (Warum nicht?)

Zufallsvariable werden sehr intensiv untersucht werden, deswegen verzichten wir auf weitere spezielle Beispiele. Wir beginnen unsere Untersuchungen mit einigen *allgemeinen Ergebnissen*. Zunächst zeigen wir einen *Charakterisierungssatz*:

**Satz 3.1.2.** *Sei  $(\Omega, \mathcal{E}, \mathbb{P})$  ein Wahrscheinlichkeitsraum.*

(i) *Ist  $C$  höchstens abzählbar, so ist eine Abbildung  $X : \Omega \rightarrow C$  genau dann eine Zufallsvariable, wenn  $\{X = c\} (= \{\omega \mid X(\omega) = c\})$  für alle  $c \in C$  zu  $\mathcal{E}$  gehört.*

(ii) *Eine Abbildung  $X : \Omega \rightarrow \mathbb{R}$  ist genau dann eine Zufallsvariable, wenn*

$$\{X \geq a\} (= \{\omega \mid X(\omega) \geq a\})$$

*für alle  $a \in \mathbb{R}$  Ereignis ist.*

**Beweis:** (i) Für die nichttriviale Richtung sei  $F \subset C$  beliebig. Die Menge  $\{X \in F\}$  ist Vereinigung der  $\{X = c\}$ , wobei  $c$  alle Punkte aus  $F$  durchläuft. Das sind höchstens abzählbar viele Mengen, und alle gehören nach Voraussetzung zu  $\mathcal{E}$ . Da  $\mathcal{E}$  eine  $\sigma$ -Algebra ist, liegt auch  $\{X \in F\}$  in  $\mathcal{E}$ .

(ii) Hier werden erstmals die in Abschnitt 1.6 hergeleiteten Beweismethoden wichtig. Damit ist der Beweis ganz kurz.

Eine Beweisrichtung ist trivial: Es ist doch  $\{X \geq a\} = \{X \in [a, +\infty[\}$ , und alle Intervalle  $[a, +\infty[$  sind Borelmengen. Deswegen ist  $\{X \geq a\}$  ein Ereignis.

Für den Beweis der Umkehrung wird angenommen, dass alle  $\{X \geq a\}$  Ereignisse sind, und es ist zu zeigen, dass dann sogar  $\{X \in B\}$  für jede Borelmenge

$B \subset \mathbb{R}$  zu  $\mathcal{E}$  gehört. Dazu wollen wir die Sätze 1.5.2 und 1.6.2 kombinieren. Wir beginnen mit einer Definition:  $\mathcal{B}_X$  soll das System derjenigen Borelmengen  $B$  von  $\mathbb{R}$  sein, für die  $\{X \in B\} \in \mathcal{E}$  gilt. Beachte dann:

- $\mathcal{B}_X$  enthält alle  $[a, +\infty[$ : Das ist die Voraussetzung.
- $\mathcal{B}_X$  ist eine  $\sigma$ -Algebra. Das folgt aus elementaren mengentheoretischen Überlegungen und der Tatsache, dass  $\mathcal{E}$  eine  $\sigma$ -Algebra ist. (Ist zum Beispiel  $B \in \mathcal{B}_X$ , so beachte, dass  $\{X \in (\mathbb{R} \setminus B)\} = \Omega \setminus \{X \in B\}$  gilt; deswegen ist auch  $\mathbb{R} \setminus B$  in  $\mathcal{B}_X$ .)
- Die  $[a, +\infty[$  erzeugen die Borelmengen von  $\mathbb{R}$  (Satz 1.5.2).

Und damit sind wir fertig, das Beweisprinzip finden Sie in Satz 1.6.2.  $\square$

(Zusatz: Genau so könnte man Varianten dieser Charakterisierung zeigen. Es reicht zum Beispiel vorauszusetzen, dass  $\{X \geq a\}$  für alle rationalen  $a$  Ereignis ist oder dass alle  $\{X > a\}$  zu  $\mathcal{E}$  gehören.)

Der Satz hat ein nützliches

**Korollar 3.1.3.** *Es sei  $\Omega \subset \mathbb{R}^n$  eine Borelmenge, die  $\sigma$ -Algebra der Ereignisse bestehe aus allen in  $\Omega$  enthaltenen Borelmengen. Ist dann  $X : \Omega \rightarrow \mathbb{R}$  stetig, so ist  $X$  eine Zufallsvariable.*

**Beweis:** Für beliebige  $a$  ist  $\{X \geq a\}$  wegen der Stetigkeit von  $X$  eine abgeschlossene Menge und deswegen eine Borelmenge.  $\square$

Wie in anderen Bereichen der Mathematik auch gelten für Zufallsvariable *Permanenzsätze*: Aus schon als Zufallsvariable erkannten Abbildungen entstehen durch die üblichen Operationen neue Beispiele.

**Satz 3.1.4.** *Es sei  $(\Omega, \mathcal{E}, \mathbb{P})$  ein Wahrscheinlichkeitsraum. Summen, Produkte, Vielfache, Suprema und punktweise Limites von Folgen von Zufallsvariablen sind wieder Zufallsvariable. Genauer: Es seien  $X, Y, X_1, X_2, \dots : \Omega \rightarrow \mathbb{R}$  Zufallsvariable. Dann sind auch die folgenden Abbildungen Zufallsvariable:*

- (i)  $cX$  für jedes  $c \in \mathbb{R}$ .
- (ii)  $X + Y$ .
- (iii)  $X \cdot Y$ .
- (iv)  $X/Y$ , falls  $Y$  nirgendwo verschwindet (falls also  $Y(\omega) \neq 0$  für alle  $\omega$  gilt).
- (v) Die Abbildung  $\sup_{n \in \mathbb{N}} X_n : \omega \mapsto \sup_{n \in \mathbb{N}} X_n(\omega)$ , falls dieses Supremum<sup>1)</sup> überall endlich ist. (Eine entsprechende Aussage gilt für  $\inf_{n \in \mathbb{N}} X_n$ .)
- (vi) Die Abbildung  $\omega \mapsto \lim_n X_n(\omega)$ , wenn für alle  $\omega$  der Limes  $\lim_n X_n(\omega)$  in  $\mathbb{R}$  existiert.

---

<sup>1)</sup>Die Definition und die wichtigsten Eigenschaften des Supremums können Sie im Anhang auf Seite 360 noch einmal nachlesen.

**Beweis:** Die Beweise sind alle ähnlich, man zeigt – mitunter etwas trickreich – dass die Urbilder aller  $[a, +\infty[$  oder aller  $]a, +\infty[$  Ereignisse sind.

(i) Ist  $c > 0$ , so ist  $\{cX > a\} = \{X > a/c\}$ . Im Fall  $c = 0$  ist  $\{cX > a\}$  entweder leer oder gleich  $\Omega$ , und im Fall  $c < 0$  ist  $\{cX > a\} = \{X < a/c\}$ , also auch Ereignis.

(ii) Hier wird wichtig, dass die rationalen Zahlen in  $\mathbb{R}$  dicht liegen. Genauer: Gilt für zwei reelle Zahlen  $x, y$  die Ungleichung  $x + y > a$ , so kann man  $x$  bzw.  $y$  zu rationalen Zahlen  $p$  bzw.  $q$  so verkleinern, dass auch noch  $p + q > a$  richtig ist. Das führt zu der Gleichung

$$\{X + Y > a\} = \bigcup_{p, q \in \mathbb{Q}, p+q>a} \{X > p\} \cap \{Y > q\},$$

und da das eine abzählbare Vereinigung von Ereignissen ist, liegt die Menge  $\{X + Y > a\}$  in  $\mathcal{E}$ .

(iii) Wir zeigen das zunächst für den Fall  $X = Y$ . Da beachten wir: Die Menge  $\{X^2 \geq a\}$  ist gleich  $\{X \geq \sqrt{a}\} \cup \{X \leq -\sqrt{a}\}$  im Fall  $a > 0$  und gleich  $\Omega$  im Fall  $a \leq 0$ , sie ist also ein Ereignis. Der allgemeine Fall wird darauf zurückgeführt: Es ist  $X \cdot Y = ((X + Y)^2 - X^2 - Y^2)/2$ , und aufgrund der schon bewiesenen Ergebnisse ist die rechts stehende Funktion eine Zufallsvariable.

(iv) Wegen (iii) und  $X/Y = X \cdot (1/Y)$  reicht es zu zeigen, dass  $1/Y$  Zufallsvariable ist. Sei zunächst  $a > 0$ . Dann ist  $\{1/Y > a\} = \{Y > 0\} \cap \{Y < 1/a\}$ . Der Fall  $a < 0$  wird ähnlich behandelt, und wenn  $a = 0$  zu untersuchen ist, beachte man, dass  $\{1/Y > 0\} = \{Y > 0\}$  gilt.

(v) Ein Supremum einer Menge von Zahlen ist genau dann größer als  $a$ , wenn eine der Zahlen größer als  $a$  ist. Das heißt

$$\{\sup_n X_n > a\} = \bigcup_n \{X_n > a\}.$$

$\{\sup_n X_n > a\}$  ist damit als abzählbare Vereinigung von Ereignissen selber ein Ereignis.

(vi) Für konvergente Folgen  $(a_n)$  gilt  $\lim a_n = \sup_{m \in \mathbb{N}} \inf_{n \geq m} a_n$ . (Das ist gerade die Gleichheit von Limes und Limes inferior.) Damit ist  $\lim_n X_n(\omega) = \sup_m \inf_{n \geq m} X_n(\omega)$  für jedes  $\omega$ , und wegen (v) ist damit alles gezeigt.  $\square$

Durch Kombination des Korollars und des Satzes kann damit eine unübersehbare Fülle von Abbildungen als Zufallsvariable erkannt werden. In diesem Sinn ist die oben angegebene „Faustregel“ zu verstehen: „Praktisch alles, was man konkret definieren kann, ist Zufallsvariable“.

## 3.2 Induzierte Wahrscheinlichkeitsräume

Wir hatten die eher technische Bedingung an Zufallsvariable („stets ist  $\{X \in F\}$  Ereignis“) damit begründet, dass die Wahrscheinlichkeiten der Mengen  $\{X \in F\}$  definiert sein sollen. Das kann nun umgesetzt werden:

**Satz 3.2.1.** Es sei  $(\Omega, \mathcal{E}, \mathbb{P})$  ein Wahrscheinlichkeitsraum,  $C$  eine Menge, auf der eine  $\sigma$ -Algebra  $\mathcal{F}$  definiert ist, und  $X : \Omega \rightarrow C$  eine Zufallsvariable. Definiert man dann  $\mathbb{P}_X : \mathcal{F} \rightarrow [0, 1]$  durch  $\mathbb{P}_X(F) := \mathbb{P}(\{X \in F\})$ , so ist  $\mathbb{P}_X$  ein Wahrscheinlichkeitsmaß.  $\mathbb{P}_X$  heißt das durch  $X$  induzierte Wahrscheinlichkeitsmaß auf  $(C, \mathcal{F})$ , und  $(C, \mathcal{F}, \mathbb{P}_X)$  wird der durch  $X$  induzierte Wahrscheinlichkeitsraum genannt.

Unter der Verteilungsfunktion  $F_X : \mathbb{R} \rightarrow [0, +\infty[$  von  $X$  verstehen wir die Verteilungsfunktion von  $\mathbb{P}_{\mathbb{P}_X}$ , d.h. es ist  $F_X(x) := \mathbb{P}(\{X \leq x\})$ .

**Beweis:** Diese Aussage folgt direkt aus der Tatsache, dass  $\mathbb{P}$  ein Wahrscheinlichkeitsmaß ist, zusätzlich muss man zwei einfache mengentheoretische Tatsachen beachten: Wegen  $\{X \in C\} = \Omega$  ist  $\mathbb{P}_X(C) = 1$ , und da für disjunkte Mengen  $F_1, F_2, \dots$  in  $\mathcal{F}$  auch die Mengen  $\{X \in F_1\}, \{X \in F_2\}, \dots$  disjunkt sind, gilt

$$\begin{aligned}\mathbb{P}_X\left(\bigcup_i F_i\right) &= \mathbb{P}(\{X \in \bigcup_i F_i\}) \\ &= \mathbb{P}\left(\bigcup_i \{X \in F_i\}\right) \\ &= \sum_i \mathbb{P}(\{X \in F_i\}) \\ &= \sum_i \mathbb{P}_X(F_i).\end{aligned}$$

□

Wir betrachten einige *elementare Beispiele*:

**1.** Ist  $X$  die Zufallsvariable „Augensumme von zwei fairen Würfeln“, so wird dadurch  $\{2, 3, \dots, 12\}$  zu einem Wahrscheinlichkeitsraum (vgl. Seite 72). Es ist zum Beispiel  $\mathbb{P}_X(\{2\}) = 1/36$ , denn  $\{X = 2\} = \{(1, 1)\}$ , und  $\mathbb{P}(\{(1, 1)\}) = 1/36$ . Ganz analog ermittelt man  $\mathbb{P}_X(4) = 3/36$ , denn  $\{X = 4\} = \{(1, 3), (2, 2), (3, 1)\}$ , und dieses Ereignis hat – da wir die Gleichverteilung betrachten – die Wahrscheinlichkeit  $3/36$ .

**2.** Sei  $(\Omega, \mathcal{E}, \mathbb{P})$  ein Wahrscheinlichkeitsraum,  $E$  ein Ereignis und  $X = \chi_E$  die Indikatorfunktion von  $E$  (vgl. Seite 73). Das auf  $\{0, 1\}$  induzierte Wahrscheinlichkeitsmaß ist die Bernoulliverteilung auf  $\{0, 1\}$  mit  $p = \mathbb{P}_X(\{1\}) = \mathbb{P}(E)$ .

**3.** Welches Maß induziert die Zufallsvariable  $X = \text{„Abstand zum Nullpunkt“}$  auf der Einheitskreisscheibe? Es ist sicher ein Wahrscheinlichkeitsmaß auf  $[0, 1]$ . Um es genauer zu bestimmen, berechnen wir die Verteilungsfunktion. Es ist  $F_X(x)$  für  $x \in [0, 1]$  die Wahrscheinlichkeit der Kreisscheibe mit Radius  $x$ , also gleich  $(\pi x^2)/\pi = x^2$ . Damit hat nach Lemma 2.4.2 das Maß  $\mathbb{P}_X$  die Dichtefunktion  $2x$ .

Im zweiten Beispiel haben wir den Bildraum  $C$  als  $\{0, 1\}$  gewählt. Man hätte natürlich auch eine beliebige Obermenge von  $\{0, 1\}$  als  $C$  deklarieren können, etwa  $\{0, 1, 2\}$ , oder  $[0, 1]$  oder  $\mathbb{R}$ . Im ersten Fall wäre ein diskreter Wahrscheinlichkeitsraum entstanden, wobei  $\{2\}$  Wahrscheinlichkeit Null hätte. Bei  $C = [0, 1]$

und  $C = \mathbb{R}$  wäre  $\mathbb{P}_X$  auf der jeweiligen  $\sigma$ -Algebra der Borelmengen definiert.  $\mathbb{P}(B)$  würde die Werte 0 bzw.  $p$  bzw.  $1-p$  bzw. 1 annehmen, je nachdem, ob gilt:  $0 \notin B$  und  $1 \notin B$ ; bzw.  $0 \notin B$  und  $1 \in B$ ; bzw.  $0 \in B$  und  $1 \notin B$ ; bzw.  $0, 1 \in B$ . Alles, was außerhalb von  $\{0, 1\}$  liegt, wird von  $\mathbb{P}_X$  sozusagen „nicht gesehen“, und deswegen untersucht man in der Regel Situationen, in denen  $C = X(\Omega)$  gilt.

Wir berechnen noch  $\mathbb{P}_X$  für unsere wichtigen Beispielklassen. Sei zunächst  $\Omega$  höchstens abzählbar, es geht also um diskrete Wahrscheinlichkeitsräume. Ist dann  $X : \Omega \rightarrow C$  surjektiv, so wird auch  $C$  höchstens abzählbar sein, und deswegen ist wieder ein diskreter Raum zu erwarten.  $\mathbb{P}_X$  ist leicht zu berechnen:

**Satz 3.2.2.** *Das induzierte Wahrscheinlichkeitsmaß  $\mathbb{P}_X$  ist im diskreten Fall durch*

$$\mathbb{P}_X(\{c\}) = \mathbb{P}(\{X = c\}) = \sum_{\substack{\omega \in \Omega \\ X(\omega) = c}} \mathbb{P}(\{\omega\})$$

(für  $c \in C$ ) definiert.

**Beweis:** Das ist klar, da die  $\mathbb{P}(E)$  durch Aufsummieren der Wahrscheinlichkeiten der  $\{\omega\}$  mit  $\omega \in E$  berechnet werden können.  $\square$

Für Räume mit Dichten ist die Situation komplizierter. Die vorstehenden Beispiele zeigen, dass alles Mögliche passieren kann: Der induzierte Wahrscheinlichkeitsraum kann diskret sein oder durch eine Dichtefunktion definiert sein. Es kann aber auch sein, dass er weder diskret ist noch eine Dichte hat: Als Beispiel betrachte man  $X : x \mapsto \max\{0, x\}$  auf  $[-1, 1]$  mit der Gleichverteilung.  $\mathbb{P}_X$ , definiert auf den Borelmengen von  $[0, 1]$ , ist dann wie folgt erklärt:  $\mathbb{P}_X([0, d])$  ist gleich  $(1 + d)/2$ , und für  $c > 0$  ist  $\mathbb{P}_X([c, d]) = (d - c)/2$ .

Es gibt aber eine spezielle Klasse von Zufallsvariablen, bei denen  $\mathbb{P}_X$  eine Dichtefunktion hat, die man sogar konkret angeben kann:

**Satz 3.2.3.** *[a, b] sei durch eine stetige Dichtefunktion f zu einem Wahrscheinlichkeitsraum gemacht worden. Weiter sei  $X : [a, b] \rightarrow [c, d]$  eine Zufallsvariable. Wir setzen voraus, dass X stetig differenzierbar und bijektiv ist und überall eine positive Ableitung hat; insbesondere ist X also streng monoton steigend, und die inverse Abbildung  $X^{-1}$  ist stetig differenzierbar.*

Dann hat  $\mathbb{P}_X$  auf  $[c, d]$  eine Dichtefunktion h, die folgendermaßen gefunden werden kann: Bestimme die inverse Abbildung  $X^{-1} : [c, d] \rightarrow [a, b]$  und setze dann

$$h(y) := f(X^{-1}(y))(X^{-1})'(y).$$

**Beweis:** Wir haben, um nicht durcheinander zu kommen, die Elemente aus  $[a, b]$  bzw. aus  $[c, d]$  mit  $x$  bzw.  $y$  bezeichnet. So kann man auch die zu  $X$  inverse Abbildung  $X^{-1}$  leicht durch Auflösen der Gleichung  $X(x) = y$  finden<sup>2)</sup>.

<sup>2)</sup>Ist etwa  $X(x) = (3 + 2\sqrt{x})/4$ , so führt  $y = (3 + 2\sqrt{x})/4$  auf  $x = (2y - 1.5)^2$ ; es ist also  $X^{-1}(y) = (2y - 1.5)^2 = 4y^2 - 6y + 2.25$ . Beim Ableiten ist das  $y$  als Variable aufzufassen, in diesem Beispiel wäre  $(X^{-1})'(y) = 8y - 6$ .

Wir definieren  $h$  wie im Satz, es ist dann zu zeigen, dass

$$\mathbb{P}_X([c', d']) = \int_{c'}^{d'} h(y) dy$$

für jedes Teilintervall  $[c', d']$  von  $[c, d]$  gilt.

Sei ein solches Intervall vorgegeben. Wir bestimmen  $a', b' \in [a, b]$  so, dass  $X(a') = c'$  und  $X(b') = d'$  gilt: setze dazu  $a' := X^{-1}(c')$  und  $b' = X^{-1}(d')$ . Nach Definition ist

$$\mathbb{P}_X([c', d']) = \mathbb{P}(\{X \in [c', d']\}) = \mathbb{P}([a', b']),$$

denn wegen der Monotonie von  $X$  gilt  $\{X \in [c', d']\} = [a', b']$ . Es ist also nachzuweisen, dass

$$\int_{c'}^{d'} h(y) dy = \int_{a'}^{b'} f(x) dx.$$

Das kann man so einsehen<sup>3)</sup>: Wähle eine Stammfunktion  $F$  zu  $f$ , es ist also  $F' = f$ , und  $\int_{a'}^{b'} f(x) dx = F(b') - F(a')$ . Nach der Kettenregel ist  $F \circ X^{-1}$  (als Funktion von  $y$  geschrieben) eine Stammfunktion zu  $h$ , und deswegen ist

$$\int_{c'}^{d'} h(y) dy = (F \circ X^{-1})(d') - (F \circ X^{-1})(c') = F(b') - F(a') = \int_{a'}^{b'} f(x) dx.$$

Damit ist alles gezeigt. □

Zu diesem Satz gibt es naheliegende *Verallgemeinerungen*.

- Man kann – bei gleichem Beweis –  $\Omega = [a, b]$  durch  $\Omega = [a, +\infty[$  oder  $\Omega = \mathbb{R}$  ersetzen.
- Bei Räumen mit Dichten haben einpunktige Mengen die Wahrscheinlichkeit Null. Wenn also die Voraussetzungen des Satzes an einer einzigen Stelle verletzt sind, sollte er immer noch gelten. So kann man zum Beispiel auch für  $X(x) = x^n$  auf  $[0, 1]$  (versehen mit der Gleichverteilung oder irgendeiner anderen Dichtefunktion) anwenden.  $X^{-1}(y) = \sqrt[n]{y}$  ist bei 0 nicht differenzierbar. Trotzdem kann man mit  $h(y) = (X^{-1})'(y) = (1/n)y^{(1-n)/n}$  die richtige Dichtefunktion für  $\mathbb{P}_X$  finden. Im Fall der Gleichverteilung ergibt sich damit  $h(y) = (1/n)y^{(1-n)/n}$ , das ist eine Funktion, die uneigentlich integrierbar mit Integral Eins ist und deswegen völlig berechtigt als Dichtefunktion verwendet werden kann.
- Wenn  $X$  differenzierbar ist und *streng monoton fällt*, gibt es analoge Ergebnisse, vgl. Übung 3.2.5.

---

<sup>3)</sup>Es folgt auch direkt aus der Formel für die Integration durch Substitution.

Es folgen zwei *Beispiele*, in denen der vorstehende Satz angewendet wird:

**1.**  $\Omega = [0, 1]$  sei mit der Gleichverteilung versehen. Wir betrachten  $X(x) = e^x$ , eine Abbildung von  $[0, 1]$  nach  $[1, e]$ . Die Voraussetzungen des Satzes sind offensichtlich erfüllt. Die Rechnung sieht dann so aus:

- $X(x) = y$  nach  $y$  auflösen, um  $X^{-1}$  zu ermitteln:  $X^{-1}(y) = x = \log y$ .
- Nach  $y$  ableiten:  $(X^{-1})'(y) = 1/y$ .
- Die Funktion  $h$  des vorigen Satzes bestimmen:  $h(y) = 1/y$ .
- Von der Variablen  $y$  zur Variablen  $x$  übergehen, falls das einem vertrauter vorkommt: Wir wissen damit, dass die Funktion  $1/x$  auf dem Intervall  $[1, e]$  die Dichtefunktion zu  $\mathbb{P}_X$  ist.

**2.**  $[1, 2]$  sei mit der Dichtefunktion  $f(x) = 3x^2/7$  versehen, und wir sind an  $X(x) := x^5$  interessiert. Es ist  $X^{-1}(y) = \sqrt[5]{y}$ , also  $(X^{-1})'(y) = 1/(5\sqrt[5]{y^4})$ . Die Dichtefunktion von  $\mathbb{P}_X$  auf dem Intervall  $X([1, 2]) = [1, 32]$  ist also

$$h(y) = f(X^{-1}(y))(X^{-1})'(y) = \frac{3}{7}(\sqrt[5]{y})^2 \left( \frac{1}{5\sqrt[5]{y^4}} \right) = \frac{3}{35\sqrt[5]{y^2}}.$$

Wer  $x$  als Variable vorzieht, kann noch  $y$  in  $x$  umtaufen: Die Dichtefunktion auf  $[1, 32]$  für  $\mathbb{P}_X$  lautet  $3/(35\sqrt[5]{x^2})$ .

*Schlussbemerkung:* Implizit sind induzierte Wahrscheinlichkeiten schon in Abschnitt 2.4 aufgetreten. Da haben wir in Satz 2.4.3 dadurch Simulationsvorschriften konstruiert, dass wir Zufallsvariable  $X$  auf  $[0, 1]$  (mit der Gleichverteilung) so definiert haben, dass der zu simulierende Wahrscheinlichkeitsraum gerade der durch  $X$  induzierte ist. Genauer:  $[0, 1]$  trägt die Dichtefunktion  $\tilde{f}(x) = 1$  und  $[a, b]$  eine beliebige Dichtefunktion  $f$ . Wir haben dann  $X : [0, 1] \rightarrow [a, b]$  so finden wollen, dass  $\mathbb{P}_X$  die Dichte  $f$  hat. Aufgrund des vorstehenden Satzes<sup>4)</sup> heißt das: Finde  $X$  so, dass  $(X^{-1})' = f$  gilt.  $X$  muss damit die inverse Abbildung zu einer Stammfunktion von  $f$  sein, und das ist eine Umformulierung von Satz 2.4.3.

### 3.3 Erwartungswert, Varianz und Streuung

In diesem Abschnitt wird es um *reellwertige* Zufallsvariable gehen:  $X : \Omega \rightarrow \mathbb{R}$ . Jede Zufallsabfrage erzeugt also eine Zahl  $X(\omega)$ , und es könnte interessant sein zu erfahren, was denn „im Mittel“ zu erwarten ist. Genauer: Wenn man „oft“ – etwa  $m$  Mal – abgefragt und für die  $X(\omega)$  die Ergebnisse  $x_1, \dots, x_m$  erhalten hat, so möchte man wissen, was sich über den Mittelwert  $(x_1 + \dots + x_m)/m$  aussagen lässt. Beispiele, wo das wichtig sein wird, sind schnell gefunden: Wie hoch wird der Gewinn im Mittel beim Lotto sein? Wie viele Telefonate pro

---

<sup>4)</sup>Das dort stehende  $f$  ist durch 1 zu ersetzen, der Faktor  $f(X^{-1}(y))$  fällt also weg.

Tag sind im Mittel für eine spezielle Nummer zu erwarten? Welche mittlere Schadenshöhe kommt im nächsten Jahr auf ein Versicherungsunternehmen zu? Und so weiter.

Wir werden diesen noch recht vagen Wunsch in der Definition „Erwartungswert von  $X$ “ präzisieren. Der Aufwand ist für Zufallsvariable auf allgemeinen Wahrscheinlichkeitsräumen recht erheblich, deswegen kümmern wir uns zunächst um die Spezialfälle, die in diesem Buch hauptsächlich untersucht werden.

Erwartungswert: diskrete endliche Räume

Das ist noch recht einfach. Bevor wir die Definition angeben, gibt es eine Motivation, warum man es genau so machen sollte. Gegeben ist ein endlicher Wahrscheinlichkeitsraum  $(\Omega, \mathcal{E}, \mathbb{P})$ , und  $X : \Omega \rightarrow \mathbb{R}$  ist vorgelegt. Die Elementarereignisse in  $\Omega$  werden als  $\omega_1, \dots, \omega_n$  bezeichnet, und nun wird  $m$  Mal abgefragt. Dabei sollte  $m$  viel größer als  $n$  sein. Wir werden dann als Ausgabe unsere Elementarereignisse erhalten:  $m_1$  Mal  $\omega_1$ ,  $m_2$  Mal  $\omega_2$ , ...,  $m_n$  Mal  $\omega_n$ ; dabei gilt  $m_1 + \dots + m_n = m$ .

Wie oft erhalten wir ein spezielles  $\omega_i$ ? Aufgrund unserer Interpretation von Wahrscheinlichkeit sollte sich der relative Anteil, also  $m_i/m$ , gut durch  $\mathbb{P}(\{\omega_i\})$  approximieren lassen.

Und wie groß ist der Mittelwert  $M$  der erzeugten  $X(\omega)$ ? Wenn  $\omega_i$  erzeugt wurde, ist  $X(\omega_i)$  zu berücksichtigen, es ist also

$$M = \frac{X(\omega_1)m_1 + \dots + X(\omega_n)m_n}{m}.$$

Und das kann man als

$$X(\omega_1)\frac{m_1}{m} + \dots + X(\omega_n)\frac{m_n}{m} \approx X(\omega_1)\mathbb{P}(\{\omega_1\}) + \dots + X(\omega_n)\mathbb{P}(\{\omega_n\})$$

schreiben, wobei die Approximation mit wachsendem  $m$  immer besser werden sollte. Die rechts stehende Summe ist also im Mittel zu erwarten, und das führt zu

**Definition 3.3.1.** Ist  $(\Omega, \mathcal{E}, \mathbb{P})$  endlich und  $X : \Omega \rightarrow \mathbb{R}$ , so versteht man unter dem Erwartungswert von  $X$  die Zahl

$$X(\omega_1)\mathbb{P}(\{\omega_1\}) + \dots + X(\omega_n)\mathbb{P}(\{\omega_n\}) \left(= \sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\{\omega\})\right).$$

Man schreibt dafür  $\mathbb{E}(X)$  („ $E$  von  $X$ “).

**Beispiele: 1.** Die Wahrscheinlichkeiten auf  $\Omega = \{1, 2, 3\}$  seien durch  $\mathbb{P}(\{1\}) := 0.2$ ,  $\mathbb{P}(\{2\}) := 0.3$  und  $\mathbb{P}(\{3\}) := 0.5$  definiert, und es sei  $X(i) := i^3$  für  $i \in \Omega$ . Dann ist

$$\mathbb{E}(X) = 0.2 \cdot 1^3 + 0.3 \cdot 2^3 + 0.5 \cdot 3^3 = 16.1.$$

(Auch wenn  $X$  nur ganzzahlige Werte annimmt, muss also  $\mathbb{E}(X)$  nicht ganzzahlig sein.)

**2.** Was ist der *Erwartungswert der Gleichverteilung* auf  $\{1, \dots, n\}$ ? Das soll bedeuten, dass wir  $\{1, \dots, n\}$  mit der Gleichverteilung versehen und  $X(i) := i$  (für  $i = 1, \dots, n$ ) untersuchen<sup>5)</sup>. Der Erwartungswert ist folglich gleich

$$(1 + 2 + \dots + n)/n = (n + 1)/2.$$

Zum Beispiel ist der Erwartungswert bei einem fairen Würfel gleich 3.5.

**3.** Der *Erwartungswert der Bernoulliverteilung* auf  $\{0, 1\}$  ist  $(1 - p) \cdot 0 + p \cdot 1 = p$ .

**Erwartungswert: diskrete abzählbare Räume**

Es ist naheliegend, die Definition zu übernehmen, die endliche Summe also durch eine unendliche zu ersetzen. Was soll aber  $\sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\{\omega\})$  für abzählbare  $\Omega$  bedeuten? Ein ähnliches Problem haben wir schon einmal im Zusammenhang mit Satz 2.1.1 besprochen: Die „ungeordnete Reihe“ ist als gewöhnliche Reihe  $\sum_{n=1}^{\infty} X(\omega_n)\mathbb{P}(\{\omega_n\})$  reeller Zahlen zu interpretieren, wobei  $\omega_1, \omega_2, \dots$  eine beliebige Durchnummerierung der Elemente von  $\Omega$  ist. Damit das wohldefiniert ist, muss allerdings sichergestellt sein, dass die Definition nicht von der zufällig gewählten Nummerierung abhängt. Oder anders ausgedrückt: Umordnungen der Reihe  $\sum_{n=1}^{\infty} X(\omega_n)\mathbb{P}(\{\omega_n\})$  müssen gegen den gleichen Wert konvergieren. Man weiß aus der Analysis, dass das (sogar genau) dann der Fall ist, wenn die Reihe absolut konvergent ist, wenn also  $\sum_{n=1}^{\infty} |X(\omega_n)\mathbb{P}(\{\omega_n\})| < \infty$  gilt. Das führt zu

**Definition 3.3.2.**  $(\Omega, \mathcal{E}, \mathbb{P})$  sei ein abzählbarer Wahrscheinlichkeitsraum, dabei wird  $\Omega$  als  $\{\omega_1, \omega_2, \dots\}$  geschrieben. Wir sagen, dass der Erwartungswert einer Zufallsvariablen  $X : \Omega \rightarrow \mathbb{R}$  existiert, wenn die Reihe  $\sum_{n=1}^{\infty} X(\omega_n)\mathbb{P}(\{\omega_n\})$  absolut konvergent ist. In diesem Fall definieren wir den Erwartungswert von  $X$  durch

$$\mathbb{E}(X) := \sum_{n=1}^{\infty} X(\omega_n)\mathbb{P}(\{\omega_n\}).$$

**Beispiele:** 1. Das erste Beispiel ist ein Gegenbeispiel. Wir behaupten, dass auf abzählbaren Räumen quasi *immer* Zufallsvariable ohne Erwartungswert existieren. Sei dazu  $\Omega$  als  $\{\omega_1, \omega_2, \dots\}$  geschrieben, wir nehmen an, dass alle Elementereignisse eine strikt positive Wahrscheinlichkeit haben. Definiert man dann  $X : \Omega \rightarrow \mathbb{R}$  durch  $X(\omega_n) := (-1)^n/\mathbb{P}(\{\omega_n\})$ , so kann

$$\mathbb{E}(X) := \sum_{n=1}^{\infty} X(\omega_n)\mathbb{P}(\{\omega_n\}) = -1 + 1 - 1 + 1 \pm \dots$$

---

<sup>5)</sup>Ganz analog sind „Erwartungswert der Binomialverteilung/Poissonverteilung/...“ zu verstehen: Wenn  $\Omega \subset \mathbb{R}$  ist, kann für  $X$  die identische Abbildung  $x \mapsto x$  betrachtet werden. Der Erwartungswert von *diesem*  $X$  ist dann der Erwartungswert des zugehörigen Wahrscheinlichkeitsraums.

nicht definiert werden.

**2.** Wir berechnen hier den *Erwartungswert der Poissonverteilung* zum Parameter  $\lambda \geq 0$ . Es geht um die Reihe

$$\sum_{n=0}^{\infty} n \frac{\lambda^n}{n!} e^{-\lambda} = \left( 1 \frac{\lambda}{1!} + 2 \frac{\lambda^2}{2!} + \dots \right) e^{-\lambda}.$$

Sie ist sicher absolut konvergent<sup>6)</sup>. Der Wert der Reihe ist auch leicht zu ermitteln: Im  $n$ -ten Summanden  $n$  kürzen, ein  $\lambda$  aus der Reihe ausklammern, Formel für die  $e$ -Reihe einsetzen. So folgt: Der Erwartungswert ist  $\lambda$ .

**3.** Auch die Rechnung für den *Erwartungswert der geometrischen Verteilung* führt auf das Problem, eine Reihe konkret auszuwerten. Diesmal geht es um die Reihe  $\sum_{n=1}^{\infty} n(1-q)q^{n-1} = (1-q)(1+2q+3q^2+\dots)$ . Nun ist  $1+2q+3q^2+\dots = 1/(1-q)^2$ , das folgt zum Beispiel durch gliedweises Differenzieren der geometrischen Reihe. Deswegen ist der Erwartungswert gleich  $1/(1-q)$ .

### Erwartungswert: Räume mit Dichten

Sei zunächst  $\Omega = [a, b]$  ein kompaktes Intervall, das durch eine stetige Dichtefunktion  $f$  zu einem Wahrscheinlichkeitsraum gemacht wurde. Weiter sei eine stetige Zufallsvariable  $X : [a, b] \rightarrow \mathbb{R}$  gegeben. Um die Formel zu motivieren, mit der wir gleich den Erwartungswert definieren werden, approximieren wir die vorliegende Situation durch einen diskreten Wahrscheinlichkeitsraum. Dazu unterteilen wir  $[a, b]$  in „sehr viele“ und „sehr kleine“ Teilintervalle  $I_1, \dots, I_n$ . Man muss nur Zwischenpunkte  $a = t_0 < t_1 < \dots < t_n = b$  wählen und  $I_k = [t_{k-1}, t_k]$  setzen ( $k = 1, \dots, n$ ), beispielsweise  $t_k = a + k(b-a)/n$  mit einem „großen“  $n \in \mathbb{N}$ . Statt die exakt ausgegebenen  $\omega$  zu betrachten und  $X(\omega)$  zu notieren, registrieren wir nur, in welchem  $I_k$  das  $\omega$  liegt und merken uns  $X(t_{k-1})$ . Das wird keinen sehr großen Unterschied ausmachen, wenn die Unterteilung fein genug ist, denn für alle  $\omega \in I_k$  wird  $X(\omega)$  aus Stetigkeitsgründen nahe bei  $X(t_{k-1})$  liegen.

Damit sind wir de facto zum Raum  $\{0, \dots, n-1\}$  übergegangen.  $\{k\}$  hat die Wahrscheinlichkeit  $\mathbb{P}(I_k) = \int_{t_{k-1}}^{t_k} f(x) dx$ , und die jetzt relevante Zufallsvariable bildet  $k$  auf  $X(t_k)$  ab. Die  $\mathbb{P}(I_k)$  lassen sich noch, ohne einen großen Fehler zu machen, durch  $f(t_{k-1})(t_k - t_{k-1})$  ersetzen, denn da  $f$  stetig und  $t_k - t_{k-1}$  „klein“ ist, kann  $f$  auf  $I_k$  durch  $f(t_{k-1})$  gut approximiert werden; dann ist das Integral leicht auszurechnen.

Zusammen: Das, was wir suchen, sollte durch

$$\sum_{k=1}^n X(t_{k-1}) f(t_{k-1})(t_k - t_{k-1})$$

zu approximieren sein, und zwar umso besser, je feiner die Unterteilung von  $[a, b]$  war. Diese Summe ist aber ein alter Bekannter aus der Analysis. Sie tritt

---

<sup>6)</sup>Z. B. wegen des Quotientenkriteriums: Der Betrag des Quotienten zweier aufeinander folgender Summanden ist  $|\lambda|/n$ , und der ist für  $n > 2|\lambda|$  kleiner als  $1/2$ .

auf, wenn man das Integral  $\int_a^b X(x)f(x) dx$  durch Riemannsummen approximiert. Und deswegen ist die folgende Definition wenig überraschend:

**Definition 3.3.3.** *f sei eine stetige Dichtefunktion auf  $[a, b]$ , und auch die Zufallsvariable  $X : [a, b] \rightarrow \mathbb{R}$  sei stetig. Dann definieren wir den Erwartungswert von  $X$  durch*

$$\mathbb{E}(X) := \int_a^b X(x)f(x) dx.$$

**Beispiele: 1.**  $[0, 1]$  sei mit der Dichte  $f(x) = (n+1)x^n$  versehen, und  $X$  sei durch  $X(x) := x^2$  definiert. Dann ist

$$\mathbb{E}(X) = \int_0^1 x^2(n+1)x^n dx = (n+1) \int_0^1 x^{n+2} dx = \frac{n+1}{n+3}.$$

**2.** Der Erwartungswert der Gleichverteilung auf  $[a, b]$  – dazu ist  $X(x) = x$  auf  $[0, 1]$  mit der Gleichverteilung zu untersuchen – ist natürlich  $(a+b)/2$ , der Mittelpunkt des Intervalls:

$$\mathbb{E}(X) = \frac{1}{b-a} \int_a^b x dx = \frac{1}{b-a} \frac{x^2}{2} \Big|_a^b = \frac{b^2 - a^2}{2(b-a)} = \frac{a+b}{2}.$$

Geht man zu allgemeineren Situationen (unbeschränkte Intervalle, stückweise stetige oder unbeschränkte Funktionen, ...) über, so sind ähnliche Überlegungen anzustellen wie im diskreten Fall beim Übergang von endlichen zu abzählbaren Räumen: Der Erwartungswert wird immer noch als  $\int_{\Omega} X(x)f(x) dx$  erklärt, vorausgesetzt, dass dieses Integral vernünftig definiert werden kann. In allen Fällen reicht es vorauszusetzen, dass  $\int_{\Omega} |X(x)|f(x) dx$  einen endlichen Wert hat, dass also  $|X|f$  absolut integrierbar ist. Auf die analytischen Feinheiten soll hier nicht eingegangen werden.

Wir setzen unsere Beispiele fort, nach den vorstehenden Bemerkungen können wir auch die Fälle  $\Omega = [0, +\infty[$  und  $\Omega = \mathbb{R}$  zulassen:

**3.** Der Erwartungswert der Exponentialverteilung zum Parameter  $\lambda$  führt auf das uneigentliche Integral  $\int_0^{+\infty} x \lambda e^{-\lambda x} dx$ . Partielle Integration verschafft uns eine Stammfunktion zu  $xe^{-\lambda x}$ , nämlich  $-e^{-\lambda x}(\lambda x + 1)/\lambda^2$ . Und deswegen ist

$$\begin{aligned} \int_0^{+\infty} x \lambda e^{-\lambda x} dx &= \lim_{r \rightarrow +\infty} \lambda \int_0^r xe^{-\lambda x} dx \\ &= \lim_{r \rightarrow +\infty} -e^{-\lambda x}(\lambda x + 1)/\lambda \Big|_0^r \\ &= \lim_{r \rightarrow +\infty} -e^{-\lambda r}(\lambda r + 1)/\lambda + 1/\lambda \\ &= 1/\lambda. \end{aligned}$$

Zusammen: Der Erwartungswert der Exponentialverteilung ist  $1/\lambda$ .

**4.** Auch der *Erwartungswert* der Normalverteilung  $N(a, \sigma^2)$  ist leicht zu ermitteln. Allgemein gilt: Ist  $x_0$  ein Symmetriepunkt des Intervalls  $\Omega$  (d.h., mit  $x_0 + x \in \Omega$  ist stets auch  $x_0 - x \in \Omega$ ) und ist die Dichtefunktion  $f$  symmetrisch um  $x_0$  (d.h. immer ist  $f(x_0 + x) = f(x_0 - x)$ ), so hat die Zufallsvariable  $X(x) = x$  den Erwartungswert  $x_0$ . Die Begründung: Ist  $x_0 = 0$ , so ist  $x \mapsto xf(x)$  eine schiefsymmetrische Funktion<sup>7)</sup>. Und da  $\Omega$  zu 0 symmetrisch ist, heben sich positive und negative Werte des Integrals gegenseitig auf. Der allgemeine Fall, wenn  $x_0$  beliebig ist, kann durch die Transformation  $x \mapsto x - x_0$  darauf zurückgeführt werden.

Bei der Normalverteilung  $N(a, \sigma^2)$  spielt  $a$  die Rolle des Symmetriepunkts, diese Zahl ist also der Erwartungswert.

#### Eigenschaften des Erwartungswerts

Der Erwartungswert ist durch eine Summe bzw. durch ein Integral definiert worden. Deswegen sind Linearitäts- und Monotonie-Eigenschaften zu erwarten:

**Satz 3.3.4.** *Es seien  $X$  und  $Y$  reellwertige Zufallsvariable, für die sich der Erwartungswert definieren lässt, und es sei  $c \in \mathbb{R}$ .*

- (i) *Auch für  $cX$  existiert der Erwartungswert, und es ist  $\mathbb{E}(cX) = c\mathbb{E}(X)$ .*
- (ii) *Auch für  $X + Y$  existiert der Erwartungswert, und es ist  $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$ .*
- (iii) *Gilt  $X(\omega) \leq Y(\omega)$  für alle  $\omega$ , so ist  $\mathbb{E}(X) \leq \mathbb{E}(Y)$ .*
- (iv) *Ist  $E$  ein Ereignis und  $X = \chi_E$  die zugehörige Indikatorfunktion, so ist  $\mathbb{E}(X) = \mathbb{P}(E)$ .*

**Beweis:** (i), (ii) und (iii) folgen unmittelbar aus den entsprechenden Eigenschaften für Summen und Integrale. (iv) ist im diskreten Fall klar, denn bei den Seiten der behaupteten Gleichung stimmen mit  $\sum_{\omega \in E} \mathbb{P}(\{\omega\})$  überein. Bei Räumen mit Dichten reduziert sich  $\int_{\Omega} X(x)f(x)dx$  auf  $\int_E f(x)dx$ , und diese Zahl ist gleich  $\mathbb{P}(E)$  (vgl. den Anfang von Abschnitt 2.2).

Wir bemerken noch, dass (i) und (ii) gerade besagen, dass  $X \mapsto \mathbb{E}(X)$  eine lineare Abbildung auf dem Raum derjenigen Zufallsvariablen ist, für die der Erwartungswert definiert werden kann.  $\square$

Im Allgemeinen ist für eine Zufallsvariable  $X$  nur  $\mathbb{P}_X$  bekannt: Man weiß, mit welchen Wahrscheinlichkeiten die Ergebnisse  $X(\omega)$  zu erwarten sind. Dass man  $\mathbb{E}(X)$  allein aus der Kenntnis von  $\mathbb{P}_X$  berechnen kann, zeigt der folgende Satz, in dem in Hinblick auf spätere Anwendungen etwas mehr gezeigt wird, als wir hier benötigen.

**Satz 3.3.5.**

- (i) *Sei  $\Omega$  ein höchstens abzählbarer Wahrscheinlichkeitsraum,  $X : \Omega \rightarrow \mathbb{R}$  und  $\phi : \mathbb{R} \rightarrow \mathbb{R}$ . Dann ist, falls der Erwartungswert existiert,  $\mathbb{E}(\phi(X)) =$*

---

<sup>7)</sup>Der Wert bei  $-x$  ist das Negative des Werts bei  $x$ .

$\sum \phi(y) \mathbb{P}_X(\{y\})$ , wobei über alle  $y \in X(\Omega)$  zu summieren ist. Insbesondere ist  $\mathbb{E}(X) = \sum_{y \in X(\Omega)} y \mathbb{P}_X(\{y\})$ .

(ii) Ist  $\Omega = [a, b]$  ein Intervall in  $\mathbb{R}$ , das durch eine Dichtefunktion  $f$  zu einem Wahrscheinlichkeitsraum gemacht wurde und  $X : [a, b] \rightarrow [c, d]$  stetig differenzierbar und bijektiv mit strikt positiver Ableitung, so gilt für jede stetig differenzierbare Funktion  $\phi : \mathbb{R} \rightarrow \mathbb{R}$ :

$$\mathbb{E}(\phi(X)) = \int_c^d \phi(y) h(y) dy;$$

dabei ist  $h$  die Dichtefunktion von  $\mathbb{P}_X$  gemäß Satz 3.2.3. Speziell für  $\phi(y) = y$  folgt  $\mathbb{E}(X) = \int_c^d y h(y) dy$ .

Ein entsprechendes Ergebnis gilt für unbeschränkte Intervalle, wenn vorausgesetzt wird, dass die auftretenden Integrale existieren.

**Beweis:** (i) Wenn man in der Definition von  $\mathbb{E}(X)$  Summanden mit gleichem  $X(\omega)$  zusammenfasst, steht das Ergebnis schon da:

$$\begin{aligned} \mathbb{E}(\phi(X)) &= \sum_{\omega \in \Omega} \phi(X(\omega)) \mathbb{P}(\{\omega\}) \\ &= \sum_{y \in X(\Omega)} \sum_{\substack{\omega \in \Omega \\ X(\omega)=y}} \phi(X(\omega)) \mathbb{P}(\{\omega\}) \\ &= \sum_{y \in X(\Omega)} \phi(y) \sum_{\substack{\omega \in \Omega \\ X(\omega)=y}} \mathbb{P}(\{\omega\}) \\ &= \sum_{y \in X(\Omega)} \phi(y) \mathbb{P}_X(\{y\}). \end{aligned}$$

Im abzählbaren Fall ist wieder wichtig, dass die auftretenden Reihen absolut konvergent sind, damit beim Umsortieren die gleichen Werte herauskommen.

(ii) Wir werden die Variable in  $[a, b]$  mit  $x$  und die in  $[c, d]$  mit  $y$  bezeichnen. Es ist  $h(y) = f(X^{-1}(y))(X^{-1})'(y)$ , und deswegen läuft die Behauptung auf

$$\int_c^d \phi(y) f(X^{-1}(y))(X^{-1})'(y) dy = \int_a^b \phi(X(x)) f(x) dx$$

hinaus. Diese Gleichheit ergibt sich aber aus der Formel für die Integration durch Substitution, wenn man im links stehenden Integral  $x := X^{-1}(y)$  substituiert.

Der Fall unbeschränkter Intervalle wird analog behandelt: Die auftretenden uneigentlichen Integrale sind Grenzwerte von Integralen über kompakte Intervalle, auf denen kann der Integrand  $yh(y)$  in  $X(x)f(x)$  transformiert werden.

□

Erwartungswert: beliebige Wahrscheinlichkeitsräume

Der Vollständigkeit halber soll noch skizziert werden, wie man bei beliebigen Wahrscheinlichkeitsräumen vorgeht<sup>8)</sup>. Gesucht ist eine Definition, die für Räume ohne Zusatzbedingungen sinnvoll ist und die in den schon behandelten Spezialfällen (diskrete Räume, Räume mit Dichten) das schon Bekannte liefert.

Für diejenigen, die schon Maß- und Integrationstheorie kennen, ist die Sache einfach: Der Erwartungswert von  $X$  ist einfach das Integral  $\int_{\Omega} X d\mathbb{P}$ . (Eine Kurzfassung dieses Zugangs findet man im Anhang auf Seite 357.) Hier definieren wir den Erwartungswert wie folgt.

Wir gehen von einem beliebigen Wahrscheinlichkeitsraum  $(\Omega, \mathcal{E}, \mathbb{P})$  und einer Zufallsvariable  $X : \Omega \rightarrow \mathbb{R}$  aus. Zunächst nehmen wir an, dass  $X(\omega) \geq 0$  für alle  $\omega$  gilt. Wir kombinieren zwei Tatsachen:

- Mal angenommen,  $X$  nimmt nur höchstens abzählbar viele Werte  $a_1, a_2, \dots$  an. Mit  $E_i := \{X = a_i\}$  hat dann  $X$  mit Wahrscheinlichkeit  $p_i := \mathbb{P}(E_i)$  den Wert  $a_i$ . Diese Situation ist nicht von einem diskreten Wahrscheinlichkeitsraum zu unterscheiden, und deswegen gibt es nur die Möglichkeit,  $\mathbb{E}(X)$  durch  $\sum_i a_i p_i$  zu definieren. (Diese Reihensumme existiert in  $[0, \infty]$ , da alle Summanden nichtnegativ sind.) Das soll aber nur dann die Definition des Erwartungswertes sein, wenn diese Reihensumme endlich ist.
- Wenn es eine vernünftige Definition des Erwartungswertes gibt, so ist die sicher monoton: aus  $X \leq Y$  folgt  $\mathbb{E}(X) \leq \mathbb{E}(Y)$ . Und das hat zur Konsequenz, dass aus  $X - \varepsilon \leq Y \leq X + \varepsilon$  stets  $\mathbb{E}(X) - \varepsilon \leq \mathbb{E}(Y) \leq \mathbb{E}(X) + \varepsilon$  und damit  $|\mathbb{E}(X) - \mathbb{E}(Y)| \leq \varepsilon$  folgt.

Wenn wir also  $X$  bis auf  $\varepsilon$  durch eine Zufallsvariable mit höchstens abzählbaren Werten approximieren können, so ist – die vielleicht noch unbekannte – Zahl  $\mathbb{E}(X)$  bis auf  $\varepsilon$  bekannt.

Diese Idee wird nun präzisiert. Es sei  $n \in \mathbb{N}$ . Wir zerlegen  $[0, \infty[$  in Teilintervalle der Länge  $1/2^n$ : Für  $k \in \{0, 1, \dots\}$  sei  $I_k^n$  das Intervall  $[k/2^n, (k+1)/2^n[$ .  $E_k^n$  soll das Ereignis  $\{X \in I_k^n\}$  sein (für  $k = 0, 1, \dots$ ), und mit  $X_n$  bezeichnen wir die Funktion, die auf  $E_k^n$  den Wert  $k/2^n$  hat. Aufgrund der Konstruktion gilt dann:  $X_n$  nimmt höchstens abzählbar viele Werte an, und es gilt  $X - 1/2^n \leq X_n \leq X$ .

Wie verhalten sich  $X_n, X_m$ , falls  $n < m$ ? Die Unterteilung ist feiner geworden, aber die Werte von  $X_m$  auf einem  $E_k^n$  liegen weiterhin in  $I_k^n$ . Anders ausgedrückt heißt das, dass  $|X_n(\omega) - X_m(\omega)| \leq 1/2^n$  für alle  $\omega$  gilt. Und das impliziert – die Endlichkeit von  $\mathbb{E}(X_n)$  vorausgesetzt –, dass  $(\mathbb{E}(X_n))_n$  eine Cauchy-Folge ist. Das führt zur

**Definition 3.3.6.**  $X$  sei nichtnegativ und so, dass  $\mathbb{E}(X_n)$  für ein  $n$  endlich ist<sup>9)</sup>. Dann ist  $(\mathbb{E}(X_n))_n$  eine Cauchy-Folge, wir definieren  $\mathbb{E}(X) := \lim_n \mathbb{E}(X_n)$ .

---

<sup>8)</sup>Wer möchte, kann gleich zu „Varianz und Streuung“ weiterblättern und die übersprungenen Seiten später lesen.

<sup>9)</sup>In diesem Fall ist übrigens  $\mathbb{E}(X_n)$  für alle  $n$  endlich. Das folgt daraus, dass  $X_m \leq X_n + 1$  für beliebige  $n, m$  gilt.

Ist  $X$  beliebig, so schreiben wir  $X = X^+ - X^-$  mit  $X^+, X^- \geq 0$ ; man kann z.B.  $X^+(\omega) := \max\{X(\omega), 0\}$  und  $X^-(\omega) := \max\{-X(\omega), 0\}$  setzen. Falls die Erwartungswerte von  $X^+$  und  $X^-$  existieren (d.h.: endliche Zahlen sind), setzen wir  $\mathbb{E}(X) := \mathbb{E}(X^+) - \mathbb{E}(X^-)$ .

Mit dieser Definition ist wirklich alles erreicht, was wir uns vorgenommen haben: Sie gilt allgemein, sie führt im Spezialfall auf Bekanntes, und auch die Ergebnisse, die wir hier nur für diskrete Räume und Räume mit Dichten nachweisen werden, gelten in der allgemeinen Version:

**Satz 3.3.7.** *Mit den vorstehenden Bezeichnungen gilt:*

- (i) Ist  $\Omega$  diskret oder ist das Wahrscheinlichkeitsmaß durch eine Dichtefunktion definiert, so stimmt die neue Definition mit der alten überein.
- (ii) Gilt  $0 \leq X \leq Y$  und existiert  $\mathbb{E}(Y)$ , so existiert auch  $\mathbb{E}(X)$  und es ist  $\mathbb{E}(X) \leq \mathbb{E}(Y)$ .
- (iii) Wenn  $\mathbb{E}(X)$  und  $\mathbb{E}(Y)$  existieren, so existiert auch  $\mathbb{E}(X + Y)$  und es gilt  $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$ . Auch ist  $\mathbb{E}(cX) = c\mathbb{E}(X)$  für  $c \in \mathbb{R}$ .

**Beweis:** (i) Angenommen,  $\Omega$  ist diskret. Für die Zahl  $\mathbb{E}(X_n)$  kommt dann das gleiche heraus, ob wir sie wie vorstehend oder gemäß Definition 3.3.2 berechnen: Das liegt an  $\mathbb{P}(\{X \in E_k^n\}) = \sum_{\omega \in E_k^n} \mathbb{P}(\{\omega\})$ . Aus der Monotonie des „diskreten“ Erwartungswertes und aus  $X_n \leq X \leq X_n + 1/2^n$  folgt dann sofort, dass  $\lim \mathbb{E}(X_n)$  mit dem Wert von  $\mathbb{E}(X)$  in Definition 3.3.2 übereinstimmen muss.

Ist  $\Omega$  ein Intervall und hat  $\mathbb{P}$  eine Dichtefunktion  $f$ , so argumentiert man wie folgt. Die  $E_k^n$  sind Borelmengen in  $\Omega$ , und es ist zu zeigen, dass

$$\sum_k \frac{k}{2^n} \mathbb{P}(E_k^n) = \sum_k \frac{k}{2^n} \int_{E_k^n} f(x) dx = \int_{\Omega} X_n(x) \cdot f(x) dx$$

gegen  $\int_{\Omega} X(x) f(x) dx$  konvergiert. Das folgt aber wieder sofort aus den Ungleichungen  $X_n \leq X \leq X_n + 1/2^n$ . (Hier haben wir ausgenutzt, dass die Rechenregeln für die Integration „einfacher“ Funktionen aus der Analysis auch allgemeiner für Borel-messbare Funktionen gelten.)

(ii) Wenn  $\mathbb{E}(Y)$  existiert, so sind die  $\mathbb{E}(Y_n)$  endlich. Da  $X_n \leq Y_n$  gilt, sind auch die  $\mathbb{E}(X_n)$  endlich, und es ist  $\mathbb{E}(X_n) \leq \mathbb{E}(Y_n)$ . So folgt

$$\mathbb{E}(X) = \lim \mathbb{E}(X_n) \leq \lim \mathbb{E}(Y_n) = \mathbb{E}(Y).$$

(iii) Wieder reicht es, sich um positive  $X, Y$  zu kümmern. Nun ist  $X_n + Y_n \leq X + Y \leq X_n + Y_n + 2/2^n$ , und wegen  $\mathbb{E}(X_n + Y_n) = \mathbb{E}(X_n) + \mathbb{E}(Y_n)$  (für diskrete Situationen ist das ja schon bekannt) folgt die Behauptung.

Der zweite Teil wird so bewiesen. Starte mit einer positiven Zufallsvariablen, einem  $c > 0$  und der Ungleichung  $X_n \leq X \leq X_n + 1/2^n$ . Dann ist auch  $cX_n \leq cX \leq cX_n + c/2^n$ . Da die Monotonie schon in (ii) bewiesen wurde, folgt

$$c\mathbb{E}(X_n) = \mathbb{E}(cX_n) \leq \mathbb{E}(cX) \leq c\mathbb{E}(X_n) + c/2^n.$$

Und das beweist

$$\mathbb{E}(cX) = \lim c\mathbb{E}(X_n) = c\mathbb{E}(X).$$

Der Fall beliebiger Zufallsvariablen und beliebiger  $c \in \mathbb{R}$  kann leicht darauf zurückgeführt werden.  $\square$

### Varianz und Streuung

Der Erwartungswert ist eine erste wichtige Größe, um Zufallsvariable zu klassifizieren. Allerdings sagt er nichts darüber aus, wie weit die einzelnen  $X(\omega)$  von  $\mathbb{E}(X)$  abweichen.

Nehmen wir zum Beispiel  $\Omega = \{-1, 1\}$  mit der Gleichverteilung und die durch  $X(x) := Kx$  definierte Zufallsvariable; dabei soll  $K > 0$  eine reelle Zahl sein. Man kann diese Situation als faires Spiel interpretieren, bei der man mit gleicher Wahrscheinlichkeit  $K$  Euro bekommt bzw.  $K$  Euro zahlen muss. Der Erwartungswert ist für jedes  $K$  gleich Null, die meisten werden sich aber vielleicht im Fall  $K = 1.000.000$  nicht auf dieses Spiel einlassen wollen.

Formal gesehen handelt es sich darum, eine „vernünftige“ Definition für den Abstand der beiden auf  $\Omega$  definierten Funktionen  $X$  und  $\omega \mapsto \mathbb{E}(X)$  zu finden. Es gibt eine fast unübersehbare Fülle von solchen Abstandsdefinitionen. In der Wahrscheinlichkeitstheorie ist der „richtige“ Abstand die *mittlere quadratische Abweichung*:

**Definition 3.3.8.** *Es sei  $(\Omega, \mathcal{E}, \mathbb{P})$  ein Wahrscheinlichkeitsraum und  $X : \Omega \rightarrow \mathbb{R}$  eine Zufallsvariable, für die der Erwartungswert  $\mathbb{E}(X)$  existiert. Wir definieren dann die Varianz von  $X$  als*

$$V(X) := \mathbb{E}((X - \mathbb{E}(X))^2),$$

*falls der Erwartungswert von  $(X - \mathbb{E}(X))^2$  erklärt werden kann.*

*Die Streuung  $\sigma(X)$  von  $X$  ist in diesem Fall die (positive) Wurzel aus der Varianz:  $\sigma(X) := \sqrt{V(X)}$ .*

Bevor wir uns um Beispiele kümmern, gibt es einige

**Bemerkungen:** 1. Aus Satz 3.3.4 (iii) folgt, dass  $V(X)$  nichtnegativ ist, und daher kann man wirklich die Wurzel ziehen. Die Wurzel ist deswegen erforderlich, damit die Maßeinheit die gleiche ist wie die bei  $X$ : Ist zum Beispiel  $X$  eine Größe, die Längen misst, etwa in Metern, so hat  $V(X)$  die Einheit Quadratmeter und  $\sigma(X)$  ist wieder eine Meterangabe.

2. Im Fall endlicher  $\Omega$  oder für kompakte Intervalle und stetige Zufallsvariable existiert die Streuung stets. Für komplizierte Situationen muss die Existenz der Varianz (und damit der Streuung) stets extra begründet werden, sie folgt nicht automatisch aus der Existenz des Erwartungswerts.

Betrachte zum Beispiel auf  $\mathbb{N}$  das durch

$$\mathbb{P}(\{1\}) = \mathbb{P}(\{2\}) = 1/4, \mathbb{P}(\{3\}) = \mathbb{P}(\{4\}) = 1/8, \mathbb{P}(\{5\}) = \mathbb{P}(\{6\}) = 1/16, \dots$$

definierte Wahrscheinlichkeitsmaß und die Zufallsvariable

$$X(1) = -X(2) = 4^{1/2}, X(3) = -X(4) = 8^{1/2}, X(5) = -X(6) = 16^{1/2}, \dots$$

Die Reihe  $\sum_n X(n)\mathbb{P}(\{n\})$  ist absolut konvergent, denn  $\sum_n |X(n)|\mathbb{P}(\{n\}) = 2 \sum_{n=2}^{\infty} (1/\sqrt{2})^n < \infty$ . Der Erwartungswert ist Null, die Varianz existiert aber nicht, denn der Versuch, den Erwartungswert von  $(X - (\mathbb{E}(X)))^2$  auszurechnen, führt auf die nicht konvergente Reihe  $1 - 1 + 1 - 1 \dots$

**3.** Wer in seinem bisherigen Studium schon Räume messbarer Funktionen kennen gelernt hat, wird bemerkt haben, dass  $\sigma(X)$  gerade die  $L^2$ -Norm der Funktion  $X - \mathbb{E}(X)$  im Raum der quadratintegrablen Funktionen auf  $(\Omega, \mathcal{E}, \mathbb{P})$  ist.

**4.** Warum wählt man ausgerechnet  $\mathbb{E}((X - \mathbb{E}(X))^2)$  als Maß für den Abstand, warum nicht  $\mathbb{E}(|X - \mathbb{E}(X)|)$  oder  $\sup_{\omega} |X(\omega) - \mathbb{E}(X)|$ ? Der Hauptgrund liegt darin, dass man sich auf diese Weise die speziellen strukturellen Eigenschaften des Raums der quadratintegrablen Funktionen zunutze machen kann. Dort ist der Abstand durch ein inneres Produkt erklärt, es ist ein euklidischer Raum. Und deswegen kann man in vielen Fällen mit Zufallsvariablen so wie mit Punkten im  $\mathbb{R}^n$  rechnen. (Mehr dazu im Anhang auf Seite 359.)

**5.** Für reelle Zahlen  $x > 0$  gilt ja: Ist  $x < 1$  (bzw.  $x > 1$ ), so wird  $x$  beim Übergang zu  $x^2$  kleiner (bzw. größer), und dieser Effekt wird für „sehr kleine“ (bzw. „sehr große“)  $x$  immer extremer. Das bedeutet für die Varianz: Abweichungen zwischen  $X$  und  $\mathbb{E}(X)$ , die kleiner als Eins sind, werden bei der Berechnung von  $V(X)$  quasi heruntergerechnet, sind sie aber größer als Eins, bekommen sie ein besonderes Gewicht.

*Beispiele:* **1.** Auf  $\Omega = \{1, 2\}$  sei  $\mathbb{P}(\{1\}) = 0.2$  und  $\mathbb{P}(\{2\}) = 0.8$ , und  $X$  sei durch  $X(1) = 2, X(2) = -1$  erklärt. Es ist  $\mathbb{E}(X) = 2 \cdot 0.2 - 1 \cdot 0.8 = -0.4$ . Folglich ist  $V(X) = (2 + 0.4)^2 \cdot 0.2 + (-1 + 0.4)^2 \cdot 0.8 = 1.44$ , und es folgt  $\sigma(X) = 1.2$ .

**2. Varianz und Streuung der Gleichverteilung** auf  $\{1, \dots, n\}$ . (Hier und in vergleichbaren Beispielen, in denen  $\Omega$  eine Teilmenge von  $\mathbb{R}$  ist, geht es um  $V(X)$  und  $\sigma(X)$  für die spezielle Zufallsvariable  $X(x) = x$ .)

Wir wissen schon, dass  $\mathbb{E}(X) = (n+1)/2$  gilt, und deswegen ist

$$\begin{aligned} V(X) &= \sum_{i=1}^n \left( i - \frac{n+1}{2} \right)^2 \frac{1}{n} \\ &= \left( \sum_{i=1}^n \left( i^2 - i(n+1) + \frac{(n+1)^2}{4} \right) \right) \frac{1}{n} \\ &= \left( \frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)^2}{2} + \frac{n(n+1)^2}{4} \right) \frac{1}{n} \\ &= \frac{n^2 - 1}{12}; \end{aligned}$$

bei der Berechnung haben wir die Formeln  $1 + 2 + \dots + n = n(n+1)/2$  und  $1^2 + 2^2 + \dots + n^2 = n(n+1)(2n+1)/6$  verwendet. Für die Streuung folgt  $\sigma(X) = \sqrt{(n^2 - 1)/12}$ .

**3. Varianz und Streuung der Bernoulliverteilung** auf  $\{0, 1\}$ . Da der Erwartungswert gleich  $p$  ist, ergibt sich für die Varianz der Wert  $(0-p)^2(1-p)+(1-p)^2p=p(1-p)$ . Die Streuung ist also  $\sqrt{p(1-p)}$ . (Sie ist folglich klein für  $p$ -Werte in der Nähe von 0 oder 1 und maximal für  $p = 1/2$ .)

Weitere Ergebnisse zu Varianz und Streuung konkreter Verteilungen sind Gegenstand der Übungsaufgaben. Hier eine Übersicht:

| Verteilung                                       | Erwartungswert | Varianz        | Streuung              |
|--|----------------|----------------|-----------------------|
| Laplaceraum $\{1, \dots, n\}$                    | $(1+n)/2$      | $(n^2 - 1)/12$ | $\sqrt{(n^2 - 1)/12}$ |
| Bernoulliverteilung                              | $p$            | $p(1-p)$       | $\sqrt{p(1-p)}$       |
| Poissonverteilung                                | $\lambda$      | $\lambda$      | $\sqrt{\lambda}$      |
| geometrische Verteilung                          | $1/(1-q)$      | $q/(1-q)^2$    | $\sqrt{q}/(1-q)$      |
| Gleichverteilung auf $[a, b]$                    | $(b-a)/2$      | $(b-a)^2/12$   | $(b-a)/\sqrt{12}$     |
| Exponentialverteilung                            | $1/\lambda$    | $1/\lambda^2$  | $1/\lambda$           |
| Normalverteilung $N(a, \sigma^2)$ <sup>10)</sup> | $a$            | $\sigma^2$     | $\sigma$              |

### Eigenschaften von Varianz und Streuung

Für Varianz und Streuung gibt es einige nützliche Formeln:

**Satz 3.3.9.** Wir setzen voraus, dass  $X$  und  $Y$  reellwertige Zufallsvariable auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{E}, \mathbb{P})$  sind und dass  $V(X)$  und  $V(Y)$  existieren.

- (i)  $V(cX) = c^2V(X)$ , und  $\sigma(cX) = |c|\sigma(X)$  für alle  $c \in \mathbb{R}$ .
- (ii)  $V(X+c) = V(X)$ , und  $\sigma(X+c) = \sigma(X)$  für alle  $c \in \mathbb{R}$ .
- (iii)  $\mathbb{E}(XY)$  existiert, und  $(\mathbb{E}(XY))^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2)$ .
- (iv)  $V(X+Y)$  existiert.
- (v) Die Varianz (und damit die Streuung) kann mit Hilfe von  $\mathbb{P}_X$  ermittelt werden: Es ist  $V(X) = \sum_{y \in X(\Omega)} (y - \mathbb{E}(X))^2 \mathbb{P}(\{y\})$  im diskreten Fall und (unter den Voraussetzungen und mit den Bezeichnungen von Satz 3.3.5)

$$V(X) = \int_c^d (y - \mathbb{E}(X))^2 h(y) dy,$$

wenn das Wahrscheinlichkeitsmaß durch eine Dichtefunktion definiert ist.

- (vi)  $V(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$ .

**Beweis:** (i) und (ii) folgen sofort aus Satz 3.3.4, und (v) ergibt sich aus Satz 3.3.5, wenn man dort  $\phi(y) := (y - \mathbb{E}(X))^2$  setzt.

Zum Beweis von (iii) erinnern wir an die Cauchy-Schwarzsche Ungleichung<sup>11)</sup> aus der linearen Algebra: Ist  $\langle \cdot, \cdot \rangle$  ein inneres Produkt auf einem  $\mathbb{R}$ -Vektorraum

<sup>10)</sup>Die Berechnung findet man in Abschnitt 8.4.

<sup>11)</sup>Vgl. den Anhang, Seite 359.

$V$ , so gilt  $\langle x, y \rangle^2 \leq \langle x, x \rangle \langle y, y \rangle$  für alle  $x, y$  in  $V$ . Wir brauchen diese Ungleichung für das innere Produkt  $\langle x, y \rangle := \sum_{i=1}^n c_i x_i y_i$  auf dem  $\mathbb{R}^n$ ; dabei sind  $c_1, \dots, c_n$  positive Zahlen:  $(\sum_i c_i x_i y_i)^2 \leq (\sum_i c_i x_i^2)(\sum_i c_i y_i^2)$ .

Für endliche diskrete Räume ergibt sich damit die Abschätzung  $(\mathbb{E}(XY))^2 \leq \mathbb{E}(X^2) \mathbb{E}(Y^2)$ , und durch Grenzübergang folgt daraus das entsprechende Ergebnis zunächst für abzählbare diskrete Räume und dann für beliebige Wahrscheinlichkeitsräume.

(iv) ist eine unmittelbare Folgerung aus (iii), denn  $((X + Y) - \mathbb{E}(X + Y))^2$  ist die Summe aus  $(X - \mathbb{E}(X))^2$ ,  $2(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))$  und  $(Y - \mathbb{E}(Y))^2$ , und für diese drei Funktionen existiert der Erwartungswert.

Die noch zu beweisende Aussage (vi) folgt schnell aus den schon gezeigten Ergebnissen:

$$\begin{aligned} V(X) &= \mathbb{E}((X - \mathbb{E}(X))^2) \\ &= \mathbb{E}(X^2 - 2X\mathbb{E}(X) + (\mathbb{E}(X))^2) \\ &= \mathbb{E}(X^2) - 2\mathbb{E}(X)\mathbb{E}(X) + (\mathbb{E}(X))^2 \\ &= \mathbb{E}(X^2) - (\mathbb{E}(X))^2. \end{aligned}$$

□

## 3.4 Elementare Kombinatorik

In diesem Abschnitt gibt es einen notwendigen Exkurs. Für Laplaceräume können doch die Wahrscheinlichkeiten als Quotient zweier Anzahlen berechnet werden, und das führt bei der Berechnung induzierter Wahrscheinlichkeiten in vielen Fällen auf das Problem, die Elemente in einer vorgegebenen Menge zu zählen. Wir werden *vier grundlegende Zählprobleme* und den *Inklusion-Exklusion-Satz* behandeln.

Sind  $N$  bzw.  $M$  endliche Mengen mit  $n$  bzw.  $m$  Elementen, so hat die Produktmenge  $n \cdot m$  Elemente. Kurz:  $\#(N \times M) = \#N \cdot \#M$ . Das kann ganz elementar mit vollständiger Induktion bewiesen werden, und ebenfalls durch Induktion folgt dann

$$\#(N_1 \times N_2 \times \cdots \times N_k) = \#N_1 \cdot \#N_2 \cdots \#N_k$$

Insbesondere ist  $\#(N^k) = (\#N)^k$ .

Damit können schon elementare Zählprobleme gelöst werden. Wenn es z.B. in einem Restaurant 4 Vorspeisen, 5 Hauptgerichte und 3 Nachspeisen gibt, so kann man  $4 \cdot 5 \cdot 3 = 60$  Menüs zusammenstellen. Und man kann  $10 \cdot 26 \cdot 26 \cdot 26 \cdot 10 \cdot 10 \cdot 10 = 175.760.000$  verschiedene Autokennzeichen des Typs Ziffer-Buchstabe-Buchstabe-Ziffer-Ziffer-Ziffer erzeugen.



Bild 3.4.1: Wie viele verschiedene Autokennzeichen?

Bei unserer systematischen Untersuchung wird es ebenfalls um Auswahlen gehen. Man wählt endlich oft aus einer endlichen Menge etwas aus. Dabei sind *zwei Aspekte des Zählens* zu unterscheiden:

- Die *Reihenfolge des Auswählens kann wichtig oder unwichtig sein*. Wichtig ist sie zum Beispiel, wenn man Buchstaben wählt und ein Wort zusammensetzt: OTTO ist etwas anderes als TOTO. In anderen Fällen spielt sie keine Rolle, so ist es völlig egal, in welcher Reihenfolge ich die sechs Zahlen auf dem Lottoschein ankreuze.
- In manchen Situationen ist es *möglich, mehrfach das gleiche Objekt zu wählen*. Wenn es etwa um aus vier Buchstaben gebildete Wörter geht, ist nicht nur VIER erlaubt, sondern auch TOTO. Es kann aber auch sein, dass Wiederholungen nicht zugelassen sind: Soll aus 20 Spielern eine elfköpfige Fußballmannschaft gebildet werden, kann jeder Einzelne nur einmal auftreten.

Insgesamt sind damit *vier elementare Zählprobleme* zu behandeln.

### Problem 1: Reihenfolge wichtig, Wiederholung möglich

*Typische Situationen:* Vierbuchstabige Wörter, vierstellige Zahlen im Hexadezimalsystem

*Die Formel:* Dieses Problem ist schon in der Einleitung zu diesem Abschnitt gelöst worden. Man kann auf genau  $n^k$  verschiedene Weisen Auswählen von  $k$  Elementen einer  $n$ -elementigen Menge bilden, wenn die Reihenfolge wichtig ist und Elemente mehrfach ausgewählt werden dürfen.



Bild 3.4.2: Einige der 456.976 „Four-letter-words“.

*Ein Beispiel:* Es gibt  $26^4 = 456.976$  „Four-letter-words“. (Davon stellt allerdings nur ein winziger Bruchteil ein Wort in irgendeiner Sprache dar.)

### Problem 2: Reihenfolge wichtig, keine Wiederholung

*Typische Situationen:* Wahl des Vorstands (Präsident, Vertreter, Schriftführer, Schatzmeister) in einem Verein. Planung einer Reise, bei der  $n$  Städte je einmal besucht werden sollen. Dreibuchstabige Wörter ohne Buchstabenwiederholung.

*Die Formel:* Es sei  $n \in \mathbb{N}$  und  $k \in \{1, \dots, n\}$ . Dann gibt es

$$n \cdot (n - 1) \cdots (n - k + 1)$$

Möglichkeiten,  $k$  Elemente aus einer  $n$ -elementigen Menge auszuwählen, wenn die Reihenfolge wichtig ist und Wiederholungen nicht zugelassen sind.

Das zeigt man so: Für das erste Element der Auswahl gibt es  $n$  Möglichkeiten, für das zweite nur noch  $n - 1$ , für das dritte  $n - 2$  usw. Im letzten Schritt, beim  $k$ -ten Element, gibt es noch  $n - k + 1$  Kandidaten zur Auswahl. Für die Gesamtzahl ist das Produkt zu bilden, so erhält man die Formel.

*Beispiele:* Es gibt  $26 \cdot 25 \cdot 24 = 15.600$  dreibuchstabige Wörter ohne Buchstabenwiederholung. Man beachte auch, dass die gesuchte Anzahl im Fall  $k = n$  gleich  $n!$  ist. So kann man zum Beispiel auf  $5! = 120$  verschiedene Weisen eine Rundreise durch fünf Städte machen.

### Problem 3: Reihenfolge unwichtig, keine Wiederholung<sup>12)</sup>

*Typische Situationen:* Lottotipps, Skatblätter, wie viele Verabschiedungen bei  $n$  Personen, ...

*Die Formel:* Es sei  $n \in \mathbb{N}$  und  $k \in \{1, \dots, n\}$ . Dann gibt es

$$\frac{n \cdot (n - 1) \cdots (n - k + 1)}{k!}$$

$k$ -elementige Teilmengen von  $\{1, \dots, n\}$ ; das entspricht den Möglichkeiten,  $k$  Elemente auszuwählen, wenn die Reihenfolge der ausgewählten Elemente unwichtig ist und Wiederholungen nicht zugelassen sind.

Um diese Formel zu beweisen, berechnen wir zunächst die Anzahl der Auswahlen unter Berücksichtigung der Reihenfolge. Im vorstehenden Problem 2 haben wir gezeigt, dass diese Anzahl gleich  $n \cdot (n - 1) \cdots (n - k + 1)$  ist. Wir identifizieren nun zwei Auswahlen, wenn sie sich nur durch die Reihenfolge der Elemente unterscheiden. Da die Anzahl der Permutationen einer  $k$ -elementigen Menge gleich  $k!$  ist, ergibt sich wirklich  $n \cdot (n - 1) \cdots (n - k + 1)/k!$ .

Man kann es auch etwas formaler machen. In der  $n \cdot (n - 1) \cdots (n - k + 1)$ -elementigen Menge der Auswahlen von  $k$  Elementen mit Berücksichtigung der Reihenfolge und ohne Wiederholung wird eine Äquivalenzrelation eingeführt: Zwei Auswahlen sollen äquivalent heißen, wenn die gleichen Elemente vorkommen. Jede Äquivalenzklasse hat  $k!$  Elemente, und folglich muss es  $n \cdot (n - 1) \cdots (n - k + 1)/k!$  Äquivalenzklassen geben.

---

<sup>12)</sup>Das ist sicher das wichtigste der Zählprobleme.

Der Ausdruck  $n \cdot (n-1) \cdots (n-k+1)/k!$  kommt sehr oft in der Kombinatorik vor, man kürzt ihn durch das Symbol  $\binom{n}{k}$  ab<sup>13)</sup>. Gesprochen wird das „ $n$  über  $k$ “. (Im Englischen sagt man übrigens – wesentlich suggestiver als im Deutschen – „ $n$  choose  $k$ “.)

*Beispiele:* Wenn sich 10 Leute voneinander verabschieden, gibt es  $\binom{10}{2} = 45$  Mal einen Händedruck, und man kann auf einem Lottoschein auf  $\binom{49}{6} = 13.983.816$  verschiedene Arten sechs Zahlen ankreuzen

#### Problem 4: Reihenfolge unwichtig, Wiederholung möglich

*Eine typische Situation:* Gegeben seien  $n$  unterscheidbare Kästen und  $k$  identisch aussehende Kugeln. Auf wie viele Weisen können die Kugeln in den Kästen platziert werden? (Das kann wirklich als Auswahlproblem uminterpretiert werden: Wenn bei der  $i$ -ten Auswahl eines Elementes aus  $\{1, \dots, n\}$  die Zahl  $j$  gezogen wurde, so stecke die  $i$ -te Kugel in Kasten  $j$ . Dann ist die Reihenfolge der  $j$ -Auswahlen unwichtig, denn am Ende interessieren nur die Kugelanzahlen in den einzelnen Kästen. Und Wiederholungen sind möglich, da man ja auch mehrere Kugeln im gleichen Kasten unterbringen kann.)

*Die Formel:* Wir behaupten, dass es genau  $\binom{n+k-1}{k}$  Möglichkeiten gibt.

Zum Beweis stellen wir uns das Kugeln-in-Kästen-Packen so vor. Wir fügen zu den  $k$  Kugeln (sie sollen alle die gleiche Farbe haben, etwa weiß) noch  $n-1$  schwarze Kugeln hinzu. Dann legen wir diese  $k+(n-1)$  Kugeln auf irgendeine Weise in einer Reihe aus. Fasst man die schwarzen Kugeln als „Trennungen“ zwischen den Kästen auf, so induziert die Kugelreihe eine Aufteilung der weißen Kugeln. Zum Beispiel würde im Fall  $n=4$  und  $k=5$  die Reihe

schwarz–weiß–weiß–schwarz–schwarz–weiß–weiß–weiß

zu der Belegung der vier Kästen mit Null bzw. 2 bzw. Null bzw. 3 Kugeln führen.

Und umgekehrt gilt das auch, man kann jede Belegung durch so eine Kugelreihe verschlüsseln. Zusammen heißt das, dass die Anzahl der Kugelaufteilungen gleich der Anzahl der Möglichkeiten ist, eine  $(n-1)$ -elementige Teilmenge aus einer Menge mit  $k+n-1$  Elementen auszuwählen. Die gesuchte Zahl ist also  $\binom{n+k-1}{n-1}$ . Und da allgemein  $\binom{a}{b} = \binom{a}{a-b}$  gilt (s.u.), kann man dafür auch  $\binom{n+k-1}{k}$  schreiben.

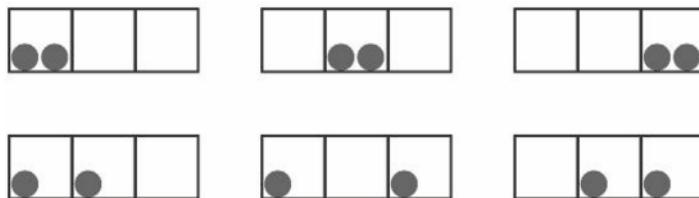


Bild 3.4.3: Zwei Kugeln in drei Kästen:  $\binom{3+2-1}{2} = 6$  Möglichkeiten.

---

<sup>13)</sup>Man nennt die  $\binom{n}{k}$  auch die *Binomialkoeffizienten*. Der Grund: Sie treten in der Entwicklung von  $(a+b)^n$  auf (s.u., Ergänzungen zu Problem 3).

*Ein weiteres Beispiel:* 15 Kugeln sollen in 3 Kästen untergebracht werden. Die Anzahl der Möglichkeiten ist  $\binom{17}{15} = \binom{17}{2} = 136$ .

### Ergänzungen zu Zählproblem 3

Es wurde schon erwähnt, dass dieses Problem das wichtigste ist. Es folgen einige ergänzende Bemerkungen.

**1.** Es ist bequem, auch den Fall  $k = 0$  zuzulassen. Da es genau eine leere Menge gibt, ist es naheliegend,  $\binom{n}{0}$  als Eins zu definieren. Diese Festsetzung ist auch wichtig, damit in der „binomischen Formel“

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

alle Terme definiert sind.

**2.** Wer auf Pünktchen verzichten möchte, kann  $\binom{n}{k}$  auch als  $n!/(k!(n - k)!)$  definieren. Das ist für Beweise praktisch, empfiehlt sich aber in der Regel nicht für konkrete Rechnungen. So ist es etwa viel leichter,  $\binom{100}{2}$  als  $100 \cdot 99 / 2! = 4950$  auszurechnen, als zunächst  $100!$ ,  $98!$  und  $2!$  zu bestimmen.

**3.** Hier sind die Binomialkoeffizienten bei der Beschreibung von Zählproblemen eingeführt worden. Das hat einige Konsequenzen:

- Die  $\binom{n}{k}$  liegen in  $\mathbb{N}$ : Aus der Formel ist das wirklich nicht zu sehen.
- Man kann eine  $k$ -elementige Teilmenge von  $\{1, \dots, n\}$  auf zwei Weisen definieren: Entweder gibt man alle  $k$  Elemente an, die dazugehören sollen, oder man legt fest, welche  $n - k$  Elemente *nicht* berücksichtigt werden. Es folgt, dass es genau so viele  $k$ -elementige wie  $(n - k)$ -elementige Teilmengen gibt, und folglich ist  $\binom{n}{k} = \binom{n}{n - k}$ . (Das ergibt sich natürlich auch sofort aus der Formel  $\binom{n}{k} = n!/(k!(n - k)!)$ .)
- Insgesamt gibt es  $2^n$  Teilmengen von  $\{1, \dots, n\}$ , und deswegen muss

$$\binom{n}{0} + \binom{n}{1} + \dots + \binom{n}{n} = 2^n$$

gelten<sup>14)</sup>.

Wesentlich aufwändiger als die bisherigen Untersuchungen ist der *Beweis des Inklusion-Exklusion-Satzes*, der nun behandelt werden wird. Es geht um eine Verallgemeinerung der Aussage<sup>15)</sup>  $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F)$  auf mehr als zwei Ereignisse: Wie kann man die Wahrscheinlichkeit von  $\bigcup_{i=1}^n E_i$  aus den Wahrscheinlichkeiten von Durchschnitten berechnen?

Um den Satz übersichtlich formulieren zu können, führen wir einige Bezeichnungen ein. Es seien  $E_1, \dots, E_n$  Ereignisse in einem Wahrscheinlichkeitsraum

<sup>14)</sup>Das kann man aber auch leicht aus der binomischen Formel für  $(1 + 1)^n$  ablesen.

<sup>15)</sup>Vgl. Satz 1.3.2(iv).

$(\Omega, \mathcal{E}, \mathbb{P})$ . Für  $k = 1, \dots, n$  bezeichnen wir mit  $\Delta_k$  die Menge der  $k$ -elementigen Teilmengen von  $\{1, \dots, n\}$ , und für  $I \in \Delta_k$  soll  $E_I$  das Ereignis  $\bigcap_{i \in I} E_i$  bedeuten. Dann gilt der folgende

**Satz 3.4.1.** (*Inklusion-Exklusion-Satz*) Es ist

$$\mathbb{P}\left(\bigcup_{i=1}^n E_i\right) = \sum_{I \in \Delta_1} \mathbb{P}(E_I) - \sum_{I \in \Delta_2} \mathbb{P}(E_I) + \sum_{I \in \Delta_3} \mathbb{P}(E_I) - \dots + (-1)^{n+1} \sum_{I \in \Delta_n} \mathbb{P}(E_I).$$

**Beweis:** Der Beweis wird mit Hilfe von Indikatorfunktionen  $\chi_E$  geführt (zur Definition vgl. Seite 73). Sei  $\omega \in \Omega$  beliebig, wir nehmen an, dass es in  $\bigcup_{i=1}^n E_i$  liegt. Mit  $r \in \{1, \dots, n\}$  bezeichnen wir die Anzahl der Mengen  $E_i$ , zu denen  $\omega$  gehört. Dann gibt es offensichtlich

- $\binom{r}{1}$  einelementige Teilmengen  $I$  von  $\{1, \dots, n\}$ , so dass  $\omega$  in  $E_I$  liegt;
- $\binom{r}{2}$  zweielementige Teilmengen  $I$  von  $\{1, \dots, n\}$  mit  $\omega \in E_I$ ;
- allgemein ist  $\binom{r}{k}$  die Anzahl der  $k$ -elementigen Teilmengen  $I$  von  $\{1, \dots, n\}$ , so dass  $\omega \in E_I$  gilt.

Das bedeutet, dass die Funktion

$$\chi := \sum_{I \in \Delta_1} \chi_{E_I} - \sum_{I \in \Delta_2} \chi_{E_I} + \sum_{I \in \Delta_3} \chi_{E_I} - \dots + (-1)^{n+1} \sum_{I \in \Delta_n} \chi_{E_I}$$

bei  $\omega$  den Wert

$$\binom{r}{1} - \binom{r}{2} + \binom{r}{3} - \dots + (-1)^{r+1} \binom{r}{r}$$

hat. Diese natürliche Zahl ist aber gleich Eins, was man sofort aus der binomischen Formel für  $(1 - 1)^r$  abliest.

Die Funktion  $\chi$  stimmt folglich auf  $\bigcup E_i$  mit der Indikatorfunktion dieser Menge überein. Für die  $\omega \notin \bigcup E_i$  gilt das auch, denn da sind beide Funktionen Null. Anders ausgedrückt:  $\chi$  ist die Indikatorfunktion der Vereinigung. Nun ist nur noch an Satz 3.3.4 zu erinnern: Aufgrund der Linearität des Erwartungswerts und der Gleichung  $\mathbb{E}(\chi_E) = \mathbb{P}(E)$  für Ereignisse  $E$  ergibt sich die Behauptung<sup>16)</sup>.  $\square$

Der Satz hat ein wichtiges

**Korollar 3.4.2.** Es seien  $E_1, \dots, E_n$  Teilmengen der endlichen Menge  $\Omega$ . Dann gilt mit den Bezeichnungen des vorigen Satzes:

$$\#\left(\bigcup_{i=1}^n E_i\right) = \sum_{I \in \Delta_1} \#(E_I) - \sum_{I \in \Delta_2} \#(E_I) + \sum_{I \in \Delta_3} \#(E_I) - \dots + (-1)^{n+1} \sum_{I \in \Delta_n} \#(E_I).$$

**Beweis:** Man braucht nur  $\Omega$  mit der Gleichverteilung zu versehen und zu beachten, dass  $\#E = \mathbb{P}(E)\#\Omega$  gilt. Die behauptete Gleichung ergibt sich also aus derjenigen des vorigen Satzes nach Multiplikation mit  $\#\Omega$ .  $\square$

---

<sup>16)</sup>Satz 3.3.4 war nur für diskrete Räume und Räume mit Dichten formuliert worden, er gilt aber allgemein.

### 3.5 Berechnung induzierter Wahrscheinlichkeiten

Die kombinatorischen Ergebnisse des vorigen Abschnitts sollen nun an einigen ausgewählten Beispielen angewendet werden.

#### Das Geburtstagsparadoxon

Es seien  $n$  und  $r$  natürliche Zahlen, wir betrachten die Gleichverteilung auf  $\Omega = \{1, \dots, n\}^r$ . Man kann sich ein zufällig erzeugtes Element auch als  $r$ -malige Abfrage einer gleichverteilten Zahl in  $\{1, \dots, n\}$  vorstellen. Wie wahrscheinlich ist es, dass diese  $r$  Zahlen alle verschieden sind<sup>17)</sup>?

Offensichtlich ist die Wahrscheinlichkeit Null, wenn  $r$  größer als  $n$  ist, und im Fall  $r = 1$  ist sie sicher gleich Eins. Wie sieht es für die restlichen  $r$  aus? Wir fixieren so ein  $r \in \{2, \dots, n\}$  und definieren  $E_{n,r} \subset \Omega$  als die Menge aller  $r$ -Tupel, in denen die Komponenten paarweise verschieden sind. Welche Wahrscheinlichkeit hat diese Menge?

Da  $\Omega$  die Gleichverteilung trägt, ist  $\mathbb{P}(E_{n,r}) = \#E_{n,r}/\#\Omega$ . Zähler und Nenner sind leicht ermittelt, es geht ja um „ $r$ -fache Auswahl aus  $\{1, \dots, n\}$ , Reihenfolge wichtig“, und im Nenner ist die Wiederholung zugelassen, im Zähler aber nicht. So folgt mit den Ergebnissen des vorigen Abschnitts

$$\mathbb{P}(E_{n,r}) = \frac{n \cdot (n-1) \cdots (n-r+1)}{n^r} = 1 \cdot \left(1 - \frac{1}{n}\right) \cdots \left(1 - \frac{n-r+1}{n}\right).$$

$1 - \mathbb{P}(E_{n,r})$  ist dann die Wahrscheinlichkeit, dass es mindestens eine Ziffernwiederholung gibt. In Abhängigkeit von  $r$  steigen diese Zahlen von 0 (bei  $r = 1$ ) bis 1 (bei  $r > n$ ). Bemerkenswert ist dabei die Tatsache, dass sie *sehr schnell* steigen. Das berühmteste Beispiel in diesem Zusammenhang ist das sogenannte *Geburtstagsparadoxon*. Hier die übliche Verkleidung des Problems:

Auf einem Fest treffen sich  $r$  zufällig zusammengewürfelte Leute. Wie groß muss  $r$  sein, dass die Wahrscheinlichkeit dafür, dass mindestens zwei am gleichen Tag des Jahres Geburtstag haben, größer als 50 Prozent ist?

Bei einer naiven Schätzung sollte  $r$  in der Nähe von 180 liegen, da das Jahr ja 365 Tage hat. Das ist aber nicht richtig, der korrekte Wert ist 23. Wer es genauer wissen möchte: In der nachstehenden Tabelle sind für  $r = 17$  bis  $r = 24$  die Zahlen  $\mathbb{P}(E_{365,r})$  und  $1 - \mathbb{P}(E_{365,r})$  aufgeführt.

| $r$                         | 17    | 18    | 19    | 20    | 21    | 22    | 23    | 24    |
|-----------------------------|-------|-------|-------|-------|-------|-------|-------|-------|
| $\mathbb{P}(E_{365,r})$     | 0.685 | 0.653 | 0.621 | 0.589 | 0.556 | 0.524 | 0.493 | 0.462 |
| $1 - \mathbb{P}(E_{365,r})$ | 0.315 | 0.347 | 0.379 | 0.411 | 0.444 | 0.476 | 0.507 | 0.538 |

<sup>17)</sup>Das kann man auch mit Hilfe von Zufallsvariablen formulieren:  $X : \Omega \rightarrow \{1, \dots, r\}$  sei die Abbildung, die jedem  $(i_1, \dots, i_r) \in \Omega$  die Anzahl der verschiedenen Zahlen in  $(i_1, \dots, i_r)$ , also  $\#\{i_1, \dots, i_n\}$ , zuordnet. Gefragt ist dann nach  $\mathbb{P}_X(\{r\})$ .

Bei  $r = 23$  steigt die Wahrscheinlichkeit für einen Doppelgeburtstag erstmals auf über 50 Prozent.

Immer dann, wenn Intuition und mathematische Wahrheit so extrem weit auseinanderliegen, spricht man von einer *Paradoxie*<sup>18)</sup>. Die trifft man in der Wahrscheinlichkeitstheorie recht häufig an, aufgrund unserer Entwicklungsgeschichte sind wir für das intuitive Erfassen von Wahrscheinlichkeiten wohl recht schlecht vorbereitet.

In Vorlesungen zur elementaren Stochastik wird oft an dem Tag, wenn es um das Geburtstagsparadoxon geht, ein „Experiment“ gemacht: Die Teilnehmer der Vorlesung kreuzen in einem Kalender ihren jeweiligen Geburtstag an, und man kann sich dann fast darauf verlassen, dass es bei einigen Tagen im Jahr zwei oder sogar mehr Kreuze gibt.

Auch bei der Fußballweltmeisterschaft 2006 in Deutschland wurden Geburtstage verglichen, und zwar für jedes Spiel die Geburtstage der 23 beteiligten Personen (22 Spieler und der Schiedsrichter). Das Ergebnis stimmte (schon fast verdächtig gut) mit der Voraussage des Geburtstagsparadoxons überein, denn in 32 von 64 Spielen gab es einen Doppelgeburtstag<sup>19)</sup>.

Ein Beispiel aus dem Finale: Die Franzosen Patrick Viera und Zinedine Zidane feiern beide am 23. Juni.

### Die hypergeometrische Verteilung

Wie wahrscheinlich sind drei Richtige im Lotto? Dieses Problem kann etwas allgemeiner so formuliert werden:

In einem Kasten befinden sich  $n$  Kugeln, und zwar  $r$  rote Kugeln und  $n - r$  weiße. Sie werden gut durchgemischt, und dann wird  $m$  Mal ohne Zurücklegen gezogen, wobei  $m \in \{1, \dots, n\}$ . Wie wahrscheinlich ist es, dass dabei genau  $k$  rote Kugeln gezogen wurden (wobei  $k = 0, \dots, m$ )? Diese Wahrscheinlichkeit soll mit  $h(k; m; r, n)$  bezeichnet werden.

Gefragt ist also nach den  $\mathbb{P}_X(\{k\})$  für  $k = 0, \dots, m$ , wobei  $X$  die auf der Menge aller möglichen Ziehungen von  $m$  Kugeln definierte Zufallsvariable ist, die die Anzahl der roten Kugeln zählt.

Das Lottobispiel kann als Spezialfall interpretiert werden. Die „Kugeln“ sind die 49 Felder des Lottoscheins, und die „roten“ Kugeln sind „die Richtigen“. Ein Lottotipp besteht dann aus dem Ziehen von sechs „Kugeln“, und alle hoffen, dass möglichst viele „rote Kugeln“ dabei sind. Vor wenigen Zeilen wurde folglich nach  $h(3; 6; 49)$  gefragt.

---

<sup>18)</sup>Auch: Paradoxon. Das Wort ist aus dem Griechischen abgeleitet. Frei übersetzt bedeutet „paradox“ etwa „entgegen der öffentlichen Meinung“.

<sup>19)</sup>Konstantin Weixelbaum hat diese Recherche damals durchgeführt.



Um die gesuchten Wahrscheinlichkeiten zu berechnen, stellen wir uns eine Ziehung als zufällige Auswahl einer  $m$ -elementigen Teilmenge aus einer  $n$ -elementigen Menge vor. Es gibt, wie auf Seite 93 gezeigt,  $\binom{n}{m}$  solche Teilmengen. Um eine mit genau  $k$  roten Kugeln zu erhalten, muss eine  $k$ -elementige Teilmenge aus den roten und eine  $(m - k)$ -elementige Teilmenge aus den weißen Kugeln gezogen werden. Vom ersten Typ gibt es  $\binom{r}{k}$ , vom zweiten  $\binom{n-r}{m-k}$ , das ergibt insgesamt  $\binom{r}{k} \binom{n-r}{m-k}$  Möglichkeiten<sup>20)</sup>. Da wir alle Auswahlen als gleichberechtigt betrachten, also einen Laplace Raum vor uns haben, erhalten wir für die gesuchten Wahrscheinlichkeiten die Formel

$$h(k, m; r, n) = \frac{\binom{r}{k} \binom{n-r}{m-k}}{\binom{n}{m}}.$$

Man spricht von der *hypergeometrischen Verteilung*, dabei handelt es sich um ein Wahrscheinlichkeitsmaß auf  $\{0, \dots, m\}$ .

→  
Programm!

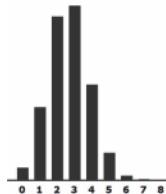


Bild 3.5.1: Die hypergeometrische Verteilung:  $h(k, 8; 10, 30); k = 0, \dots, 8$ .

Es ist nun leicht, sämtliche Lottowahrscheinlichkeiten auszurechnen:

- Wahrscheinlichkeit für 0 Richtige:  $h(0, 6; 6, 49) = \binom{6}{0} \binom{43}{6} / \binom{49}{6} = 0.436 \dots$
- Wahrscheinlichkeit für 1 Richtige:  $h(1, 6; 6, 49) = \binom{6}{1} \binom{43}{5} / \binom{49}{6} = 0.413 \dots$
- Wahrscheinlichkeit für 2 Richtige:  $h(2, 6; 6, 49) = \binom{6}{2} \binom{43}{4} / \binom{49}{6} = 0.132 \dots$
- Wahrscheinlichkeit für 3 Richtige:  $h(3, 6; 6, 49) = \binom{6}{3} \binom{43}{3} / \binom{49}{6} = 0.018 \dots$
- Wahrscheinlichkeit für 4 Richtige:  $h(4, 6; 6, 49) = \binom{6}{4} \binom{43}{2} / \binom{49}{6} = 0.001 \dots$
- Wahrscheinlichkeit für 5 Richtige:  $h(5, 6; 6, 49) = \binom{6}{5} \binom{43}{1} / \binom{49}{6} \approx 2 \cdot 10^{-5}$
- Wahrscheinlichkeit für 6 Richtige:  

$$h(6, 6; 6, 49) = \binom{6}{6} \binom{43}{0} / \binom{49}{6} = 1/13.983.816 \approx 7 \cdot 10^{-8}.$$

<sup>20)</sup>Ist  $k > r$ , so ist  $\binom{r}{k}$  eigentlich nicht definiert. Es ist sinnvoll, in diesem Fall  $\binom{r}{k} := 0$  zu setzen, da es in einer  $r$ -elementigen Menge dann keine  $k$ -elementige Teilmenge gibt. Die gleiche Bemerkung gilt für  $\binom{n-r}{m-k}$  in Fällen, in denen  $m - k > n - r$  gilt.

Von den deprimierend kleinen Wahrscheinlichkeiten der interessanten Gewinne können wir uns nur schwer eine Vorstellung machen. Es ist hilfreich, eine Übersetzung zu Hilfe zu nehmen (so wie auch zum Beispiel in der Astronomie, wenn man die Größenverhältnisse im Planetensystem visualisieren möchte). Die Lieblingsbeispiele des Autors sind die folgenden:

*Sechs Richtige und das Telefon:* Stellen Sie sich vor, dass Sie eine bestimmte wildfremde Person in Ihrer Stadt anrufen wollen. Zum Beispiel den Herrn, der seinen Schirm im Bus vergessen hat oder den Kommilitonen / die Kommilitonin, der/die heute erstmals in der Vorlesung war. Sie nehmen am Abend Ihr Telefon/Handy und wählen in Ihrer Stadt eine 7-stellige Nummer, indem Sie einfach siebenmal zufällig auf die Tasten drücken. Die Wahrscheinlichkeit, dass das die richtige Nummer ist, ist  $10^{-7}$ , und das ist noch deutlich mehr als die Wahrscheinlichkeit für sechs Richtige.



Bild 3.5.2: Veranschaulichung von Lotto-Wahrscheinlichkeiten

*Der Stab an der Autobahn.* Wählen Sie eine Autobahnstrecke von etwa 140 Kilometern Länge (z.B. Berlin–Cottbus) und lassen Sie irgendwo an dieser Strecke von einem Helfer einen ein Zentimeter breiten Stab in den rechten Seitenstreifen stecken. Nun werden Sie – auf dem Beifahrersitz – mit verbundenen Augen und offenem Seitenfenster diese Strecke entlanggefahren. In Ihrer Hand haben Sie ein Centstück, das sollen Sie irgendwann unterwegs aus dem Fenster werfen. Die Wahrscheinlichkeit, dass Sie dabei den Stab treffen, ist ziemlich genau gleich der Sechser-Wahrscheinlichkeit (denn 140 Kilometer entsprechen 14 Millionen Zentimetern).

Die Wahrscheinlichkeiten der hypergeometrischen Verteilung sollen noch etwas detaillierter diskutiert werden. Wir beginnen mit der trivialen Beobachtung, dass die  $h(k, m; r, n)$ ,  $k = 0, \dots, m$ , Wahrscheinlichkeiten sind, deswegen gilt  $\sum_{k=0}^n h(k, m; r, n) = 1$ . Wenn man die oben ermittelten Werte einsetzt und die Gleichung mit  $\binom{n}{m}$  multipliziert, wird daraus

$$\sum_{k=0}^m \binom{r}{k} \binom{n-r}{m-k} = \binom{n}{m}.$$

Es ist schlecht vorstellbar, wie man diese Identität auf ähnlich elegante Weise direkt aus der Definition der Binomialkoeffizienten herleiten könnte.

Für welche  $k$  sind die größten  $h(k, m; r, n)$  zu erwarten? Als Beispiel betrachten wir den Fall  $n = 1000, r = 500$ : Die Hälfte der Kugeln ist rot. Wenn

dann  $m = 10$  Kugeln gezogen werden, ist es plausibel anzunehmen, dass es am wahrscheinlichsten ist, dass davon die Hälfte rot ist, bei  $k = 5$  sollte also das Maximum liegen.

Leider lässt sich das aus der expliziten Formel nicht direkt ablesen, aber ein Trick hilft weiter. Wir führen für die nächsten Zeilen ein neues Symbol  $\square$  ein, das wird der Platzhalter für eine der Relationen  $=, <, >$  sein. Es wird gleich eine Rolle spielen, dass für  $a, b \in \mathbb{R}$  und  $c > 0$  aus  $a \square b$  stets  $ac \square bc$  folgt.

Nun der Trick: Wir studieren die Quotienten aufeinanderfolgender Wahrscheinlichkeiten, also die Zahlen  $h(k, m; r, n)/h(k + 1, m; r, n)$ . Bemerkenswerterweise lässt sich sehr viel kürzen, man erhält

$$\frac{h(k, m; r, n)}{h(k + 1, m; r, n)} = \frac{((n - r) - (m - k) + 1)(k + 1)}{(m - k)(r - k)}.$$

Um die Abhängigkeit von  $k$  zu analysieren, formen wir die Relation

$$h(k, m; r, n)/h(k + 1, m; r, n) \square 1$$

so um, dass  $k$  isoliert ist. Es ergibt sich

$$k \square \frac{m(r + 1) + (r - n - 1)}{n + 2}.$$

Die rechte Seite ist von  $k$  unabhängig. Abgesehen von Extremfällen (wenn die rechte Seite negativ oder größer als  $m$  ist) ist damit für wachsende  $k$  das Symbol  $\square$  zunächst durch „<“, einmal evtl. durch „=“ (wenn die rechte Seite ganzzahlig ist) und danach durch „>“ zu ersetzen. Das bedeutet, dass die  $h(k, m; r, n)$  als Funktion von  $k$  zunächst wachsen, dass dann möglicherweise zwei gleiche Werte auftreten und dass die Folge dann fällt.

Wo geht das Steigen in das Fallen über? Wenn die Zahlen  $n$  und  $r$  genügend groß sind und  $n - r$  gegen  $mr$  vernachlässigbar ist, kann die rechte Seite durch  $rm/n$  approximiert werden. Das größte  $h(k, m; r, n)$  ist also bei einem  $k$  zu erwarten, für das  $k/m \approx r/n$  gilt. Das stimmt mit der obigen Vermutung überein, dass nämlich der Anteil der roten Kugeln in der Stichprobe am wahrscheinlichsten dem Anteil der roten Kugeln in der Gesamtauswahl entspricht.

Mit dem gleichen Verfahren kann man auch die Variation anderer Parameter studieren. Wenn zum Beispiel  $k, m$  und  $n$  fest sind, kann man nach dem  $r$  fragen, für das  $h(k, m; r, n)$  maximal ist. Eine kurze Rechnung zeigt: Es gilt

$$\frac{h(k, m; r, n)}{h(k, m; r - 1, n)} = \frac{r}{r - k} \frac{n - r - m - k - 1}{n - r - 1},$$

und dieser Ausdruck ist genau dann  $\square 1$ , wenn  $k(n+1)/m \square r$ . Der Maximalwert wird damit bei  $[k(n+1)/m]$  erreicht<sup>21)</sup>. Vernachlässigt man das „Abschneiden“ und approximiert man  $n + 1$  durch  $n$ , so heißt das, dass die größte Wahrscheinlichkeit bei  $n(k/m)$  zu erwarten ist, also dann, wenn der Anteil der roten Kugeln im Kasten gleich dem in der Stichprobe ist.

---

<sup>21)</sup>Für eine reelle Zahl  $x$  bezeichnet  $[x]$  die größte ganze Zahl  $z$  mit  $z \leq x$ .

An dieser Stelle soll die Gelegenheit genutzt werden, ein erstes Verfahren aus der *Statistik* zu erläutern. Es handelt sich um eine Formalisierung des „gesunden Menschenverstands“:

### Maximum-likelihood-Schätzungen

Zur Motivation stellen wir uns vor, dass auf dem Tisch zwei Kartenstapel von jeweils 20 Karten liegen. Der eine, Stapel 1, enthält fünf rote und 15 schwarze Karten, und bei Stapel 2 ist es umgekehrt. Nun wird einer der Stapel entfernt, uns ist nicht bekannt, ob nun Stapel 1 oder Stapel 2 auf dem Tisch liegt. Es wird gemischt, wir ziehen eine Karte, sie ist rot. Und danach sollen wir eine Vermutung äußern, aus welchem der beiden Stapel wir wohl gezogen haben. Klar, dass wir auf Stapel 2 tippen, denn da war die Wahrscheinlichkeit für „rot“ wesentlich größer.

Etwas allgemeiner und abstrakter kann das Verfahren so formulieren. Es ist  $\Omega = \{1, \dots, n\}$  gegeben, außerdem gibt es einen „Vorrat“  $\mathbb{P}_1, \dots, \mathbb{P}_k$  von Wahrscheinlichkeitsmaßen auf  $\Omega$ . Irgendjemand wählt ein  $\mathbb{P}_i$  aus (das  $i$  wird uns aber nicht verraten),  $\Omega$  wird damit zu einem Wahrscheinlichkeitsraum. Es wird einmal abgefragt, das Ergebnis sei ein  $m \in \Omega$ , und wir sollen schätzen, welches  $\mathbb{P}_i$  wohl ausgewählt worden war. Eine naheliegende Lösung ist dann die *maximum-likelihood-Schätzung*: Suche dasjenige  $i$  (das hoffentlich eindeutig bestimmt ist), für das  $\mathbb{P}_i(\{m\})$  so groß wie möglich ist, also das Maß, das mit der höchsten Wahrscheinlichkeit das konkret erzeugte  $m$  liefert hätte.

(„Maximum likelihood“ heißt übrigens so viel wie „größte Wahrscheinlichkeit“. Allgemein akzeptierte Vorschläge für eine deutsche Bezeichnung hat es noch nicht gegeben.)

Im täglichen Leben argumentieren wir übrigens ähnlich: Wenn zum Beispiel in dem Mietshaus, in dem Sie wohnen, am Abend schon wieder einmal nicht abgeschlossen wurde, verdächtigen alle diejenige Familie, die das schon öfter so gemacht hat.

Doch zurück zur hypergeometrischen Verteilung. Da haben wir gerade nachgewiesen, dass  $[k(n+1)/m]$  eine Maximum-likelihood-Schätzung für  $r$  bei bekannten  $k, m, n$  ist. Zum Beispiel schätzen wir  $r$  als 400, wenn  $n = 1000$  ist und in einer Stichprobe von 10 Kugeln vier rote dabei waren (denn  $[4 \cdot 1000/10] = 400$ ). Ganz analog zeigt man, dass  $[rm/k]$  eine Maximum-likelihood-Schätzung für  $n$  bei bekannten  $r, m, k$  ist. ( $n/r$  soll also möglichst nahe bei  $m/k$  sein.)

Eine beliebte Verkleidung des Verfahrens sieht so aus. Sie haben einen Fischteich geerbt und wollen wissen, wie viele Karpfen drin sind. Dazu fangen Sie 10 Karpfen, markieren sie und setzen sie wieder aus. Nach ein paar Tagen fangen Sie noch einmal Fische, diesmal sind es 20 Karpfen, und davon sind 2 markiert. Hier ist  $n$  die unbekannte Anzahl der Karpfen,  $r = 10$ ,  $k = 2$  und  $m = 20$ . Die Schätzung für  $n$  ist also  $[20 \cdot 10/2] = 100$ .

Wir wollen noch den *Erwartungswert der hypergeometrischen Verteilung* ausrechnen: Mit wie vielen roten Kugeln kann man im Mittel rechnen? Das Ergebnis

wird – das ist ja auch plausibel –  $m \cdot r/n$  sein, für den Beweis betrachten wir geeignete Zufallsvariable.

Dazu modifizieren wir den Wahrscheinlichkeitsraum etwas.  $\Omega'$  soll aus allen möglichen Ziehungen (ohne Zurücklegen) von  $m$  Kugeln bestehen, wobei die Kugeln nummeriert sind und es jetzt auf die Reihenfolge ankommen soll.  $\Omega'$  hat  $n \cdot (n - 1) \cdots (n - m + 1)$  Elemente, diese Menge wird mit der Gleichverteilung versehen. Auf  $\Omega'$  werden Zufallsvariable  $X_1, \dots, X_m$  definiert, und zwar ist  $X_i$  bei einem  $\omega$  gleich Eins, wenn die  $i$ -te ausgewählte Kugel rot ist und Null sonst. Damit zählt  $X := X_1 + \cdots + X_m$  die Anzahl der roten Kugeln in  $\omega$ , und folglich ist  $\mathbb{P}_X$  die hypergeometrische Verteilung.

Wie ist der Erwartungswert eines speziellen  $X_i$ ? Für  $i = 1$  ist diese Zahl offensichtlich gleich  $r/n$ , denn *das* ist der Anteil der  $\omega$ , bei denen die erste gezogene Kugel rot ist. Nun folgt ein Symmetrieargument. Fixiert man  $i$  und betrachtet man auf  $\Omega'$  die Abbildung, durch die der erste und der  $i$ -te Eintrag vertauscht werden, so ist das eine Bijektion. Folglich haben  $\{X_1 = 1\}$  und  $\{X_i = 1\}$  die gleiche Anzahl von Elementen, diese Mengen haben also die gleiche Wahrscheinlichkeit. Damit ist der Erwartungswert von  $X_i$  ebenfalls gleich  $r/n$ , und aus der Linearität des Erwartungswerts folgt (erwartungsgemäß)

$$\mathbb{E}(X) = \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_m) = \frac{m \cdot r}{n}.$$

### Das Übereinstimmungsparadoxon

An zwei Beispielen soll am Ende dieses Abschnitts noch demonstriert werden, wie man den Inklusion-Exklusion-Satz anwenden kann: Er eignet sich zur Berechnung von Wahrscheinlichkeiten für Vereinigungen, wenn man etwas über die Wahrscheinlichkeiten der Durchschnitte weiß.

Zunächst geht es um *Permutationen* auf der Menge  $\{1, \dots, n\}$ , wir notieren sie als  $n$ -Tupel, die aus allen Zahlen in  $\{1, \dots, n\}$  gebildet werden. Manchmal kommt es in einer Permutation vor, dass irgendein  $i$  an der  $i$ -ten Stelle steht, z.B. die 3 in  $(2, 4, 3, 1)$  oder sogar alle Zahlen in  $(1, 2, 3, 4)$ . Es ist aber nicht immer so, in  $(2, 1, 4, 3)$  etwa bleibt kein Element fest. Wie sind die entsprechenden Wahrscheinlichkeiten für eine zufällige Permutation?

Sei  $\Omega$  die Menge aller Permutationen. Sie hat  $n!$  Elemente, diese Zahl kann schon für mäßig große  $n$  gigantisch sein. Wir bezeichnen für  $i = 1, \dots, n$  mit  $E_i$  die Menge derjenigen Permutationen, bei denen  $i$  fest bleibt. Die Frage „Mit welcher Wahrscheinlichkeit bleibt ein  $i$  fest“ läuft dann darauf hinaus, die Wahrscheinlichkeit von  $\bigcup_{i=1}^n E_i$  zu bestimmen.

Sei (mit den Bezeichnungen von Satz 3.4.1)  $k \in \{1, \dots, n\}$  und  $I \in \Delta_k$  eine  $k$ -elementige Teilmenge von  $\{1, \dots, n\}$ . Dann hat  $E_I$  genau  $(n - k)!$  Elemente, denn  $k$  Elemente bleiben fest und die restlichen können beliebig permutiert werden. Die Wahrscheinlichkeit von  $E_I$  ist folglich  $\#E_I/\#\Omega = (n - k)!/n!$ . Da

es  $\binom{n}{k}$  solche  $I$  gibt, erhalten wir die Formel

$$\begin{aligned}\mathbb{P}\left(\bigcup_i E_i\right) &= \binom{n}{1} \frac{(n-1)!}{n!} - \binom{n}{2} \frac{(n-2)!}{n!} + \cdots + (-1)^{n+1} \binom{n}{n} \frac{(n-n)!}{n!} \\ &= 1 - \frac{1}{2!} + \frac{1}{3!} - \cdots + (-1)^{n+1} \frac{1}{n!}.\end{aligned}$$

Manchen wird diese Summe bekannt vorkommen, sie ist beinahe eine Partialsumme der schnell konvergierenden Reihe für  $e^{-1} = 1 + (-1)/1! + (-1)^2/2! + (-1)^3/3! \dots$ . Mit dieser Beobachtung folgt

$$\begin{aligned}\mathbb{P}\left(\bigcup_i E_i\right) &= 1 - \left(1 - 1 + \frac{1}{2!} - \frac{1}{3!} + \cdots + (-1)^{n+1} \frac{1}{n!}\right) \\ &\approx 1 - e^{-1} \\ &= 0.632\dots\end{aligned}$$

Die Wahrscheinlichkeit für eine Übereinstimmung ist also überraschend hoch. Viele hätten einen winzigen Wert erwartet, wegen dieser Diskrepanz spricht man vom *Übereinstimmungsparadoxon*.

*Illustrationen* gibt es reichlich. Ein (wohl nicht mehr ganz zeitgemäßer) Klassiker ist die *verwirrte Sekretärin*, die ganz zufällig  $n$  persönliche Briefe in  $n$  beschriftete Umschläge steckt. Da ist die Wahrscheinlichkeit über 63 Prozent, dass mindestens ein Brief im richtigen Umschlag landet.

Der Autor kann ein *weiteres Beispiel* beisteuern. Für ein improvisiertes Theaterstück sollten bei einem großen Fest Paare zufällig zusammen gewürfelt werden: Die Damen schrieben ihren Namen auf einen Zettel, und jeder der Herren zog einen davon. Erst nach mehreren Versuchen war es so, dass niemand mit seinem/ihrem Partner zusammengelost wurde.

Das Ganze eignet sich auch für eine *Kneipenwette*. Man sortiere (zum Beispiel) die Kreuz-Karten aus einem Skatspiel aus und gebe sie jemandem zum Mischen. Man selbst bekommt die verdeckten acht Karten und legt sie dann einzeln – von oben anfangend – mit der Bildseite nach oben auf den Tisch, dabei sagt man laut „Sieben, Acht, …, König, As“. Die Wette: Irgendwann kommt genau die Karte, die gerade genannt wird. Es sollte nicht schwer sein, jemanden zu finden, der dagegen wettet. Und die eigene Gewinnchance ist fast zwei Drittel.

### Das Popstarbildchen-Problem

Hier folgt die zweite Anwendung von Satz 3.4.1. Wir beginnen mit einer abstrakten Formulierung. In einem großen Kasten befinden sich jeweils  $n$  Kugeln mit den Farben  $F_1, \dots, F_s$ , also insgesamt  $n \cdot s$  Kugeln. Es wird kräftig gemischt und dann wird  $m$ -mal ohne Zurücklegen gezogen (wobei  $m \geq s$  und  $m \leq n \cdot s$  sein soll). Wie wahrscheinlich ist es, dass von jeder Farbe eine Kugel dabei ist? Naheliegende Illustrationen sind:

- Wie wahrscheinlich ist es, dass ein Skatblatt Kreuz-, Pik-, Herz- und Karokarten enthält? (Hier ist  $n = 8$ ,  $s = 4$ ,  $r = 10$ .)



Bild 3.5.3: Sind alle Farben enthalten?

- Eine Firma versteckt in ihren Popcorn-Schachteln Bilder von Popstars (oder Fußballern, oder ...). Eine vollständige Serie besteht aus  $s$  Popstars, und jeder einzelne ist in  $n$  Schachteln zu finden. Man kauft  $m(\geq s)$  Schachteln, kann man sich auf einen vollständigen Satz freuen? Oder: Wie viele sollte man kaufen, damit die Sammlung mit mindestens 50 Prozent Wahrscheinlichkeit vollständig ist?

$\Omega$  soll die Menge der Auswahlen von  $m$  Kugeln aus dem Kasten bezeichnen, dafür gibt es  $\binom{s \cdot n}{m}$  Möglichkeiten, die alle die gleiche Wahrscheinlichkeit haben. Diesmal bezeichnen wir für  $i = 1, \dots, s$  mit  $E_i$  diejenigen Auswahlen von  $m$  Kugeln, bei denen die Farbe  $F_i$  nicht vorkommt.  $\bigcup_{i=1}^s E_i$  sind dann die Auswahlen, bei denen mindestens eine Farbe fehlt.

Für  $I \subset \{1, \dots, s\}$  setzen wir wieder  $E_I := \bigcap_{i \in I} E_i$ : Da fehlen alle  $F_i$ ,  $i \in I$ . Die  $m$  Kugeln wurden also aus den andersfarbigen Kandidaten gewählt, dafür gibt es  $\binom{n(s-\#I)}{m}$  Möglichkeiten. Es ergibt sich mit Satz 3.4.1 für die Wahrscheinlichkeit, dass alle Farben dabei sind, also für  $1 - \mathbb{P}(\bigcup_{i=1}^s E_i)$ , der Wert

$$1 - \sum_{i=1}^s (-1)^{i+1} \frac{\#E_i}{\#\Omega} \\ = 1 - \sum_{i=1}^s (-1)^{i+1} \binom{s}{i} \frac{[n(s-i)][n(s-i)-1] \cdots [n(s-i)-m+1]}{(ns)(ns-1) \cdots (ns-m+1)}.$$

Wenn  $n$  groß gegen  $m$  und  $s$  ist, kann man den Zähler (bzw. den Nenner) des  $i$ -ten Summanden durch  $(n(s-i))^r$  (bzw.  $(ns)^r$ ) approximieren, und man erhält die etwas übersichtlichere Näherungsformel

$$1 - \sum_{i=1}^s (-1)^{i+1} \binom{s}{i} \left(1 - \frac{i}{s}\right)^m$$

für die gesuchte Wahrscheinlichkeit.

## 3.6 Ergänzungen

Zufallsvariable, für die der Erwartungswert nicht existiert

Unsere Herleitung des Erwartungswertes ging von der Annahme aus, dass Wahrscheinlichkeiten als relative Erfolgshäufigkeiten („ $\omega \in E!$ “) interpretiert werden können. Das wird in Abschnitt 8.3 auch präzisiert und bewiesen werden. In der Definition spielte die Existenz von Reihensummen und Integralen eine wichtige Rolle, und man kann sich fragen, ob das eine notwendige Voraussetzung ist.

Wir betrachten als Beispiel die *Cauchyverteilung*, sie hat auf  $\mathbb{R}$  die Dichtefunktion  $1/(\pi(1+x^2))$ . Die Zufallsvariable  $X : \mathbb{R} \rightarrow [0, +\infty[, X(x) := x$  hat dann keinen Erwartungswert, denn das Integral  $\int_{\mathbb{R}} |x|/(1+x^2) dx$  ist nicht endlich.

Trotzdem wäre natürlich denkbar, dass sich Zufallsabfragen von  $X$  „ausmitteln“, denn die Dichtefunktion ist symmetrisch. Gehen die Mittelwerte der Abfragen gegen Null?

Das ist nicht der Fall. Da die Dichtefunktion der Cauchyverteilung sehr langsam abfällt, zum Beispiel wesentlich langsamer als die der Standard-Normalverteilung, sind bei Abfragen immer einmal wieder sehr große positive und negative Werte zu erwarten. Das führt dazu, dass sich die Folge  $(x_1 + \dots + x_n)/n$  auch bei großen  $n$  nicht „stabilisiert“: Es gibt keine Zahl  $z$ , so dass bei immer neuen Abfragen dieser Mittelwert in der Regel nicht weit entfernt von  $z$  liegt.

Im folgenden Bild sind für jeweils etwa 1000 Simulationen  $x_1, x_2, \dots$  die Mittelwerte  $(x_1 + \dots + x_n)/n$  aufgetragen. Bei der Normalverteilung (oben) kann man glauben, dass Konvergenz gegen den Erwartungswert Null vorliegt, bei der Cauchyverteilung (unten) gibt es immer wieder Sprünge<sup>22)</sup>. Mehr zu diesem Phänomen findet man in Kapitel 8.3 auf Seite 242.

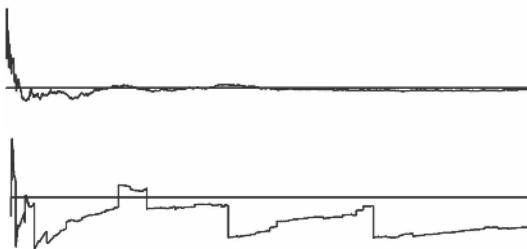


Bild 3.6.1: Verhalten der Mittelwerte bei Normalverteilung und Cauchyverteilung.

---

<sup>22)</sup>Die Cauchyverteilung kann man übrigens so simulieren: Es werden standardnormalverteilte Zahlen  $x, y$  erzeugt, dann wird  $x/y$  ausgegeben. Für eine Begründung reicht der Platz in einer Fußnote nicht.

**Das Petersburger Paradoxon**

Stellen Sie sich vor, dass man Ihnen das folgende Spiel anbietet. Sie zahlen eine gewisse Spielgebühr  $S$ , dann geht es los:

- Eine faire Münze wird geworfen. Wenn Sie Kopf zeigt, bekommen Sie einen Euro, und das Spiel ist zu Ende. Andernfalls geht das Spiel weiter.
- Die Münze wird noch einmal geworfen. Wenn Sie Kopf zeigt, bekommen Sie 2 Euro, und das Spiel ist zu Ende. Andernfalls wird das Spiel fortgesetzt.
- Und so weiter. In der nächsten Runde wäre der Gewinn 4 Euro, dann 8, 16 usw.

Das hört sich für Sie attraktiv an, immerhin könnte man eine gigantische Summe gewinnen, wenn die Münze erst nach vielen Durchgängen Kopf zeigt.

Das Problem ist allerdings, dass wir noch nicht erfahren haben, wie groß der Spieleinsatz  $S$  ist.

*Bevor Sie weiterlesen: Für welchen Einsatz  $S$  würden Sie das Spiel spielen? 50 Euro? 100 Euro?*

Aus Sicht des Veranstalters sollte der Einsatz mindest so groß sein wie der Erwartungswert des Gewinns. Der ist leicht auszurechnen: Mit Wahrscheinlichkeit  $1/2$  gibt es einen Euro, dazu mit Wahrscheinlichkeit  $1/4$  zwei Euro, usw. Insgesamt sind das  $\sum_{k=1}^{\infty} 2^{k-1}/2^k = +\infty$  Euro. Das ist eine ganze Menge, dieses Spiel sollte man aus Sicht des Veranstalters also besser nicht anbieten. Dass der Einsatz so gewaltig sein muss, widerspricht der Intuition. Sicher hätten die meisten schon protestiert, wenn einige hundert Euro verlangt worden wären. Dieses „Petersburger Paradoxon“ zeigt, dass man im Fall eines unendlichen Erwartungswertes merkwürdige Phänomene antreffen kann.

**Die hypergeometrische Verteilung: Mehr als zwei Kugelfarben**

Unsere Überlegungen zur hypergeometrischen Verteilung sind leicht zu verallgemeinern. Es seien  $n$  Kugeln in einem Kasten, und zwar  $f_1$  mit der Farbe  $F_1$ ,  $f_2$  mit der Farbe  $F_2$ , ...,  $f_l$  mit der Farbe  $F_l$ . (Es ist also  $f_1 + \dots + f_l = n$ .)

Nun werden  $m$  Kugeln ohne Zurücklegen gezogen. Wie wahrscheinlich ist es, dass  $k_1$  Kugeln der Farbe  $F_1$ , ...,  $k_l$  Kugeln der Farbe  $F_l$  dabei sind? Das muss nur ausgerechnet werden, wenn jeweils  $0 \leq k_i \leq f_i$  gilt, denn andernfalls ist die Wahrscheinlichkeit Null.

Die Lösung ist einfach. Es gibt  $\binom{n}{m}$  Auswählen, und um zu dem richtigen Ergebnis zu kommen, müssen  $k_1$  aus den  $f_1$  Kugeln mit der Farbe  $F_1$  gewählt werden ( $\binom{f_1}{k_1}$  Möglichkeiten), dazu  $k_2$  aus den  $f_2$  Kugeln mit der Farbe  $F_2$  ( $\binom{f_2}{k_2}$  Möglichkeiten), usw. So erhalten wir als Wert für die gesuchte Wahrscheinlichkeit die Zahl

$$\frac{\binom{f_1}{k_1} \cdots \binom{f_l}{k_l}}{\binom{n}{m}}.$$

Der Spezialfall  $n = 2$  entspricht der vorstehend behandelten hypergeometrischen Verteilung.

Wahrscheinlichkeitserzeugende Funktionen

Wahrscheinlichkeitsmaße auf diskreten Räumen lassen sich manchmal durch eine einzige Funktion beschreiben:

**Definition 3.6.1.** Es sei  $\mathbb{P}$  ein Wahrscheinlichkeitsmaß auf  $\mathbb{N}_0$ . Definiere eine Funktion  $\phi_{\mathbb{P}}$  durch  $\phi_{\mathbb{P}}(x) := \sum_{n=0}^{\infty} \mathbb{P}(\{n\})x^n$ . Da die Beträge der Koeffizienten  $\mathbb{P}(\{n\})$  durch Eins beschränkt sind, existiert die Reihensumme mindestens für die  $x$  mit  $|x| < 1$ . (Manchmal sind auch alle  $x$  zulässig, sicher zum Beispiel dann, wenn nur endlich viele  $\mathbb{P}(\{n\})$  von Null verschieden sind.)  
 $\phi_{\mathbb{P}}$  heißt die zu  $\mathbb{P}$  gehörige wahrscheinlichkeitserzeugende Funktion<sup>23)</sup>.

Ist  $\phi(x) = \sum_{n=0}^{\infty} a_n x^n$  eine Potenzreihe mit positivem Konvergenzradius, so sind die  $a_n$  durch  $\phi$  eindeutig bestimmt, denn  $a_n = \phi^{(n)}(0)/n!$  für alle  $n$ . Hier heißt das, dass  $\mathbb{P}$  durch  $\phi_{\mathbb{P}}$  eindeutig bestimmt ist, oder anders ausgedrückt: Die Abbildung  $\mathbb{P} \mapsto \phi_{\mathbb{P}}$  ist injektiv.

Hier einige Beispiele:

1. Ist  $\mathbb{P}(\{n\}) = 0$  für  $n \geq n_0$ , so ist  $\phi_{\mathbb{P}}$  ein Polynom. Ist zum Beispiel  $\mathbb{P}$  die Gleichverteilung auf  $\{0, \dots, n-1\}$ , so ist  $\phi_{\mathbb{P}}(x) = (1 + x + \dots + x^{n-1})/n$ .
2. Für die Poissonverteilung zum Parameter  $\lambda$  ergibt sich

$$\begin{aligned}\phi_{\mathbb{P}}(x) &= \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} x^n e^{-\lambda} \\ &= \sum_{n=0}^{\infty} \frac{(\lambda x)^n}{k!} e^{-\lambda} \\ &= e^{-\lambda(1-x)}.\end{aligned}$$

3. Für die geometrische Verteilung zum Parameter  $q$  erhalten wir

$$\begin{aligned}\phi_{\mathbb{P}}(x) &= \sum_{n=1}^{\infty} (1-q)q^{n-1}x^n \\ &= x(1-q)(1+(qx)+(qx)^2+(qx)^3+\dots) \\ &= \frac{x(1-q)}{1-qx}.\end{aligned}$$

Das alles hätte schon in den Ergänzungen zu Kapitel 1 stehen können. Es wird aber erst hier aufgeführt, weil es erst nach Kapitel 2 möglich ist, auf den Zusammenhang zu Erwartungswerten einzugehen.

---

<sup>23)</sup>Oder auch kürzer *erzeugende Funktion*.

Ist  $\mathbb{P}$  ein Wahrscheinlichkeitsmaß auf  $\mathbb{N}_0$  und bezeichnet  $X$  die Zufallsvariable  $n \mapsto n$ , so ist doch  $\mathbb{E}(X) = \sum_{n=0}^{\infty} n\mathbb{P}(\{n\})$ , falls die rechts stehende Reihe absolut konvergent ist. Der gleiche Wert ergibt sich aber auch für  $\phi'_{\mathbb{P}}(1)$ , den Wert der Ableitung von  $\phi_{\mathbb{P}}$  bei Eins. (Wir wollen annehmen, dass der Konvergenzradius der Potenzreihe, durch die  $\phi_{\mathbb{P}}$  definiert ist, größer als Eins ist. Dann kann  $\phi_{\mathbb{P}}$  nämlich gliedweise differenziert werden.) Im Fall der Poissonverteilung etwa ist die Ableitung gleich  $\lambda e^{-\lambda(1-x)}$ , und der Wert bei Eins ist wirklich gleich  $\lambda$ , dem Erwartungswert.

In den nachstehenden Kapiteln wird die erzeugende Funktion allerdings nur eine untergeordnete Rolle spielen.

## 3.7 Verständnisfragen

### Zu Abschnitt 3.1

#### *Sachfragen*

**S1:** Was ist eine Zufallsvariable?

**S2:** Warum muss man die eher technische Bedingung an die Urbilder der Elemente der Bild- $\sigma$ -Algebra fordern?

**S3:** Durch welche einfache Bedingung sind reellwertige Zufallsvariable charakterisiert?

**S3:** Was versteht man unter der Indikatorfunktion eines Ereignisses?

#### *Methodenfragen*

**M1:** Nachprüfen können, ob eine vorgelegte Abbildung eine Zufallsvariable ist.

**M2:** Beweise im Zusammenhang mit Zufallsvariablen führen können.

### Zu Abschnitt 3.2

#### *Sachfragen*

**S1:** Wie wird durch eine Zufallsvariable ein neuer Wahrscheinlichkeitsraum definiert?

**S2:** Wie kann man das neue Maß im diskreten Fall und in manchen Fällen für Räume mit Dichten berechnen?

#### *Methodenfragen*

**M1:** Das induzierte Maß im diskreten Fall berechnen können.

**M2:** Bei Räumen mit Dichtefunktionen die Dichtefunktion für das induzierte Maß bestimmen können, wenn die entsprechenden Voraussetzungen erfüllt sind.

### Zu Abschnitt 3.3

#### *Sachfragen*

**S1:** Welche Größe soll durch den Begriff *Erwartungswert* präzisiert werden?

**S2:** Wie ist der Erwartungswert definiert: für diskrete Räume? für Räume mit Dichten? für beliebige Wahrscheinlichkeitsräume?

**S3:** Warum muss man bei der Definition von  $\mathbb{E}(X)$  im diskreten Fall verlangen, dass die Reihe  $\sum_{n=1}^{\infty} X(n)\mathbb{P}(\{n\})$  absolut konvergent ist, warum reicht Konvergenz nicht aus?

**S4:** Welchen Erwartungswert haben Gleichverteilung, Bernoulliverteilung, Poissonverteilung, geometrische Verteilung, Exponentialverteilung und Normalverteilung?

**S5:** Was sind Varianz und Streuung einer Zufallsvariablen?

**S6:** Wie kann man den Erwartungswert von  $X$  durch  $\mathbb{P}_X$  ausdrücken?

*Methodenfragen*

**M1:** Erwartungswerte und Varianzen für konkrete Zufallsvariable berechnen können.

**M2:** Einfache Eigenschaften zu Erwartungswerten beweisen können.

### Zu Abschnitt 3.4

*Sachfragen*

**S1:** Wie viele  $k$ -elementige Teilmengen einer  $n$ -elementigen Menge gibt es?

**S2:** Welche vier fundamentalen Zählprobleme spielen in der elementaren Kombinatorik eine wichtige Rolle? Wie lauten die entsprechenden Formeln?

**S3:** In welchen Situationen wird der Inklusion-Exklusion-Satz angewendet?

*Methodenfragen*

**M1:** Entscheiden können, welches der vier Zählprobleme vorliegt (Wiederholung erlaubt? Reihenfolge wichtig?) und die entsprechenden Rechnungen durchführen können.

**M2:** Den Inklusion-Exklusion-Satz anwenden können.

### Zu Abschnitt 3.5

*Sachfragen*

**S1:** Was versteht man unter dem Geburtstagsparadoxon?

**S2:** Welche Wahrscheinlichkeiten werden mit der hypergeometrischen Verteilung beschrieben?

**S3:** Was ist eine maximum-likelihood-Schätzung?

**S4:** Was versteht man unter dem Übereinstimmungsparadoxon?

*Methodenfragen*

**M1:** Die Paradoxien erklären können.

**M2:** Maximum-likelihood-Schätzungen ermitteln können.

**M3:** Die Formel für die hypergeometrische Verteilung herleiten können.

## 3.8 Übungsaufgaben

### Zu Abschnitt 3.1

**Ü3.1.1** Es seien  $X_n, n = 1, 2, \dots$  reellwertige Zufallsvariable (dabei seien die reellen Zahlen mit den Borelmengen als  $\sigma$ -Algebra versehen). Dann sind die folgenden Mengen Ereignisse:

- a) Die  $\omega$ , bei denen mindestens ein  $X_n$  positiv ist.
- b) Die  $\omega$ , bei denen alle  $X_n$  positiv sind.
- c) Die  $\omega$ , bei denen die Folge  $(X_n(\omega))_n$  beschränkt ist.

**Ü3.1.2** Es sei  $X$  eine reellwertige Zufallsvariable mit der Eigenschaft, dass für jedes  $c$  die Wahrscheinlichkeit  $\mathbb{P}(\{X \geq c\})$  entweder 0 oder 1 ist. Zeigen Sie, dass  $X$  dann „im Wesentlichen konstant“ sein muss. Genauer: Es gibt ein  $c_0$ , so dass  $\mathbb{P}(\{X = c_0\}) = 1$ .

**Ü3.1.3** Der  $\mathbb{R}^2$  sei mit den Borelmengen versehen, und  $(\Omega, \mathcal{E}, \mathbb{P})$  sei ein Wahrscheinlichkeitsraum. Eine vektorwertige Abbildung  $X : \Omega \rightarrow \mathbb{R}^2$  sei durch  $X(\omega) := (X_1(\omega), X_2(\omega))$  definiert, wobei  $X_1, X_2 : \Omega \rightarrow \mathbb{R}$ .

Beweisen Sie:  $X$  ist genau dann Zufallsvariable, wenn  $X_1$  und  $X_2$  Zufallsvariablen sind.

**Ü3.1.4**  $X$  sei eine reellwertige Zufallsvariable auf  $\Omega$ . Definiere  $Y : \Omega \rightarrow \mathbb{R}$  durch  $Y(\omega) := X(\omega)$  für  $|X(\omega)| < 1$  und  $Y(\omega) := 0$  sonst. Zeigen Sie, dass auch  $Y$  eine Zufallsvariable ist.

**Ü3.1.5** Sei  $X$  eine Zufallsvariable. Beweisen Sie:  $X$  ist genau dann fast sicher konstant, wenn es keine Zahl  $a$  mit der Eigenschaft

$$\mathbb{P}(\{X < a\}) > 0 \text{ und } \mathbb{P}(\{X > a\}) > 0$$

gibt. Dabei heißt  $X$  fast sicher konstant, wenn es ein Ereignis  $N$  mit  $\mathbb{P}(N) = 0$  und eine Zahl  $c$  so gibt, dass  $X(\omega) = c$  für  $\omega \notin N$ .

Tipp: In der schwierigeren Beweisrichtung könnte man es mit der Konstanten  $c := \sup\{a \mid P(\{X < a\}) = 0\}$  versuchen. Zeigen Sie zuerst, dass das wirklich eine reelle Zahl ist (Teil 1: Es kann nicht sein, dass  $\mathbb{P}(\{X < a\}) = 0$  für alle  $a$  gilt; Teil 2: Es gibt ein  $a$  mit  $\mathbb{P}(\{X < a\}) = 0$ .)

### Zu Abschnitt 3.2

**Ü3.2.1**  $(\Omega, \mathcal{E}, \mathbb{P})$  modelliere das Werfen von drei fairen Würfeln.  $(\Omega, \mathcal{E}, \mathbb{P})$  ist also  $\{1, \dots, 6\}^3$  mit der Gleichverteilung. Bestimmen Sie  $\mathbb{P}_X$  für die Zufallsvariable  $X : (i, j, k) \rightarrow i - j + k$ .

**Ü3.2.2**  $\Omega = [1, 27]$  sei mit der Dichte  $f(x) = cx$  versehen ( $c$  ist eine geeignete Konstante). Eine Zufallsvariable  $X : \Omega \rightarrow \mathbb{R}$  sei durch  $X(x) := \sqrt[3]{x}$  definiert. Bestimmen Sie die Dichte von  $\mathbb{P}_X$  und berechnen Sie damit  $\mathbb{P}(\{X \in [1, 2]\})$ .

**Ü3.2.3** Bei den neuen ALDI-Computern hat es eine Panne gegeben. Der Zufallsgenerator wirft nicht, wie beabsichtigt, in  $[0, 1]$  gleichverteilte Zufallszahlen aus, sondern Zahlen, die gemäß der Dichte  $f(x) := 2x$  (auf  $[0, 1]$ ) verteilt sind. Die

Computer werden zu einem Spottpreis verkauft. Stellen Sie die `random`-Funktion durch ein geeignetes Unterprogramm wieder her!

(Genauer heißt das: Sie sollen eine Zufallsvariable  $X : [0, 1] \rightarrow [0, 1]$  so finden, dass  $\mathbb{P}_X$  die Gleichverteilung auf  $[0, 1]$  ist. Falls Sie übrigens während des Beweises eine Funktion  $g$  brauchen, für die (für alle  $x$ )  $g'(x)g(x) = 0.5$  gilt, so versuchen Sie es doch mit  $g(x) = \sqrt{x}$ .)

**Ü3.2.4**  $X$  sei exponentialverteilt zum Parameter  $\lambda$ . Zeigen Sie, dass  $cX$  für  $c > 0$  ebenfalls exponentialverteilt ist. Wie groß ist der zugehörige Parameter?

**Ü3.2.5** Beweisen Sie ein zu Satz 3.2.3 analoges Ergebnis für streng monoton fallende Zufallsvariable.

**Ü3.2.6** Es sei  $f : [-1, 1] \rightarrow [0, +\infty[$  eine stetige Dichtefunktion. Weiter sei  $X : [-1, 1] \rightarrow \mathbb{R}$  sei eine stetig differenzierbare Zufallsvariable. Wir setzen voraus, dass  $X$  auf dem Teilintervall  $[0, 1]$  streng monoton steigt. Bestimmen Sie in Analogie zu Satz 2.4.2 eine Dichtefunktion für  $\mathbb{P}_X$ . Berechnen Sie diese Dichtefunktion konkret für den Fall  $f(x) = (1+x)/2$  und  $X(x) = x^2$ .

**Ü3.2.7**  $F_X$  sei die Verteilungsfunktion einer reellwertigen Zufallsvariablen  $X$ . Zeigen Sie:

- $F_X$  besitzt höchstens abzählbar unendlich viele Unstetigkeitsstellen.
- $F_X$  ist unstetig bei einem  $x \in \mathbb{R}$  genau dann, wenn  $P_X(\{x\}) > 0$  gilt.

### Zu Abschnitt 3.3

**Ü3.3.1** In einem Kasino wird nach folgenden Regeln mit drei Würfeln gespielt: Ein Spieler bekommt 1000 Euro für drei Sechsen, 100 Euro für zwei Sechsen und 10 Euro für eine Sechs. In allen anderen Fällen gibt es gar nichts. Welchen Mindesteinsatz wird der Kasinobetreiber verlangen, wenn er nicht draufzahlen will?

**Ü3.3.2** Bestimmen Sie die Varianz und die Streuung der Poissonverteilung.

**Ü3.3.3** Bestimmen Sie die Varianz und die Streuung der geometrischen Verteilung.

**Ü3.3.4** Bestimmen Sie die Varianz und die Streuung der Gleichverteilung auf dem Intervall  $[a, b]$ .

**Ü3.3.5** Bestimmen Sie die Varianz und die Streuung der Exponentialverteilung.

**Ü3.3.6** Es sei  $X : \Omega \rightarrow \mathbb{R}$ . Zeigen Sie, dass  $\mathbb{E}(X)$  genau dann existiert, wenn  $\sum_n \mathbb{P}(X \geq n) < \infty$  gilt.

Zusatz: Ist  $X : \Omega \rightarrow \{0, 1, 2, \dots\}$ , so gilt sogar  $\mathbb{E}(X) = \sum_n \mathbb{P}(X \geq n)$  im Fall  $\sum_n \mathbb{P}(X \geq n) < \infty$ .

**Ü3.3.7** Ein Stab der Länge 1 werde zufällig in zwei Teile zerbrochen (Bruchstelle gleichverteilt).

- Wie groß ist der Erwartungswert für  $X$ , wobei  $X$  die Länge des kleineren der beiden Stücke misst?
- Berechnen Sie auch den Erwartungswert des Quotienten “kürzeres Stück durch längeres Stück”.
- In diesem Aufgabenteil sei die Bruchstelle gleichverteilt in  $[0.5 - c, 0.5 + c]$ ,

wobei  $0 \leq c \leq 0.5$ . Die Zufallsvariable  $X$  sei wie in „a)“ definiert. Für welche Werte von  $c$  ist der Erwartungswert von  $X$  größer als 0.4?

### Zu Abschnitt 3.4

**Ü3.4.1** Es werden  $n$  Kugeln auf gut Glück auf  $n$  Fächer verteilt. Wie groß ist die Wahrscheinlichkeit dafür, dass genau ein Fach leer bleibt?

**Ü3.4.2** Eine endliche Folge  $(x_1, y_1), \dots, (x_n, y_n)$  in  $\mathbb{Z}^2$  soll ein *Treppenweg* genannt werden, wenn stets entweder  $(x_{k+1}, y_{k+1}) = (x_k, y_k) + (0, 1)$  gilt oder  $(x_{k+1}, y_{k+1}) = (x_k, y_k) + (1, 0)$ , wenn es also immer jeweils einen Schritt nach rechts oder nach oben geht.

Wie viele Treppenwege gibt es von  $(0, 0)$  nach  $(20, 30)$ ?

**Ü3.4.3** Wie viele verschiedene neunbuchstabige Wörter kann man aus den Buchstaben  $A, B, C, E, E, H, I, K, O, R, S, S, T, T, T, W, Z$  bilden? Jeder Buchstabe dieses Vorrats darf dabei nur einmal verwendet werden. („STOCHASTIK“ ist also zugelassen, „STOCKKORB“ aber nicht.)

**Ü3.4.4** Beweisen Sie die Formel  $\binom{n+1}{k} = \binom{n}{k} + \binom{n}{k-1}$  für  $k \leq n$ .

### Zu Abschnitt 3.5

**Ü3.5.1** Diskutiere die Wahrscheinlichkeiten für das neue Lotto:  
„5 aus 35, mit zwei Zusatzzahlen“.

Es werden also aus 35 nummerierten Kugeln 5 „Richtige“ und zwei Zusatzzahlen gezogen, und gefragt sind die interessierenden Wahrscheinlichkeiten („5 Richtige“, „4 Richtige mit Zusatzzahl“, „3 Richtige mit 2 Zusatzzahlen“, usw.).

**Ü3.5.2** Sei  $\sigma$  eine Zufallspermutation der Zahlen  $1, \dots, n$ . Weiter sei  $k$  eine natürliche Zahl,  $k \leq n$ . Gefragt ist nach der Wahrscheinlichkeit, dass in  $\sigma$  genau  $k$  Elemente festbleiben.

Schließen Sie aus der Formel, dass diese Wahrscheinlichkeit für festes  $k$  und genügend große  $n$  durch  $1/(k!e)$  approximiert werden kann.

Bei zufälligem Zusammenwürfeln von Tanzpaaren ist es also ziemlich wahrscheinlich, dass nicht nur ein Paar sondern gleich mehrere in der alten Zusammensetzung antreten.

**Ü3.5.3** In Italien spielen sie „6 aus 90“. Berechnen Sie die Wahrscheinlichkeit für 6 Richtige und denken Sie sich eine originelle Illustration aus, um diese geringe Wahrscheinlichkeit zu veranschaulichen.

**Ü3.5.4** Wie groß ist die Wahrscheinlichkeit, beim Lotto (6 aus 49) zwei aufeinanderfolgende Zahlen zu ziehen?

Tipp: Um die Anzahl der günstigen Ausgänge zu bestimmen, ist es hilfreich, die Elementarereignisse als 6-Tupel  $(n_1, \dots, n_6)$ ,  $n_i \in \{1, \dots, 49\}$  mit  $n_1 < n_2 < \dots < n_6$ , aufzufassen. Wie lassen sich dann die günstigen Ausgänge beschreiben? Finden Sie ein  $N \in \mathbb{N}$  mit  $N < 49$ , so dass die gesuchte Anzahl gerade die Anzahl aller Ziehungen „6 aus  $N$ “ ist.

**Ü3.5.5** Wir betrachten  $\mathbb{N}_0$ , versehen mit allen Poissonverteilungen. Ein (unbekannter) Parameter  $\lambda$  wird ausgewählt, dann wird einmal abgefragt. Das Ergebnis sei  $n$ . Finden Sie eine maximum-likelihood-Schätzung für  $\lambda$ .

# Kapitel 4

## Bedingte Wahrscheinlichkeiten

Auch dieses Kapitel hat ein Leitmotiv:

„Informationen verändern Wahrscheinlichkeiten“.

Was das genau bedeuten soll, präzisieren wir in *Abschnitt 4.1*. Wenn Informationen keine Auswirkungen haben, sprechen wir von *Unabhängigkeit*, das ist einer der wichtigsten Begriffe der Wahrscheinlichkeitstheorie. In *Abschnitt 4.2* werden bedingte Wahrscheinlichkeiten umgekehrt, genauer wird das durch die *Bayes-Formel* quantifiziert. Wieder einmal wird sich zeigen, dass unser „Bauchgefühl“ bei Wahrscheinlichkeiten recht unzuverlässig ist, das gilt ganz besonders in diesem Zusammenhang.

In *Abschnitt 4.3* und in *Abschnitt 4.4* wird das Thema „Unabhängigkeit“ fortgesetzt. Dazu machen wir uns zunächst klar, was es bedeutet, Informationen zu kombinieren, um definieren zu können, was Unabhängigkeit für mehr als zwei Ereignisse bedeutet. Dann übertragen wir die Definitionen von Ereignissen auf Zufallsvariable.

Es wird für theoretische Untersuchungen von fundamentaler Bedeutung sein zu wissen, dass man Zufallsabfragen beliebig oft „voneinander unbeeinflusst“ wiederholen kann. Was das genau bedeutet, kann nun präzisiert werden, und dass es geht, wird in *Abschnitt 4.5* bewiesen.

Und was nutzt das alles? Falls Unabhängigkeit vorliegt, gibt es die Möglichkeit, manche der Wahrscheinlichkeiten zu berechnen, über die man sonst keine konkreten Aussagen machen kann. Die entsprechenden Ergebnisse stehen in *Abschnitt 4.6*. Das Kapitel schließt in den *Abschnitten 4.7 und 4.8* mit Verständnisfragen und Übungsaufgaben.

## 4.1 Bedingte Wahrscheinlichkeiten: die Idee

Stellen Sie sich vor, dass wir gespannt auf den Ausgang eines Zufallsexperiments warten, zum Beispiel darauf, ob die Augensumme von zwei fairen Würfeln mindestens 7 ist. Es gibt unter den 36 gleichverteilten Möglichkeiten für den Ausgang des zwei-Würfel-werfen-Versuchs genau 21, für die das zutrifft:  $(6, 1), (6, 2), (5, 2)$  und 18 andere. Die Wahrscheinlichkeit, mit der das Gewünschte eintritt, ist also  $21/36 = 7/12$ .

Angenommen nun, wir bekommen – bevor das Ergebnis bekanntgegeben wird – die Information, dass der erste Würfel eine 5 zeigt. Jetzt sind unter den gleichverteilten  $(5, i)$ ,  $i = 1, \dots, 6$ , diejenigen interessant, für die  $5 + i \geq 7$  gilt, das ist unter den 6 möglichen Fällen gerade für  $i = 2, \dots, 6$  der Fall. Die Wahrscheinlichkeit, dass die Augensumme mindestens 7 ist, hat sich damit von  $7/12$  auf  $5/6$  erhöht.

Ganz anders würde es aussehen, wenn uns die Information „Der erste Würfel zeigt eine 1“ erreicht hätte. Jetzt führt nur noch eine von den 6 noch möglichen Fällen  $(1, i)$ ,  $i = 1, \dots, 6$ , zum Ziel, die Wahrscheinlichkeit ist nur noch  $1/6$ .

Allgemeiner lässt sich die Situation so beschreiben. Wir warten auf die Abfrage eines Wahrscheinlichkeitsraums: Ein  $\omega$  wird erzeugt, und wir interessieren uns für das Ergebnis  $\omega \in A$ , wobei  $A$  ein Ereignis ist. Die Wahrscheinlichkeit ist dann  $\mathbb{P}(A)$ .

Bevor wir das  $\omega$  sehen, verrät uns ein „Insider“, dass  $\omega$  in  $B$  liegt, wobei  $B$  ebenfalls ein Ereignis ist. Dadurch verändert sich die Situation schlagartig: Wir sind nicht mehr in  $(\Omega, \mathcal{E}, \mathbb{P})$ , sondern in der „Spur“ dieses Wahrscheinlichkeitsraums auf  $B$ . Elementarereignisse sind die Elemente aus  $B$ , Ereignisse sind die  $E \in \mathcal{E}$ , die in  $B$  enthalten sind, und die neuen Wahrscheinlichkeiten sind durch  $\mathbb{P}_B(E) := \mathbb{P}(E)/\mathbb{P}(B)$  für diese  $E$  definiert<sup>1)</sup>. Statt „ $\omega \in A$ ?“ interessiert uns jetzt die Antwort auf die Frage „ $\omega \in A \cap B$ ?“ in diesem neuen Raum.

Das führt zu der folgenden Definition, in der „Wahrscheinlichkeit von  $A$ , wenn schon bekannt ist, dass  $B$  eingetreten ist“ präzisiert wird:

**Definition 4.1.1.** Es sei  $(\Omega, \mathcal{E}, \mathbb{P})$  ein Wahrscheinlichkeitsraum, und  $A, B$  seien Ereignisse, wobei  $\mathbb{P}(B) > 0$  vorausgesetzt wird. Dann ist die bedingte Wahrscheinlichkeit von  $A$  unter der Voraussetzung  $B$  durch die Zahl

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

definiert. Man sagt auch kürzer „bedingte Wahrscheinlichkeit von  $A$  unter  $B$ “, und das Symbol  $\mathbb{P}(A|B)$  wird „ $P$  von  $A$  unter  $B$ “ ausgesprochen.

Zur Illustration betrachten wir einige Beispiele:

**1.** Beim obigen Würfelsversuch war  $A = \{(i, j) \in \Omega \mid i + j \geq 7\}$  und  $B = \{(5, j) \mid j = 1, \dots, 6\}$ . Dann hat  $A \cap B$  fünf Elemente, und  $B$  hat sechs Elemente, es

---

<sup>1)</sup>Das setzt natürlich voraus, dass  $\mathbb{P}(B)$  positiv ist.

folgt

$$\mathbb{P}(A | B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{5/36}{6/36} = \frac{5}{6}.$$

In der Variante „Der erste Würfel zeigt eine Eins“ ist  $B = \{(1, j) \mid j = 1, \dots, 6\}$ , und da  $A \cap B$  nur ein Element hat (nämlich  $(1, 6)$ ), beträgt die bedingte Wahrscheinlichkeit nur noch  $1/6$ .

Es sind auch noch extremere Situationen denkbar: Die Information „Beide Augenzahlen sind kleiner als 4“ reduziert  $\mathbb{P}(A | B)$  auf Null, und bei „Beide Augenzahlen sind größer als 3“ steigt die bedingte Wahrscheinlichkeit auf 1.

**2.** Wir betrachten noch einmal das Würfeln mit zwei Würfeln, diesmal interessieren wir uns für das Ereignis  $A = \{(i, j) \mid i = j\}$ : Stimmen die Augenzahlen überein?

Es ist  $\mathbb{P}(A) = 1/6$ , was nutzt uns die Information  $B$ : „Der erste Würfel zeigt eine 3“? Eine leichte Rechnung zeigt, dass auch  $\mathbb{P}(A | B)$  gleich  $1/6$  ist, die Kenntnis von  $\omega \in B$  hat also unsere Erwartung für  $A$  nicht verändert. Das kann also auch passieren, dieser Spezialfall wird eine besonders wichtige Rolle spielen.

**3.** Bedingte Wahrscheinlichkeiten können natürlich auch in komplizierteren Situationen untersucht werden. Zur Übung versehen wir  $[0, 1]$  mit dem durch die Dichtefunktion  $f(x) = 2x$  induzierten Wahrscheinlichkeitsmaß und fragen nach  $\mathbb{P}(A | B)$  für  $A = [0.2, 0.6]$  und  $B = [0.5, 0.9]$ . Es ist  $A \cap B = [0.5, 0.6]$ , und so folgt

$$\mathbb{P}(A | B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\int_{0.5}^{0.6} 2x \, dx}{\int_{0.5}^{0.9} 2x \, dx} = \frac{0.36 - 0.25}{0.81 - 0.25} = \frac{11}{56}.$$

**4.** Ziehung der Lottozahlen, Sie sitzen gebannt vor dem Fernseher. Die ersten drei gezogenen Zahlen haben Sie auf Ihrem Tippschein angekreuzt. Bevor die restlichen Zahlen gezogen werden, möchten Sie wissen, wie es denn nach dem bisher günstigen Verlauf um die Chancen für einen Hauptgewinn steht. Gefragt ist also nach der bedingten Wahrscheinlichkeit für sechs Richtige ( $= A$ ) unter der Voraussetzung „mindestens drei Richtige“ ( $= B$ ). Wegen  $A \subset B$  ist  $\mathbb{P}(A \cap B) = \mathbb{P}(A) = 1/13.983.816$ . Und  $\mathbb{P}(B)$  ist die Summe der Wahrscheinlichkeiten für 3, 4, 5 und 6 Richtige. So folgt

$$\mathbb{P}(A | B) = \frac{\binom{6}{3} \binom{43}{3} + \binom{6}{4} \binom{43}{2} + \binom{6}{5} \binom{43}{1} + 1}{\binom{49}{6}} \approx 3.7 \cdot 10^{-6}.$$

Das ist zwar fast 50 Mal besser als  $\mathbb{P}(A)$ , aber leider immer noch deprimierend niedrig.

Im nächsten Satz sind einige Eigenschaften der neuen Definition zusammengestellt:

**Satz 4.1.2.** *Es seien  $A, B, A_1, A_2, \dots$  Ereignisse in einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{E}, \mathbb{P})$ , wobei  $\mathbb{P}(B) > 0$ . Dann gilt:*

(i) Sind  $A$  und  $B$  disjunkt, so ist  $\mathbb{P}(A | B) = 0$ , und aus  $A \supset B$  folgt  $\mathbb{P}(A | B) = 1$ . Insbesondere ist stets  $\mathbb{P}(\emptyset | B) = 0$  und  $\mathbb{P}(\Omega | B) = 1$ .

(ii)  $\mathbb{P}(\Omega \setminus A | B) = 1 - \mathbb{P}(A | B)$ .

(iii) Sind die  $A_1, A_2, \dots$  paarweise disjunkt, so gilt

$$\mathbb{P}\left(\bigcup_n A_n | B\right) = \sum_n \mathbb{P}(A_n | B).$$

**Beweis:** Alle Aussagen folgen sofort aus den Eigenschaften von Wahrscheinlichkeitsmaßen. Als Beispiel beweisen wir (ii):

$$\begin{aligned} \mathbb{P}(\Omega \setminus A | B) &= \frac{\mathbb{P}((\Omega \setminus A) \cap B)}{\mathbb{P}(B)} \\ &= \frac{\mathbb{P}(B \setminus (A \cap B))}{\mathbb{P}(B)} \\ &= \frac{\mathbb{P}(B) - \mathbb{P}(A \cap B)}{\mathbb{P}(B)} \\ &= 1 - \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \\ &= 1 - \mathbb{P}(A | B). \end{aligned}$$

□

**Bemerkung:** Der Satz kann so interpretiert werden, dass  $A \mapsto \mathbb{P}(A | B)$  ein Wahrscheinlichkeitsmaß auf der Spur- $\sigma$ -Algebra  $\mathcal{E}_B := \{E \in \mathcal{E} \mid E \subset B\}$  definiert. So wird  $(B, \mathcal{E}_B)$  zu einem Wahrscheinlichkeitsraum.

Die Definition von bedingten Wahrscheinlichkeiten kann auch nach der Wahrscheinlichkeit des Durchschnitts aufgelöst werden:  $\mathbb{P}(A \cap B) = \mathbb{P}(A | B) \mathbb{P}(B)$ . Als Beispiel, wie man das einsetzen kann, betrachten wir einen Kasten, in dem fünf rote und zehn weiße Kugeln sind. Wir ziehen zweimal ohne Zurücklegen und fragen nach der Wahrscheinlichkeit, dass die erste Kugel rot und die zweite weiß ist. Mit  $B = \text{„erste Kugel rot“}$  und  $A = \text{„zweite Kugel weiß“}$  sind wir also an  $\mathbb{P}(A \cap B)$  interessiert.

Nun ist sicher  $\mathbb{P}(B) = 1/3$ , und  $\mathbb{P}(A | B)$  wird so ermittelt. Wenn  $B$  eingetreten ist, sind im Kasten noch 4 rote und 10 weiße Kugeln. Die Wahrscheinlichkeit für „weiß“ ist also  $10/14$ . Zusammen ergibt das:  $\mathbb{P}(A \cap B) = (1/3)(10/14) = 5/21$ .

Diese Idee kann man auch iterieren. Zum Beispiel folgt bei drei Ereignissen  $A, B, C$  aus

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

dass  $\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A | B \cap C) \mathbb{P}(B | C) \mathbb{P}(C)$ . Ähnlich leicht lässt sich herleiten, dass, für Ereignisse  $A_1, \dots, A_k$ , die Wahrscheinlichkeit von  $A_k \cap \dots \cap A_1$  mit

$$\mathbb{P}(A_k | A_{k-1} \cap \dots \cap A_1) \mathbb{P}(A_{k-1} | A_{k-2} \cap \dots \cap A_1) \dots \mathbb{P}(A_2 | A_1) \mathbb{P}(A_1)$$

übereinstimmt. Das lässt sich ausnutzen, um Wahrscheinlichkeitsräume durch *Wahrscheinlichkeitsbäume* zu definieren. So ein Baum besteht aus einer „Wurzel“ (ganz links), von dort geht es mit gewissen Wahrscheinlichkeiten zu den nächsten Verzweigungen. Und wenn man dort angekommen ist, wird gemäß der ebenfalls angegebenen Wahrscheinlichkeiten auf den nächsten „Ästen“ der Weg fortgesetzt.

Hier ist ein sehr einfaches Beispiel:

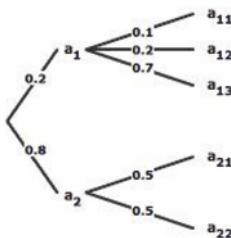


Bild 4.1.1: Ein Wahrscheinlichkeitsbaum.

Mit Wahrscheinlichkeit 0.2 bzw. 0.8 geht es nach  $a_1$  bzw.  $a_2$ . Von  $a_1$  aus gibt es drei Möglichkeiten der Fortsetzung (nach  $a_{11}, a_{12}, a_{13}$ ), die haben die Wahrscheinlichkeiten 0.1, 0.2, 0.7. Und von  $a_2$  geht es mit gleicher Wahrscheinlichkeit nach  $a_{21}$  und  $a_{22}$ .

Der zugehörige Wahrscheinlichkeitsraum besteht aus allen „Spaziergängen“ von links nach rechts, und die Wahrscheinlichkeiten entstehen durch Multiplikation der Wahrscheinlichkeiten auf den Kanten. Zum Beispiel ist die Wahrscheinlichkeit für den Weg  $a_2, a_{22}$  gleich  $0.8 \cdot 0.5 = 0.4$ . (Die Kantenwahrscheinlichkeiten stehen nämlich für bedingte Wahrscheinlichkeiten, es wurde die vor wenigen Zeilen hergeleitete Formel angewendet.)

Das Ganze kann natürlich viel komplizierter aussehen. Wir wollen hier nicht näher darauf eingehen, da dieser Zugang zu Wahrscheinlichkeitsräumen in diesem Buch keine Rolle spielen wird.

Für manche Situationen ist das wirklich der beste Weg. Wenn man zum Beispiel aus einem Kasten, in dem drei rote und fünf weiße Kugeln sind, dreimal ohne Zurücklegen zieht, so kann man auf diese Weise zum Beispiel die Frage beantworten: „Mit welcher Wahrscheinlichkeit ist die dritte gezogene Kugel weiß?“.

Dazu wird der folgende Wahrscheinlichkeitsbaum betrachtet:

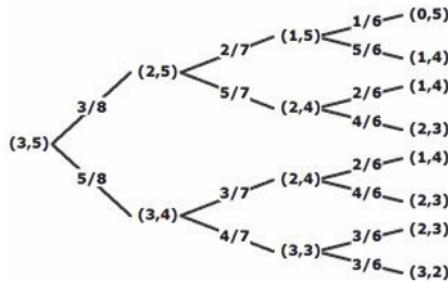


Bild 4.1.2: Drei rote und fünf weiße Kugeln: der zugehörige Wahrscheinlichkeitsbaum.

(Die Zahlen in Klammern geben an, wie viele rote und weiße Kugeln noch zur Verfügung stehen, und an den Ästen findet man die Übergangswahrscheinlichkeiten. Die Zahl  $2/7$  an der oberen rechten Kante kommt zum Beispiel so zustande: Es sind noch 7 Kugeln verfügbar, und davon sind 2 rot. Daher ist die Wahrscheinlichkeit für „rot“ im nächsten Zug, also für den Übergang zu  $(1,5)$ , gleich  $2/7$ .)

Die acht möglichen Ergebnisse, drei Kugeln zu ziehen, sind

$$rrr, rrw, rwr, rww, wrr, wrw, wwr, www.$$

Die zugehörigen Wahrscheinlichkeiten können aus dem Baum durch Multiplikation der Kan tenwahrscheinlichkeiten abgelesen werden<sup>2)</sup>:

$$\frac{6}{336}, \frac{30}{336}, \frac{30}{336}, \frac{60}{336}, \frac{30}{336}, \frac{60}{336}, \frac{60}{336}, \frac{60}{336}.$$

Und um die am Anfang gestellte Frage zu beantworten, müssen noch die Wahrscheinlichkeiten derjenigen Elementarereignisse addiert werden, bei denen die dritte gezogene Kugel weiß ist:

$$\mathbb{P}(\text{"Dritte Kugel weiß"}) = \frac{30 + 60 + 60 + 60}{336} = \frac{210}{336} \left(= \frac{5}{8}\right).$$

Wir kommen auf das Beispiel 2 von Seite 117 zurück. Da war  $\mathbb{P}(A|B) = \mathbb{P}(A)$ , das „Insiderwissen“, dass  $\omega \in B$  ist, hatte also keine Konsequenzen. Setzt man die Definition ein, so ist diese Gleichheit gleichwertig zu  $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ , und nach dieser Umformulierung muss man nicht mehr  $\mathbb{P}(B) > 0$  voraussetzen. Das führt zu der wichtigen

**Definition 4.1.3.** Zwei Ereignisse  $A, B$  in einem Wahrscheinlichkeitsraum hei ßen unabhängig, wenn  $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$  gilt.

Wir werden uns mit diesem Begriff in den nächsten Abschnitten ausführlich beschäftigen, hier sammeln wir nur einige erste Beobachtungen:

**1.** Unabhängigkeit ist eine symmetrische Eigenschaft:  $\mathbb{P}(A|B) = \mathbb{P}(A)$  ist gleichwertig zu  $\mathbb{P}(B|A) = \mathbb{P}(B)$ . Das ist etwas überraschend, wenn man an unsere Motivation denkt, denn warum sollte „Die Information  $\omega \in B$  ist für unsere Erwartung von  $A$  folgenlos“ äquivalent zu der entsprechenden Aussage sein, bei der  $A$  und  $B$  die Rollen vertauscht haben?

<sup>2)</sup>Zum Beispiel ist  $\mathbb{P}(rrr) = (3/8)(2/7)(1/6) = 6/336$ .

**2.** Die leere Menge und  $\Omega$  sind von allen  $A \in \mathcal{E}$  unabhängig. Allgemeiner gilt: Ist  $\mathbb{P}(B) \in \{0, 1\}$ , so ist  $B$  von allen  $A \in \mathcal{E}$  unabhängig. Im Fall  $\mathbb{P}(B) = 0$  ist nämlich auch  $\mathbb{P}(A \cap B) = 0$ , und aus  $\mathbb{P}(B) = 1$  folgt  $\mathbb{P}(A \cap B) = \mathbb{P}(A)$ . (Warum?)

Die Umkehrung ist auch richtig: Ist  $B$  ein Ereignis, das von allen  $A \in \mathcal{E}$  unabhängig ist, so folgt  $\mathbb{P}(B) \in \{0, 1\}$ . (Denn  $B$  ist dann insbesondere von sich selbst unabhängig. Das impliziert  $\mathbb{P}(B) = \mathbb{P}(B \cap B) = \mathbb{P}(B) \cdot \mathbb{P}(B)$ , und nun ist nur noch zu beachten, dass die Gleichung  $x^2 = x$  nur die Lösungen  $x = 0$  und  $x = 1$  hat.)

Aus den definierenden Eigenschaften von Wahrscheinlichkeitsmaßen lassen sich leicht Folgerungen für unabhängige Ereignisse ziehen.

**Satz 4.1.4.** *Es sei  $A$  ein Ereignis in einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{E}, \mathbb{P})$ . Dann gilt*

- (i)  *$A$  ist von  $\Omega$  und von der leeren Menge unabhängig.*
- (ii) *Ist  $A$  unabhängig von  $B$ , so ist  $A$  auch von  $\Omega \setminus B$  unabhängig.*
- (iii) *Sind  $B_1, B_2, \dots$  paarweise disjunkte Ereignisse und ist  $A$  von  $B_i$  für jedes  $i$  unabhängig, so ist  $A$  auch von  $\bigcup_i B_i$  unabhängig.*

Das bedeutet gerade, dass das System der von  $A$  unabhängigen Ereignisse ein Dynkinsystem ist<sup>3)</sup>.

**Beweis:** Es sind nur einige elementare Identitäten für Mengen zu beachten:

- (i)  $A \cap \Omega = A$  und  $A \cap \emptyset = \emptyset$  implizieren

$$\mathbb{P}(A \cap \Omega) = \mathbb{P}(A) = \mathbb{P}(A) \mathbb{P}(\Omega), \quad \mathbb{P}(A \cap \emptyset) = 0 = \mathbb{P}(A) \mathbb{P}(\emptyset);$$

- (ii) aus  $A \cap (\Omega \setminus B) = A \setminus (A \cap B)$  folgt

$$\begin{aligned} \mathbb{P}(A \cap (\Omega \setminus B)) &= \mathbb{P}(A \setminus (A \cap B)) \\ &= \mathbb{P}(A) - \mathbb{P}(A \cap B) \\ &= \mathbb{P}(A) - \mathbb{P}(A) \mathbb{P}(B) \\ &= \mathbb{P}(A)(1 - \mathbb{P}(B)) \\ &= \mathbb{P}(A) \mathbb{P}(\Omega \setminus B); \end{aligned}$$

- (iii) und  $A \cap (\bigcup_i B_i) = \bigcup_i A \cap B_i$  ergibt

$$\begin{aligned} \mathbb{P}\left(A \cap \left(\bigcup_i B_i\right)\right) &= \mathbb{P}\left(\bigcup_i A \cap B_i\right) \\ &= \sum_i \mathbb{P}(A \cap B_i) \end{aligned}$$

---

<sup>3)</sup>Vgl. Definition 1.6.3.

$$\begin{aligned}
 &= \sum_i \mathbb{P}(A) \mathbb{P}(B_i) \\
 &= \mathbb{P}(A) \sum_i \mathbb{P}(B_i) \\
 &= \mathbb{P}(A) \mathbb{P}\left(\bigcup_i B_i\right)
 \end{aligned}$$

(dabei wurde ausgenutzt dass mit den  $B_i$  auch die  $A \cap B_i$  paarweise disjunkt sind).  $\square$

### **Bedingte Wahrscheinlichkeiten und „gesunder Menschenverstand“.**

An verschiedenen Stellen wurde in den vorigen Kapiteln schon klar, dass die Evolution uns nicht gut auf intuitives Erfassen von Wahrscheinlichkeiten vorbereitet hat. Die in diesem Abschnitt behandelten Konzepte bilden eine *bemerkenswerte Ausnahme*.

Sicher viele Dutzend Male am Tag verändern sich unsere qualitativen Einschätzungen von Wahrscheinlichkeiten durch neue Informationen. Das geschieht unbewusst und in Sekundenbruchteilen. Hier eine winzige Auswahl:

- Sie fahren auf der Autobahn, und der Wetterbericht hatte einen trockenen Tag vorausgesagt. Plötzlich kommen Ihnen aber viele Fahrzeuge mit eingeschalteten Scheinwerfern entgegen: Erwartet Sie in wenigen Minuten ein Wolkenbruch?
- Sie haben am Wochenende einen netten Menschen kennen gelernt. Könnte sich daraus eine Freundschaft entwickeln? Die Wahrscheinlichkeit steigt sicher, wenn er/sie Ihr Lieblingshobby auch faszinierend findet, Ihre Ansichten in vielen Punkten teilt usw.
- Sie haben vergessen, einen Fahrschein zu kaufen, mit welcher Wahrscheinlichkeit wird das gut gehen? Wenn an der nächsten Station zwei Typen zusteigen, die wie Kontrolleure aussehen, sind Sie nicht mehr so optimistisch.

Und mit dem Thema „Unabhängigkeit“ verhält es sich ähnlich. Uns ist intuitiv klar, dass die meisten Informationen Wahrscheinlichkeiten *nicht* verändern. Im vorstehenden zweiten Beispiel etwa ist es für die gerade interessierende Wahrscheinlichkeit völlig unerheblich, welche Hausnummer er/sie hat, wann der Geburtstag ist, mit welchem Buchstaben der Name anfängt usw.

## 4.2 Der Satz von Bayes

In diesem Abschnitt sollen bedingte Wahrscheinlichkeiten „umgekehrt“ werden. Dazu benötigen wir den

**Satz 4.2.1.** (*Satz von der totalen Wahrscheinlichkeit*) Es sei  $(\Omega, \mathcal{E}, \mathbb{P})$  ein Wahrscheinlichkeitsraum, und  $B_1, \dots, B_n$  seien paarweise disjunkte Ereignisse mit positiver Wahrscheinlichkeit, für die  $\bigcup_{i=1}^n B_i = \Omega$  gilt. Für jedes Ereignis  $A$  gilt dann

$$\mathbb{P}(A) = \mathbb{P}(A | B_1) \mathbb{P}(B_1) + \dots + \mathbb{P}(A | B_n) \mathbb{P}(B_n).$$

**Beweis:** Man muss nur ausnutzen, dass  $A$  die disjunkte Vereinigung der  $A \cap B_i$  ist:

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}\left(\bigcup_{i=1}^n A \cap B_i\right) \\ &= \sum_{i=1}^n \mathbb{P}(A \cap B_i) \\ &= \sum_{i=1}^n \frac{\mathbb{P}(A \cap B_i) \mathbb{P}(B_i)}{\mathbb{P}(B_i)} \\ &= \sum_{i=1}^n \mathbb{P}(A | B_i) \mathbb{P}(B_i).\end{aligned}$$

□

Zur Illustration dieses Satzes betrachten wir die folgende Situation. Es wird gewürfelt, und wenn das Ergebnis eine der Zahlen 1, 2, 3, 4 ist, geht es mit Kartenstapel 1 weiter (der enthält 2 rote und 8 schwarze Karten), andernfalls mit Kartenstapel 2 (5 rote und 5 schwarze Karten). Es wird gemischt, dann wird eine Karte gezogen. Mit welcher Wahrscheinlichkeit ist eine rote Karte zu erwarten?

Bemerkenswerterweise muss man den hier relevanten Wahrscheinlichkeitsraum  $(\Omega, \mathcal{E}, \mathbb{P})$  gar nicht im Detail beschreiben. Man definiert einfach  $B_1 = \text{„Es wird eine Zahl in } \{1, 2, 3, 4\} \text{ gewürfelt“}$ ,  $B_2 = \text{„Es wird eine 5 oder eine 6 gewürfelt“}$ , und  $A$  ist das Ereignis „Es wird eine rote Karte gezogen“. Wir sind an  $\mathbb{P}(A)$  interessiert, und da die Zahlen  $\mathbb{P}(B_1) = 2/3$ ,  $\mathbb{P}(B_2) = 1/3$ ,  $\mathbb{P}(A | B_1) = 1/5$ ,  $\mathbb{P}(A | B_2) = 1/2$  leicht zu bestimmen sind, folgt

$$\mathbb{P}(A) = \mathbb{P}(A | B_1) \mathbb{P}(B_1) + \mathbb{P}(A | B_2) \mathbb{P}(B_2) = \frac{1}{5} \cdot \frac{2}{3} + \frac{1}{2} \cdot \frac{1}{3} = \frac{3}{10}.$$

Nach dieser Vorbereitung kann einer der bekanntesten Sätze der Wahrscheinlichkeitstheorie bewiesen werden. Er besagt, dass man die  $\mathbb{P}(A | B_i)$  aus den  $\mathbb{P}(B_i | A)$  berechnen kann:



Bayes

**Satz 4.2.2.** (Satz von Bayes<sup>4)</sup>) Es sei  $(\Omega, \mathcal{E}, \mathbb{P})$  ein Wahrscheinlichkeitsraum, und  $B_1, \dots, B_n$  seien paarweise disjunkte Ereignisse mit positiver Wahrscheinlichkeit, für die  $\bigcup_{i=1}^n B_i = \Omega$  gilt. Ist dann  $A$  ein Ereignis, so gilt für jedes  $i_0 \in \{1, \dots, n\}$  die Bayesformel:

$$\mathbb{P}(B_{i_0} | A) = \frac{\mathbb{P}(A | B_{i_0}) \mathbb{P}(B_{i_0})}{\sum_{i=1}^n \mathbb{P}(A | B_i) \mathbb{P}(B_i)}.$$

Ist  $B$  ein Ereignis, so dass  $B$  und  $\Omega \setminus B$  positive Wahrscheinlichkeit haben, so kann die vorstehende Formel mit  $n = 2$ ,  $B_1 = B$  und  $B_2 = \Omega \setminus B$  angewendet werden. Es ergibt sich

$$\mathbb{P}(B | A) = \frac{\mathbb{P}(A | B) \mathbb{P}(B)}{\mathbb{P}(A | B) \mathbb{P}(B) + \mathbb{P}(A | \Omega \setminus B) (1 - \mathbb{P}(B))}.$$

Manchmal werden Ereignisse als Aussagen interpretiert. Dann ist es suggestiv,  $\neg B$  („nicht  $B$ “) statt  $\Omega \setminus B$  zu schreiben.

**Beweis:** Der Beweis folgt sofort aus der Definition bedingter Wahrscheinlichkeiten und dem vorstehenden Satz:

$$\begin{aligned} \mathbb{P}(B_{i_0} | A) &= \frac{\mathbb{P}(A \cap B_{i_0})}{\mathbb{P}(A)} \\ &= \frac{(\mathbb{P}(A \cap B_{i_0}) / \mathbb{P}(B_{i_0})) \mathbb{P}(B_{i_0})}{\mathbb{P}(A)} \\ &= \frac{\mathbb{P}(A | B_{i_0}) \mathbb{P}(B_{i_0})}{\sum_{i=1}^n \mathbb{P}(A | B_i) \mathbb{P}(B_i)}. \end{aligned}$$

□

Es ist nicht schwer, Beispiele zur Illustration dieses Satzes zu finden.

**1.** Wir nehmen noch einmal das Beispiel auf, das wir vor wenigen Zeilen behandelt haben. Wir stellen nun die Frage: Wir ziehen eine rote Karte, mit welcher Wahrscheinlichkeit hatten wir es mit Kartenstapel 1 zu tun? Wir wollen also wissen, wie groß  $\mathbb{P}(B_1 | A)$  ist. Eine qualitative Vermutung ist schwierig, denn in Stapel 2 sind zwar mehr rote Karten, allerdings wurde der auch nur mit der geringeren Wahrscheinlichkeit gewählt. Der Satz von Bayes hilft weiter, danach ist  $\mathbb{P}(B_1 | A)$  gleich

$$\frac{\mathbb{P}(A | B_1) \mathbb{P}(B_1)}{\mathbb{P}(A | B_1) \mathbb{P}(B_1) + \mathbb{P}(A | B_2) \mathbb{P}(B_2)} = \frac{(1/5) \cdot (2/3)}{(1/5) \cdot (2/3) + (1/2) \cdot (1/3)} = \frac{4}{9}.$$

Ganz entsprechend folgt  $\mathbb{P}(B_2 | A) = 5/9$ .

---

<sup>4)</sup>Thomas Bayes, 1702 bis 1761. Er war Pfarrer in England, nebenbei beschäftigte er sich mit mathematischen Problemen. Sein Hauptwerk zur Wahrscheinlichkeitsrechnung, in dem natürlich auch das Bayes-Theorem zu finden ist, erschien 1764.

**2.** Hier ist ein berühmtes Anwendungsbeispiel aus der *Medizin*. Es geht um Krankheiten und Tests, zunächst lernen wir einige Vokabeln kennen. Es wird um irgendeine spezielle Krankheit  $K$  gehen, die kann ganz beliebig sein. (Es muss ja nicht gleich Krebs oder Aids sein ...). Mit einem Test soll festgestellt werden, ob eine Person diese Krankheit hat.

- Unter *Prävalenz* versteht man den Anteil der Personen in der Bevölkerung, die diese Krankheit haben. In der Regel ist das eine sehr kleine Zahl. Bedeutet das Ereignis  $B$ , dass eine Person  $K$  hat, so ist die Prävalenz also die Zahl  $\mathbb{P}(B)$ .
- Die *Sensitivität* ist eine Güteeigenschaft des Tests, durch den geprüft werden soll, ob  $K$  vorliegt. Es sei  $A$  das Ereignis, dass er positiv ausfällt, d.h. die Person als krank klassifiziert. Die Sensitivität ist dann als  $\mathbb{P}(A | B)$  definiert: Mit welcher Wahrscheinlichkeit wird eine kranke Person als solche erkannt? Es ist anzustreben, dass diese Zahl sehr nahe bei Eins liegt.
- Es fehlt noch die *Spezifität*, sie ist als  $\mathbb{P}(\neg A | \neg B)$  erklärt: Das ist die Wahrscheinlichkeit, dass eine gesunde Person auch das Test-Ergebnis „gesund“ erzielt. Auch diese Zahl sollte nahe bei Eins liegen.

Für die Bayes-Formel ist allerdings  $\mathbb{P}(A | \neg B)$  interessant, also die Wahrscheinlichkeit für einen „Fehlalarm“ (= Person gesund, Test positiv). Sie ist leicht zu ermitteln, sie ist gleich  $1 - \mathbb{P}(\neg A | \neg B)$ .

Hier ein Beispiel: Prävalenz 0.003, Sensitivität 0.98, Spezifität 0.99. Wie groß ist dann  $\mathbb{P}(B | A)$ : Muss man sich bei einem positiven Testergebnis Sorgen machen? Mit der Bayes-Formel erhalten wir für diese bedingte Wahrscheinlichkeit den Wert

$$\frac{\mathbb{P}(A | B) \mathbb{P}(B)}{\mathbb{P}(A | B) \mathbb{P}(B) + \mathbb{P}(A | \neg B) (1 - \mathbb{P}(B))} = \frac{0.98 \cdot 0.003}{0.98 \cdot 0.003 + 0.01 \cdot 0.997} \approx 0.09.$$

Trotz hoher Prävalenz und Sensitivität ist die Wahrscheinlichkeit, dass eine positiv getestete Person wirklich krank ist, also noch nicht einmal 10 Prozent. Das ist intuitiv nicht gut nachzuvollziehen, der mathematische Grund liegt in der niedrigen Prävalenz.

Es ist zu hoffen, dass Mediziner diese Zusammenhänge kennen, wenn sie ihren Patienten ein positives Testergebnis mitteilen müssen.

Stellt man sich alles geometrisch vor, ist das Ergebnis nicht mehr ganz so überraschend.  $\Omega$ , die Population, wird als Rechteck dargestellt, und  $B$  (die kranken Personen) ist dann eine „sehr kleine“ Teilmenge (der kleine dunkle Kreis im folgenden Bild).  $A$  („Test positiv“) ist ebenfalls klein, dieses Ereignis ist durch den hellen Kreis dargestellt.  $A \cap B$  ist eine große Teilmenge von  $B$ , denn der Test erkennt  $B$  recht zuverlässig.  $A$  ragt nur wenig in  $\Omega \setminus B$  hinein, denn wenige gesunde Personen werden fälschlich als positiv diagnostiziert. Trotz dieser Bedingungen kann der Anteil von  $A \cap B$  in  $A$ , also  $\mathbb{P}(B | A)$ , sehr klein sein.



Bild 4.2.1: Eine Erklärung für die unerwartet kleine Wahrscheinlichkeit  $\mathbb{P}(B | A)$ .

**3.** Ab 1990 fand ein Problem aus der Wahrscheinlichkeitsrechnung ein breites öffentliches Interesse: das *Ziegenproblem*. Es geht um folgende Situation:

Am Ende einer Spielshow hat der siegreiche Kandidat die Chance, ein Auto zu gewinnen. Er steht vor drei Türen, hinter einer ist das Auto, hinter den beiden anderen sind Ziegen. Das Auto gehört ihm nur dann, wenn er die richtige Tür wählt.

Er entscheidet sich für Tür 1. Der Quizmaster öffnet Tür 3, dahinter meckert eine der Ziegen. Dann bietet er dem Kandidaten an, seine Entscheidung noch einmal zu überdenken, also zu Tür 2 zu wechseln.

Das Problem: *Sollte man dem Kandidaten raten, auf dieses Angebot einzugehen?*

Falls Sie die Lösung noch nicht kennen, können Sie ja vor dem Weiterlesen über das Problem nachdenken<sup>5)</sup>.

*Bevor* der Quizmaster etwas sagt, ist die Gewinnwahrscheinlichkeit 1/3. Um zu analysieren, wie es nach dem Öffnen von Tür 3 und dem Wechsel-Angebot aussieht, führen wir einige Bezeichnungen ein.  $G_i$  soll für  $i = 1, 2, 3$  das Ereignis sein, dass der Gewinn hinter Tür  $i$  steht, und mit  $Q_j$  bezeichnen wir das Ereignis, dass der Quizmaster Tür  $j$  geöffnet hat.

Wir sind an  $\mathbb{P}(G_2 | Q_3)$  interessiert. Die bedingten Wahrscheinlichkeiten, die wir dafür in die Bayes-Formel einsetzen müssen, sehen so aus:

- $\mathbb{P}(Q_3 | G_2) = 1$ . Denn wenn der Gewinn hinter Tür 2 ist, *muss* Tür 3 geöffnet werden. (Tür 1 wurde ja vom Kandidaten gewählt.)
- $\mathbb{P}(Q_3 | G_1) = \frac{1}{2}$ . Hier hat der Quizmaster zwei Möglichkeiten. Wir nehmen an, dass er eine von beiden mit gleicher Wahrscheinlichkeit wählt.
- $\mathbb{P}(Q_3 | G_3) = 0$ . Er wird ja nicht verraten, wo das Auto steht, die Wahl von Tür 3 scheidet aus.

---

<sup>5)</sup>Auch wenn Sie total falsch liegen, werden Sie in guter Gesellschaft sein, denn damals haben sich auch viele Mathematiker von einer falschen Intuition leiten lassen.

Wir können auch annehmen, dass  $\mathbb{P}(G_1) = \mathbb{P}(G_2) = \mathbb{P}(G_3) = 1/3$  gilt. Damit ergibt sich

$$\begin{aligned}\mathbb{P}(G_2 | Q_3) &= \frac{\mathbb{P}(Q_3 | G_2) \mathbb{P}(G_2)}{\mathbb{P}(Q_3 | G_1) \mathbb{P}(G_1) + \mathbb{P}(Q_3 | G_2) \mathbb{P}(G_2) + \mathbb{P}(Q_3 | G_3) \mathbb{P}(G_3)} \\ &= \frac{1 \cdot (1/3)}{(1/2) \cdot (1/3) + 1 \cdot (1/3) + 0 \cdot (1/3)} \\ &= \frac{2}{3}.\end{aligned}$$

Ganz analog folgt  $\mathbb{P}(G_1 | Q_3) = 1/3$ , und das zeigt: Wechseln ist dringend zu empfehlen, denn die Gewinnwahrscheinlichkeit verdoppelt sich.

Die Analyse könnte noch vertieft werden: Was ist, wenn der Gewinn mit unterschiedlichen Wahrscheinlichkeiten hinter den drei Türen platziert wird? Wie sieht es aus, wenn der Quizmaster in der Situation  $G_1$  nicht mit gleicher Wahrscheinlichkeit Tür 2 oder Tür 3 öffnet, wie wir es angenommen haben, als wir  $\mathbb{P}(Q_3 | G_1) = 1/2$  postuliert haben? ... In allen Fällen zeigt sich, dass Wechseln vorzuziehen ist, mindestens werden die Chancen nicht schlechter. Das soll aber hier nicht im Detail ausgeführt werden, mehr als die Kenntnis der Bayes-Formel ist für eine ausführlichere Diskussion nicht erforderlich.

Als Variante könnte man den Kandidaten vor *vier* Türen stellen. Wieder gibt es nur einen Gewinn, und der Quizmaster öffnet nach der Wahl des Kandidaten *zwei* dieser Türen. Er darf wechseln, ist das empfehlenswert? Das ist als Zweipersonen-Spiel im Wissenschaftsmuseum „Exploratorium“ in San Francisco aufgebaut:



Bild 4.2.2: Das Ziegenproblem im „Exploratorium“ in San Francisco.

### Bayes-Formel und Intuition

Im vorigen Abschnitt wurde darauf hingewiesen, dass wir eine bemerkenswert gute Intuition für bedingte Wahrscheinlichkeiten haben. Für die Einschätzung von  $\mathbb{P}(B | A)$  bei bekanntem  $\mathbb{P}(A | B)$  sind wir offensichtlich nicht gut vorbereitet, wie die vorstehenden Beispiele – besonders das aus der Medizin – zeigen.

Der Grund ist wahrscheinlich, dass wir bei der intuitiven Bestimmung von  $\mathbb{P}(B | A)$  den Einfluss von  $\mathbb{P}(A | B)$  überschätzen und  $\mathbb{P}(B)$  zu wenig berücksichtigen.

Am Ende dieses Abschnitts soll noch eine weitere Anwendung des Satzes von der totalen Wahrscheinlichkeit vorgestellt werden: „Erfolg macht sicher!“.

Wir stellen uns  $n$  Kästen vor, wobei  $n$  „sehr groß“ ist. Alle enthalten rote und weiße Kugeln, und zwar sind im  $i$ -ten Kasten  $i$  rote und  $n - i$  weiße. Ein  $i$  wird gleichverteilt ausgewählt, und wir ziehen mehrfach aus dem entsprechenden Kasten. (Nach dem Ziehen wird die gezogene Kugel wieder zurückgelegt.) Das  $i$  ist uns nicht bekannt, aber es ist intuitiv klar, dass bei größeren  $i$  mehr rote Kugeln zu erwarten sind.

Und nun die Frage: Mal angenommen, wir haben  $k$ -mal gezogen, und es war immer eine rote Kugel. Wie wahrscheinlich ist es, dass auch die nächste rot sein wird? Dazu definieren wir  $B_i$  als das Ereignis, dass der  $i$ -te Kasten ausgesucht wurde. Die Wahrscheinlichkeit, dass dann  $k$ -mal „rot“ erscheint, ist  $(i/n)^k$ , denn es handelt sich um ein  $k$ -fach wiederholtes Bernoulliexperiment mit Erfolgswahrscheinlichkeit  $i/n$ . (Wir haben bei den Wiederholungen Unabhängigkeit angenommen, so dass sich die Wahrscheinlichkeiten multiplizieren.) Aufgrund von Satz 4.2.1 ist damit die Wahrscheinlichkeit für „Es werden  $k$ -mal rote Kugeln gezogen“ gleich  $\sum_{i=1}^n (i/n)^k (1/n)$ . Diese Summe entsteht auch, wenn man  $[0, 1]$  in  $n$  gleiche Teile teilt und das Integral über  $x^k$  durch Riemannsummen approximiert. Deswegen ist sie gut durch  $\int_0^1 x^k dx = 1/(k+1)$  zu approximieren. Und unsere Frage kann nun auch beantwortet werden:

$$\mathbb{P}(k+1 \text{ Erfolge} | k \text{ Erfolge}) = \frac{\mathbb{P}(k+1 \text{ Erfolge})}{\mathbb{P}(k \text{ Erfolge})} \approx \frac{1/(k+2)}{1/(k+1)} = \frac{k+1}{k+2}.$$

Das entspricht der Intuition: Wenn man „oft“ hintereinander eine rote Kugel gezogen hat, so hat man sicher einen Kasten mit „vielen“ roten Kugeln vor sich, und deswegen sollte auch der nächste Versuch mit hoher Wahrscheinlichkeit zum gleichen Ergebnis führen.

Auch im Alltagsleben schätzen wir Wahrscheinlichkeiten manchmal auf diese Weise ein: Wenn man „sehr oft“ in seinem Lieblingsrestaurant gut gegessen hat (oder erfolgreich den Computer gestartet hat, oder schwarz gefahren ist, ohne erwischt zu werden, oder ...), wird es wohl auch beim nächsten Mal so sein.

Das ist mathematisch allerdings nicht zu rechtfertigen, denn unsere Herleitung der Formel setzte voraus, dass die Wahrscheinlichkeit für „Erfolg (= rote

Kugel ziehen)“ zufällig und gleichverteilt aus  $\{1/n, \dots, n/n\}$  gewählt wurde, und so etwas kann bei den Beispielen des vorigen Absatzes sicher nicht angenommen werden.

Trotzdem waren derartige Schlüsse in der Frühzeit der Wahrscheinlichkeitsrechnung durchaus salonfähig, so etwa bei der Aussage, dass morgen früh mit überwältigender Wahrscheinlichkeit die Sonne aufgehen wird. Heute ist man mit derartigen stochastischen Prognosen etwas vorsichtiger.

### 4.3 Unabhängigkeit für mehr als zwei Ereignisse

Zur Erinnerung: Zwei Ereignisse in einem Wahrscheinlichkeitsraum heißen *unabhängig*, wenn die Information dass das eine eingetreten ist, die Wahrscheinlichkeit für das andere nicht verändert (vgl. Definition 4.1.3). Nun nehmen wir an, dass es um drei Ereignisse  $A, B, C$  geht, dass wir auf das Ergebnis „ $\omega \in A?$ “ der Zufallsausgabe warten und über Informationen verfügen, ob  $\omega \in B$  bzw.  $\omega \in C$  eingetreten ist. Wenn  $B$  und  $C$  jeweils für sich die Wahrscheinlichkeit von  $A$  nicht verändern, wenn also sowohl  $A, B$  als auch  $A, C$  unabhängig sind, kann es dann beim Kombinieren der  $B-C$ -Information anders aussehen? Das ist tatsächlich der Fall:

*Die Kombination von Informationen führt manchmal zu neuen Informationen.*

Das ist für Krimi-Leser natürlich keine Überraschung, Beispiele lassen sich aber auch leicht im Alltagsleben finden. Hier ein Beispiel:

#### Der Junge auf dem Reiterhof

Zwei Mädchen, 13 und 14 Jahre alt, sind auf dem Reiterhof angekommen. Dort gibt es auch einen Jungen, dessen Alter  $a$  schwer zu schätzen ist. Wie könnte man herausbekommen, ob er schon über 14 ist, wenn man – das wäre ja peinlich – eine direkte Frage vermeiden möchte?

Die beiden denken sich die folgende Strategie aus. Irgendwann stellen sie ihm die wirklich harmlos wirkende Frage „Wie lange reitest Du eigentlich schon?“ (Antwort: 4 Jahre), und ein paar Tage später, ebenfalls unverdächtig, wollen sie wissen „Wie alt warst Du denn, als Du angefangen hast zu reiten?“ (Antwort: 12).

Bezeichnet man mit  $A$  die Information „ $\{a > 14\}?$ “ (um die geht es ja hier), so ist es dafür völlig belanglos zu wissen, dass er seit 4 Jahren reitet. Für sich genommen würde es auch nichts nutzen zu erfahren, dass er mit 12 angefangen hat. Nimmt man aber beides zusammen, ist alles klar.

Solche Phänomene gibt es in der Wahrscheinlichkeitsrechnung ebenfalls. Das wohl einfachste Beispiel besteht aus dem Werfen von zwei Münzen. Die Ereignisse  $A, B$  und  $C$  seien durch „Das Ergebnis ist (Kopf, Zahl) oder umgekehrt“, „Die

erste Münze zeigt 'Kopf'" und „Die zweite Münze zeigt 'Zahl'" definiert. Dann ist  $A$  unabhängig von  $B$  und von  $C$ , aber nicht von  $B \cap C$ , denn  $\mathbb{P}(A | B \cap C) = 1 \neq 1/2 = \mathbb{P}(A)$ .

Diese Beobachtung soll nun in mehreren Schritten präzisiert werden. Wer nur am Endergebnis interessiert ist, kann gleich bis zur Definition 4.3.4 vorblättern.

$A$  unabhängig von  $B_1, \dots, B_n$

Wir wollen präzisieren, was es bedeutet, dass Informationen über den Ausgang der Abfrage „ $\omega \in B_i$ ?“ ( $i = 1, \dots, n$ ) keine Auswirkungen auf unsere Erwartung für „ $\omega \in A$ ?“, also auf  $\mathbb{P}(A)$ , haben. Bekanntlich kann man Informationen kombinieren. Wenn man über die „ $\omega \in B_i$ ?“ Bescheid weiß, weiß man auch – zum Beispiel – ob  $\omega \in (B_1 \cap B_7) \cup (B_{14} \cap B_{122})$ . Allgemein: Die Information „ $\omega \in B_i$ ?“ für  $i = 1, \dots, n$  impliziert die Information „ $\omega \in C$ ?“ für alle  $C$  in  $\sigma(B_1, \dots, B_n)$ , der von  $B_1, \dots, B_n$  erzeugten  $\sigma$ -Algebra<sup>6)</sup>. Das motiviert die folgende

**Definition 4.3.1.**  *$A, B_1, \dots, B_n$  seien Ereignisse in einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{E}, \mathbb{P})$ . Wir sagen, dass  $A$  von  $B_1, \dots, B_n$  unabhängig ist, wenn gilt:  $A$  ist unabhängig von allen  $C \in \sigma(B_1, \dots, B_n)$ .*

Nun kann  $\sigma(B_1, \dots, B_n)$  sehr viele Mengen enthalten. Deswegen ist es gut zu wissen, dass man sich nur um Durchschnitte der  $B_i$  kümmern muss. Der Beweis wird unter Verwendung der in Abschnitt 1.6 hergeleiteten Techniken recht einfach sein<sup>7)</sup>.

**Satz 4.3.2.** *Mit den Bezeichnungen der vorigen Definition sind die folgenden Aussagen äquivalent:*

- (i)  $A$  ist unabhängig von  $B_1, \dots, B_n$ .
- (ii) Für beliebige  $k \leq n$  und beliebige  $i_1, \dots, i_k$  mit  $1 \leq i_1 < \dots < i_k \leq n$  sind die Ereignisse  $A$  und  $B_{i_1} \cap \dots \cap B_{i_k}$  unabhängig.

**Beweis:** Die Implikation „(i) $\Rightarrow$ (ii)“ ist klar, denn alle Durchschnitte liegen in der erzeugten  $\sigma$ -Algebra. Für die andere Beweisrichtung setzen wir (ii) voraus und betrachten das Mengensystem

$$\mathcal{D}_A := \{B \in \mathcal{E} \mid A, B \text{ sind unabhängig}\}.$$

Das ist nach Satz 4.1.4 ein Dynkinsystem. Es enthält nach Voraussetzung alle Mengen des Typs  $B_{i_1} \cap \dots \cap B_{i_k}$ . Und da das ein durchschnittsstabiles Mengensystem ist, stimmt das von den  $B_1, \dots, B_n$  erzeugte Dynkinsystem  $\mathcal{D}(B_1, \dots, B_n)$  mit  $\sigma(B_1, \dots, B_n)$  überein (Satz 1.6.4). Damit gilt  $\sigma(B_1, \dots, B_n) \subset \mathcal{D}_A$ , und das ist gerade die Aussage (i).  $\square$

---

<sup>6)</sup>Vgl. Satz 1.4.1.

<sup>7)</sup>Die werden ab jetzt immer wieder vorkommen. Wer Abschnitt 1.6 beim ersten Lesen ausgelassen hat, sollte sich jetzt genauer damit auseinandersetzen.

$A_1, \dots, A_n$  unabhängig

Nun seien  $A_1, \dots, A_n$  Ereignisse, und es soll präzisiert werden, was es bedeutet, dass Informationen über gewisse  $A_i$  keine Konsequenz für die Wahrscheinlichkeiten der restlichen  $A$ 's haben. Wir wollen dazu fordern:

- $A_1$  ist unabhängig von  $A_2, A_3, \dots, A_n$ , und
- $A_2$  ist unabhängig von  $A_1, A_3, \dots, A_n$ , und
- $\dots$ , und
- $A_n$  ist unabhängig von  $A_1, A_2, \dots, A_{n-1}$ .

**Lemma 4.3.3.** *Die vorstehenden Bedingungen sind äquivalent dazu, dass*

$$\mathbb{P}(A_{i_1} \cap \dots \cap A_{i_k}) = \mathbb{P}(A_{i_1}) \cdots \mathbb{P}(A_{i_k})$$

für alle  $k = 2, \dots, n$  und alle  $1 \leq i_1 < \dots < i_k \leq n$  gilt.

**Beweis:** Sei zunächst die Bedingung des Lemmas erfüllt. Wir zeigen, dass dann  $A_1$  von  $A_2, \dots, A_n$  unabhängig ist. (Die anderen Aussagen beweist man analog.) Wir wollen den vorigen Satz anwenden und geben dazu Indizes  $2 \leq i_1 < \dots < i_r \leq n$  vor: Es ist  $\mathbb{P}(A_1 \cap (A_{i_1} \cap \dots \cap A_{i_r})) = \mathbb{P}(A_1) \mathbb{P}(A_{i_1} \cap \dots \cap A_{i_r})$  zu zeigen.

Die linke Seite ist aufgrund der Voraussetzung gleich  $\mathbb{P}(A_1) \mathbb{P}(A_{i_1}) \cdots \mathbb{P}(A_{i_r})$ . Wenn man nur die  $i_1, \dots, i_r$  betrachtet, liefert die Voraussetzung auch die Gleichung  $\mathbb{P}(A_{i_1} \cap \dots \cap A_{i_r}) = \mathbb{P}(A_{i_1}) \cdots \mathbb{P}(A_{i_r})$ , und damit ist alles gezeigt.

Nun seien die vor dem Lemma zusammengestellten Bedingungen erfüllt, und  $1 \leq i_1 < \dots < i_k \leq n$  seien vorgegeben: Wir sollen  $\mathbb{P}(A_{i_1} \cap \dots \cap A_{i_k}) = \mathbb{P}(A_{i_1}) \cdots \mathbb{P}(A_{i_k})$  beweisen.

Nun ist  $A_{i_1}$  nach Voraussetzung unabhängig von den  $A_i$  mit  $i \neq i_1$ , aus dem vorigen Satz folgt deswegen

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

Und so geht es weiter:  $\mathbb{P}(A_{i_2} \cap \dots \cap A_{i_k})$  wird zu  $\mathbb{P}(A_{i_2}) \mathbb{P}(A_{i_3} \cap \dots \cap A_{i_k})$  umgeformt, bis nach endlich vielen Schritten  $\mathbb{P}(A_{i_1} \cap \dots \cap A_{i_k})$  als  $\mathbb{P}(A_{i_1}) \cdots \mathbb{P}(A_{i_k})$  geschrieben werden kann.  $\square$

Das Ergebnis des Lemmas nutzen wir für die folgende

**Definition 4.3.4.** *Es sei  $(\Omega, \mathcal{E}, \mathbb{P})$  ein Wahrscheinlichkeitsraum.*

(i) *Ereignisse  $A_1, \dots, A_n$  heißen unabhängig, wenn*

$$\mathbb{P}(A_{i_1} \cap \dots \cap A_{i_k}) = \mathbb{P}(A_{i_1}) \cdots \mathbb{P}(A_{i_k})$$

für alle  $k = 2, \dots, n$  und alle  $1 \leq i_1 < \dots < i_k \leq n$  gilt.

(ii) *Ist  $\mathcal{E}_0 \subset \mathcal{E}$  eine Teilmenge der Ereignisse, so heißt  $\mathcal{E}_0$  unabhängig, wenn jede endliche Teilstamme unabhängig ist.*

Bevor wir uns um etwas anspruchsvollere Ergebnisse im Zusammenhang mit dieser neuen Definition kümmern, gibt es einige *Bemerkungen und Beispiele*:

**1.** Es sei  $n \in \mathbb{N}$ , wir versehen  $\Omega = \{0, 1\}^n$  mit der Gleichverteilung. Definiert man dann  $A_i := \{(x_1, \dots, x_n) \mid x_i = 1\}$  für  $i = 1, \dots, n$ , so sind die  $A_1, \dots, A_n$  unabhängig: Für  $1 \leq i_1 < \dots < i_k \leq n$  besteht doch  $A_{i_1} \cap \dots \cap A_{i_k}$  aus denjenigen  $n$ -Tupeln, bei denen an den Stellen  $i_1, \dots, i_k$  eine Eins steht. Das trifft für genau  $2^{n-k}$  Elemente zu, und deswegen ist die Wahrscheinlichkeit des Durchschnitts gleich  $2^{n-k}/2^n = 2^{-k}$ . Da jedes  $A_i$  Wahrscheinlichkeit  $1/2$  hat, stimmt diese Zahl mit  $\mathbb{P}(A_{i_1}) \cdots \mathbb{P}(A_{i_k})$  überein.

**2.** Jede Teilmenge einer Menge unabhängiger Ereignisse ist offensichtlich ebenfalls unabhängig. Insbesondere sind in einer unabhängigen Menge  $\mathcal{E}_0$  von Ereignissen je zwei Ereignisse unabhängig. Man sagt dann auch, dass die Ereignisse in  $\mathcal{E}_0$  *paarweise unabhängig* sind.

**3.** Um zu ermitteln, ob  $A_1, \dots, A_n$  unabhängig sind, muss man sehr viele Gleichungen nachprüfen: Gilt für  $i_1 < i_2$  stets  $\mathbb{P}(A_{i_1} \cap A_{i_2}) = \mathbb{P}(A_{i_1}) \mathbb{P}(A_{i_2})$ ? Gilt eine entsprechende Gleichung bei der Auswahl von  $3, 4, \dots, n$  Ereignissen? Da es  $\binom{n}{k}$   $k$ -elementige Teilmengen von  $\{1, \dots, n\}$  gibt, ergibt sich die folgende Anzahl:

$$\binom{n}{2} + \binom{n}{3} + \dots + \binom{n}{n} = (1+1)^n - n - 1 = 2^n - n - 1.$$

Sie wächst also mit  $n$  exponentiell.

**4.** Es lässt sich auch nichts einsparen: Selbst wenn alle echten Teilmengen von  $\{A_1, \dots, A_n\}$  unabhängig sind, kann man noch nicht Unabhängigkeit für  $\{A_1, \dots, A_n\}$  garantieren. Als Beispiel betrachten wir wieder die  $A_1, \dots, A_n$  aus dem vorstehenden Beispiel 1 und zusätzlich das Ereignis  $A_0$ , das alle diejenigen  $(x_1, \dots, x_n)$  enthält, bei denen die Anzahl der Einsen unter den  $x_i$  eine gerade Zahl ist. Dann gilt

- Je  $n$  Ereignisse unter den  $A_0, \dots, A_n$  sind unabhängig.

Dazu fehlt noch die Behandlung von Teilmengen des Typs  $A_0, A_{i_1}, A_{i_k}$ , wobei  $k \leq n-1$ . Sei zunächst  $k$  gerade. Um dann in  $A_{i_1} \cap \dots \cap A_{i_k}$  ein Element von  $A_0$  zu finden, muss man eine gerade Anzahl von Einsen für die Indizes  $i$  aussuchen, die nicht in  $\{i_1, \dots, i_k\}$  liegen. Nun hat aus Symmetriegründen exakt die Hälfte der Teilmengen einer  $m$ -elementigen Menge eine gerade Elementanzahl, und folglich gibt es in der Menge  $A_0 \cap A_{i_1} \cap \dots \cap A_{i_k}$  genau  $2^{n-k-1}$  Elemente. Die Wahrscheinlichkeit dieses Durchschnitts ist folglich  $2^{n-k-1}/2^n = 1/2^{k+1}$ . Diese Zahl stimmt mit  $\mathbb{P}(A_0) \mathbb{P}(A_{i_1}) \cdots \mathbb{P}(A_{i_k})$  überein, denn alle Ereignisse haben Wahrscheinlichkeit  $1/2$ .

Für ungerade  $k$  argumentiert man ganz ähnlich.

- Die Familie  $A_0, A_1, \dots, A_n$  ist *nicht* unabhängig.

Das Produkt der Wahrscheinlichkeiten ist  $1/2^{n+1}$ , der Durchschnitt hat aber Wahrscheinlichkeit 0 für ungerades und  $1/2^n$  für gerades  $n$ .

**5.** Als weiteres Beispiel betrachten wir das mit der Gleichverteilung versehene Intervall  $[0, 1]$ . Prüfen Sie zur Übung nach, dass

$$A_1 := [0, 0.5], \quad A_2 := [0, 0.25] \cup [0.5, 0.75] \quad \text{und}$$

$$A_3 := [0, 0.125] \cup [0.25, 0.375] \cup [0.5, 0.625] \cup [0.75, 0.875]$$

unabhängig sind.

**6.** Manchmal kann man das Verfahren auch umkehren: Da gibt man Wahrscheinlichkeiten  $p_i \in ]0, 1[$  für die  $A_i$  vor und möchte ein Wahrscheinlichkeitsmaß so finden, dass die Ereignisse unabhängig sind. Das wird sicher nicht immer gehen, zum Beispiel dann nicht, wenn der Schnitt über alle  $A_i$  leer ist. Man hat aber nur eine Wahl,  $\mathbb{P}(A_{i_1} \cap \dots \cap A_{i_k})$  muss als  $p_{i_1} \cdots p_{i_k}$  definiert werden. Sind zum Beispiel die  $A_1, \dots, A_n$  wie im vorstehenden Beispiel 1, so führt dieses Verfahren zu dem folgenden Wahrscheinlichkeitsmaß: Für die Berechnung der Wahrscheinlichkeit der einpunktigen Menge  $\{(x_1, \dots, x_n)\}$  bilde ein Produkt. Dabei ist der  $i$ -te Faktor  $p_i$ , wenn  $x_i = 1$  gilt, und  $1 - p_i$  sonst.

Im Zusammenhang mit dem Thema „Unabhängigkeit“ gibt es zwei Fallen, die unbedingt zu vermeiden sind:

### Zwei verbreitete Irrtümer:

Unabhängigkeit für mehr als zwei Ereignisse wird eine wichtige Rolle spielen. Die Definition ist etwas technisch, es ist verführerisch, sie etwas zu „vereinfachen“.

*Irrtum 1:* „ $A_1, \dots, A_n$  unabhängig“ ist gleichwertig dazu, dass je zwei  $A_i, A_j$  für  $i \neq j$  unabhängig sind.

Das ist **falsch!** Das ist nur die paarweise Unabhängigkeit. Die folgt zwar aus der Unabhängigkeit, impliziert sie aber nicht. (Wie das vorstehende Beispiel 4 zeigt, reicht nicht einmal die Unabhängigkeit aller  $(n - 1)$ -elementigen Teilstufen).

*Irrtum 2:* „ $A_1, \dots, A_n$  unabhängig“ heißt, dass  $\mathbb{P}(A_1 \cap \dots \cap A_n)$  mit  $\mathbb{P}(A_1) \cdots \mathbb{P}(A_n)$  übereinstimmt.

Auch das ist **falsch!** Andernfalls wäre jede Familie von Ereignissen unabhängig, bei denen ein  $A_i$  Wahrscheinlichkeit Null hat.

Vielelleicht ist es zur Vermeidung des ersten Irrtums hilfreich, auf die formalen *Parallelen zur linearen Unabhängigkeit* in der linearen Algebra hinzuweisen. Auch da ist ja definiert, dass eine Familie von Vektoren linear unabhängig ist, wenn jede endliche Familie diese Eigenschaft hat. Auch bleibt lineare Unabhängigkeit beim Übergang zu Teilmengen erhalten, und niemand käme auf die Idee, aus „je zwei Vektoren in  $\{x_1, \dots, x_n\}$  sind linear unabhängig“ zu schließen, dass auch die Gesamtheit linear unabhängig ist.

„Unabhängigkeit“ in der Wahrscheinlichkeitsrechnung ist eine bemerkenswert effektive mathematische Präzision der etwas vagen Intuition, dass Ereignisse „nichts miteinander zu tun haben, wenn es um die Einschätzung von Wahrscheinlichkeiten geht“. Um sich davon zu überzeugen, dass dieser Ansatz richtig ist, ist immer wieder nachzuprüfen, ob unsere Erwartungen an diesen Begriff auch wirklich beweisbare Ergebnisse sind.

Hier ein Beispiel. Wenn  $A, B, C$  unabhängig sind, so sollten auch  $A$  und  $B \cup C$  unabhängig sein, denn das Kombinieren von irrelevanten Informationen kann ja keine neuen Erkenntnisse liefern. Das ist tatsächlich der Fall. Wir zeigen dazu ein sehr allgemeines Ergebnis, der Beweis wird unter Verwendung der Techniken aus Abschnitt 1.6 bemerkenswert elegant sein:

**Satz 4.3.5.**  $A_1, \dots, A_n$  seien unabhängige Ereignisse in einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{E}, \mathbb{P})$ , und  $\Delta_1, \dots, \Delta_r$  seien nicht leere und paarweise disjunkte Teilmengen von  $\{1, \dots, n\}$ . Ist dann für  $j = 1, \dots, r$  jeweils  $C_j$  ein Element in der von den  $A_i$  mit  $i \in \Delta_j$  erzeugten  $\sigma$ -Algebra, so sind  $C_1, \dots, C_r$  unabhängig.

**Beweis:** Die Formulierung ist recht technisch. Inhaltlich bedeutet sie: Man kann eine Familie unabhängiger Ereignisse in disjunkte Teilstücke zerlegen und dann innerhalb jeder Teilstücke eine beliebige Mengenoperation durchführen. Was dann herauskommt, wird wieder eine unabhängige Familie sein. Zerlegt man etwa  $\{A, B, C, D, E, F\}$  in  $\{A, B, C\}$ ,  $\{D\}$  und  $\{E, F\}$ , so garantiert der Satz, dass – zum Beispiel – die Ereignisse  $(A \cap B) \setminus C$ ,  $\Omega \setminus D$  und  $E \cup F$  unabhängig sind.

Nun zum Beweis. Zunächst betrachten wir die Menge  $\mathcal{D}_1$  aller Ereignisse  $C \in \sigma(\{A_i \mid i \in \Delta_1\})$ , so dass die Familie  $C$  zusammen mit den  $A_i$  mit  $i \notin \Delta_1$  unabhängig ist. Das ist ein Dynkinsystem, dazu muss man nur den Beweis von Satz 4.1.4 kopieren. Außerdem enthält  $\mathcal{D}_1$  nach Voraussetzung alle endlichen Schnitte der  $A_i$  mit  $i \in \Delta_1$ , und das ist ein schnitt-stabiler Erzeuger der  $\sigma$ -Algebra  $\sigma(\{A_i \mid i \in \Delta_1\})$ . Aus Satz 1.6.4 können wir dann schließen, dass alle  $C \in \sigma(\{A_i \mid i \in \Delta_1\})$  in  $\mathcal{D}_1$  liegen.

Fixiere nun ein beliebiges  $C_1$  in dieser  $\sigma$ -Algebra und betrachte als  $\mathcal{D}_2$  alle diejenigen  $C \in \sigma(\{A_i \mid i \in \Delta_2\})$ , so dass  $C_1, C$  zusammen mit den  $A_i$  für die  $i \notin \Delta_1 \cup \Delta_2$  eine unabhängige Familie bilden. Das ist ein Dynkinsystem, das alle endlichen Schnitte der  $A_i$  mit  $i \in \Delta_2$  enthält. Diese Schnitte bilden einen schnitt-stabilen Erzeuger von  $\sigma(\{A_i \mid i \in \Delta_2\})$ , und folglich stimmt  $\mathcal{D}_2$  mit dieser  $\sigma$ -Algebra überein.

Wähle darin ein beliebiges  $C_2$  und fahre analog zu den ersten Schritten fort. Nach  $r$  Durchgängen steht die Behauptung da.  $\square$

## 4.4 Unabhängigkeit für Zufallsvariable

Das Konzept „Unabhängigkeit“ soll nun auf Zufallsvariable übertragen werden. Das ist ein sehr wichtiger Schritt, denn erst dadurch können wir präzisieren, was wir eigentlich in Abschnitt 1.1 gemeint haben, als davon die Rede war, dass man „auch nach vielen Versuchen den Ausgang eines Zufallsexperiments nicht mit Sicherheit voraussagen“ kann.

Das, was wir für Ereignisse behandelt haben, soll dazu auf Zufallsvariable übertragen werden. Wir betrachten zunächst nur zwei auf einem Wahrscheinlichkeitsraum  $\Omega$  definierte reellwertige Zufallsvariable  $X, Y$ . Angenommen, wir warten auf das, was mit  $Y$  passiert (zum Beispiel: Ist  $Y(\omega) > 2?$ ), und wir

wissen, welches Ergebnis bei  $X$  herauskam: Für die relevanten  $B \subset \mathbb{R}$  lässt sich entscheiden, ob  $X(\omega) \in B$  gilt oder nicht. Sinnvollerweise wird man dann davon sprechen, dass  $X, Y$  unabhängig sind, wenn die  $X$ -Information keine Konsequenzen für die  $Y$ -Wahrscheinlichkeiten hat und umgekehrt. Etwas formaler – indem wir für die „relevanten Teilmengen von  $\mathbb{R}$ “ die Borelmengen einsetzen – heißt das:  $X, Y$  sollen unabhängig genannt werden, wenn für alle Borelmengen  $B, C \subset \mathbb{R}$  die Ereignisse  $\{X \in B\}$  und  $\{Y \in C\}$  unabhängig sind, wenn also stets

$$\mathbb{P}(\{X \in B\} \cap \{Y \in C\}) = \mathbb{P}(\{X \in B\}) \mathbb{P}(\{Y \in C\})$$

gilt. Das ist eine recht schwerfällige Bedingung, sie lässt sich aber vereinfachen:

**Lemma 4.4.1.** *Die beiden folgenden Aussagen sind äquivalent:*

- (i) *Für beliebige Borelmengen  $B, C \subset \mathbb{R}$  sind  $\{X \in B\}$  und  $\{Y \in C\}$  unabhängig.*
- (ii) *Für alle  $a, b \in \mathbb{R}$  sind die Ereignisse  $\{X \geq a\}$  und  $\{Y \geq b\}$  unabhängig.*

**Beweis:** „(i)  $\Rightarrow$  (ii)“ gilt trivialerweise, denn die  $[a, \infty[$  und  $[b, \infty[$  sind als abgeschlossene Mengen Borelmengen. Für die andere Beweisrichtung wird wieder einmal die Dynkin-Technik ausgenutzt.

Wir nehmen an, dass die Bedingung in (ii) erfüllt ist und betrachten alle Borelmengen  $B$ , so dass  $\{X \in B\}$  von allen  $\{Y \geq b\}$  unabhängig ist. Das ist, wie leicht zu sehen, ein Dynkinsystem, das nach Voraussetzung alle  $[a, \infty[$  enthält. Da diese Intervalle einen schnitt-stabilen Erzeuger der Borelmengen bilden (Satz 1.5.2), folgt aus Satz 1.6.4, dass  $\{X \in B\}$  für beliebige Borelmengen von allen  $\{Y \geq b\}$  unabhängig ist.

Fixiere nun eine Borelmenge  $B$  und betrachte alle Borelmengen  $C$ , für die  $\{X \in B\}$  von  $\{Y \in C\}$  unabhängig ist. Das ist wieder ein Dynkinsystem, und da wegen des ersten Beweisschritts alle  $[b, \infty[$  dazugehören, sind aufgrund des gleichen Arguments wie im ersten Schritt alle Borelmengen in dieser Familie. Das beweist die Behauptung.  $\square$

**Definition 4.4.2.** *Es seien  $X, Y : \Omega \rightarrow \mathbb{R}$  Zufallsvariable auf dem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{E}, \mathbb{P})$ . Sie heißen unabhängig, wenn die äquivalenten Bedingungen des vorstehenden Lemmas erfüllt sind.*

Wie bei Ereignissen auch handelt es sich um die Präzisierung von etwas, mit dem wir intuitiv ganz gut umgehen können. Als Beispiel stellen wir uns vor, dass es um Informationen – die entsprechen den Zufallsvariablen – über eine zufällig ausgewählte Person geht. Die Informationen über die Hausnummer und das Monatseinkommen sind dann sicher unabhängig. Das trifft aber nicht zu, wenn es um das Monatseinkommen und die Größe der Wohnung in Quadratmetern geht: Bei einem guten Einkommen kann man sich eine größere Wohnung leisten.

Manchmal ist das Thema sogar von großer gesellschaftspolitischer Brisanz: Erhöht exzessive Nutzung von brutalen Computerspielen die Gewaltbereitschaft? Das ist umstritten. Sind Zigarettenkonsum und Lebenserwartung unabhängig? Nein, sicher nicht.

Als mathematisches *Beispiel* dafür, dass Unabhängigkeit so etwas wie „keine gegenseitige Beeinflussung“ ausdrückt, betrachten wir das Werfen von zwei Würfeln. Das haben wir durch  $\Omega = \{1, 2, 3, 4, 5, 6\} \times \{1, 2, 3, 4, 5, 6\}$  modelliert, wobei dieser Raum mit der Gleichverteilung versehen wurde.

Wir betrachten die Zufallsvariablen  $X(i, j) := i$  und  $Y(i, j) := j$ , also die Augenzahlen auf dem ersten bzw. auf dem zweiten Würfel. Da der erste Würfel nichts vom zweiten weiß, sollten das unabhängige Zufallsvariable sein. Und wirklich: Für beliebige  $a, b \in \{1, 2, 3, 4, 5, 6\}$  (nur diese Zahlen müssen hier betrachtet werden) gibt es

1.  $(6 - a + 1) \cdot 6$  Elemente in  $\Omega$  mit  $X(i, j) \geq a$ ;
2.  $6 \cdot (6 - b + 1)$  Elemente in  $\Omega$  mit  $Y(i, j) \geq b$ ;
3.  $(6 - a + 1) \cdot (6 - b + 1)$  Elemente in  $\Omega$  mit  $X(i, j) \geq a$  und  $Y(i, j) \geq b$ .

Damit ist stets

$$\mathbb{P}(\{X \geq a, Y \geq b\}) = (6 - a + 1)(6 - b + 1)/36 = \mathbb{P}(\{X \geq a\}) \mathbb{P}(\{Y \geq b\}),$$

und  $X, Y$  sind folglich unabhängig.

Wenn wir allerdings statt  $Y$  die Zufallsvariable  $Z(i, j) := i + j$  (also die Augensumme) untersuchen, sollten  $X$  und  $Z$  *nicht* unabhängig sein: Zum Beispiel lässt ein großes  $X$  ein großes  $Z$  erwarten. Unsere Definition erkennt das. So sind etwa  $\{X \geq 6\}$  und  $\{Z \geq 3\}$  nicht unabhängig, denn die zugehörigen Wahrscheinlichkeiten sind

$$\mathbb{P}(\{X \geq 6\}) = \frac{1}{6}, \quad \mathbb{P}(\{Z \geq 3\}) = \frac{35}{36}, \quad \mathbb{P}(\{X \geq 6, Z \geq 3\}) = \frac{6}{36}.$$

Und die dritte Zahl ist nicht das Produkt der ersten beiden.

Wir beweisen noch zwei Sätze für Situationen, in denen  $\Omega$  diskret ist oder in denen die auftretenden Zufallsvariablen eine Dichte haben.

**Satz 4.4.3.** ( $\Omega, \mathcal{E}, \mathbb{P}$ ) sei ein diskreter Wahrscheinlichkeitsraum. Dann sind zwei Zufallsvariable  $X, Y : \Omega \rightarrow \mathbb{R}$  genau dann unabhängig, wenn

$$\mathbb{P}(\{X = c\} \cap \{Y = d\}) = \mathbb{P}(\{X = c\}) \mathbb{P}(\{Y = d\})$$

für alle  $c$  im Bildbereich von  $X$  und alle  $d$  im Bildbereich von  $Y$  gilt.

**Beweis:** Im Fall der Unabhängigkeit ist die Bedingung des Satzes sicher erfüllt, denn die einpunktigen Mengen  $\{c\}$  und  $\{d\}$  sind Borelmengen.

Sind umgekehrt  $B, C$  Borelmengen in  $\mathbb{R}$ , so rechnen wir wie folgt (dabei ist  $B_X$  die höchstens abzählbare Menge der Elemente von  $B$ , die im Bild von  $X$

liegen, und  $C_Y$  ist analog definiert):

$$\begin{aligned}\mathbb{P}(\{X \in B\} \cap \{Y \in C\}) &= \sum_{b \in B_X, c \in C_Y} \mathbb{P}(\{X = b\} \cap \{Y = c\}) \\ &= \sum_{b \in B_X, c \in C_Y} \mathbb{P}(\{X = b\}) \mathbb{P}(\{Y = c\}) \\ &= \left( \sum_{b \in B_X} \mathbb{P}(\{X = b\}) \right) \left( \sum_{c \in C_Y} \mathbb{P}(\{Y = c\}) \right) \\ &= \mathbb{P}(\{X \in B\}) \mathbb{P}(\{Y \in C\}).\end{aligned}$$

(Die Reihenrechnungen sind deswegen legitim, weil wir es mit absolut konvergenten Reihen zu tun haben.) Das beweist die Behauptung.  $\square$

Für die Charakterisierung der Unabhängigkeit durch Dichten benötigen wir *zwei Vorbereitungen*.

*Vorbereitung 1:* Sind  $X, Y : \Omega \rightarrow \mathbb{R}$  Zufallsvariable, so kann man doch beide Abbildungen zu einer einzigen vektorwertigen Abbildung zusammenfassen: Einem  $\omega$  wird  $(X(\omega), Y(\omega)) \in \mathbb{R}^2$  zugeordnet. Für diese Abbildung schreiben wir  $(X, Y)$ .

Es handelt sich wieder um eine Zufallsvariable: Für beliebige  $a, b$  sind die Mengen  $\{X \geq a\} \cap \{Y \geq b\}$  nach Voraussetzung Ereignisse, und die Mengen  $[a, +\infty[ \times [b, +\infty[$  erzeugen die Borelmengen des  $\mathbb{R}^2$  (vgl. Abschnitt 1.5). Folglich ist nach Satz 1.6.1  $\{(X, Y) \in B\}$  für jede Borelmenge  $B \subset \mathbb{R}^2$  ein Ereignis.

Es folgt, dass durch  $(X, Y)$  ein Wahrscheinlichkeitsmaß auf den Borelmengen des  $\mathbb{R}^2$  definiert wird. Es ist das induzierte Maß  $\mathbb{P}_{(X, Y)}$ , die Wahrscheinlichkeit einer Borelmenge  $B \subset \mathbb{R}^2$  ist  $\mathbb{P}(\{(X, Y) \in B\})$ .

*Vorbereitung 2:* Um sinnvoll untersuchen zu können, ob das eben eingeführte Maß eine Dichte hat, werden wir einige elementare Tatsachen aus der Integrationstheorie verwenden<sup>8)</sup>. Wir geben eine stetige Funktion  $\phi : \mathbb{R}^2 \rightarrow [0, +\infty[$  und ein Rechteck  $R = [a, b] \times [c, d]$  vor. Über das Doppelintegral von  $\phi$  über  $R$  sollte man dann wissen:

- $\int \int_R \phi(x, y) dx dy$  kann man sich als das Volumen der zwischen  $R$  und dem Graphen von  $\phi$  eingeschlossenen Menge vorstellen, also das Volumen von

$$\{(x, y, z) \in \mathbb{R}^3 \mid (x, y) \in R, 0 \leq z \leq \phi(x, y)\}.$$

Ist zum Beispiel  $\phi$  eine konstante Funktion, so ergibt sich ein Quader.

- Dieses Integral kann durch *iterierte Integration* berechnet werden:

$$\int \int_R \phi(x, y) dx dy = \int_c^d \left( \int_a^b \phi(x, y) dx \right) dy$$

---

<sup>8)</sup>Vgl. dazu auch den Anhang zur analysis, Seite 361.

Dabei sind dann nur Integrale über Funktionen in einer Veränderlichen zu bestimmen.

Sei etwa  $\phi(x, y) := 6xy^2$  und  $R = [0, 1] \times [0, 2]$ . Zunächst ist – bei festgehaltenem  $y$  – das „innere Integral“  $\int_0^1 6xy^2 dx$  auszurechnen.  $y$  wird wie eine Konstante behandelt, wir erhalten den Wert  $3y^2$ .

Und nun ist  $\int_0^2 3y^2 dy$  zu bestimmen. Auch das ist leicht, das Endergebnis lautet 8.

- Für uns wird es wieder reichen, mit Integrationsbereichen arbeiten zu können, die ein Rechteck sind. Man kann aber im Rahmen einer etwas fortgeschrittenen Integrationstheorie allgemeiner  $\int \int_B \phi(x, y) dx dy$  für Borelmengen  $B$  im  $\mathbb{R}^2$  erklären. Für beschränkte  $B$  geht das immer, im Allgemeinen muss man wie im Eindimensionalen Zusatzbedingungen stellen.

Nun kombinieren wir die beiden Vorbereitungen. Wir sagen, dass eine Funktion  $\phi : \mathbb{R}^2 \rightarrow [0, +\infty[$  die *gemeinsame Dichtefunktion* von  $(X, Y)$  ist, wenn

$$\mathbb{P}_{(X,Y)}(B) = \int \int_B \phi(x, y) dx dy$$

für alle Borelmengen  $B$  im  $\mathbb{R}^2$  gilt. Weil zwei Maße schon dann übereinstimmen, wenn sie auf einem schnitt-stabilen Erzeuger die gleichen Werte liefern (vgl. Satz 1.6.5), ist die vorstehende Bedingung dazu gleichwertig, dass für Rechtecke stets

$$\mathbb{P}(\{X \in [a, b]\} \times \{Y \in [c, d]\}) = \int_c^d \left( \int_a^b \phi(x, y) dx \right) dy$$

gilt.

Der nächste Satz zeigt, dass man Unabhängigkeit beim Vorliegen von Dichtefunktionen leicht feststellen kann:

**Satz 4.4.4.** *X und Y seien Zufallsvariable auf  $\Omega$ , so dass  $\mathbb{P}_X$  und  $\mathbb{P}_Y$  eine stetige Dichtefunktion haben. Es gibt also stetige  $g_X, g_Y : \mathbb{R} \rightarrow [0, +\infty[$ , so dass  $\mathbb{P}_X([a, b]) = \int_a^b g_X(x) dx$  und  $\mathbb{P}_Y([c, d]) = \int_c^d g_Y(y) dy$  für beliebige Intervalle gilt. Dann sind äquivalent:*

- X, Y sind unabhängig.*
- Die Zufallsvariable  $(X, Y)$  hat die durch  $\phi(x, y) = g_X(x)g_Y(y)$  definierte gemeinsame Dichtefunktion.*

**Beweis:** Wir setzen (i) voraus und definieren  $\phi$  wie in (ii). Für beliebige Rechtecke ist dann

$$\begin{aligned} \mathbb{P}_{(X,Y)}([a, b] \times [c, d]) &= \mathbb{P}(\{X \in [a, b]\} \cap \{Y \in [c, d]\}) \\ &= \mathbb{P}(\{X \in [a, b]\}) \mathbb{P}(\{Y \in [c, d]\}) \\ &= \left( \int_a^b g_X(x) dx \right) \left( \int_c^d g_Y(y) dy \right) \\ &= \int_c^d \int_a^b \phi(x, y) dx dy. \end{aligned}$$

(Dabei haben wir die Unabhängigkeit beim Übergang zur zweiten Zeile ausgenutzt.) Das zeigt, dass  $\phi$  wirklich Dichtefunktion für  $\mathbb{P}_{(X,Y)}$  ist.

Ist (ii) erfüllt, so lässt sich die eben durchgeführte Rechnung umkehren:

$$\begin{aligned}\mathbb{P}_{(X,Y)}([a,b] \times [c,d]) &= \int_c^d \int_a^b \phi(x,y) dx dy \\ &= \left( \int_a^b g_X(x) dx \right) \left( \int_c^d g_Y(y) dy \right) \\ &= \mathbb{P}(\{X \in [a,b]\}) \mathbb{P}(\{Y \in [c,d]\}),\end{aligned}$$

und das beweist die Unabhängigkeit.  $\square$

Wir verallgemeinern nun Definition 4.4.2 auf mehr als zwei Zufallsvariable. Auch das wird auf die Unabhängigkeit von Ereignissen zurückgeführt:

**Definition 4.4.5.** Es seien  $X_1, \dots, X_n$  reellwertige Zufallsvariable auf dem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{E}, \mathbb{P})$ . Sie heißen unabhängig, wenn gilt: Für beliebige reelle Zahlen  $a_1, \dots, a_n$  sind die Ereignisse

$$\{X_1 \geq a_1\}, \dots, \{X_n \geq a_n\}$$

unabhängig.

Zusatz: Eine beliebige Familie von Zufallsvariablen heißt unabhängig, wenn jede endliche Teilfamilie diese Eigenschaft hat.

Lemma 4.4.1 hat ein Analogon:

**Lemma 4.4.6.** Die folgenden Aussagen sind äquivalent:

- (i)  $X_1, \dots, X_n$  sind unabhängig.
- (ii) Für Borelmengen  $B_1, \dots, B_n \subset \mathbb{R}$  sind  $\{X_1 \in B_1\}, \dots, \{X_n \in B_n\}$  unabhängig.

**Beweis:** Für den Beweis kann die Beweisstrategie für Lemma 4.4.1 kopiert werden: Betrachte alle Borelmengen  $B_1$ , so dass für beliebige  $a_2, \dots, a_n$  die Ereignisse  $\{X_1 \in B_1\}, \{X_2 \geq a_2\}, \dots, \{X_n \geq a_n\}$  unabhängig sind. Das ist ein Dynkinsystem, das nach Voraussetzung den schnitt-stabilen Erzeuger  $\{\{a, +\infty[ \mid a \in \mathbb{R}\}$  aller Borelmengen enthält. Also gehören alle Borelmengen zu diesem System. Dann wird  $B_1$  fixiert und man betrachtet alle  $B_2$ , so dass die Ereignisse  $\{X_1 \in B_1\}, \{X_2 \in B_2\}, \{X_3 \geq a_3\}, \dots, \{X_n \geq a_n\}$  für alle  $a_3, \dots, a_n$  unabhängig sind. Es ergibt sich wie im ersten Schritt, dass alle  $B_2$  zugelassen sind. Und so weiter.  $\square$

Am Ende dieses Abschnitts zeigen wir noch, dass das „Bündeln“ von unabhängigen Zufallsvariablen wieder zu unabhängigen Zufallsvariablen führt. Es ist zum Beispiel plausibel, dass bei unabhängigen  $X, Y, Z$  auch  $X, Y + Z$  unabhängig sind.

Um dieses Analogon zu Satz 4.3.5 beweisen zu können, benötigen wir eine Vorbereitung:

Wie viel Information steckt in Zufallsvariablen?

Mal angenommen,  $(\Omega, \mathcal{E}, \mathbb{P})$  ist ein Wahrscheinlichkeitsraum und  $Y_1, \dots, Y_m : \Omega \rightarrow \mathbb{R}$  sind Zufallsvariable. Wenn der Zufall ein  $\omega$  erzeugt hat und wir die  $Y_1(\omega), \dots, Y_m(\omega)$  kennen, so ist doch für jede Menge  $B \subset \mathbb{R}^m$  die Frage beantwortbar, ob das  $m$ -Tupel  $(Y_1(\omega), \dots, Y_m(\omega))$  in  $B$  liegt oder nicht. Insbesondere gilt das für Borelmengen  $B$ .

Bemerkenswerterweise bilden die Mengen  $\{\omega \mid (Y_1(\omega), \dots, Y_m(\omega)) \in B\}$  eine Teil- $\sigma$ -Algebra von  $\mathcal{E}$ , der  $\sigma$ -Algebra der Ereignisse, wenn  $B$  alle Borelmengen durchläuft. Dabei handelt es sich um die *kleinste Teil- $\sigma$ -Algebra* von  $\mathcal{E}$ , für die  $Y_1, \dots, Y_m$  Zufallsvariable sind.

Begründung: Sei  $\mathcal{E}_Y$  das Mengensystem dieser  $\{\omega \mid (Y_1(\omega), \dots, Y_m(\omega)) \in B\}$ , wobei  $B$  alle Borelmengen durchläuft. Für Borelmengen  $B' \subset \mathbb{R}$  ist jede Menge  $\{Y_i \in B'\}$  in  $\mathcal{E}_Y$ , da man sie als  $\{(Y_1, \dots, Y_n) \in B_1 \times \dots \times B_n \mid Y_i \in B'\}$  schreiben kann, wobei  $B_j := \mathbb{R}$  für  $j \neq i$  und  $B_i := B'$ . Damit sind alle  $Y_i$  Zufallsvariable in Bezug auf  $\mathcal{E}_Y$ . Umgekehrt: Ist  $\mathcal{E}' \subset \mathcal{E}$  eine  $\sigma$ -Algebra, so dass alle  $\{Y_i \in B'_i\}$  zu  $\mathcal{E}'$  gehören, so liegen auch die  $\{(Y_1, \dots, Y_m) \in B'_1 \times \dots \times B'_m\}$  als Schnitt derartiger Mengen in  $\mathcal{E}'$ . Und da die von den  $B'_1 \times \dots \times B'_m$  erzeugte  $\sigma$ -Algebra<sup>9)</sup> aus allen Borelmengen des  $\mathbb{R}^m$  besteht, gilt  $\mathcal{E}_Y \subset \mathcal{E}'$ . Die  $\sigma$ -Algebra  $\mathcal{E}_Y$  ist also wirklich kleinstmöglich.

$\mathcal{E}_Y$  kann übrigens sehr unterschiedlich sein. Sind alle  $Y_i$  konstant, so besteht  $\mathcal{E}_Y$  nur aus  $\Omega$  und der leeren Menge. Es kann aber auch sein, dass *alle* Ereignisse herauskommen. Zum Beispiel dann, wenn  $\Omega$  endlich viele Elemente hat und ein  $Y_i$  injektiv ist.

Aus diesem Grund heißt sie *die von  $Y_1, \dots, Y_m$  erzeugte  $\sigma$ -Algebra*. Man schreibt dafür  $\sigma(Y_1, \dots, Y_m)$ .

„Bündeln“ erhält die Unabhängigkeit:

**Satz 4.4.7.**  *$X_1, \dots, X_n$  seien reellwertige unabhängige Zufallsvariable auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{E}, \mathbb{P})$ , und es seien  $\Delta_1, \dots, \Delta_m$  nichtleere und paarweise disjunkte Teilmengen von  $\{1, \dots, n\}$ . Sind dann*

$$E_1 \in \sigma(\{X_i \mid i \in \Delta_1\}), \dots, E_m \in \sigma(\{X_i \mid i \in \Delta_m\}),$$

so sind  $E_1, \dots, E_m$  unabhängig.

**Beweis:** Der Beweis nutzt das in diesem Abschnitt schon mehrfach verwendete Dynkinargument aus. Zunächst betrachtet man die  $E_1 \in \sigma(\{X_i \mid i \in \Delta_1\})$ , für die die Familie  $E_1$  zusammen mit den  $\{X_i \in C_i\}, i \in \Delta_2 \cup \dots \cup \Delta_m$  für beliebige Borelmengen  $C_i \subset \mathbb{R}$  unabhängig ist. Die Gesamtheit dieser  $E_1$  ist ein Dynkinsystem, und nach Voraussetzung enthält es die endlichen Schnitte der  $\{X_i \in C_i\}$  für Borelmengen  $C_i \subset \mathbb{R}$  und  $i \in \Delta_1$ . Das ist ein schnittstabilier Erzeuger von  $\sigma(\{X_i \mid i \in \Delta_1\})$ , und deswegen sind *alle*  $E_1$  in dieser  $\sigma$ -Algebra zulässig. Weiter geht es damit, dass  $E_1$  fixiert wird und man alle  $E_2 \in \sigma(\{X_i \mid i \in \Delta_2\})$  betrachtet, für die  $E_1, E_2$  zusammen mit den  $\{X_i \in C_i\}, i \in \Delta_3 \cup \dots \cup \Delta_m$  für beliebige  $C_i$  unabhängig ist. Und so weiter, nach  $m$  Schritten ist die Behauptung bewiesen.  $\square$

Die für uns wichtigste Konsequenz steht in

---

<sup>9)</sup>Die  $B'_i$  durchlaufen dabei die eindimensionalen Borelmengen.

**Satz 4.4.8.** Die  $X_i$  und die  $\Delta_j$  seien wie im vorstehenden Satz, dabei habe  $\Delta_j$   $n_j$  Elemente. Ohne Einschränkung sei  $\Delta_j = \{r_j, \dots, r_{j+1} - 1\}$  für geeignete Zahlen  $1 = r_1 < \dots < r_m < r_{m+1} \leq n+1$  und  $j = 1, \dots, m$ . Es sind Funktionen  $f_j : \mathbb{R}^{n_j} \rightarrow \mathbb{R}$  vorgegeben, sie sollen die Eigenschaft haben, dass  $\{f_j \in C\}$  für jede Borelmenge  $C$  in  $\mathbb{R}$  eine Borelmenge im  $\mathbb{R}^{n_j}$  ist<sup>10)</sup>.

Wenn man, für  $j = 1, \dots, m$ , die  $X_i$  mit  $i \in \Delta_j$  in  $f_j$  einsetzt, entsteht eine neue Funktion  $Y_j : \Omega \rightarrow \mathbb{R}$ :

$$Y_j(\omega) := f_j(X_{r_j}(\omega), \dots, X_{r_{j+1}-1}(\omega)).$$

$Y_j$  ist wegen der Voraussetzung an  $f_j$  eine Zufallsvariable. Die Behauptung:  $Y_1, \dots, Y_m$  sind unabhängig.

**Beweis:**  $C_1, \dots, C_m$  seien Borelmengen in  $\mathbb{R}$ . Nach Voraussetzung ist  $B_j := \{f_j \in C_j\}$  Borelmenge in  $\mathbb{R}^{n_j}$  für  $j = 1, \dots, m$  und folglich liegen die durch

$$E_j := \{\omega \mid (X_{r_j}(\omega), \dots, X_{r_{j+1}-1}(\omega)) \in B_j\}$$

definierten Mengen in  $\sigma(X_{r_j}, \dots, X_{r_{j+1}-1})$ .

Da die  $E_1, \dots, E_m$  nach Satz 4.4.7 unabhängig sind und da  $\{Y_j \in C_j\} = E_j$  für alle  $j$  gilt, ist alles gezeigt.  $\square$

Damit sind wir am Ziel: Ohne Verlust der Unabhängigkeit kann beliebig zusammengefasst werden, wobei (mindestens) alle stetigen Operationen angewendet werden dürfen. Wichtig ist nur, dass jedes einzelne  $X_i$  bei höchstens einer der neuen Zufallsvariablen verwendet wurde. Wenn also etwa  $X, Y, Z, W$  unabhängig sind, so weiß man:

- $X^2, \sin(Y), 3Z$  sind unabhängig;
- $X + Y, Z - \sqrt{Z^2 + W^2}$  sind unabhängig
- ...

Es kann aber – z.B. – nicht garantiert werden, dass  $X^2 + Y, 2Y + Z + W$  unabhängig sind, denn  $Y$  kommt in beiden Funktionen vor.

Es sollte noch bemerkt werden, dass Satz 4.3.5 Spezialfall des vorstehenden Ergebnisses ist: Man muss es nur für die Zufallsvariablen  $\chi_{A_1}, \dots, \chi_{A_n}$  anwenden.

## 4.5 Der „Klonsatz“

Dieser Abschnitt ist von großer theoretischer Bedeutung: Die hier behandelten Ergebnisse dienen erstens dazu, den Zusammenhang Zufallsautomat  $\leftrightarrow$  Wahr-

<sup>10)</sup>Solche Funktionen heißen *Borelfunktionen*. Die Faustregel: Alles, was man sinnvoll definieren kann, ist Borelfunktion. Für uns wird es reichen zu wissen, dass stetige Funktionen Borelfunktionen sind. Vgl. Korollar 3.1.3.

scheinlichkeitsraum besser verstehen zu können, und zweitens sind sie eine unerlässliche Voraussetzung für spätere Untersuchungen zum Verhalten des Zufalls im Unendlichen<sup>11)</sup>.

Zentrales Ergebnis ist der (in diesem Buch so genannte) nachstehende „Klonsatz“ 4.5.1: Man kann Zufallsvariable in beliebiger Anzahl „vervielfältigen“, so dass die Kopien eine unabhängige Familie bilden:

**Satz 4.5.1.** *Gegeben seien ein Wahrscheinlichkeitsraum  $(\Omega, \mathcal{E}, \mathbb{P})$  und eine Zufallsvariable  $X : \Omega \rightarrow \mathbb{R}$ .*

*Dann gibt es einen Wahrscheinlichkeitsraum  $(\Omega_\infty, \mathcal{E}_\infty, \mathbb{P}_\infty)$  und darauf definierte Zufallsvariable  $X_1, X_2, \dots : \Omega_\infty \rightarrow \mathbb{R}$  mit den folgenden Eigenschaften:*

*(i) Die  $X_n$  sind wie  $X$  verteilt, d.h. für die induzierten Maße auf  $\mathbb{R}$  gilt  $\mathbb{P}_X = \mathbb{P}_{X_n}$  für alle  $n$ .*

*(ii)  $X_1, X_2, \dots$  sind unabhängig.*

**Beweis:** *Schritt 1: Die Definition von  $\Omega_\infty$ :* Wir definieren  $\Omega_\infty$  als die Menge aller Folgen  $(\omega_k)_k$  in  $\Omega$ . Ist  $\Omega$  zum Beispiel der Raum, mit dem wir den fairen Würfel modellieren, so besteht  $\Omega_\infty$  aus allen Folgen, in denen nur die Zahlen 1, 2, 3, 4, 5, 6 vorkommen. Man kann sich  $\Omega_\infty$  in diesem Fall als die Menge der möglichen Ergebnisse bei einer unendlichen Folge von Würfelwürfen vorstellen<sup>12)</sup>.

*Schritt 2: Die Definition von  $\mathcal{E}_\infty$ :* Ereignisse entsprechen doch erlaubten Fragen über den Ausgang von Zufallsexperimenten. Und sicher sollen Fragen des Typs „War der 10. Würfelwurf eine 4?“ oder „War der dritte Wurf eine gerade Zahl und waren die Ergebnisse bei den Würfeln 17 und 22 größer als 2?“ zulässig sein.

Um das zu präzisieren, gehen wir so vor: Zunächst definieren wir für  $m \in \mathbb{N}$  und Ereignisse  $E_1, \dots, E_m \in \mathcal{E}$  eine Teilmenge  $F_{E_1, \dots, E_m}$  von  $\Omega_\infty$  durch

$$F_{E_1, \dots, E_m} := \{(\omega_k)_k \in \Omega_\infty \mid \omega_k \in E_k \text{ für alle } k = 1, \dots, m\},$$

und dann erklären wir  $\mathcal{E}_\infty$  als die von

$$\{F_{E_1, \dots, E_m} \mid m \in \mathbb{N}, E_1, \dots, E_m \in \mathcal{E}\}$$

erzeugte  $\sigma$ -Algebra. Damit können Fragen wie die des vorigen Abschnitts als Ereignisse interpretiert werden. Es ist aber viel mehr richtig: Da  $\mathcal{E}_\infty$  eine  $\sigma$ -Algebra ist, sind auch Vereinigungen mit „und“, „oder“ und „nicht“ zugelassen, sofern man sich dabei auf höchstens abzählbar viele Ereignisse beschränkt. Im Würfelbeispiel etwa kann man die Menge der  $(\omega_k)$  betrachten, bei denen für gerade  $k$  stets  $\omega_k = 2$  oder für  $k \geq 10.000$  stets  $\omega_k \geq 4$  gilt. Auch das ist ein Element von  $\mathcal{E}_\infty$ .

---

<sup>11)</sup> Mehr dazu im grauen Kasten am Ende dieses Abschnitts.

<sup>12)</sup> Schon für diesen Fall ist  $\Omega_\infty$  eine gigantisch große Menge, sie hat überabzählbar viele Elemente.

Als Faustregel kann man sich merken, dass alles, was man über eine Folge von Zufallsergebnissen sinnvollerweise erfahren möchte, einem Element aus  $\mathcal{E}_\infty$  entspricht.

*Schritt 3: Die Definition der  $X_n$ :* Der Hintergedanke bei der Definition von  $\Omega_\infty$  war doch, jedes einzelne Element  $(\omega_k)$  als eine Folge von Zufallsabfragen in  $(\Omega, \mathcal{E}, \mathbb{P})$  zu interpretieren. Die entsprechenden  $X$ -Ergebnisse sind dann  $X(\omega_1), X(\omega_2), \dots$ . Deswegen ist es naheliegend, die  $X_n$  für  $n = 1, 2, \dots$  wie folgt zu definieren:

$$X_n : \Omega_\infty \rightarrow \mathbb{R}, \quad X_n((\omega_k)_k) := X(\omega_n).$$

Diese Abbildungen sind dann Zufallsvariable auf  $(\Omega_\infty, \mathcal{E}_\infty, \mathbb{P}_\infty)$ , unabhängig davon, wie das Wahrscheinlichkeitsmaß  $\mathbb{P}_\infty$  definiert werden wird. Die Begründung: Ist  $a \in \mathbb{R}$ , so definiere  $E_k := \Omega$  für  $k = 1, \dots, n-1$  und  $E_n := \{X \geq a\}$ . Die  $E_1, \dots, E_n$  liegen in  $\mathcal{E}$ , und es gilt

$$\{X_n \geq a\} = F_{E_1, \dots, E_n} \in \mathcal{E}_\infty.$$

*Schritt 4: Die Definition von  $\mathbb{P}_\infty$ :* Zunächst kümmern wir uns darum, wie dieses noch unbekannte Wahrscheinlichkeitsmaß auf Ereignissen des Typs  $F_{E_1, \dots, E_m}$  definiert sein muss. Es soll doch zwei Eigenschaften haben:

- Es soll  $\mathbb{P}_X = \mathbb{P}_{X_n}$  für alle  $n$  gelten. Das heißt: Ist  $B \subset \mathbb{R}$  eine Borelmenge, so muss  $\mathbb{P}(\{X \in B\}) = \mathbb{P}_\infty(F_{E_1, \dots, E_n})$  sein, wobei  $E_1 = \dots = E_{n-1} := \Omega$  und  $E_n := \{X \in B\}$ .
- Die  $X_n$  sollen unabhängig sein. Das bedeutet: Für Borelmengen  $B_1, \dots, B_n$  in  $\mathbb{R}$  muss stets

$$\mathbb{P}_\infty\left(\bigcap_{i=1}^n \{X_i \in B_i\}\right) = \mathbb{P}_\infty(\{X_1 \in B_1\}) \cdots \mathbb{P}_\infty(\{X_n \in B_n\})$$

gelten<sup>13)</sup>. Definiert man  $E_i := \{X \in B_i\}$  für  $i = 1, \dots, n$ , so heißt das gerade, dass

$$\mathbb{P}_\infty(F_{E_1, \dots, E_n}) = \mathbb{P}(E_1) \cdots \mathbb{P}(E_n)$$

gelten muss.

Das können wir so zusammenfassen: Wenn für  $\mathbb{P}_\infty$  stets die Formel  $\mathbb{P}_\infty(F_{E_1, \dots, E_n}) = \mathbb{P}(E_1) \cdots \mathbb{P}(E_n)$  gilt, wobei die  $E_i$  von der Form  $\{X \in B\}$  sind, so sind alle im Satz behaupteten Aussagen bewiesen.

Gibt es so ein  $\mathbb{P}_\infty$ ? Wie oft in der Mathematik nimmt man das, was gelten soll, als Ausgangspunkt für eine Definition. Wir lassen dazu alle  $E_i \in \mathcal{E}$  zu (nicht nur die von der Form  $\{X \in B\}$ ) und definieren  $\mathbb{P}_\infty$  für Ereignisse des Typs  $F_{E_1, \dots, E_m}$  durch  $\mathbb{P}(E_1) \cdots \mathbb{P}(E_n)$ . Es ist dann noch zu zeigen:

a)  $\mathbb{P}_\infty$  ist dadurch auf diesen Mengen wohldefiniert<sup>14)</sup>. Damit ist gemeint: Ist eine Teilmenge von  $\Omega_\infty$  auf verschiedene Weise als  $F_{E_1, \dots, E_m}$  dargestellt, so ergibt

<sup>13)</sup>Vgl. Übungsaufgabe Ü4.4.5.

<sup>14)</sup>Den meisten wird das spitzfindig vorkommen, es muss aber wirklich nachgewiesen werden.

sich trotzdem unter  $\mathbb{P}_\infty$  die gleiche Zahl. Als Beispiel betrachten wir  $F_{E_1, E_2}$ . Diese Menge kann auch – zum Beispiel – als  $F_{E_1, E_2, \Omega, \Omega}$  geschrieben werden, denn  $\omega_3, \omega_4 \in \Omega$  gilt ja immer. Rechnet man die zugehörige Wahrscheinlichkeit auf der Grundlage von  $F_{E_1, E_2}$  aus, so ergibt sich  $\mathbb{P}(E_1)\mathbb{P}(E_2)$ . Geht man von der Darstellung  $F_{E_1, E_2, \Omega, \Omega}$  aus, gelangt man zu  $\mathbb{P}(E_1)\mathbb{P}(E_2)\mathbb{P}(\Omega)\mathbb{P}(\Omega)$ . Wegen  $\mathbb{P}(\Omega) = 1$  sind beide Zahlen gleich. Da auch für allgemeine  $F_{E_1, \dots, E_m}$  die einzige Mehrdeutigkeit wie die in diesem Beispiel ist, ist  $\mathbb{P}_\infty$  wohldefiniert.

b) *Die Definition von  $\mathbb{P}_\infty$  kann von der Menge der  $F_{E_1, \dots, E_m}$  so auf die erzeugte  $\sigma$ -Algebra, also auf  $\mathcal{E}_\infty$ , erweitert werden, dass sich ein Wahrscheinlichkeitsmaß ergibt.*

Das ist wirklich ein schwieriger Punkt. Das Problem besteht darin, dass beim Übergang von einem Mengensystem zur erzeugten  $\sigma$ -Algebra viele Mengen zu berücksichtigen sind, die man nicht explizit beschreiben kann. Dieses Problem kann aber gelöst werden: Erstens durch einen tiefliegenden allgemeinen Satz (den *Satz von Carathéodory*), der besagt, dass man unter gewissen Bedingungen eine auf einem Mengensystem definierte Abbildung zu einem Wahrscheinlichkeitsmaß auf der erzeugten  $\sigma$ -Algebra fortsetzen kann. Und dann muss man zweitens noch zeigen, dass die fraglichen Bedingungen in der vorliegenden Situation erfüllt sind (Satz vom Produktmaß). Ein Beweis dieser Ergebnisse würde den Rahmen dieses Buches sprengen, deswegen verweise ich hier nur auf die Literatur zur Maßtheorie (Anhang, Seite 356).  $\square$

### Zur Bedeutung des „Klonsatzes“:

*Erstens* können wir nun präzisieren, was mit „Wir fragen einen Wahrscheinlichkeitsraum  $(\Omega, \mathcal{E}, \mathbb{P})$  mehrfach ab“ in Kapitel 1 gemeint war. Man muss nur von  $(\Omega, \mathcal{E}, \mathbb{P})$  zu  $(\Omega_\infty, \mathcal{E}_\infty, \mathbb{P}_\infty)$  übergehen und die Elemente von  $\Omega_\infty$  als mögliches Ergebnis einer Folge von Abfragen interpretieren. Das  $n$ -te Element von  $(\omega_k)_k$  entspricht dem Ergebnis im  $n$ -ten Versuch, die Wahrscheinlichkeiten sind so, wie sie durch  $\mathbb{P}$  vorgegeben sind, und die einzelnen Abfragen „haben nichts miteinander zu tun“, denn sie sind unabhängig.

Und *zweitens* werden wir in späteren Kapiteln untersuchen, was passiert, wenn wir „sehr oft“ Abfragen  $\omega$  aus einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{E}, \mathbb{P})$  erzeugen und untersuchen, was sich dann über das mittlere Verhalten der so entstehenden  $X(\omega)$  aussagen lässt. Der „Klonsatz“ ist das theoretische Fundament für derartige Untersuchungen.

In Abschnitt 8.4 werden wir noch eine etwas allgemeinere Variante des Satzes benötigen:

**Satz 4.5.2.** *Gegeben seien Wahrscheinlichkeitsräume  $(\Omega_n, \mathcal{E}_n, \mathbb{P}_n)$  und Zufallsvariable  $Y_n : \Omega_n \rightarrow \mathbb{R}$  für  $n = 1, 2, \dots$ . Dann gibt es einen Wahrscheinlichkeitsraum  $(\Omega_\infty, \mathcal{E}_\infty, \mathbb{P}_\infty)$  und darauf definierte Zufallsvariable  $X_1, X_2, \dots : \Omega_\infty \rightarrow \mathbb{R}$  mit den folgenden Eigenschaften:*

(i) Die  $X_n$  sind wie  $Y_n$  verteilt, d.h. für die induzierten Maße auf  $\mathbb{R}$  gilt  $\mathbb{P}_{Y_n} = \mathbb{P}_{X_n}$  für alle  $n$ .

(ii)  $X_1, X_2, \dots$  sind unabhängig.

**Beweis:** Der Beweis ist völlig analog zum Beweis von Satz 4.5.1.  $\Omega_\infty$  ist diesmal die Menge der Folgen  $(\omega_k)$ , für die  $\omega_k \in \Omega_k$  für alle  $k$  gilt. Die Zufallsvariablen  $X_n$  sind durch  $X_n((\omega_k)) := Y_n(\omega_n)$  definiert.

Satz 4.5.1 entspricht dem Spezialfall, in dem alle  $(\Omega_n, \mathcal{E}_n, \mathbb{P}_n)$  gleich einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{E}, \mathbb{P})$  sind und alle  $Y_n$  mit einer Zufallsvariablen  $X : \Omega \rightarrow \mathbb{R}$  übereinstimmen.  $\square$

Wir können jetzt auch präzisieren, was wir mit den „wiederholten Abfragen eines Zufallsautomaten“ in Abschnitt 1.1 eigentlich gemeint haben. Es sollten natürlich *unabhängige* Abfragen sein, und  $n$ -maliges unabhängiges Abfragen aus einem Wahrscheinlichkeitsraum entspricht gerade einer einmaligen Abfrage aus dem Produktraum: Den haben wir im Klonsatz – sogar für unendlich viele Abfragen – implizit schon kennen gelernt. Im nächsten Satz formulieren wir eine weitere Variante, die in Kapitel 10 gebraucht werden wird.

**Satz 4.5.3.** *Es sei  $(\Omega, \mathcal{E}, \mathbb{P})$  ein Wahrscheinlichkeitsraum und  $n \in \mathbb{N}$ . Dann gibt es eine  $\sigma$ -Algebra  $\mathcal{E}_n$  auf  $\Omega^n$  und ein Wahrscheinlichkeitsmaß  $\mathbb{P}^n$  auf  $(\Omega^n, \mathcal{E}_n)$ , so dass gilt: Definiert man  $X_i : \Omega^n \rightarrow \Omega$  als die  $i$ -te Projektion (also durch  $(\omega_1, \dots, \omega_n) \mapsto \omega_i$ ) für  $i = 1, \dots, n$ , so sind die  $X_1, \dots, X_n$  unabhängige Zufallsvariable und die induzierten Wahrscheinlichkeitsmaße  $\mathbb{P}_{X_i}$  stimmen alle mit  $\mathbb{P}$  überein.*

*Kurz: Man kann auch einen Wahrscheinlichkeitsraum  $n$ -fach „klonen“.*

**Beweis:** Eine vereinfachte Variante des Beweises des Klonsatzes 4.5.1 führt zum Ziel:

- Für  $E_1, \dots, E_n \in \mathcal{E}$  sei  $F_{E_1, \dots, E_n}$  die Menge

$$F_{E_1, \dots, E_n} := \{(\omega_1, \dots, \omega_n) \in \Omega^n \mid \omega_i \in E_i \text{ für alle } i = 1, \dots, n\}.$$

- $\mathcal{E}_n$  soll die von diesen Mengen erzeugte  $\sigma$ -Algebra auf  $\Omega^n$  bezeichnen.
- Nun müssen wir wieder ein Ergebnis aus der Maßtheorie übernehmen: Danach gibt es ein Wahrscheinlichkeitsmaß  $\mathbb{P}^n$  auf  $(\Omega^n, \mathcal{E}^n)$ , so dass stets

$$\mathbb{P}^n(F_{E_1, \dots, E_n}) = \mathbb{P}(E_1) \cdots \mathbb{P}(E_n)$$

gilt. Dieses Maß wird das  $n$ -fache *Produktmaß* genannt.

Es ist dann klar, dass die Projektionen  $X_1, \dots, X_n$  die behaupteten Eigenschaften haben.  $\square$

## 4.6 Folgerungen aus der Unabhängigkeit

Wenn  $X, Y : \Omega \rightarrow \mathbb{R}$  Zufallsvariable sind, für die der Erwartungswert existiert, so gilt  $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$ , es gibt aber keine allgemein gültige Formel, durch die  $\mathbb{E}(X \cdot Y)$  durch  $\mathbb{E}(X)$  und  $\mathbb{E}(Y)$  ausgedrückt werden kann. So ist sicher *nicht* im Allgemeinen,  $\mathbb{E}(X \cdot Y) = \mathbb{E}(X)\mathbb{E}(Y)$ , als Gegenbeispiel kann man irgendein  $X$  mit  $\mathbb{E}(X) = 0$  und  $\mathbb{E}(X^2) > 0$  wählen und  $X = Y$  setzen.

Im Fall der Unabhängigkeit gilt aber der

**Satz 4.6.1.** *Sind  $X, Y : \Omega \rightarrow \mathbb{R}$  unabhängige Zufallsvariable mit existierendem Erwartungswert, so existiert auch der Erwartungswert von  $X \cdot Y$ , und es gilt  $\mathbb{E}(X \cdot Y) = \mathbb{E}(X)\mathbb{E}(Y)$ .*

**Beweis:** Wir führen die Aussage auf den Spezialfall zurück, dass  $X$  und  $Y$  Indikatorfunktionen sind. Dabei wird wichtig, dass der Erwartungswert der Indikatorfunktion  $\chi_E$  gleich  $\mathbb{P}(E)$  ist. Für unabhängige Ereignisse  $E, F$  lässt sich deswegen so rechnen:

$$\mathbb{E}(\chi_E \cdot \chi_F) = \mathbb{E}(\chi_{E \cap F}) = \mathbb{P}(E \cap F) = \mathbb{P}(E)\mathbb{P}(F) = \mathbb{E}(\chi_E)\mathbb{E}(\chi_F).$$

Etwas weiter unten ist der Beweis für allgemeine Wahrscheinlichkeitsräume zu finden. Für alle, denen diskrete Räume allgemein genug sind, gibt es davor den (einfacheren) Beweis für diese Situation.

*Fall 1:  $\Omega$  diskret.* Zunächst soll  $X, Y \geq 0$  gelten. Wähle Zahlen  $a_i, b_j \geq 0$ , so dass der Bildbereich von  $X$  (bzw.  $Y$ ) die Menge  $\{a_1, a_2, \dots\}$  (bzw.  $\{b_1, b_2, \dots\}$ ) ist und setze  $E_i = \{X = a_i\}$  (bzw.  $F_j := \{Y = b_j\}$ ). Nach Voraussetzung sind stets  $E_i, F_j$  unabhängig, und  $\mathbb{E}(X)$  (bzw.  $\mathbb{E}(Y)$ ) ist die absolut konvergente Reihe  $\sum_i a_i \mathbb{P}(E_i)$  (bzw.  $\sum_j b_j \mathbb{P}(F_j)$ ). Die Funktion  $X \cdot Y$  kann als  $\sum_{i,j} a_i b_j \chi_{E_i \cap F_j}$  geschrieben werden, folglich gilt

$$\begin{aligned} \mathbb{E}(X \cdot Y) &= \sum_{i,j} a_i b_j \mathbb{E}(\chi_{E_i \cap F_j}) \\ &= \sum_{i,j} a_i b_j \mathbb{P}(E_i \cap F_j) \\ &= \sum_{i,j} a_i b_j \mathbb{P}(E_i) \mathbb{P}(F_j) \\ &= (\sum_i a_i \mathbb{P}(E_i)) (\sum_j b_j \mathbb{P}(F_j)) \\ &= \mathbb{E}(X)\mathbb{E}(Y). \end{aligned}$$

Die vorletzte Umformulierung ist deswegen gerechtfertigt, weil die auftretenden Reihen absolut konvergent sind.

Falls  $X$  und  $Y$  auch negative Werte annehmen, schreibe  $X = X^+ - X^-$  und  $Y = Y^+ - Y^-$  mit  $X^+, X^-, Y^+, Y^- \geq 0$ . Die Behauptung folgt dann aus dem vorstehenden Beweisteil und der Linearität des Erwartungswerts.

*Fall 2: Allgemeine Wahrscheinlichkeitsräume*<sup>15)</sup>. Wie im Beweis für Fall 1 gezeigt, reicht es, sich auf den Fall  $X, Y \geq 0$  zu konzentrieren. Zur Berechnung der Erwartungswerte gehen wir wie in Definition 3.3.6 vor:

$$E_k^n := \{X \in I_k^n\}, \quad F_l^n := \{Y \in I_l^n\}, \quad X_n = \sum_{k \geq 0} \frac{k}{2^n} \chi_{E_k^n}, \quad Y_n = \sum_{l \geq 0} \frac{l}{2^n} \chi_{F_l^n}.$$

Dabei ist wichtig zu bemerken, dass  $E_k^n, F_l^n$  für alle  $k, l$  unabhängig sind, denn  $E_k^n \in \sigma(X)$  und  $F_l^n \in \sigma(Y)$ . Deswegen folgt wie im diskreten Fall  $\mathbb{E}(X_n Y_n) = \mathbb{E}(X_n)\mathbb{E}(Y_n)$ .

Insbesondere hat  $X_n Y_n$  einen endlichen Erwartungswert, und da aus  $X_n \leq X \leq X_n + 1/2^n$  und  $Y_n \leq Y \leq Y_n + 1/2^n$  die Ungleichung

$$X_n Y_n \leq XY \leq X_n Y_n + X_n/2^n + Y_n/2^n + 2/2^n$$

folgt, existiert nach Satz 3.3.7(ii) auch  $\mathbb{E}(XY)$ . Außerdem gilt

$$\mathbb{E}(X_n)\mathbb{E}(Y_n) = \mathbb{E}(X_n Y_n) \leq \mathbb{E}(XY) \leq \mathbb{E}(X_n)\mathbb{E}(Y_n) + (\mathbb{E}(X_n) + \mathbb{E}(Y_n))/2^n + 2/2^n.$$

Beachtet man noch, dass  $\lim \mathbb{E}(X_n) = \mathbb{E}(X)$  und  $\lim \mathbb{E}(Y_n) = \mathbb{E}(Y)$ , so ergibt sich  $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ .  $\square$

Die wichtigste Konsequenz dieses Ergebnisses steht in

**Satz 4.6.2.** *Es seien  $X_1, \dots, X_n : \Omega \rightarrow \mathbb{R}$  Zufallsvariable, für die Erwartungswert und Varianz existieren.*

(i) *Sind für beliebige  $i, j$  mit  $i \neq j$  die Zufallsvariablen  $X_i, X_j$  unabhängig<sup>16)</sup>, so gilt*

$$V(X_1 + \dots + X_n) = V(X_1) + \dots + V(X_n).$$

(ii) *Es folgt: Sind  $X_1, \dots, X_n$  unabhängig und wie eine Zufallsvariable  $X$  mit existierendem Erwartungswert und existierender Streuung verteilt<sup>17)</sup>, so gilt*

$$\sigma\left(\frac{X_1 + \dots + X_n}{n}\right) = \frac{\sigma(X)}{\sqrt{n}}.$$

Das ist das Wurzel- $n$ -Gesetz.

**Beweis:** Ohne Einschränkung können wir annehmen, dass alle Erwartungswerte Null sind: Andernfalls gehe man zu  $Y_i := X_i - \mathbb{E}(X_i)$  über, dabei bleiben die Voraussetzungen an die Unabhängigkeit erhalten (vgl. Satz 4.4.7).

<sup>15)</sup> Vielleicht vermissen Sie hier eine Diskussion von Räumen mit Dichten. Darauf wird hier verzichtet, denn der Beweis wäre nicht kürzer als der Beweis für die allgemeine Situation.

<sup>16)</sup> Man sagt dann auch, dass sie *paarweise unabhängig* sind.

<sup>17)</sup> D.h., es ist  $\mathbb{P}_{X_i} = \mathbb{P}_X$  für alle  $i$ .

(i) Zunächst ist zu beachten, dass wegen Satz 3.3.9 die Varianz von  $X_1 + \dots + X_n$  existiert. Und unter der Voraussetzung  $\mathbb{E}(X_i) = 0$  ist wirklich

$$\begin{aligned} V(X_1 + \dots + X_n) &= \mathbb{E}((X_1 + \dots + X_n)^2) \\ &= \mathbb{E}\left(\sum_i X_i^2 + 2 \sum_{i,j} X_i X_j\right) \\ &= \sum_i V(X_i) + 2 \sum_{i,j} \mathbb{E}(X_i X_j) \\ &= \sum_i V(X_i) + 2 \sum_{i,j} \mathbb{E}(X_i) \mathbb{E}(X_j) \\ &= \sum_i V(X_i). \end{aligned}$$

Dabei wurde im vorletzten Beweisschritt Satz 4.6.1 ausgenutzt.

(ii) Die Varianzen hängen nur von der Verteilung ab, und mit  $V(cY) = c^2 V(Y)$  folgt

$$\begin{aligned} V\left(\frac{X_1 + \dots + X_n}{n}\right) &= \frac{1}{n^2} \sum_i V(X_i) \\ &= \frac{nV(X)}{n^2} \\ &= \frac{V(X)}{n}. \end{aligned}$$

Und nun ist nur noch die Wurzel zu ziehen.  $\square$

Für zwei unabhängige Zufallsvariable besagt die Formel in Teil (i), dass

$$\sigma^2(X + Y) = \sigma^2(X) + \sigma^2(Y)$$

gilt. Das erinnert formal an den *Satz von Pythagoras*. Der Zusammenhang ist nicht nur oberflächlich. Wirklich kann man Zufallsvariable mit Erwartungswert Null und existierender Varianz als Elemente eines Vektorraums auffassen, in dem ein inneres Produkt  $\langle X, Y \rangle := \mathbb{E}(XY)$  definiert ist<sup>18)</sup>. Aus der Unabhängigkeit folgt dann  $\langle X, Y \rangle = 0$ . In dieser Interpretation entspricht  $\sigma(X)$  der „Länge“ von  $X$ , und Unabhängigkeit impliziert Orthogonalität.

---

<sup>18)</sup>Für diejenigen, die schon eine Vorlesung zur Integrationstheorie oder zur Funktionalanalysis gehört haben: Es geht um den Hilbertraum der quadratintegrierbaren Funktionen auf  $(\Omega, \mathcal{E}, \mathbb{P})$  mit Integral Null.

**Der Zufall verschwindet bei Überlagerung:** Das Wurzel- $n$ -Gesetz kann als „gegenseitige Auslöschung von Zufallseinflüssen“ interpretiert werden. Wir hatten doch Varianz als mittlere Abweichung vom Erwartungswert eingeführt. Da  $(X_1 + \dots + X_n)/n$  den gleichen Erwartungswert wie  $X$  hat, wissen wir jetzt, dass der Mittelwert von  $n$  unabhängigen  $X$ -Messungen bei großem  $n$  viel weniger von  $\mathbb{E}(X)$  abweicht als eine einmalige Abfrage. Und das kann sogar quantifiziert werden: Wenn etwa die Abweichung auf ein Zehntel fallen soll, muss man zu Mittelwerten von 100 Messungen übergehen.

Es gibt noch ein zweites wichtiges Thema dieses Abschnitts: Was lässt sich über die Verteilung von Summen unabhängiger Zufallsvariablen sagen? Wir werden sehen, dass diese Verteilung dann nur von der Verteilung der Summanden abhängt. (Das ist für allgemeine Zufallsvariable nicht zu erwarten: Auch wenn  $X$  und  $-X$  die gleiche Verteilung haben können, werden sich  $X + X$  und  $X - X$  sehr unterschiedlich verhalten.)

Hier zunächst ein Ergebnis für Situationen, in denen die induzierte Wahrscheinlichkeitsmaße diskret sind:

**Satz 4.6.3.** *Es sei  $(\Omega, \mathcal{E}, \mathbb{P})$  ein Wahrscheinlichkeitsraum, und  $X, Y$  seien unabhängige Zufallsvariable, die jeweils höchstens abzählbar viele Werte annehmen. Dann ist*

$$\mathbb{P}_{X+Y}(\{z\}) = \sum_{x,y \text{ mit } x+y=z} \mathbb{P}_X(\{x\}) \mathbb{P}_Y(\{y\})$$

für alle  $z \in \mathbb{R}$ .

Wir weisen noch auf den Spezialfall hin, dass die  $X(\omega)$  und die  $Y(\omega)$  in  $\mathbb{N}_0$  liegen<sup>19)</sup>. Setzt man in diesem Fall  $p_k = \mathbb{P}(\{X = k\})$  und  $q_l := \mathbb{P}(\{Y = l\})$ , so folgt

$$\mathbb{P}\{X + Y = n\} = p_0 q_n + p_1 q_{n-1} + \dots + p_n q_0$$

für alle  $n \in \mathbb{N}_0$ .

**Beweis:** Das Ergebnis folgt leicht daraus, dass die Ereignisse  $\{X = x\}$  und  $\{Y = y\}$  unabhängig sind:

$$\begin{aligned} \mathbb{P}(X + Y = z) &= \sum_{x,y \text{ mit } X=x, Y=y, x+y=z} \mathbb{P}(\{X = x, Y = y\}) \\ &= \sum_{x,y \text{ mit } X=x, Y=y, x+y=z} \mathbb{P}_X(\{x\}) \mathbb{P}_Y(\{y\}). \end{aligned}$$

(Übrigens sind hier und in der Formulierung des Satzes bei den Summen jeweils nur höchstens abzählbar viele Summanden zu berücksichtigen.)  $\square$

Allgemein nennt man die Folge  $(p_0 q_0, p_0 q_1 + p_1 q_0, p_0 q_2 + p_1 q_1 + p_2 q_0, \dots)$  die *diskrete Faltung* der Folgen  $(p_0, p_1, \dots)$  und  $(q_0, q_1, \dots)$ , auch dann, wenn es nicht um Wahrscheinlichkeitstheorie geht.

<sup>19)</sup>Das tritt typischerweise dann auf, wenn etwas gezählt wird.

### Morgens beim Bäcker

Sie haben Appetit auf Croissants, Sie würden gern drei kaufen. Beim Bäcker sind noch zwei Kunden vor Ihnen, und im Regal sind noch sieben Croissants vorrätig. Wie wahrscheinlich ist es, dass Sie noch die gewünschten drei Stück bekommen?

Dazu muss man natürlich das Croissant-Kaufverhalten der Kunden in dieser Gegend kennen: Die Wahrscheinlichkeiten für den Kauf von 0 (bzw. 1, 2, 3, 4) Croissants sind 0.3 (bzw. 0.2, 0.3, 0.1, 0.1).

Nun kann gerechnet werden:  $X$  und  $Y$  bezeichnen die Anzahl der von den Kunden vor Ihnen gekauften Croissants, und wir wollen wissen, wie groß  $\mathbb{P}(\{X + Y \leq 4\})$  ist (denn im Fall  $X + Y \leq 4$  bleiben für Sie noch drei übrig). Es ist  $p_0 = q_0 = 0.3$ ,  $p_1 = q_1 = 0.2$ ,  $p_2 = q_2 = 0.3$ ,  $p_3 = q_3 = 0.1$ ,  $p_4 = q_4 = 0.1$ , und damit ist

$$\mathbb{P}(\{X + Y = 0\}) = 0.3 \cdot 0.3 = 0.09,$$

$$\mathbb{P}(\{X + Y = 1\}) = 0.3 \cdot 0.2 + 0.2 \cdot 0.3 = 0.12,$$

$$\mathbb{P}(\{X + Y = 2\}) = 0.3 \cdot 0.3 + 0.2 \cdot 0.2 + 0.3 \cdot 0.3 = 0.22,$$

$$\mathbb{P}(\{X + Y = 3\}) = 0.3 \cdot 0.1 + 0.2 \cdot 0.3 + 0.3 \cdot 0.2 + 0.1 \cdot 0.3 = 0.18,$$

$$\mathbb{P}(\{X + Y = 4\}) = 0.3 \cdot 0.1 + 0.2 \cdot 0.1 + 0.3 \cdot 0.3 + 0.1 \cdot 0.2 + 0.1 \cdot 0.3 = 0.19.$$

Das ergibt

$$\mathbb{P}(\{X + Y \leq 4\}) = 0.09 + 0.12 + 0.22 + 0.18 + 0.19 = 0.80.$$

Die Chancen stehen also gar nicht so schlecht!

Wir wollen auch ein Beispiel noch einmal aufgreifen, das in diesem Buch schon mehrfach auftrat: die *Augensumme von zwei Würfeln*. Sind  $X, Y$  die Augenzahlen auf den beiden Würfeln, so ist die Verteilung der Zufallsvariablen  $X + Y$  die diskrete Faltung der Gleichverteilung mit sich selber (wenn wir die Würfel als fair voraussetzen). Man erhält  $\mathbb{P}(\{X + Y = 2\}) = (1/6)(1/6) = 1/36$ ,  $\mathbb{P}(\{X + Y = 3\}) = (1/6)(1/6) + (1/6)(1/6) = 2/36$  usw.

Man kann die Formel in Satz 4.6.3 auch *iterieren*. Sind  $X_1, \dots, X_n$  unabhängige Zufallsvariable, die jeweils höchstens abzählbar viele Werte annehmen und möchte man die Verteilung von  $X_1 + \dots + X_n$  berechnen, so kann man zunächst die Verteilung von  $X_1 + X_2$  bestimmen, das Ergebnis mit der Verteilung von  $X_3$  falten, die neue Verteilung mit der Verteilung von  $X_4$  falten usw.<sup>20)</sup> Das kann schnell unhandlich werden, aber im Fall von nur endlich vielen Bildwerten ist das auch für große  $n$  mit Computerhilfe kein Problem.

---

<sup>20)</sup>An dieser Stelle ist an Satz 4.4.8 zu erinnern. Er garantiert, dass auch  $X_1 + X_2, X_3$  unabhängig sind, der vorstehende Satz also auch für den zweiten Schritt angewendet werden darf. Entsprechend sind  $X_1 + X_2 + X_3, X_4$  unabhängig usw.

Hier ein Beispiel. Wie ist denn die Augensumme von  $n$  fairen Würfeln verteilt? Dazu ist immer wieder mit der Gleichverteilung auf  $\{1, 2, 3, 4, 5, 6\}$  zu falten. In den nachstehenden Bildern ist das Ergebnis für  $n = 2$ ,  $n = 4$ ,  $n = 8$  und  $n = 16$  durch Balkendiagramme dargestellt<sup>21)</sup>:

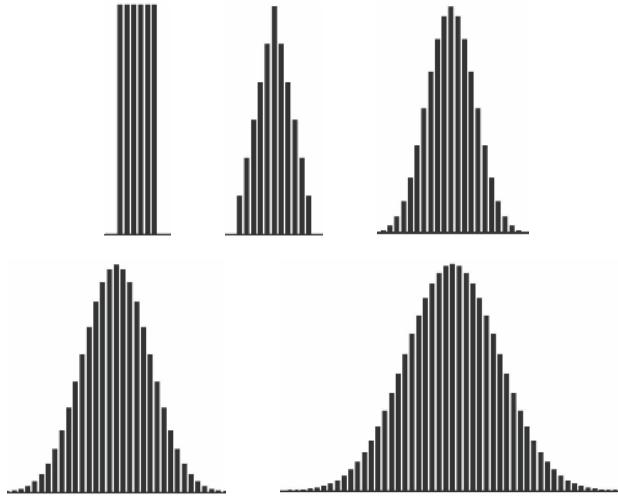


Bild 4.6.1: Der Würfel (Summen von 1, 2, 4, 8, und 16 unabhängigen Kopien).

Die Bilder erinnern an den Graphen der Standardnormalverteilung (vgl. Seite 51). In Abschnitt 9.6 werden wir einsehen, dass das kein Zufall ist, denn eines der überraschendsten Ergebnisse der Wahrscheinlichkeitstheorie besagt, dass beim Aufsummieren von unabhängigen Zufallsvariablen so gut wie immer Verteilungen entstehen, die „so ungefähr“ wie die Normalverteilung aussehen.

Dass die Wahrscheinlichkeiten für Summen aus  $n$  identisch verteilter Zufallsvariablen mit größer werdendem  $n$  immer „Glockenkurven-ähnlicher“ werden, ist ein universelles Phänomen. Hier ist es noch einmal für den Wahrscheinlichkeitsraum  $\Omega = \{0, 1, 2\}$  mit  $\mathbb{P}(\{0\}) = 3/12$ ,  $\mathbb{P}(\{1\}) = 1/12$ ,  $\mathbb{P}(\{2\}) = 8/12$  dargestellt:

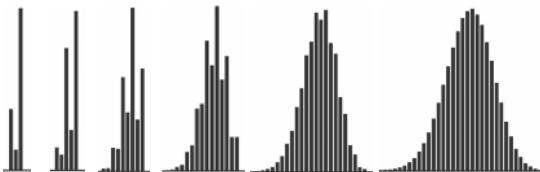


Bild 4.6.2: Summen von 1, 2, 4, 8, 16 und 32 unabhängigen Kopien dieses Raums.

Auch für dieses sehr unsymmetrische Beispiel gibt es unverkennbar so etwas wie eine „Konvergenz gegen die Glockenkurve“, doch ist die „Konvergenz“ deutlich langsamer als im Würfelbeispiel.

<sup>21)</sup> Achtung: Die Bilder der Wahrscheinlichkeiten sind so skaliert, dass die jeweils größten Wahrscheinlichkeiten die gleiche Höhe haben. Die Summe der Balkenhöhen ist also sehr unterschiedlich.

→  
Programm!

Als weiteres Beispiel für diskrete Faltungen behandeln wir die *Poissonverteilung*. Es sei  $X$  poissonverteilt zum Parameter  $\lambda$  und  $Y$  poissonverteilt zum Parameter  $\mu$ , und  $X, Y$  seien unabhängig. Dann folgt für  $n \in \mathbb{N}_0$ :

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

$X + Y$  ist also wieder poissonverteilt, der neue Parameter ist  $\lambda + \mu$ .

Das entsprechende Ergebnis für Räume mit Dichten ist aufwändiger zu formulieren<sup>22)</sup>. Wir beginnen mit einer Vorbereitung aus der Analysis:

**Definition 4.6.4.** Es seien  $f, g : \mathbb{R} \rightarrow [0, \infty[$  Wahrscheinlichkeitsdichten. Dann definieren wir eine neue Funktion  $f * g : \mathbb{R} \rightarrow [0, \infty[$  („ $f$  gefaltet mit  $g$ “) durch

$$(f * g)(x) := \int_{-\infty}^{\infty} f(x-t)g(t) dt.$$

Diese Funktion heißt die Faltung von  $f$  und  $g$ .

Sind  $f$  und/oder  $g$  nur auf einem Teilintervall  $\Omega$  von  $\mathbb{R}$  definiert, so werden die Dichten außerhalb von  $\Omega$  durch die Nullfunktion fortgesetzt, bevor man  $f * g$  berechnet.

### Bemerkungen und Beispiele:

1. Es gibt Dichten  $f, g$ , bei denen  $(f * g)(x)$  für manche  $x$  nicht definiert ist, weil das Integral unendlich wird. In unseren Beispielen sind die Dichten aber immer beschränkt. Dann ist der Integrand durch ein Vielfaches einer Dichtefunktion abschätzbar und folglich integrierbar.

Analytische Feinheiten werden bei unseren Untersuchungen keine Rolle spielen, denn die auftretenden Funktionen werden (wenigstens stückweise) stetig sein.

---

<sup>22)</sup>Auf den Fall beliebiger Wahrscheinlichkeitsräume werden wir in diesem Buch nicht eingehen, der technische Aufwand wäre zu groß.

**2.** Durch die Substitution  $t \mapsto x - t$  wird klar, dass stets  $(f * g)(x)$  gleich  $(g * f)(x)$  ist. Und wenn man bei der Definition allgemeinere Funktionen als nur Dichtefunktionen zulässt, sind auch andere Eigenschaften der Faltung schnell einzusehen:

$$f * (g_1 + g_2) = f * g_1 + f * g_2, (cf) * g = c(f * g) \text{ usw.}$$

**3.** In vielen Fällen muss bei der Berechnung von  $(f * g)(x)$  nur über ein beschränktes Intervall integriert werden. Ist nämlich  $f$  außerhalb von  $[a, b]$  und  $g$  außerhalb von  $[c, d]$  gleich Null und ist  $x \in \mathbb{R}$ , so verschwindet  $f(t)g(x - t)$  für alle  $t$ , die nicht in  $[a, b] \cap [x - d, x - c]$  liegen. Anders ausgedrückt: Es muss nur über das (möglicherweise leere) Intervall  $[a, b] \cap [x - d, x - c]$  integriert werden.

Dazu ein Beispiel, wir wollen die Gleichverteilung auf  $[0, 1]$  mit sich selber falten. Die Funktion  $f(t)g(x - t)$  ist Eins für  $t \in [0, 1] \cap [x - 1, x]$  und Null sonst.  $(f * g)(x)$  ist also die Länge von  $[0, 1] \cap [x - 1, x]$ :

$$(f * g)(x) = \begin{cases} 0 & : x \leq 0 \\ x & : x \in [0, 1] \\ 2 - x & : x \in [1, 2] \\ 0 & : x \geq 2. \end{cases}$$

**4.** Die konkrete Berechnung von  $f * g$  ist meist recht mühsam. Nur selten treten geschlossen auswertbare Integrale auf, in vielen Fällen muss man mit der numerischen Auswertung zufrieden sein.

Nach diesem Analysis-Exkurs geht es nun wieder um Wahrscheinlichkeits-theorie. Das Analogon zu Satz 4.6.3 für Räume mit Dichten steht in

**Satz 4.6.5.**  *$X, Y$  seien unabhängige reellwertige Zufallsvariable auf dem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{E}, \mathbb{P})$ , und für  $\mathbb{P}_X$  bzw.  $\mathbb{P}_Y$  sollen Dichtefunktionen  $f$  bzw.  $g$  existieren. Dann ist  $f * g$  Dichtefunktion für  $\mathbb{P}_{X+Y}$ .*

**Beweis:** Es ist doch zu zeigen, dass für beliebige Intervalle  $[a, b]$  die Formel

$$\mathbb{P}(\{X + Y \in [a, b]\}) = \int_a^b (f * g)(x) dx$$

gilt. Zum Beweis kombinieren wir die folgenden Tatsachen:

- Die gemeinsame Verteilung von  $X$  und  $Y$ , also das Maß  $\mathbb{P}_{(X,Y)}$ , hat auf  $\mathbb{R}^2$  die Dichtefunktion  $(x, y) \mapsto \Phi(x, y) := f(x)g(y)$  (Satz 4.4.4).
- $(X + Y)(\omega)$  liegt (trivialerweise) genau dann in  $[a, b]$ , wenn  $(X(\omega), Y(\omega))$  zu  $\Delta_{a,b} := \{(s, t) \mid a \leq s + t \leq b\}$  gehört<sup>23)</sup>. Deswegen ist

$$\mathbb{P}(\{X + Y \in [a, b]\}) = \int \int_{\Delta_{a,b}} f(x)g(y) dy dx.$$

---

<sup>23)</sup>  $\Delta_{a,b}$  ist der Streifen in der Ebene zwischen den Geraden  $t \mapsto a - t$  und  $t \mapsto b - t$ .

- Den Transformationssatz für Gebietsintegrale (vgl. den Anhang zur Analysis, Seite 361).

Wir wenden den Transformationssatz so an, dass bei der iterativen Integration die Faltung von  $f$  und  $g$  auftritt. Dazu muss  $\Delta_{a,b}$  geschickt parametrisiert werden. Wir definieren  $\Delta'_{a,b} := \{(s,t) \mid s \in \mathbb{R}, t \in [a,b]\}$  und  $\Psi : \Delta'_{a,b} \rightarrow \Delta_{a,b}$  durch  $(s,t) \mapsto (s, t-s)$ . Das ist eine Bijektion, und die Jacobi-Determinante ist Eins. Wegen des Transformationssatzes ist dann

$$\begin{aligned}\mathbb{P}(\{X + Y \in [a,b]\}) &= \int \int_{\Delta_{a,b}} f(x)g(y) dy dx \\ &= \int \int_{\Delta'_{a,b}} \Phi \circ \Psi(s,t) ds dt \\ &= \int \int_{\Delta'_{a,b}} f(s)g(t-s) ds dt \\ &= \int_a^b \left( \int_{-\infty}^{\infty} f(s)g(t-s) ds \right) dt \\ &= \int_a^b (f * g)(t) dt.\end{aligned}$$

Damit ist der Satz bewiesen.  $\square$

Wie im Fall diskreter Situationen kann auch der vorstehende Satz mehrfach angewendet werden, um eine Dichtefunktion von  $X_1 + \dots + X_n$  für unabhängige Zufallsvariable zu ermitteln, bei denen die  $\mathbb{P}_{X_i}$  eine Dichte haben. Nachfolgend sind als Beispiel die Dichten einiger Summen von unabhängigen Gleichverteilungen in  $[0, 1]$  skizziert. Die Dichte von zwei Summanden haben wir ja schon vor dem Satz berechnet, hier sieht man außerdem noch die Dichten für 4, 8 und 16 Summanden:

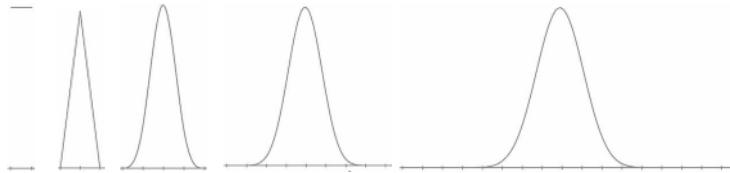


Bild 4.6.3: Dichte der Summen von 1, 2, 4, 8 und 16 unabhängigen Kopien der Gleichverteilung.

→ Es ist wieder auffällig, dass die Dichten der Dichte einer Normalverteilung stark ähneln. Mehr dazu findet man in Abschnitt 9.4.

## 4.7 Verständnisfragen

### Zu Abschnitt 4.1

*Sachfragen*

**S1:** Wie ist die bedingte Wahrscheinlichkeit  $\mathbb{P}(A | B)$  definiert?

**S2:** Was ist ein Wahrscheinlichkeitsbaum?

**S3:** Was bedeutet die Aussage, dass die Ereignisse  $A, B$  unabhängig sind? Was heißt es erstens inhaltlich, und wie lautet die zugehörige Formel?

*Methodenfragen*

**M1:** Bedingte Wahrscheinlichkeiten berechnen können.

**M2:** Wahrscheinlichkeitsräume mit Hilfe eines Wahrscheinlichkeitsbaums analysieren können.

**M3:** Feststellen können, ob zwei Ereignisse unabhängig sind.

**M4:** Elementare Beweise im Zusammenhang mit bedingten Wahrscheinlichkeiten und Unabhängigkeit führen können.

### Zu Abschnitt 4.2

*Sachfragen*

**S1:** Was besagt der „Satz von der totalen Wahrscheinlichkeit“?

**S2:** Wie lautet die Bayes-Formel?

**S3:** Welches paradoxe Phänomen gibt es, wenn man den Satz von Bayes für die Frage anwendet, ob man sich bei einem positiven Test auf eine Krankheit gleich große Sorgen machen sollte.

**S4:** Worum geht es beim Ziegenproblem?

*Methodenfragen*

**M1:** Den Satz von der totalen Wahrscheinlichkeit anwenden können.

**M2:** Ergebnisse rund um die Bayes-Formel beweisen können.

### Zu Abschnitt 4.3

*Sachfragen*

**S1:** Wann sagt man, dass  $A_1, \dots, A_n$  unabhängig sind? Was heißt es inhaltlich, wie lauten die Formeln?

**S2:** Wie verhalten sich die Aussagen „ $A_1, \dots, A_n$  sind unabhängig“ und „Die  $A_1, \dots, A_n$  sind paarweise unabhängig“ zueinander?

**S3:** Wie verhalten sich die Aussagen „ $A_1, \dots, A_n$  sind unabhängig“ und „Es gilt  $\mathbb{P}(A_1 \cap \dots \cap A_n) = \mathbb{P}(A_1) \cdots \mathbb{P}(A_n)$ “ zueinander?

*Methodenfragen*

**M1:** Die Unabhängigkeit einer Familie von Ereignissen nachweisen können.

**Zu Abschnitt 4.4***Sachfragen***S1:** Was bedeutet es, dass Zufallsvariable  $X_1, \dots, X_n$  unabhängig sind?**S2:** Wie kann man das nachprüfen?**S3:** Wie ist Unabhängigkeit von  $X, Y$  durch die Dichtefunktionen von  $X, Y$  und  $(X, Y)$  charakterisierbar?*Methodenfragen***M1:** Unabhängigkeit für Familien von Zufallsvariablen nachprüfen können.**M2:** Einfache Beweise zum Thema „Unabhängigkeit“ führen können.**Zu Abschnitt 4.5***Sachfragen***S1:** Was besagt der Klonsatz?*Methodenfragen***M1:** Die Wahrscheinlichkeiten „einfacher“ Ereignisse im unendlichen Produktraum  $(\Omega_\infty, \mathcal{E}_\infty, \mathbb{P}_\infty)$  berechnen können.**Zu Abschnitt 4.6***Sachfragen***S1:** Was weiß man über den Erwartungswert eines Produkts von Zufallsvariablen im Fall der Unabhängigkeit?**S2:** Was folgt daraus für die Varianz unabhängiger Summen?**S3:** Was ist das Wurzel- $n$ -Gesetz?**S4:** Was versteht man unter der diskreten Faltung von zwei Wahrscheinlichkeitsmaßen auf  $\mathbb{N}_0$ ?**S5:** Wie ist die Faltung zweier Funktionen definiert und welche Rolle spielt diese Definition, wenn man die Dichte einer Summe von zwei Wahrscheinlichkeitsmaßen ausrechnen möchte?*Methodenfragen***M1:** Die diskrete Falteung berechnen können.**M2:** Die Faltung von Dichten bestimmen können.

## 4.8 Übungsaufgaben

**Zu Abschnitt 4.1****Ü4.1.1** Das Intervall  $[0, 1]$  sei mit der Gleichverteilung versehen. Finden Sie alle Intervalle  $[a, b]$ , so dass  $[a, b]$  und  $[0, 0.5]$  unabhängig sind.**Ü4.1.2** Wir betrachten die Poissonverteilung auf  $\{0, 1, \dots\}$  zum Parameter  $\lambda$ . Bestimmen Sie  $\mathbb{P}(A | B)$  für  $A = \{2, 3, 4\}$ ,  $B = \{3, 4, 5\}$ .

**Ü4.1.3** Bestimmen Sie eine stetige Dichte  $f$  auf  $[0, 1]$  so, dass  $[0, 0.5]$  und  $[0.4, 0.8]$  unabhängig sind.

**Ü4.1.4** Aus einem vollständigen Skatspiel wurde das Kreuz As entfernt. Die restlichen Karten seien mit der Gleichverteilung versehen.

a) Berechnen Sie die bedingte Wahrscheinlichkeit  $\mathbb{P}(\text{König} \mid \text{Kreuz})$ .

b) Zeigen Sie, dass es nur auf triviale Weise möglich ist, zwei unabhängige Ereignisse  $A, B$  zu finden. (Das soll bedeuten:  $A, B$  sind genau dann unabhängig, wenn  $A$  oder  $B$  die leere Menge oder der ganze Raum ist.)

**Ü4.1.5** In einem Kasten befinden sich 4 weiße, 6 rote und 10 grüne Kugeln. Es wird zweimal ohne Zurücklegen gezogen. Bestimmen Sie mit Hilfe eines Wahrscheinlichkeitsbaumes die folgenden Wahrscheinlichkeiten:

- a) Die Wahrscheinlichkeit, dass zwei gleichfarbige Kugeln gezogen wurden.
- b) Die Wahrscheinlichkeit, dass keine grüne Kugel gezogen wurde.

**Ü4.1.6** Die Menge  $\{1, \dots, n\}$  sei mit der Gleichverteilung versehen, dabei sei  $n > 1$ . Zeigen Sie:  $n$  ist genau dann eine Primzahl, wenn aus „ $A, B$  Ereignisse,  $A, B$  unabhängig“ stets folgt, dass eine der Mengen leer oder gleich  $\{1, \dots, n\}$  ist.

**Ü4.1.7** Herr A. hat seine Bachelorarbeit geschrieben. Er bittet zwei Freunde darum, nach Rechtschreibfehlern zu suchen. Sie lesen die Arbeit unabhängig voneinander. Der erste findet  $k_1$ , der andere  $k_2$  Fehler, dabei gibt es  $k$  Fehler, die beide angestrichen haben. Ermitteln Sie daraus einen Schätzwert für die wirkliche Fehleranzahl  $k_0$ .

Tipp: Sei  $k_0$  die Anzahl der wirklichen Fehler. Wenn dann der erste Freund Fehler mit Wahrscheinlichkeit  $p_1$  findet, so sollte  $k_1 \approx k_0 p_1$  gelten. Und so weiter. (Die Unabhängigkeit ist auch noch zu berücksichtigen: Wie wahrscheinlich ist es, dass beide den gleichen Fehler finden?) So erhält man drei Gleichungen für  $p_1, p_2, k_0$ , und damit ist  $k_0$  leicht zu bestimmen.

## Zu Abschnitt 4.2

**Ü4.2.1** Drei Kästen  $K_1, K_2, K_3$  enthalten gut durchmischt schwarze und weiße Kugeln. Es enthalte

$K_1$ : 2 schwarze und 4 weiße Kugeln;

$K_2$ : 3 schwarze und 5 weiße Kugeln;

$K_3$ : 1 schwarze und 3 weiße Kugeln.

a) Aus Kasten  $K_3$  wird dreimal mit Zurücklegen gezogen. Wie groß ist die Wahrscheinlichkeit, dass die erste Kugel weiß, die zweite schwarz und die dritte wieder weiß ist?

b) Nun wird zunächst einer der Kästen zufällig ausgewählt (jeder mit Wahrscheinlichkeit  $1/3$ ), aus dem dann einmal gezogen wird. Mit welcher Wahrscheinlichkeit liefert das eine weiße Kugel?

c) Wenn eine Ziehung wie in „b“ eine weiße Kugel liefert, mit welcher Wahrscheinlichkeit wurde dann im ersten Schritt Kasten  $K_2$  gewählt?

**Ü4.2.2** In einem Laden ist eine Alarmanlage eingebaut, die im Falle eines Einbruchs mit Wahrscheinlichkeit 0.99 die Polizei alarmiert. In einer Nacht ohne Einbruch wird mit Wahrscheinlichkeit 0.002 Alarm ausgelöst. (Eine Maus berührt die Anlage o. ä.) Die Einbruchswahrscheinlichkeit für eine Nacht ist 0.0005. Die Anlage hat gerade Alarm gegeben. Man berechne die Wahrscheinlichkeit, dass gerade ein Einbruch stattfindet.

**Ü4.2.3** Es geht um das Problem auf Seite 128 („Erfolg macht sicher“). Dort wurden die Kästen gleichverteilt ausgesucht, d. h. der  $i$ -te Kasten wurde mit Wahrscheinlichkeit  $1/n$  gewählt.

a) Diskutieren Sie die Variante, bei der der  $i$ -te Kasten mit Wahrscheinlichkeit  $2i/[n(n+1)]$  gewählt wird ( $i = 1, \dots, n$ ); es werden also mit größerer Wahrscheinlichkeit Kästen mit „vielen“ roten Kugeln ausgewählt.

Wie ist in diesem Fall die bedingte Wahrscheinlichkeit, nach  $k$  Erfolgen noch einen weiteren zu haben? Geben Sie die exakte Lösung an sowie eine geeignete Näherung mit Hilfe einer Integration.

b) Diesmal sei die Wahrscheinlichkeit für den  $i$ -ten Kasten  $2(n-i+1)/[n(n+1)]$ . Bestimmen Sie auch für diese Situation die bedingte Wahrscheinlichkeit, dass es nach  $k$  Erfolgen noch einen weiteren gibt (Formel und Integralapproximation).

**Ü4.2.4** Diskutieren Sie das Ziegenproblem in der Variante mit  $k$  Türen, wobei  $k > 2$ : Es gibt einen Gewinn, der Kandidat wählt, der Quizmaster öffnet  $k - 2$  Türen, und der Kandidat darf noch einmal wechseln. Würden dadurch seine Gewinnchancen steigen? Und wenn ja, um wie viel?

**Ü4.2.5** Falls Sie am Sonntagvormittag das Radio bei einem zufällig gewählten Sender einstellen, erklingt mit 20 Prozent Wahrscheinlichkeit Orgelmusik, an den anderen Vormittagen nur mit 2 Prozent Wahrscheinlichkeit.

Nun hatten Sie in den Semesterferien ein etwas unstrukturiertes Leben, die Tage der Woche waren alle gleichberechtigt. An irgendeinem Vormittag wachten Sie auf, und beim Radio-Einschalten – der Sender wurde zufällig gewählt – erklang Orgelmusik. Mit welcher Wahrscheinlichkeit war das ein Sonntag?

**Ü4.2.6** Hier geht es um das Umkehren bedingter Wahrscheinlichkeiten. Es sei  $A$  das Ereignis „der Patient hat die Krankheit  $K$ “, und  $B$  bedeutet, dass ein Test auf  $K$  positiv verlaufen ist. Finden Sie – mit Begründung – eine Formel für  $P(\neg A|\neg B)$ , also die Wahrscheinlichkeit dafür, dass der Patient  $K$  nicht hat, wenn das Testergebnis negativ ist. Diese Formel soll die Zahlen  $P(A)$ ,  $P(B|A)$  und  $P(B|\neg A)$  enthalten.

Berechnen Sie danach  $P(\neg A|\neg B)$  für die konkreten Werte

$$P(A) = 0.04, \quad P(B|A) = 0.99, \quad P(B|\neg A) = 0.05.$$

**Ü4.2.7** (Das diskrete Interview): In einer Urne sind 1000 Zettel mit Fragen. Auf 500 Zetteln steht „Ist die letzte Ziffer Ihrer Personalausweisnummer gerade?“, auf den anderen 500 eine Frage  $F$ , die mit „ja“ oder „nein“ zu beantworten ist, auf die aber vor Zeugen nur ungern eine ehrliche Antwort gegeben wird (Beispiele sollten jedem selbst einfallen). Um zu erfahren, mit welcher Wahrscheinlichkeit  $p$  die Interviewten  $F$  mit „ja“ beantworten, wird so verfahren:

Der Interviewte zieht einen Zettel und antwortet; dabei weiß der Interviewer nicht, welcher Zettel gezogen wurde. Wie kann man nun  $p$  schätzen, wenn viele Interviews geführt wurden?

### Zu Abschnitt 4.3

**Ü4.3.1**  $\Omega := \{1, \dots, n\}^m$  sei mit der Gleichverteilung versehen. Wir definieren  $E_{k,l} \subset \Omega$  durch  $E_{k,l} := \{(x_1, \dots, x_m) \mid x_k = l\}$  für  $k = 1, \dots, m$  und  $l = 1, \dots, n$ . Was ist die maximale Anzahl, die Sie aus diesen  $m \cdot n$  Mengen wählen können, um unabhängige Ereignisse zu erhalten?

**Ü4.3.2** Seien  $A, B, C$  Ereignisse (in irgendeinem Wahrscheinlichkeitsraum). Falls  $A$  und  $B$  unabhängig sind, so folgt aus  $A \cap B \subset C \subset A \cup B$ , dass  $\mathbb{P}(A \cap C) \geq \mathbb{P}(A)\mathbb{P}(C)$  gilt.

**Ü4.3.3** Es sei  $(\Omega, \mathcal{E}, \mathbb{P})$  ein Wahrscheinlichkeitsraum, und  $A, B$  seien unabhängige Ereignisse. Muss dann  $\{E \in \mathcal{E} \mid A, B, E \text{ sind unabhängig}\}$  eine Teil- $\sigma$ -Algebra von  $\mathcal{E}$  sein?

**Ü4.3.4**  $A, B, C$  seien Ereignisse. Man weiß, dass sie unabhängig sind und man kennt die Wahrscheinlichkeiten. Kann man aus diesen Informationen die Wahrscheinlichkeit von  $A \cup B \cup C$  ermitteln?

### Zu Abschnitt 4.4

**Ü4.4.1** Beweisen Sie: Sind  $X_1, \dots, X_n$  reellwertige unabhängige Zufallsvariable, so sind auch  $Y, X_1, \dots, X_n$  unabhängig, wenn  $Y$  konstant ist.

**Ü4.4.2** ( $X_n$ ) sei eine Folge unabhängiger identisch verteilter Zufallsvariablen. Wie groß ist die Wahrscheinlichkeit, dass die Folge  $(X_n)$  von irgendeiner Stelle an monoton fallend ist?

**Ü4.4.3** Mit  $\Omega$  wie in Aufgabe 4.3.1 definieren wir  $X_k : \Omega \rightarrow \mathbb{R}$  durch  $(x_1, \dots, x_m) \mapsto x_k$  für  $k = 1, \dots, m$ . Zeigen Sie, dass die  $X_1, \dots, X_m$  unabhängig sind.

**Ü4.4.4**  $X_1, X_2, \dots : \Omega \rightarrow \mathbb{R}$  seien unabhängige Zufallsvariable, und alle  $P_{X_k}$  seien die Gleichverteilung auf  $[0, 1]$ . Beweisen Sie, dass das Ereignis

$$\{\omega \mid \text{für alle } k \text{ ist } X_k(\omega) \leq 0.999\}$$

Wahrscheinlichkeit Null hat.

**Ü4.4.5**  $X_1, \dots, X_n$  seien reellwertige Zufallsvariable. Zeigen Sie, dass Sie genau dann unabhängig sind, wenn

$$\mathbb{P}(\{X_1 \in B_1, \dots, X_n \in B_n\}) = \mathbb{P}(\{X_1 \in B_1\}) \cdots \mathbb{P}(\{X_n \in B_n\})$$

für beliebige Borelmengen  $B_1, \dots, B_n$  gilt.

Es ist also nicht erforderlich, die entsprechende Gleichung für Indizes  $1 \leq i_1 < i_2 \dots < i_k \leq n$  nachzuprüfen (vgl. „Irrtum 2“ auf Seite 133).

### Zu Abschnitt 4.5

**Ü4.5.1** Mit den Bezeichnungen von Satz 4.5.1 gilt:

- a)  $\{(\omega_1, \omega_2, \dots) \mid X_n(\omega_n) < n \text{ für alle } n\}$  gehört zu  $\mathcal{E}_\infty$ .

b)  $\{(\omega_1, \omega_2, \dots) \mid \left( (X_1(\omega_1) + X_n(\omega_n))/n \right)_n \text{ ist konvergent}\}$  gehört ebenfalls zur  $\sigma$ -Algebra  $\mathcal{E}_\infty$ .

**Ü4.5.2** Wir verwenden wieder die Bezeichnungen von Satz 4.5.1. Es sei  $\mathbb{P}_X$  die Gleichverteilung in  $\{1, \dots, 6\}$ , es geht also um die Modellierung einer Folge von Würfelwürfen. Berechnen Sie die Wahrscheinlichkeit von

$$\{(X_1 \geq 2) \text{ und } ((X_4 = 6) \text{ oder } (X_5 \leq 3))\}.$$

### Zu Abschnitt 4.6

**Ü4.6.1** Es seien  $X, Y : \Omega \rightarrow \mathbb{R}$  Zufallsvariable, für die der Erwartungswert existiert. Dann kann es sein, dass das Produkt  $XY$  keinen Erwartungswert hat.

**Ü4.6.2** Seien  $X_1, X_2, X_3$  unabhängige auf  $[0, 1]$  gleichverteilte Zufallsvariable. Wie ist dann  $X_1 + X_2 + X_3$  verteilt? (Es soll die Gleichung der Dichtefunktion ermittelt werden.)

**Ü4.6.3**  $X, Y$  seien unabhängige, auf dem gleichen Wahrscheinlichkeitsraum definierte reellwertige Zufallsvariable. Die zugehörigen Verteilungsfunktionen  $F_X, F_Y$  mögen stetig differenzierbar sein; wir wissen schon, dass dann die Ableitungen gerade die Dichtefunktionen zu  $P_X, P_Y$  sind.

a) Zeigen Sie, dass  $Z := \max\{X, Y\}$  ebenfalls eine Dichtefunktion hat. Drücken Sie dazu die Verteilungsfunktion von  $Z$  durch  $F_X$  und  $F_Y$  aus und bestimmen Sie daraus durch Ableiten die Dichtefunktion von  $Z$ . Als Beispiel soll berechnet werden, nach welcher Dichtefunktion

`max(random,random)` (also das Maximum zweier unabhängigen Abfragen des Zufallsgenerators) verteilt ist.

b) Das gleiche Problem, diesmal für das Minimum.

**Ü4.6.4** Es gibt Zufallsvariable  $X, Y$  mit existierendem Erwartungswert, die *nicht* unabhängig sind, so dass die Erwartungswerte von  $X, Y, XY$  existieren und  $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$  gilt. (Unabhängigkeit ist also keine notwendige Bedingung für diese Gleichheit.)

## Teil III

# Binomial- und Exponentialverteilung

# Kapitel 5

## Die Binomialverteilung

Im dritten Teil dieses Buches werden zwei wichtige konkrete Klassen von Wahrscheinlichkeitsräumen etwas ausführlicher behandelt. Wir beginnen mit der Binomialverteilung.

Der einfachste (nichttriviale) Wahrscheinlichkeitsraum ist der Bernoulliraum, bei dem  $\Omega = \{0, 1\}$  ist. Dabei wird die 1 als „Erfolg“ interpretiert, und die Zahl  $p = \mathbb{P}(\{1\})$  heißt die *Erfolgswahrscheinlichkeit*.

Die Frage, mit welchen Wahrscheinlichkeiten bei  $n$  unabhängigen Abfragen 0 bzw. 1 bzw.  $\dots$   $n$  Erfolge zu erwarten sind, führt auf die *Binomialverteilung*. Sie wird in *Abschnitt 5.1* eingeführt. Danach zeigen wir in *Abschnitt 5.2*, dass die *hypergeometrische Verteilung* unter gewissen Voraussetzungen durch die Binomialverteilung approximiert werden kann. In *Abschnitt 5.3* wird der Zusammenhang zur *Poissonverteilung* beschrieben. Sie war in *Abschnitt 2.1* als Beispiel für einen diskreten Wahrscheinlichkeitsraum eingeführt worden. Nun bekommt sie auch eine inhaltliche Bedeutung als Verteilung, durch die man die unabhängige Überlagerung seltener Ereignisse beschreiben kann.

Am Ende des letzten Kapitels haben wir an Beispielen gesehen, dass die Verteilungen, die beim Aufsummieren unabhängiger Zufallsvariablen entstehen, der Normalverteilung ähneln können. Für die Binomialverteilung gibt es den klassischen *Satz von de Moivre-Laplace*, durch den der Zusammenhang zur „Glockenkurve“ genauer beschrieben wird. Diesen Satz beweisen wir in *Abschnitt 5.4*. Verständnisfragen und Übungsaufgaben findet man danach in *Abschnitt 5.5* und *Abschnitt 5.6*.

## 5.1 Binomialverteilung: Definition

Es sei  $n \in \mathbb{N}$  und  $p \in [0, 1]$ . Wir konstruieren einen Wahrscheinlichkeitsraum  $(\Omega_n, \mathcal{P}(\Omega_n), \mathbb{P}_n)$  und unabhängige Zufallsvariable  $X_1, \dots, X_n : \Omega_n \rightarrow \{0, 1\}$ , so dass  $\mathbb{P}(\{X_i = 1\}) = p$  (und folglich  $\mathbb{P}(\{X_i = 0\}) = 1 - p$ ) für alle  $i$  gilt. Anders ausgedrückt heißt das, dass die  $X_i$  unabhängige Kopien einer bernoulliverteilten Zufallsvariablen  $X$  mit Erfolgswahrscheinlichkeit  $p$  sind.

Einen solchen Raum  $(\Omega, \mathcal{E}, \mathbb{P})$  könnte man sich mit Hilfe des „Klonsatzes“ 4.5.1 verschaffen. In dieser einfachen Situation ist aber auch eine explizite Konstruktion möglich: Setze  $\Omega_n := \{0, 1\}^n$ , definiere  $X_i(x_1, \dots, x_n) := x_i$  und erkläre das Wahrscheinlichkeitsmaß auf den einpunktigen Mengen  $\{(x_1, \dots, x_n)\}$  durch

$$p^{x_1 + \dots + x_n} (1 - p)^{n - x_1 - \dots - x_n}.$$

(„ $p$  hoch Anzahl der Einsen in  $(x_1, \dots, x_n)$ “ ist mit „ $1 - p$  hoch Anzahl der Nullen in  $(x_1, \dots, x_n)$ “ zu multiplizieren.)

Dann hat  $X_1 + \dots + X_n$  eine besondere Bedeutung, diese Zufallsvariable zählt die Anzahl der Erfolge. Wie wahrscheinlich sind  $k$  Erfolge, d.h. wie groß ist  $\mathbb{P}_n(\{X_1 + \dots + X_n = k\})$ ? Diese Zahl kann leicht berechnet werden:

**Satz 5.1.1.** *Die Wahrscheinlichkeit für genau  $k$  Erfolge bei  $n$  unabhängigen Versuchen mit Erfolgswahrscheinlichkeit  $p$  ist gleich*

$$b(k, n; p) := \binom{n}{k} p^k (1 - p)^{n - k} \quad (k = 0, 1, \dots, n).$$

**Beweis:** Sei  $k \in \{0, \dots, n\}$ . Wir fragen doch nach  $\mathbb{P}_n(E_k)$ , wobei  $E_k$  die Menge derjenigen Elemente aus  $\Omega_n$  ist, die genau  $k$  Einsen enthalten. Für jedes  $(x_1, \dots, x_n) \in E_k$  ist nach Definition  $\mathbb{P}_n(\{(x_1, \dots, x_n)\}) = p^k (1 - p)^{n - k}$ , und  $E_k$  hat  $\binom{n}{k}$  Elemente, da es  $\binom{n}{k}$  Teilmengen von  $\{1, \dots, n\}$  gibt, die genau  $k$  Elemente enthalten. Es folgt  $\mathbb{P}_n(E_k) = b(k, n; p)$  wie behauptet.  $\square$

**Definition 5.1.2.** *Das durch die  $b(\cdot, n; p)$  auf  $\{0, \dots, n\}$  definierte Wahrscheinlichkeitsmaß heißt die Binomialverteilung.*

Hier einige typische Beispiele:

**1.** Wie wahrscheinlich ist es, dass man genau zwei Mal drei Richtige hat, wenn man fünf Mal Lotto spielt? Die Antwort: Die Erfolgswahrscheinlichkeit für drei Richtige ist  $0.018\dots$  (vgl. Seite 99), folglich ist die gesuchte Wahrscheinlichkeit gleich  $b(2, 5; 0.018\dots) \approx 0.0031$ .

**2.** Werden bei fünfmaligem Würfeln mindestens drei Sechsen dabei sein? Die Wahrscheinlichkeit dafür ist

$$b(3, 5; 1/6) + b(4, 5; 1/6) + b(5, 5; 1/6) \approx 0.035.$$

Die  $b(k, n; p)$  sollen nun etwas genauer analysiert werden. Es ist zum Beispiel plausibel, dass diese Wahrscheinlichkeiten  $b(k, n; p)$  für solche  $k$  am größten sind,

für die  $k/n$  nahe bei  $p$  liegt. Ist zum Beispiel  $p = 0.5$ , so sollte die Anzahl der Erfolge nahe bei  $n/2$  liegen.

Um das nachzuprüfen, nutzen wir wieder die Idee, die wir auch bei der Diskussion der hypergeometrischen Verteilung in Abschnitt 3.5 erfolgreich angewendet haben. Dabei beschränken wir uns auf den Fall  $0 < p < 1$ , denn für  $p = 0$  und  $p = 1$  ist alles klar.

Wir berechnen  $b(k, n; p)/b(k + 1, n; p)$ , dieser Quotient hat den Wert  $(k + 1)(1 - p)/((n - k)p)$ . Er ist folglich kleiner (bzw. größer) als Eins, wenn  $(k + 1)(n + 1) < p$  (bzw.  $> p$ ) gilt. Das bedeutet, dass die  $b(k, n; p)$  mit wachsendem  $k$  zunächst steigen und danach wieder fallen. Das Maximum ist bei  $[p \cdot (n + 1)]$  zu erwarten<sup>1)</sup>. Dabei kann es vorkommen, dass  $[p \cdot (n + 1)]$  gleich 0 oder gleich  $n$  ist: Im ersten Fall sind die  $b(k, n; p)$  monoton fallend, im zweiten monoton steigend.

Hier einige Skizzen:

→  
Programm!

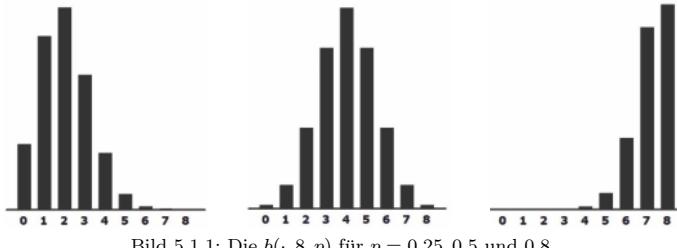


Bild 5.1.1: Die  $b(\cdot, 8, p)$  für  $p = 0.25, 0.5$  und  $0.8$ .

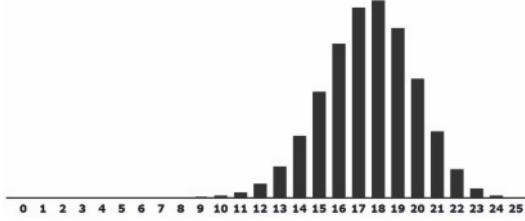


Bild 5.1.2: Die  $b(\cdot, 25, 0.7)$ .

Wieder wird man an die Normalverteilung erinnert. Eine genauere Diskussion dieses Phänomens findet man in Abschnitt 5.4.

Wir berechnen noch *Erwartungswert und Streuung*:

**Satz 5.1.3.** Es sei  $n \in \mathbb{N}$  und  $p \in [0, 1]$ .

- (i) Der Erwartungswert der zugehörigen Binomialverteilung ist  $np$ .
- (ii) Die Varianz ist  $np(1 - p)$ , die Streuung ist folglich  $\sqrt{np(1 - p)}$ .

<sup>1)</sup>Zur Erinnerung:  $[x]$  ist die größte ganze Zahl, die kleiner oder gleich  $x$  ist.

**Beweis:** (i) Direktes Einsetzen würde zu dem etwas unhandlichen Ausdruck  $\sum_{k=0}^n k \cdot b(k, n; p)$  führen. Eleganter geht es mit der Überlegung, dass die Binomialverteilung doch die Verteilung von  $X_1 + \dots + X_n$  ist (Bezeichnungen wie zu Beginn des Kapitels). Jedes  $X_i$  ist aber bernoulliverteilt zur Erfolgswahrscheinlichkeit  $p$ , der Erwartungswert ist also gleich  $p$ . Aus der Additivität des Erwartungswertes folgt dann die Behauptung.

(ii) Auch hier arbeiten wir mit  $X_1 + \dots + X_n$ , und diesmal ist noch an Satz 4.6.2 zu erinnern: Danach sind die Varianzen bei unabhängigen Zufallsvariablen additiv. Und da eine Bernoulliverteilung Varianz  $p(1 - p)$  hat, ist schon alles gezeigt.  $\square$

Der Ausgangspunkt, der zur Binomialverteilung führte, war die Frage: Wie wahrscheinlich sind  $k$  Erfolge bei  $n$  unabhängigen Versuchen? Man kann das auch *umkehren*. Dann wird das  $k$  festgehalten, und gefragt ist – bei vorgegebenem  $n \geq k$  – nach der Wahrscheinlichkeit, dass bei  $n$  Versuchen erstmals  $k$  Erfolge registriert wurden:

**Satz 5.1.4.** Es sei  $p \in ]0, 1]$ , und  $k, n$  seien natürliche Zahlen mit  $n \geq k$ . Dann ist die Wahrscheinlichkeit, bei unabhängigen Bernoulliexperimenten im  $n$ -ten Versuch erstmals  $k$  Erfolge zu haben, durch

$$b^-(n, k; p) := \binom{n-1}{k-1} p^k (1-p)^{n-k}$$

gegeben. Diese Wahrscheinlichkeitsverteilung auf  $\{k, k+1, k+2, \dots\}$  heißt die negative Binomialverteilung.

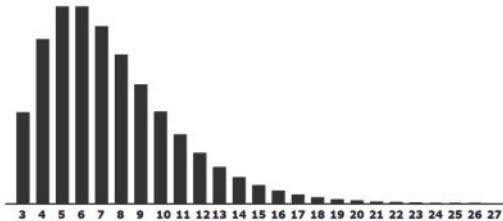


Bild 5.1.3: Die  $b^-(\cdot, 3; 0.4)$ .

**Beweis:** Diesmal gehen wir von einem Bernoulliraum  $(\Omega, \mathcal{E}, \mathbb{P})$  (Erfolgswahrscheinlichkeit  $p$ ) aus und betrachten  $(\Omega_\infty, \mathcal{E}_\infty, \mathbb{P}_\infty)$  wie in Satz 4.5.1. Damit haben wir eine Folge  $X_1, X_2, \dots$  von unabhängigen Bernoulliexperimenten zur Verfügung.

Sei  $E_{k,n}$  die Menge der 0-1-Folgen  $(x_1, x_2, \dots)$  in  $\Omega$ , bei denen es  $k$  Einsen in  $(x_1, \dots, x_n)$ , aber nur  $k-1$  Einsen in  $(X_1, \dots, X_{n-1})$  gibt. Gefragt ist doch nach  $\mathbb{P}_\infty(E_{k,n})$ .

Man kann  $E_{k,n}$  als  $\{X_1 + \dots + X_{n-1} = k-1\} \cap \{X_n = 1\}$  schreiben. Dadurch sieht man sofort, dass  $E_{k,n}$  ein Ereignis ist, und außerdem kann man

wegen der Unabhängigkeit von  $X_1 + \dots + X_{n-1}$  und  $X_n$  sofort die gesuchte Wahrscheinlichkeit ausrechnen: Die Wahrscheinlichkeit von  $\{X_1 + \dots + X_{n-1} = k-1\}$  ist doch die Wahrscheinlichkeit für genau  $k-1$  Erfolge in  $n-1$  Versuchen, also gleich  $b(k-1, n-1; p)$ , und  $\mathbb{P}(\{X_n = 1\}) = p$ . Wir erhalten den Wert

$$b(k-1, n-1; p)p = \binom{n-1}{k-1} p^{k-1} (1-p)^{n-1-(k-1)} p = b^-(n, k; p).$$

Noch ein Nachtrag: Damit  $\{k, k+1, \dots\}$  durch die  $b^-(\cdot, k; p)$  wirklich zu einem Wahrscheinlichkeitsraum wird, ist noch nachzuprüfen, dass die Summe der  $b^-(n, k; p) = 1$  für  $n = k, k+1, \dots$  gleich Eins ist. Oder anders ausgedrückt: Mit Wahrscheinlichkeit Eins wird es *irgendwann* einmal  $k$  Erfolge geben. Wir zeigen dazu etwas mehr: Mit Wahrscheinlichkeit Eins wird es unendlich oft einen Erfolg geben. Das Komplementär-Ereignis ist „Nur endlich viele Erfolge“, und das ist die Vereinigung der Ereignisse  $F_m$ , die durch „Kein Erfolg mehr nach dem  $m$ -ten Versuch“ definiert sind. Nun haben alle  $F_m$  Wahrscheinlichkeit 0, denn  $F_m$  liegt für beliebige  $r$  in  $\{X_{m+1} = X_{m+2} = \dots = X_{m+r} = 0\}$ , einem Ereignis mit Wahrscheinlichkeit  $(1-p)^r$ . (Hier wird wichtig, dass die  $(X_i)$  unabhängig sind und dass wir  $p > 0$  vorausgesetzt haben.) Aus  $\mathbb{P}_\infty(F_m) = 0$  folgt  $\mathbb{P}_\infty(\bigcup_m F_m) = 0$  wie behauptet, und damit ist alles gezeigt.  $\square$

Im Fall  $k = 1$ , wenn man also auf den *ersten Erfolg* wartet, ist  $b^-(n, 1; p) = p(1-p)^{n-1}$ . Das ist die *geometrische Verteilung* zum Parameter  $q = 1 - p$  aus Abschnitt 2.1, die auf diese Weise eine inhaltliche Interpretation bekommt. Den Erwartungswert haben wir auf Seite 82 schon ausgerechnet, er ist gleich  $1/(1-q) = 1/p$ . Deswegen ist es plausibel, dass der Erwartungswert für allgemeines  $k$  gleich  $k/p$  ist, denn  $k$  Erfolge zu haben bedeutet  $k$  Mal einen Erfolg zu haben. Das ist auch richtig, aber für einen vollständigen Beweis müsste man noch nachprüfen, dass die Situation nach dem ersten Erfolg identisch mit der Ausgangssituation ist. Diese technischen Feinheiten sollen hier nicht ausgeführt werden.

Beispiele, in denen die negative Binomialverteilung auftritt, lassen sich leicht angeben: Der Erwartungswert, bis man drei Mal eine Sechs gewürfelt hat, ist  $3/(1/6) = 18$ ; die Wahrscheinlichkeit, dass das genau im fünften Versuch tritt, ist  $b^-(5, 3; 1/6) = \binom{4}{2}(1/6)^3(1-1/6)^2 \approx 0.019$ ; und wenn ein Versicherungsunternehmen wissen möchte, ob es bis zum Monatsende schon zehn Schadensfälle gegeben haben wird, wird auch mit dieser Verteilung gerechnet.

## 5.2 Approximation der hypergeometrischen Verteilung

Sind in einem Kasten  $r$  rote und  $n-r$  weiße Kugeln und zieht man (ohne Zurücklegen)  $m$  Kugeln, so können die Wahrscheinlichkeiten  $h(k, m; r, n)$  dafür, dass genau  $k$  rote ausgewählt werden, durch die hypergeometrische Verteilung

beschrieben werden. Die entsprechenden Formeln haben wir in Abschnitt 3.5 hergeleitet.

Mal angenommen, es ist  $n = 1000$ ,  $r = 400$  und wir wollen viermal ziehen. Dann sollte es doch keinen großen Unterschied machen, ob wir die Kugeln *mit* oder *ohne* Zurücklegen auswählen: Die Wahrscheinlichkeiten für die Kugelfarbe beim jeweils nächsten Zug werden sich nur unwesentlich verändert haben, egal was wir schon gezogen haben.

Ziehen mit Zurücklegen ist aber viel einfacher zu behandeln, denn es geht um unabhängige Bernoulli-Experimente mit Erfolgswahrscheinlichkeit  $r/n$ , und die fraglichen Wahrscheinlichkeiten können mit der Binomialverteilung berechnet werden.

Falls unsere Einschätzung richtig ist, sollte also die hypergeometrische Verteilung unter geeigneten Voraussetzungen durch Binomialverteilung approximiert werden können. Das ist tatsächlich der Fall:

**Satz 5.2.1.** *Es seien  $k, m$  natürliche Zahlen mit  $0 \leq k \leq m$ , und  $p \in ]0, 1[$ . Weiter seien  $r_l, n_l \in \mathbb{N}$  für  $l = 1, 2, \dots$ , und es gelte  $\lim_{l \rightarrow \infty} r_l = \lim_{l \rightarrow \infty} n_l = +\infty$  sowie  $\lim_{l \rightarrow \infty} r_l/n_l = p$ . Dann ist  $\lim_{l \rightarrow \infty} h(k, m; r_l, n_l) = b(k, m; p)$ .*

**Beweis:** Nach Voraussetzung konvergiert  $(r_l/n_l)^k$  für  $l \rightarrow \infty$  gegen  $p^k$ , und

$$(1 - r_l/n_l)^{m-k} = ((n_l - r_l)/n_l)^{m-k}$$

konvergiert gegen  $(1 - p)^{m-k}$ . Das impliziert:

$$\begin{aligned} h(k, m; r_l, n_l) &= \frac{\binom{r_l}{k} \binom{n_l - r_l}{m - k}}{\binom{n_l}{m}} \\ &= \left( \frac{m!}{k!(m-k)!} \right) (r_l(r_l-1)\cdots(r_l-k+1)) \times \\ &\quad \times \frac{(n_l - r_l)(n_l - r_l - 1)\cdots((n_l - r_l) - (m - k) + 1)}{n_l(n_l - 1)\cdots(n_l - m + 1)}. \end{aligned}$$

Der Quotient aus

$$(r_l(r_l-1)\cdots(r_l-k+1))((n_l - r_l)(n_l - r_l - 1)\cdots((n_l - r_l) - (m - k) + 1))$$

und  $n_l(n_l - 1)\cdots(n_l - m + 1)$  konvergiert gegen den gleichen Wert wie der Quotient aus  $(r_l/n_l)^k ((n_l - r_l)/n_l)^{m-k}$ : Das sieht man dadurch ein, dass man aus Zähler und Nenner jeweils  $n_l^m$  ausklammert.

Zusammen heißt das:

$$h(k, m; r_l, n_l) \rightarrow \binom{m}{k} p^k (1 - p)^{m-k} = b(k, m; p).$$

□

Das Ergebnis kann als „Faustregel für Anwender“ so formuliert werden:

Vorgegeben seien natürliche Zahlen  $k, m, r, l$  mit  $k \leq m, r \leq n$ . Es gelte:

- $k$  ist klein gegen  $r$ .
- $m$  ist klein gegen  $n$ .
- $m - k$  ist klein gegen  $n - r$ .

Dann ist  $h(k, m; r, n) \approx b(k, m; r/n)$ , d.h. „Auswahl ohne Zurücklegen“ kann durch  $m$ -maliges Abfragen mit Erfolgswahrscheinlichkeit  $r/n$  approximiert werden.

Die  $h(\cdot, m; r, n)$  liegen unter den gegebenen Voraussetzungen nahe bei den  $b(\cdot, k; r/n)$ . Die Erwartungswerte beider Verteilungen stimmen übrigens auch ohne Zusatzbedingungen überein: Für beide Verteilungen ergibt sich  $mr/n$  (Seite 102 und Satz 5.1.3).

Die Güte der Approximation soll noch durch ein numerisches Beispiel illustriert werden. Wir betrachten den Fall  $n = 60$ ,  $r = 25$  und  $m = 4$ :

|                   | $k = 0$ | $k = 1$ | $k = 2$ | $k = 3$ | $k = 4$ |
|-------------------|---------|---------|---------|---------|---------|
| $h(k, 4; 25, 40)$ | 0.107   | 0.336   | 0.336   | 0.165   | 0.026   |
| $b(k, 4; 25/40)$  | 0.116   | 0.331   | 0.354   | 0.167   | 0.030   |

Der absolute Fehler ist jeweils kleiner als ein Prozent. Für die meisten Anwendungen sollte es also nichts ausmachen, den exakten Wert der hypergeometrischen Verteilung durch die Approximation zu ersetzen..

### 5.3 Approximation durch die Poissonverteilung

Zur Erinnerung: Die Poissonverteilung haben wir in Abschnitt 2.1 als Beispiel für ein Wahrscheinlichkeitsmaß auf  $\mathbb{N}_0$  eingeführt. Die Wahrscheinlichkeit einkommender Mengen  $\{k\}$  ist dabei durch  $p(k; \lambda) := \lambda^k e^{-\lambda} / k!$  definiert, wobei  $\lambda$  ein positiver Parameter ist. Wir wissen schon, dass der Erwartungswert gleich  $\lambda$  ist (Seite 82), und es ist nicht sehr schwer zu sehen, dass die Varianz den gleichen Wert hat; die Streuung ist folglich  $\sqrt{\lambda}$  (vgl. Übungsaufgabe 3.3.3).

Und wenn man sich ein ungefähres Bild der  $p(k; \lambda)$  machen möchte, kann man wieder den Quotiententricks anwenden:  $p(k; \lambda) / p(k+1; \lambda)$  ist gleich  $(k+1)/\lambda$ , und das bedeutet, dass die  $p(k; \lambda)$  zunächst steigen und dann – in der Nähe von  $\lambda$  – fallen.

In diesem Abschnitt soll erläutert werden, in welchen Situationen diese Verteilung auftritt.

Zur Motivation beginnen wir mit einem Bernoulliexperiment, wobei die Erfolgswahrscheinlichkeit „sehr klein“ sein soll. „Erfolg“ könnte etwa bedeuten, sechs Richtige im Lotto zu haben; in diesem Fall ist  $p = 1/13.983.816$ . Wenn

man sich dann für die Anzahl der Erfolge bei  $n$  Versuchen interessiert („Wie oft sechs Richtige in 10 Jahren?“), so ist klar, dass Wahrscheinlichkeiten für große Erfolgsanzahlen praktisch nicht sehr relevant sein werden. Zum Beispiel wird die Wahrscheinlichkeit für „140 Mal sechs Richtige in 10 Jahren<sup>2)</sup>“ von Null praktisch nicht zu unterscheiden sein.

Also noch einmal:  $p$  ist „sehr klein“,  $n$  ist „groß“, und nur „die ersten“  $k = 0, 1, \dots$  interessieren uns. Was lässt sich dann über  $b(k, n; p)$  aussagen? Um zeigen zu können, wie die Poissonverteilung ins Spiel kommt, kombinieren wir die folgenden Tatsachen:

- Es ist  $\lim_{n \rightarrow \infty} (1 + x/n)^n = e^x$ . Die Konvergenz ist gleichmäßig auf jedem beschränkten Intervall. (Das ist ein Ergebnis aus der Analysis.)
- Ist  $p$  „sehr klein“ und  $k$  „nicht zu groß“, so ist  $(1 - p)^k \approx 1$ .
- Wenn  $k$  klein gegen  $n$  ist, darf  $n(n-1) \cdots (n-k+1)$  durch  $n^k$  approximiert werden.

Und dann kann man so rechnen:

$$\begin{aligned} b(k, n; p) &= \binom{n}{k} p^k (1-p)^{n-k} \\ &= \frac{n(n-1) \cdots (n-k+1)}{k!} p^k (1-p)^{n-k} \\ &\approx \frac{n^k}{k!} p^k (1-p)^{n-k} \\ &= \frac{(np)^k}{k!} \frac{(1-p)^n}{(1-p)^k} \\ &\approx \frac{(np)^k}{k!} (1-p)^n \\ &= \frac{(np)^k}{k!} \left(1 - \frac{np}{n}\right)^n \\ &\approx \frac{(np)^k}{k!} e^{-np} \\ &= p(k; np). \end{aligned}$$

Die Binomialverteilung kann also durch die Poissonverteilung zum Parameter  $np$  approximiert werden.

Die letzte Approximation sieht ein bisschen gewagt aus, da ja nirgendwo ein Ausdruck der Form  $(1+x/n)^n$  steht. Genauer müsste man so argumentieren:  $\varepsilon > 0$  und ein Intervall  $[-R, R]$  seien vorgegeben. Wähle  $n$  so groß, dass  $e^x$  für alle  $x \in [-R, R]$  bis auf  $\varepsilon$  durch  $(1+x/n)^n$  approximiert werden kann<sup>3)</sup>. Dann ist unsere Approximation gerechtfertigt, wenn wir uns auf Fälle beschränken, bei denen es um dieses  $n$  geht und  $p$  so klein ist, dass  $x := -np \in [-R, R]$ .

---

<sup>2)</sup>Also bei 520 Versuchen.

<sup>3)</sup>So ein  $n$  gibt es wegen der Gleichmäßigkeit der Konvergenz auf  $[-R, R]$ .

### Die Eulersche Zahl $e$ in der Wahrscheinlichkeitstheorie?

Es gibt verschiedene Möglichkeiten, die Zahl  $e = 2.7181\dots$  in der Analysis einzuführen. Zum Beispiel als  $\lim(1+1/n)^n$  oder als  $\exp(1)$ , wobei  $\exp$  die eindeutig bestimmte Lösung des Anfangswertproblems  $y' = y$ ,  $y(0) = 1$  bezeichnet. Beim ersten Ansatz ist die Motivation die „stetige Verzinsung“, beim zweiten sind es Wachstumsprozesse.

Doch keiner der Ansätze hat auf den ersten Blick etwas mit Wahrscheinlichkeiten zu tun. Deswegen ist es schon bemerkenswert, dass uns diese Zahl hier immer wieder begegnen wird.

Die mathematisch präzise Formulierung der vorstehenden Überlegungen sieht dann so aus:

**Satz 5.3.1.** *Es sei  $(p_n)_{n \in \mathbb{N}}$  eine Folge in  $[0, 1]$ , für die  $(n \cdot p_n)_{n \in \mathbb{N}}$  gegen eine Zahl  $\lambda \in [0, +\infty] \subset \mathbb{R}$  konvergiert. Dann gilt für jedes  $k \in \mathbb{N}_0$ :*

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

**Beweis:** Wir fixieren ein  $k \in \mathbb{N}_0$ . Wähle ein Intervall  $[-R, R]$ , das  $\lambda$  im Innern enthält. Für große  $n$  ist dann  $n \cdot p_n \in [-R, R]$ , außerdem konvergiert die Funktionenfolge  $(1+x/n)^n$  gleichmäßig auf diesem Intervall gegen die Funktion  $e^x$ . Mit  $n \cdot p_n \rightarrow \lambda$  folgt  $(1 - (n \cdot p_n)/n)^n \rightarrow e^{-\lambda}$ .

Da  $(n \cdot p_n)$  beschränkt ist, muss  $(p_n)_{n \in \mathbb{N}}$  eine Nullfolge sein. Damit konvergiert  $1/(1-p_n)^k$  gegen 1, denn  $1/(1-x)^k$  ist an der Stelle  $x = 0$  stetig. Nun sind wir auch schon fertig, denn es gilt (wie vorstehend gezeigt)

$$b(k, n; p_n) = \frac{(np_n)^k}{k!} \cdot \left(1 - \frac{np_n}{n}\right)^n \cdot \frac{1}{(1-p_n)^k}.$$

Rechts steht damit das Produkt dreier konvergenter Folgen, die gegen  $\lambda^k/k!$  bzw.  $e^{-\lambda}$  bzw. 1 konvergieren. Der Limes ist also  $p(k; \lambda)$  wie behauptet.  $\square$

Wie im vorigen Abschnitt bei der hypergeometrischen Verteilung wollen wir uns auch hier durch ein numerisches Beispiel von der Güte der Approximation überzeugen. Es sei  $n = 20$  und  $p = 1/20$ . Die  $b(k, n; p)$  sollten dann für kleine  $k$  nahe bei den  $p(k; 1)$  liegen. Wirklich ist

|                  | $k = 0$ | $k = 1$ | $k = 2$ | $k = 3$ |
|------------------|---------|---------|---------|---------|
| $b(k, 20; 1/20)$ | 0.3584  | 0.3773  | 0.1887  | 0.0596  |
| $p(k; 1)$        | 0.3679  | 0.3679  | 0.1839  | 0.0613  |

Und wieder ist der Fehler geringer als ein Prozent.

Wie kann man diese Ergebnisse denn nun nutzen? Dazu die

**Faustregel: Wann wird die Poissonverteilung angewendet?**

Gegeben sei eine Situation, bei der erstens etwas gezählt wird und bei der man zweitens das Ergebnis als unabhängige Überlagerung von sehr vielen Bernoulli-Experimenten mit sehr kleiner Wahrscheinlichkeit interpretieren kann. Typisches Beispiel: Wie viele Anrufe gehen bei einer Autovermietung zwischen 10 und 11 Uhr ein?

Dann sagt das vorstehende Ergebnis, dass die fraglichen Wahrscheinlichkeiten Poissonverteilt zu einem geeigneten Parameter  $\lambda$  sind.

Doch wie groß ist  $\lambda$ ? Hier muss man sich noch einmal daran erinnern, dass die Poissonverteilung den Erwartungswert  $\lambda$  hat (Seite 82) und dass der Erwartungswert die mittlere Häufigkeit modelliert. Hier heißt das: Wenn im Mittel 25 Anrufe eingehen, so sollte man die Poissonverteilung zum Parameter  $\lambda = 25$  verwenden.

Und dann sind viele naheliegende Fragen zu beantworten: Wie wahrscheinlich ist es, dass es weniger als 5 Anrufe sind? (Antwort:  $\sum_{k=0}^4 p(k; 25)$ .) Oder mehr als 26? (Antwort:  $1 - \sum_{k=0}^{26} p(k; 25)$ ) ...

Es folgen einige konkrete *Beispiele*:

**1.** Wenn man einen Artikel oder ein Buch schreibt, schleichen sich manchmal Fehler ein: Buchstabendreher, „das“ statt „dass“ oder umgekehrt usw. Das passt bei jedem Wort mit einer gewissen winzigen Wahrscheinlichkeit, und deswegen ist es gerechtfertigt anzunehmen, dass die Anzahl der Fehler pro Seite Poissonverteilt ist. Wenn man nun die ersten Seiten sehr sorgfältig gelesen hat und im Mittel auf 2.4 Fehler gestoßen ist, wird man mit dem Parameter  $\lambda = 2.4$  weiterrechnen. Wenn man dann irgendeine Seite aufschlägt, mit welcher Wahrscheinlichkeit wird die fehlerfrei sein? Um das zu beantworten, muss man nur  $p(0, 2.4) = e^{-2.4} = 0.091$  ausrechnen: Nur mit etwa 9 Prozent Wahrscheinlichkeit wird alles stimmen.

**2.** Wie kann man erreichen, dass eine Seite mit 99 Prozent Wahrscheinlichkeit fehlerfrei ist? Wie sorgfältig muss man da arbeiten? Für das fragliche  $\lambda$ , das die Fehler zählt, muss also  $p(0, \lambda) = 0.99$  gelten. Das führt auf  $e^{-\lambda} = 0.99$ , und durch Logarithmieren folgt  $\lambda = 0.010$ .

Mal angenommen, es gibt 400 Wörter pro Seite. Die Fehlerwahrscheinlichkeit  $p$  bei einem einzelnen Wort soll also die Bedingung  $\lambda = p \cdot n = p \cdot 400$  erfüllen.  $p$  darf also höchstens  $0.010/400 = 0.000025$  sein, um das gewünschte Ergebnis zu erzielen. Das ist ziemlich unrealistisch, aber man kann ja noch Korrekturlesen.

Hier noch einige weitere Beispiele, bei denen die Poissonverteilung zum Modellieren verwendet werden kann:

- Die Anzahl der Unfälle auf dem Berliner Ring am Ostersonntag.
- Die Anzahl der Rosinen in einem Rosinenbrötchen.
- Die Anzahl der erhaltenen Weihnachtspostkarten.

- Die Anzahl der Vokabeln pro Seite, die Sie beim Lesen eines französischen Romans nachschlagen müssen<sup>4)</sup>.
- Anzahl der Webfehler auf einem Meter Stoff.
- ... und so weiter.

## 5.4 Der Satz von de Moivre-Laplace

Gegenstand dieses Abschnitts ist einer der historisch ersten Grenzwertsätze der Wahrscheinlichkeitstheorie. Er besagt, dass die Binomialverteilung in einem zu präzisierenden Sinn durch die Standardnormalverteilung beschrieben werden kann. Dass da wohl ein Zusammenhang besteht, ist offensichtlich: Die Wahrscheinlichkeiten der Binomialverteilung erinnern wirklich stark an die „Glockenkurve“ (vgl. zum Beispiel Bild 5.1.1).

Das Problem, das wir behandeln werden, ist das folgende: Wie groß ist die Wahrscheinlichkeit, dass – bei  $n$  Versuchen und einer Erfolgswahrscheinlichkeit  $p$  – die Anzahl der Erfolge zwischen zwei Zahlen  $\alpha, \beta \in \{0, \dots, n\}$  liegt. Dazu muss  $\sum_{k=\alpha}^{\beta} b(k; n; p)$  bestimmt werden:

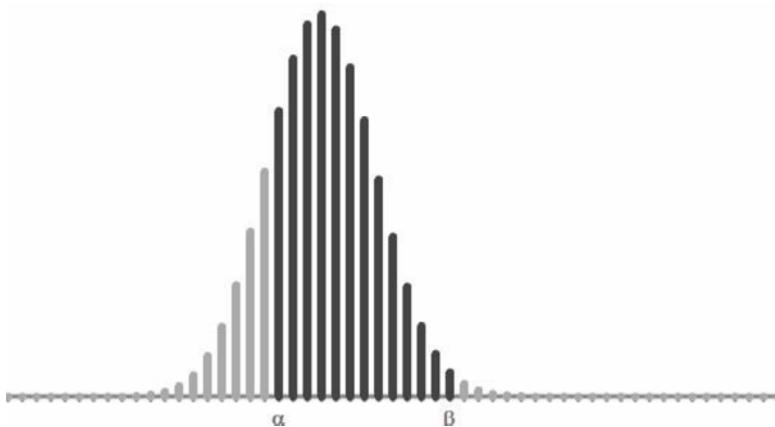


Bild 5.4.1: Wie groß ist die Summe der dunkel skizzierten Wahrscheinlichkeiten?

Um diese Frage mit der Normalverteilung in Verbindung zu bringen, werden wir wie folgt verfahren:

- Zunächst beweisen wir ein Hilfsergebnis: die Formel für das *Wallisprodukt*. Sie ist dafür verantwortlich, dass man auch die Kreiszahl  $\pi$  in der Wahrscheinlichkeitstheorie antrifft.

---

<sup>4)</sup>Das gilt allerdings nur dann, wenn Sie schon ganz gut Französisch können und nur bei den ganz ausgefallenen Wörtern Hilfe brauchen.

- Zur Berechnung der bei der Binomialverteilung auftretenden Binomialkoeffizienten wäre es doch günstig, eine handliche Formel für die Fakultäten (also die  $n!$ ,  $k!$ ,  $(n - k)!$ ) zur Verfügung zu haben. Genau das leistet die *Stirlingformel*.
- Danach kann eine gute Näherung für die  $b(k, n; p)$  hergeleitet werden. Und die führt dazu, dass Wahrscheinlichkeiten für „Die Anzahl der Erfolge liegt zwischen  $\alpha$  und  $\beta$ “ in guter Näherung als Riemannsummen für die Berechnung des Integrals der Dichtefunktion der Standardnormalverteilung über ein geeignetes Intervall interpretiert werden können. *Das* ist dann der *Satz von de Moivre-Laplace*.

Dieser Satz ist übrigens ein Spezialfall des zentralen Grenzwertsatzes, den wir in Abschnitt 9.6 beweisen werden. Erwartungsgemäß wird der Beweis des allgemeinen Ergebnisses weit aufwändiger sein.

### Das Wallisprodukt

Überraschenderweise gibt es einen Zusammenhang zwischen der Kreiszahl  $\pi$  und Fakultäten:

**Lemma 5.4.1.** (*Wallis-Produkt*) Es gilt

$$\lim_{n \rightarrow \infty} \frac{2^{4n}(n!)^4}{(2n)!(2n+1)!} = \frac{\pi}{2}.$$

**Beweis:** Um einen Eindruck davon zu bekommen, mit welch gigantisch großen Zahlen wir es in diesem Beweis zu tun haben, betrachten wir als Beispiel den Fall  $n = 20$ :

- $2^{4 \cdot 20} (20!)^4 \approx 0.423 \cdot 10^{98}$ .
- $(2 \cdot 20)!(2 \cdot 20 + 1)! \approx 0.273 \cdot 10^{98}$ .

Der Quotient ist  $1.551\dots$ , das ist schon recht nahe an  $\pi/2 = 1.571\dots$

Nun zum Beweis. Definiere  $s_n$  als  $\int_0^{\pi/2} \sin^n x \, dx$ . Für diese Zahlen gilt dann  $s_n = ((n-1)/n)s_{n-2}$ , das beweist man so:

1. Es ist  $[(\sin^{n-1} x) \cos x]' = (n-1)(\sin^{n-2} x) \cos^2 x - \sin^n x$ .
2. Damit ist

$$\int_0^{\pi/2} [(n-1)(\sin^{n-2} x) \cos^2 x - \sin^n x] = (\sin^{n-1} x) \cos x \Big|_0^{\pi/2} = 0.$$

3. Wenn man noch die Identität  $\cos^2 x = 1 - \sin^2 x$  ausnutzt, ergibt sich  $(n-1)s_{n-2} - (n-1)s_n - s_n = 0$ , und das ist die Behauptung.

Dadurch ist es möglich, die Berechnung der  $s_n$  rekursiv auf die Bestimmung von  $s_0$  (für gerade  $n$ ) und  $s_1$  (für ungerade  $n$ ) zurückzuführen. Wir erhalten:

$$\begin{aligned} s_{2n} &= \frac{2n-1}{2n} \frac{2n-3}{2n-2} \cdots \frac{1}{2} s_0 \\ &= \frac{2n-1}{2n} \frac{2n-3}{2n-2} \cdots \frac{1}{2} \frac{\pi}{2} \\ s_{2n+1} &= \frac{2n}{2n+1} \frac{2n-2}{2n-1} \cdots \frac{2}{3} s_1 \\ &= \frac{2n}{2n+1} \frac{2n-2}{2n-1} \cdots \frac{2}{3}. \end{aligned}$$

Wir setzen noch  $A_n := s_{2n}/s_{2n+1}$ , und es wird gleich wichtig werden, dass  $A_n \rightarrow 1$ . Es gilt  $A_n \geq 1$ , denn wegen  $\sin^{2n+1} x \leq \sin^{2n} x$  (für  $x \in [0, \pi/2]$ ) ist  $s_{2n+1} \leq s_{2n}$ . Außerdem ist

$$s_{2n+1} = \frac{2n}{2n+1} s_{2n-1} \geq \frac{2n}{2n+1} s_{2n},$$

und das impliziert  $A_n \leq (2n+1)/(2n) = 1 + 1/(2n)$ .

Damit haben wir bewiesen, dass die Folge  $(A_n)_n$  gegen Eins konvergiert. Um den Beweis abzuschließen, muss man nur noch die obigen Formeln für die  $s_{2n}$  und die  $s_{2n+1}$  in  $A_n$  einsetzen und den entstehenden Ausdruck ein bisschen umformen: Zum Beispiel kann statt  $(2n)(2n-2) \cdots 2$  auch  $2^n n!$  geschrieben werden. So ergibt sich

$$A_n = \frac{\pi/2}{[2^{4n}(n!)^4]/[(2n)!(2n+1)!]},$$

und da  $\lim A_n = 1$  gilt, ist die Formel für das Wallis-Produkt bewiesen.  $\square$

### Die Stirlingformel

Es ist doch  $n!$  das Produkt aus  $n$  Faktoren, die von 1 nach  $n$  wachsen. Mit welchem  $\alpha$  sollte man denn  $n$  zu  $n/\alpha$  verkleinern, um  $n!$  möglichst gut als  $n$ -te Potenz von  $n/\alpha$  schreiben zu können? Zum „Experimentieren“ betrachten wir einmal den Fall  $n = 30$ . Es ist

$$30! = 265252859812191058636308480000000,$$

und nun versuchen wir es mit verschiedenen  $\alpha$ . Für  $\alpha = 2$  erhalten wir

$$\frac{30!}{(30/2)^{30}} = 0.00138\dots$$

$(30/2)^{30}$  ist also viel größer als  $30!$ : Wir müssen stärker verkleinern. Im nächsten Versuch wählen wir  $\alpha = 3$ . Dann ist

$$\frac{30!}{(30/3)^{30}} = 265.252\dots$$

Das war wohl etwas zu viel. Die beste Wahl wird folglich zwischen  $\alpha = 2$  und  $\alpha = 3$  liegen.

Die Stirlingsche Formel besagt, dass die Eulersche Zahl  $e = 2.7182818\dots$  die optimale Wahl für  $\alpha$  ist: Bis auf eine vergleichsweise kleine Korrektur ist  $n!$  gleich  $(n/e)^n$ :

**Satz 5.4.2.** (*Stirlingformel*)

Es ist  $n! \approx \sqrt{2\pi n} \cdot (n/e)^n$ . Genauer gilt:

$$1 < \frac{n!}{\sqrt{2\pi n} \cdot (n/e)^n} < e^{\frac{1}{12n}}.$$

**Beweis:** Wenn die Ungleichung gezeigt ist, ist auch bewiesen, dass der relative Fehler bei der Approximation von  $n!$  durch  $\sqrt{2\pi n} \cdot (n/e)^n$  mit  $n \rightarrow \infty$  gegen Null geht, denn die  $e$ -Funktion ist stetig und  $e^0 = 1$

Bevor wir mit dem Beweis beginnen, wollen wir die Formel am obigen Beispiel  $n = 30$  testen. Da ist der Quotient aus  $30!$  und  $\sqrt{60\pi}(30/e)^{30}$  gleich  $1.0027815362\dots$ , der relative Fehler der Approximation ist also nur etwa drei Promille. Die Abschätzung des Satzes sagt voraus, dass er nicht größer als  $e^{1/12 \cdot 30} = 1.0027816393\dots$  ist. Das ist nur ganz unwesentlich mehr als der wirkliche Wert, die Abschätzung scheint also recht gut zu sein. Als weiteres Beispiel testen wir  $n = 100$ : Da lauten die Zahlen für den Quotienten (bzw. für  $e^{1/12n}$ )  $1.000833677872005$  bzw.  $1.000833680652026$ , der Fehler ist damit kleiner als ein Promille, und die zweite (die Abschätzung) ist nur um eine Winzigkeit größer als die erste. Und das für Quotienten im Bereich gigantisch großer Zahlen in der Größenordnung von  $10^{158}$ .

Zum Beweis der Stirlingformel definieren wir Zahlen  $c_n$  durch

$$c_n := n! \left( \frac{e}{n} \right)^n \frac{1}{\sqrt{n}}.$$

*Behauptung 1:* Es ist stets  $c_{n+1} < c_n < e^{\frac{1}{12} \left[ \frac{1}{n} - \frac{1}{n+1} \right]} c_{n+1}$ .

Dazu untersuchen wir die Quotienten

$$\frac{c_{n+1}}{c_n} = \frac{e}{\left(1 + \frac{1}{n}\right)^n} \sqrt{\frac{1}{1 + \frac{1}{n}}}.$$

Wir rechnen den Logarithmus dieser Zahl aus und ersetzen den Logarithmus durch die Reihenentwicklung<sup>5)</sup>:

$$\begin{aligned} \log\left(\frac{c_{n+1}}{c_n}\right) &= 1 - \left(n + \frac{1}{2}\right) \log\left(1 + \frac{1}{n}\right) \\ &= 1 - \left(n + \frac{1}{2}\right) \left(\frac{1}{n} - \frac{1}{2n^2} + \frac{1}{3n^3} - \frac{1}{4n^4} \pm \dots\right) \\ &= -\alpha_2 \frac{1}{n^2} + \alpha_3 \frac{1}{n^3} - \alpha_4 \frac{1}{n^4} \pm \dots; \end{aligned}$$

<sup>5)</sup>D. h.  $\log(1+x) = x - x^2/2 + x^3/3 \pm \dots$  für  $|x| < 1$ .

dabei ist  $\alpha_k = (k-1)/[(2k)(k+1)]$ .

Die Summanden der Beträge fallen, stets ist  $\alpha_k/n^k > \alpha_{k+1}/n^{k+1}$ : Das lässt sich leicht ausrechnen, indem man die Ungleichung dadurch umformt, dass man mit dem Produkt der Nenner multipliziert und vereinfacht. Und deswegen weiß man bei jeder Partialsumme, ob der Wert eher zu groß oder eher zu klein ist. Insbesondere folgt

$$\log\left(\frac{c_{n+1}}{c_n}\right) < -\alpha_2 \frac{1}{n^2} + \alpha_3 \frac{1}{n^3} = -\frac{1}{12n^2} + \frac{1}{12n^3} < 0$$

sowie

$$\log\left(\frac{c_{n+1}}{c_n}\right) > -\alpha_2 \frac{1}{n^2} + \alpha_3 \frac{1}{n^3} - \alpha_4 \frac{1}{n^4} = -\frac{1}{12n^2} + \frac{1}{12n^3} - \frac{3}{40n^4}.$$

Es wird günstig sein, noch weiter abzuschätzen: Stets ist

$$-\frac{1}{12n^2} + \frac{1}{12n^3} - \frac{3}{40n^4} > -\frac{1}{12}\left(\frac{1}{n} - \frac{1}{n+1}\right).$$

Auch das ist leicht durch elementare Umformungen – multipliziere die Ungleichung mit  $120n^4(n+1)$  und sortiere – zu verifizieren.

Insgesamt haben wir damit

$$0 > \log\left(\frac{c_{n+1}}{c_n}\right) > -\frac{1}{12}\left(\frac{1}{n} - \frac{1}{n+1}\right)$$

gezeigt, und so folgt

$$1 > \frac{c_{n+1}}{c_n} > e^{-\frac{1}{12}\left[\frac{1}{n} - \frac{1}{n+1}\right]}.$$

Das ist (im Wesentlichen) die Behauptung.

*Behauptung 2: Es ist stets  $c_{n+k} < c_n < e^{\frac{1}{12}\left[\frac{1}{n} - \frac{1}{n+k}\right]} c_{n+k}$ .*

Zum Beweis muss man nur die Formel für  $k = 1$  mehrfach anwenden. So ist etwa

$$\begin{aligned} c_n &< e^{\frac{1}{12}\left[\frac{1}{n} - \frac{1}{n+1}\right]} c_{n+1} \\ &< e^{\frac{1}{12}\left[\frac{1}{n} - \frac{1}{n+1}\right]} e^{\frac{1}{12}\left[\frac{1}{n+1} - \frac{1}{n+2}\right]} c_{n+2} \\ &= e^{\frac{1}{12}\left[\frac{1}{n} - \frac{1}{n+2}\right]} c_{n+2}. \end{aligned}$$

(Um diesen Teleskopsummen-Trick zu ermöglichen, hatten wir im vorigen Beweisteil noch weiter abgeschätzt).

Nun fixieren wir  $n$  und lassen  $k$  gegen  $\infty$  gehen. Da die  $c_n$  monoton fallen und positiv sind, existiert  $c = \lim_n c_n$ , und aus den Ungleichungen in Behauptung 2 schließen wir, dass  $c < c_n \leq e^{\frac{1}{12n}} c$  gilt. Insbesondere ist  $c > 0$ .

Welchen Wert hat  $c$ ? Dazu berechnen wir  $c_n^4/c_{2n}^2$ . Wenn man die Definition einsetzt, folgt

$$\frac{c_n^4}{c_{2n}^2} = \frac{2^{4n}(n!)^4}{(2n)!(2n+1)!} \cdot \frac{2(2n+1)}{n}.$$

Die rechte Seite ist – bis auf den Faktor  $2(2n+1)/n$  – das Wallisprodukt, konvergiert also mit  $n \rightarrow \infty$  gegen  $2\pi$ . Die rechte Seite geht gegen  $c^2$ , und so folgt  $c = \sqrt{2\pi}$ .

Nun ist nur noch in der Ungleichung  $c < c_n \leq e^{\frac{1}{12n}}c$  durch  $c$  zu teilen und der Wert für  $c$  einzusetzen. Damit ist dann die Behauptung bewiesen.  $\square$

Mit vergleichsweise geringem Aufwand ist ein wesentlich genaueres Ergebnis zu erzielen. Dazu verfeinert man die obige Abschätzung  $\log(c_{n+1}/c_n) < 0$  durch Hinzunahme weiterer Summanden zu

$$\log\left(\frac{c_n}{c_{n+1}}\right) < -\frac{1}{12n^2} + \frac{1}{12n^3} - \frac{3}{40n^4} + \frac{1}{15n^5}.$$

Die rechte Seite kann weiter durch

$$-\frac{1}{12}\left(\frac{1}{n} - \frac{1}{n+1}\right) + \frac{1}{120}\left(\frac{1}{n^2} - \frac{1}{(n+1)^2}\right)$$

abgeschätzt werden, und mit der gleichen Idee wie eben gelangt man zu

$$e^{\frac{1}{12n} - \frac{1}{120n^2}} < \frac{n!}{\sqrt{2\pi n} \cdot (n/e)^n} < e^{\frac{1}{12n}}.$$

Dieses Ergebnis, das wesentlich genauer ist als die Stirlingformel, wurde auch schon durch die numerischen Beispiele zu Beginn des Beweises qualitativ illustriert.

Die Approximation der  $b(k, n; p)$

Jetzt sind wir in der Lage, Fakultäten durch Produkte zu approximieren. Wenn man das für die Wahrscheinlichkeiten bei der Binomialverteilung anwendet, erhält man

$$\begin{aligned} b(k, n; p) &= \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \\ &\approx \frac{n^n \sqrt{2\pi n} e^{-n}}{k^k \sqrt{2\pi k} e^{-k} (n-k)^{n-k} \sqrt{2\pi(n-k)} e^{-(n-k)}} p^k (1-p)^{n-k} \\ &= \sqrt{\frac{n}{2\pi k(n-k)}} \left(\frac{np}{k}\right)^k \left(\frac{n(1-p)}{n-k}\right)^{n-k}. \end{aligned}$$

Das sieht zwar nicht besonders übersichtlich aus, wir erinnern uns aber daran, dass die  $b(k, n; p)$  für die  $k$  in der Nähe des Erwartungswertes  $np$  am größten sind.

Deswegen wählen wir  $np$  als neuen Entwicklungspunkt. Dazu definieren wir  $\delta_k$  durch  $k = np + \delta_k$  und schreiben die vorstehend berechnete Approximation für  $b(k, n; p)$  als Funktion von  $\delta_k$ :

$$\begin{aligned} b(k, n; p) &\approx \sqrt{\frac{n}{2\pi(np+\delta_k)(n(1-p)-\delta_k)}} \times \\ &\quad \times \frac{1}{\left(1 + \frac{\delta_k}{np}\right)^{np+\delta_k} \left(1 - \frac{\delta_k}{n(1-p)}\right)^{n(1-p)-\delta_k}}. \end{aligned}$$

Rechts steht ein Produkt aus zwei Faktoren, die wir für die nächsten Zeilen mit  $F_1$  und  $F_2$  bezeichnen.  $F_1$  approximieren wir durch

$$\sqrt{\frac{1}{2\pi np(1-p)}};$$

$\delta_k$  wurde also durch Null ersetzt. Das ist sicher für diejenigen  $k$  gerechtfertigt, für die  $\delta_k$  klein gegen  $np$  und klein gegen  $n(1-p)$  ist. (Für die anderen  $k$  sind nur winzige Wahrscheinlichkeiten zu erwarten.)

Für die Behandlung von  $F_2$  verwenden wir die gleiche Technik wie beim Beweis der Stirlingformel. Wir kümmern uns also zuerst um  $\log F_2$ , erinnern uns bei der Auswertung von  $\log(1 + \frac{\delta_k}{np})$  und  $\log(1 - \frac{\delta_k}{n(1-p)})$  an die Reihenentwicklung  $\log(1+x) = x - x^2/2 + x^3/3 \mp \dots$  und sortieren nach Potenzen von  $\delta_k$ . Die Überraschung: Der lineare Term in  $\delta_k$  verschwindet, die Entwicklung beginnt mit

$$\log F_2 = -\frac{\delta_k^2}{2np(1-p)} + \dots$$

Und daraus schließen wir:

$$b(k, n; p) \approx \frac{1}{\sqrt{2\pi np(1-p)}} e^{-\delta_k^2/2np(1-p)}.$$

Um eine Vorstellung von der Güte der Approximation zu bekommen, sind nachstehend einige  $b(k, n; p)$  für den Fall  $n = 20$ ,  $p = 0.4$  zusammen mit der Approximation tabelliert:

|                 | exakt   | approximativ |
|-----------------|---------|--------------|
| $b(5, 20; 0.4)$ | 0.07465 | 0.07131      |
| $b(6, 20; 0.4)$ | 0.12441 | 0.12004      |
| $b(7, 20; 0.4)$ | 0.16588 | 0.16408      |
| $b(8, 20; 0.4)$ | 0.17971 | 0.18208      |
| $b(9, 20; 0.4)$ | 0.15974 | 0.16408      |

Der absolute Fehler beträgt also in allen Fällen nur einige Promille.

### Der Satz von de Moivre-Laplace

Nun fehlt nur noch eine letzte Überlegung.

Angenommen, es ist  $\phi : \mathbb{R} \rightarrow \mathbb{R}$  eine stetige Funktion und das Intervall  $[a, b]$  ist in  $n$  gleiche Teile geteilt:  $x_0 = a, x_1 = a + h, \dots, x_n = a + n \cdot h = b$ , wobei  $h = (b - a)/n$ . Wenn dann in irgendeinem Zusammenhang der Ausdruck

$$\sum_{i=0}^n \phi(x_i)h$$

auftritt, so kann er – falls  $n$  nicht zu klein ist – durch  $\int_{a-h/2}^{b+h/2} \phi(x) dx$  approximiert werden.

Das ist sofort einzusehen, wenn man die Summe als geeignete Rechtecksumme interpretiert:

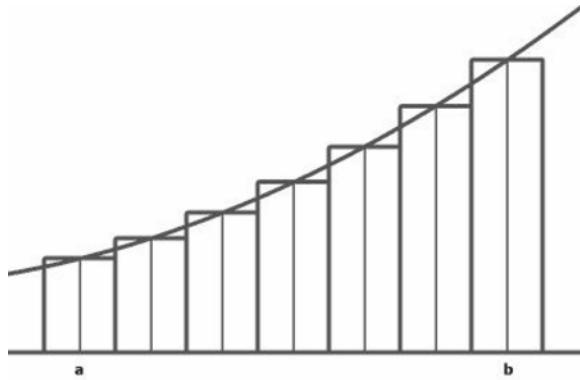


Bild 5.4.2: Balkenbreite  $h$ ,  $x_i = a + h \cdot i$ : Dann ist  $\sum \phi(x_i)h \approx \int_{a-h/2}^{b+h/2} \phi(x) dx$ .

Und genau so eine Situation liegt vor, wenn wir bei vorgegebenen Zahlen  $\alpha, \beta$  mit  $\alpha < \beta$  an  $\sum_{k=\alpha}^{k=\beta} b(k, n; p)$  interessiert sind und unsere Approximation einsetzen. Dann ist doch

$$\sum_{k=\alpha}^{k=\beta} b(k, n; p) \approx \frac{1}{\sqrt{2\pi np(1-p)}} \sum_{k=\alpha}^{k=\beta} e^{-\delta_k^2/2np(1-p)}.$$

Um das als Integralapproximation zu interpretieren, definieren wir

$$x(t) := (t - np)/\sqrt{np(1-p)}, \quad h := 1/\sqrt{np(1-p)}.$$

Mit  $\phi(x) := e^{-x^2/2}/\sqrt{2\pi}$  ist dann

$$\begin{aligned} \frac{1}{\sqrt{2\pi np(1-p)}} \sum_{k=\alpha}^{k=\beta} e^{-\delta_k^2/2np(1-p)} &= h \left( \phi(x(\alpha)) + \phi(x(\alpha) + h) + \dots + \phi(x(\beta)) \right) \\ &\approx \int_{x(\alpha-1/2)}^{x(\beta+1/2)} \phi(x) dx \end{aligned}$$

Und damit haben wir den folgenden Satz bewiesen:

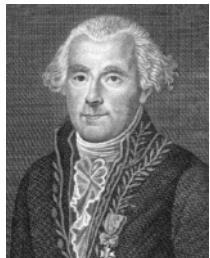
**Satz 5.4.3.** (*Satz von de Moivre-Laplace*<sup>6)</sup>) Gegeben seien  $p \in ]0, 1[$ ,  $n \in \mathbb{N}$  und  $0 \leq \alpha < \beta \leq n$ . Dann kann  $\sum_{k=\alpha}^{k=\beta} b(k, n; p)$ , also die Wahrscheinlichkeit, dass die Erfolgsanzahl bei  $n$  unabhängigen Bernoulliexperimenten mit Erfolgswahrscheinlichkeit  $p$  zwischen  $\alpha$  und  $\beta$  liegt, durch

$$\frac{1}{\sqrt{2\pi}} \int_{x(\alpha-1/2)}^{x(\beta+1/2)} e^{-x^2/2} dx$$

approximiert werden. Dabei ist  $x(t) := (t - np)/\sqrt{np(1-p)}$  für  $t \in \mathbb{R}$ .

---

<sup>6)</sup>Pierre-Simon Laplace, 1749 bis 1827. Mitglied der französischen Akademie der Wissenschaften seit 1773. Wichtige Beiträge in vielen Gebieten, insbesondere Analysis, Astronomie und Wahrscheinlichkeitsrechnung. Unter Napoleon hatte er auch politische Ämter inne.



Laplace

Es folgen einige *Bemerkungen und Beispiele*:

1. Konkrete Rechnungen zeigen, dass die Approximation im Satz von de Moivre<sup>7)</sup>-Laplace bewerkenswert gut ist. Die Faustregel unter Anwendern besagt, dass man die Formel immer dann problemlos verwenden kann, wenn die Zahl  $(\max\{|\alpha - np|, |\beta - np|\})^3/n^2$  „klein“ ist.
2. Für große  $n, \alpha, \beta$  kann man für die Integrationsgrenzen statt  $x(\alpha - 1/2)$  und  $x(\beta + 1/2)$  auch  $x(\alpha)$  und  $x(\beta)$  einsetzen; der Unterschied wird vernachlässigbar sein.
3. Wir müssen uns noch darum kümmern, wie man zu den Werten für die im Satz auftretenden Integrale kommt. Eine geschlossene Integration ist ja nicht möglich.

Wenn man ein Computerprogramm zur Stochastik zur Verfügung hat, sind diese Zahlen leicht abrufbar. Man kann sich aber auch mit einer *Tabelle zur Normalverteilung* behelfen, das ist seit mehreren Jahrhunderten das übliche Verfahren. In diesem Buch finden Sie diese Tabelle im Anhang auf Seite 364. Um sie zu nutzen, muss man nur wissen:

- Sie enthält für „ausreichend viele“  $x$  die Zahl

$$\phi(x) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-x^2/2} dx$$

auf vier Stellen genau. Zum Beispiel ist  $\phi(-1.24) = 0.1075$  oder  $\phi(2.39) = 0.9916$ . Wenn  $\phi$  an Stellen benötigt wird, die nicht in der Tabelle stehen, muss linear interpoliert werden. Zum Beispiel ist

$$\phi(1.613) = \phi(1.61) + 0.3(\phi(1.62) - \phi(1.61)) = 0.9463 + 0.3 \cdot 0.011 = 0.9466.$$

- Für  $a < b$  ist  $(1/\sqrt{2\pi}) \int_a^b e^{-x^2/2} dx = \phi(b) - \phi(a)$ . Zum Beispiel ist

$$\int_{-1.24}^{2.39} b(1/\sqrt{2\pi})e^{-x^2/2} dx = 0.9916 - 0.1075 = 0.8841$$

4. Angenommen, wir wollen 3000 Mal eine faire Münze werfen. Wie wahrscheinlich ist es, dass die Anzahl von „Kopf“ zwischen 1490 und 1520 liegen wird? Zur Berechnung einer Approximation ist der vorige Satz mit  $p = 0.5$ ,  $n = 3000$ ,  $\alpha = 1490$  und  $\beta = 1520$  anzuwenden. Es ist dann

$$x(\alpha - 0.5) = \frac{1490 - 0.5 - 1500}{\sqrt{3000 \cdot 0.5 \cdot 0.5}} = -0.384 \text{ und}$$

$$x(\beta + 0.5) = \frac{1520 + 0.5 - 1500}{\sqrt{3000 \cdot 0.5 \cdot 0.5}} = 0.748.$$

---

<sup>7)</sup>Abraham de Moivre, 1667 bis 1754. Er wuchs in Frankreich auf, musste aber wegen der Hugenottenverfolgungen nach England ziehen. Sein Hauptwerk zur Wahrscheinlichkeitstheorie („The Doctrine of Chance“) war sehr einflussreich.



de Moivre

Mit Tabellenhilfe erhalten wir

$$\begin{aligned}\frac{1}{\sqrt{2\pi}} \int_{-0.384}^{0.748} e^{-x^2/2} dx &= \phi(0.748) - \phi(-0.384) \\ &= 0.7728 - 0.3505 \\ &= 0.4223.\end{aligned}$$

Zum Vergleich: Der exakte Wert ist 0.4222, die Approximation sollte also für alle praktischen Fälle ausreichen. Überraschend ist, dass die Erfolgsanzahl mit über 40 Prozent Wahrscheinlichkeit in dem vergleichsweise winzigen Intervall  $\{1490, \dots, 1520\}$  zu finden sind. Immerhin gibt es ja positive Wahrscheinlichkeiten für alle Anzahlen zwischen 0 und 3000.

**5.** In diesem Beispiel geht es um ein Bernoulliexperiment mit  $p = 0.2$ . Wenn man das 400 Mal abfragt, sind im Mittel  $0.2 \cdot 400 = 80$  Erfolge zu erwarten. Wie wahrscheinlich ist es, dass es mindestens 100 Erfolge gibt?

Als praktische „Verkleidung“ könnte man sich eine Telefonzentrale vorstellen, die 400 Anschlüsse versorgt. Jeder einzelne Anschluss ist mit Wahrscheinlichkeit 0.2 aktiv, und wir wollen wissen, ob eine Kapazität von 100 Weiterleitungen ausreicht: Mit welcher Wahrscheinlichkeit wird es gleichzeitig mehr als 100 Vermittlungswünsche geben?

Die Rechnung ist einfach:

$$\begin{aligned}x(100 - 0.5) &= \frac{100 - 0.5 - 0.2 \cdot 400}{\sqrt{400 \cdot 0.2 \cdot 0.8}} = 2.43, \\ x(400 + 0.5) &= \frac{400 + 0.5 - 0.2 \cdot 400}{\sqrt{400 \cdot 0.2 \cdot 0.8}} = 40.06.\end{aligned}$$

die gesuchte Wahrscheinlichkeit ist also durch

$$\phi(40.06) - \phi(2.43) = 1 - 0.9925 = 0.0075$$

zu approximieren. Das ist weniger als ein Prozent, 100 Telefonleitungen sollten also ausreichen. (Der exakte Wert ist übrigens 0.0086. Obwohl der relative Fehler also deutlich größer ist als im vorigen Beispiel, ist die Approximation praktisch immer noch gut brauchbar.)

**6.** Das Phänomen aus dem vorstehenden Beispiel 4 soll noch näher untersucht werden: Wie stark ist die Erfolgsanzahl um den Erwartungswert herum konzentriert? Zum Beispiel kommt bei 6000 Würfelwürfen die Drei im Mittel 1000 Mal vor. In welchem Bereich um diesen Wert herum darf man denn das konkrete Ergebnis mit mindestens 90 Prozent Wahrscheinlichkeit erwarten? Anders gefragt: Für welches  $k$  ist bei der zugehörigen Binomialverteilung

$$b(1000-k, 6000; 1/6) + b(1000-k+1, 6000; 1/6) + \dots + b(1000+k, 6000; 1/6) \approx 0.9.$$

Wenn wir mit dem Satz von de Moivre-Laplace approximieren, muss für das unbekannte  $k$  gelten:

$$\phi(x_k) - \phi(-x_k) = 0.9, \text{ wobei } x_k = \frac{k + 0.5}{\sqrt{6000 \cdot (\frac{1}{6}) \cdot (\frac{1}{6})}}.$$

Da aus Symmetriegründen  $\phi(-x_k) = 1 - \phi(x_k)$  gilt, bedeutet das  $2\phi(x_k) - 1 = 0.9$ , d.h.  $\phi(x_k) = 0.95$ . Aus der Tafel der Werte der Normalverteilung lesen wir damit  $x_k = 1.65$  ab. Und nun ist nur noch

$$1.65 = \frac{k + 0.5}{\sqrt{6000 \cdot (\frac{1}{6}) \cdot (\frac{1}{6})}}$$

nach  $k$  aufzulösen. Wir erhalten  $k = 47.13\dots$ , und da  $k$  ganzzahlig sein soll, lautet das Endergebnis  $k = 48$ : Mit über 90 Prozent Wahrscheinlichkeit liegt die Anzahl der gewürfelten Dreien bei 6000 Versuchen zwischen 952 und 1048.

## 5.5 Verständnisfragen

### Zu Abschnitt 5.1

#### *Sachfragen*

**S1:** Welche Problemstellung liegt der Binomialverteilung zugrunde?

**S2:** Warum wurde sie nicht schon in Kapitel 2 vorgestellt?

**S3:** Wie lautet die Formel für die Wahrscheinlichkeiten der Binomialverteilung?

**S4:** Welche Problemstellung führt zur negativen Binomialverteilung?

**S5:** Wie lautet die Formel für die Wahrscheinlichkeiten der negativen Binomialverteilung.

#### *Methodenfragen*

**M1:** Einfache Probleme im Zusammenhang mit der Binomialverteilung – zum Beispiel maximum-likelihood-Schätzungen für die auftretenden Parameter – behandeln können.

### Zu Abschnitt 5.2

#### *Sachfragen*

**S1:** Unter welchen Voraussetzungen kann man die hypergeometrische Verteilung durch eine Binomialverteilung approximieren?

### Zu Abschnitt 5.3

#### *Sachfragen*

**S1:** In welchen Situationen ist die Poissonverteilung eine gute Approximation an die Binomialverteilung? Wie muss man den Parameter  $\lambda$  der Poissonverteilung dann wählen?

**S2:** Warum heißt die Poissonverteilung auch die Wahrscheinlichkeit der seltenen Ereignisse?

**S3:** Nennen Sie drei typische Beispiele, in denen die Modellierung durch eine Poissonverteilung sinnvoll ist. Welchen Wert sollte man in diesen Fällen für den Parameter  $\lambda$  einsetzen?

**S4:** Wie kommt es, dass die Eulersche Zahl  $e$  hier auftritt?

*Methodenfragen*

**M1:** Approximationen der Binomialverteilung durch die Poissonverteilung berechnen können.

**M2:** Entscheiden können, ob eine Modellierung durch die Poissonverteilung ge- rechtfertigt ist. Das zugehörige  $\lambda$  ermitteln können.

### Zu Abschnitt 5.4

*Sachfragen*

**S1:** Was besagt die Stirlingformel?

**S2:** Was versteht man unter dem Wallisprodukt?

**S3:** Was besagt der Satz von de Moivre-Laplace?

*Methodenfragen*

**M1:** Den Satz von de Moivre-Laplace anwenden können.

## 5.6 Übungsaufgaben

### Zu Abschnitt 5.1

**Ü5.1.1** Finden Sie eine maximum-likelihood-Schätzung für  $p$  bei gegebenen  $n, k$  und eine für  $n$  bei gegebenem  $p, k$  im Fall der Binomialverteilung. (Ausführlich: Wenn  $n$  und  $k$  fest vorgegeben sind, für welches  $p$  wird dann  $b(n, k; p)$  maximal? Wenn  $p$  und  $k$  fest vorgegeben sind, für welches  $n$  wird dann  $b(n, k; p)$  maximal?)

**Ü5.1.2** Ein Graphologe wird getestet. Ihm werden acht Paare von Schriftproben vorgelegt, jeweils von einem Arzt und einem Juristen geschrieben. Der Graphologe soll eingestellt werden, wenn er in mindestens sechs Fällen die richtige Zuordnung herausfindet. Wenn die fachliche Erfahrung des Graphologen so beschaffen ist, dass er im Mittel in 80% der Fälle richtig liegt, wie groß ist die Wahrscheinlichkeit, dass er nach diesem Test eingestellt wird?

**Ü5.1.3** Beweisen Sie, dass stets  $b(k, n; p) = b(n - k, n; 1 - p)$  gilt.

**Ü5.1.4**  $X$  und  $Y$  seien unabhängige Zufallsvariable.  $X$  (bzw.  $Y$ ) sei binomialverteilt zu den Parametern  $n$  und  $p$  (bzw.  $m$  und  $p$ ). Zeigen Sie, dass  $X + Y$  unter diesen Voraussetzungen binomialverteilt zu den Parametern  $n + m$  und  $p$  ist.

**Zu Abschnitt 5.2**

**Ü5.2.1** Überzeugen Sie sich durch Nachrechnen davon, dass die Voraussetzungen, unter denen wir die hypergeometrische Verteilung durch eine Binomialverteilung approximiert haben, wesentlich sind:  $k$  klein gegen  $r$ ,  $m$  klein gegen  $n$ ,  $m - k$  klein gegen  $n - r$ .

**Zu Abschnitt 5.3**

**Ü5.3.1** Im Mittel bekommt Herr H. 5 Weihnachtskarten. Wie wahrscheinlich ist es, dass er in diesem Jahr weniger als 2 bekommt?

**Ü5.3.2** Im Mittel gibt es im Büro des Professors P. zwischen 10 und 11 Uhr vier Anrufe. Wie wahrscheinlich ist es, dass es an einem speziellen Tag in dieser Zeit höchstens einen Anruf gibt?

**Ü5.3.3**  $X$  und  $Y$  seien poissonverteilt. Beweisen Sie, dass  $X+Y$  nicht notwendig ebenfalls poissonverteilt sein muss. (Falls  $X, Y$  unabhängig sind, stimmt das: vgl. Seite 152).

**Zu Abschnitt 5.4**

**Ü5.4.1** Die Zufallsvariable  $X$  sei binomialverteilt zum Parameter  $p = 0.4$  auf  $\{0, \dots, 1000\}$ .

a) Wie groß ist der Erwartungswert von  $X$ ?

b) Bestimmen Sie  $\mathbb{P}(300 < X \leq 600)$  approximativ mit Hilfe des Satzes von de Moivre-Laplace.

c) Finden Sie (wieder mit dem Satz von de Moivre-Laplace) ein möglichst kleines  $\alpha \in \mathbb{N}$ , so dass  $P(X \in [\mathbb{E}(X) - \alpha, \mathbb{E}(X) + \alpha]) \geq 0.8$  ist.

**Ü5.4.2** Zwei Spieler A und B werfen eine Münze 10.001 Mal. Zeigt sich dabei „Kopf“ öfter als „Zahl“, so gewinnt A, anderenfalls B. Spieler A, dem der Ausgang zu ungewiss ist, versucht, die Münze so zu fälschen, dass sie mit Wahrscheinlichkeit  $p > 1/2$  Kopf zeigt. Wie groß muss  $p$  sein, um mit einer Wahrscheinlichkeit von 95% zu gewinnen?

(Es reicht, die Wahrscheinlichkeiten approximativ mit dem Satz von de Moivre-Laplace zu bestimmen.)

# Kapitel 6

## Die Exponentialverteilung

In diesem Kapitel machen wir uns Gedanken über das Warten. Die zugehörigen Wahrscheinlichkeiten können sehr unterschiedlich sein. Eine besondere Rolle spielen Wartezeiten, bei denen „man nie genau weiß, wie lange es noch dauern wird“. Was das genau bedeutet, wird in *Abschnitt 6.1* als „gedächtnislose Wartezeiten“ präzisiert. Es wird dort auch gezeigt, dass die fragliche Eigenschaft genau bei den *Exponentialverteilungen* erfüllt ist.

Es ist leicht, Beispiele dafür zu finden, dass Summen, Maxima und Minima von Wartezeiten interessant sein können. Im Fall gedächtnisloser Wartezeiten gibt es explizite Formeln, die beweisen wir in *Abschnitt 6.2*. Schließlich zeigen wir in *Abschnitt 6.3*, dass die diskrete Variante des gedächtnislosen Wartens auf die geometrische Verteilung führt. Das wird auch eine Interpretation des gedächtnislosen Wartens in kontinuierlicher Zeit als „Warten auf den ersten Erfolg bei fast verschwindender Erfolgswahrscheinlichkeit“ ermöglichen.

In den *Abschnitten 6.4 und 6.5* gibt es dann noch Verständnisfragen und Übungsaufgaben.

### 6.1 Gedächtnislose Wartezeiten

Es gibt zahlreiche Beispiele im täglichen Leben, in denen wir auf etwas warten: Wir warten auf den nächsten Regen; oder darauf, dass die nächste U-Bahn kommt; oder darauf, dass das Telefon bei unserem besten Freund bzw. unserer besten Freundin nicht mehr besetzt ist usw.

Doch der Charakter des Wartens kann ganz unterschiedlich sein:

- Beispiel 1, der zuverlässige Handwerker. Er hat zugesagt, irgendwann zwischen 5 und 6 zu kommen. Um 5 Uhr erwarten wir ihn gleichverteilt innerhalb der nächsten Stunde, wir kalkulieren eine mittlere Wartezeit von 30 Minuten ein. Wenn er um 5.30 Uhr noch nicht da ist, schrumpft der Erwartungswert der Wartezeit auf 15 Minuten.

Kurz: Die Zeitspanne, die wir schon gewartet haben, verändert die Erwartung für das zukünftige Warten.

- Beispiel 2, die besetzte Telefonleitung. Sie rufen bei Ihrer Bank (oder Krankenkasse, Finanzamt, Gehaltsstelle, ...) an, es ist besetzt. Im Mittel wird die Leitung nach drei Minuten wieder frei sein, aber im Einzelfall kann man es nie genau wissen. Je nachdem, wie aufwändig die Beratung im Einzelfall ist, kann es auf unvorhersehbare Weise schneller oder langsamer gehen.

Kurz: Auch wenn man schon eine Weile gewartet hat, verändert das die Prognose für die noch zu erwartende Wartezeit nicht.

Das wollen wir nun präzisieren. Zunächst fassen wir den Begriff „Wartezeit“ recht weit: Jede auf irgendeinem Wahrscheinlichkeitsraum definierte Zufallsvariable  $T : \Omega \rightarrow [0, +\infty[$  kann als Wartezeit interpretiert werden. Damit kann man sehr unterschiedliche Wartezeit-Phänomene beschreiben. Zum Beispiel führt der vor wenigen Zeilen beschriebene Handwerker auf ein  $T$ , bei dem das durch  $T$  induzierte Wahrscheinlichkeitsmaß  $\mathbb{P}_T$  die Gleichverteilung auf  $[0, 1]$  ist: In diesem Fall misst  $T$  wirklich eine Zeit. Beim Warten auf die erste Sechs beim Würfeln ist  $T$  eine Anzahl (die Wahrscheinlichkeiten  $\mathbb{P}(\{T = n\})$  werden wir in Abschnitt 6.3 ausrechnen).  $T$  kann aber auch eine Länge sein, z.B. auf der Autobahn der Abstand zum nächsten Pannenfahrzeug oder auf einer Stoffbahn der Abstand zum nächsten Webfehler.

Die Tatsache, dass – wie im Telefonbeispiel – schon absolviertes vergebbliches Warten keinen Einfluss auf die zukünftigen Wahrscheinlichkeiten der Wartezeit hat, kann unter Verwendung des Begriffs der bedingten Wahrscheinlichkeit wie folgt präzisiert werden:

**Definition 6.1.1.** Es sei  $T$  eine Wartezeit, d.h.  $T : \Omega \rightarrow [0, +\infty[$  ist eine beliebige Zufallsvariable. Wir sagen, dass  $T$  gedächtnislos ist, wenn

$$\mathbb{P}(\{T \geq t\}) = \mathbb{P}(\{T \geq s + t\} \mid \{T \geq s\})$$

für beliebige  $s, t \geq 0$  gilt.

In Worten: Die Wahrscheinlichkeit, dass wir – ab sofort gerechnet – noch mindestens  $t$  Zeiteinheiten warten müssen, ist identisch mit der Wahrscheinlichkeit, dass wir noch weitere  $t$  Zeiteinheiten vor uns haben, wenn schon  $s$  Zeiteinheiten mit Warten vergangen sind.

Setzt man die Formel für bedingte Wahrscheinlichkeiten ein, so folgt sofort, dass  $T$  genau dann gedächtnislos ist, wenn stets

$$\mathbb{P}(\{T \geq s + t\}) = \mathbb{P}(\{T \geq s\}) \mathbb{P}(\{T \geq t\})$$

gilt.

Welche Wartezeiten haben diese Eigenschaft? Zur Vorbereitung der Charakterisierung beweisen wir ein analytisches

**Lemma 6.1.2.** Es sei  $F : [0, +\infty[ \rightarrow [0, +\infty[$  eine Funktion. Wir setzen voraus:

(i)  $F(0) = 1$ , und  $F$  ist monoton fallend.

(ii)  $\lim_{t \rightarrow \infty} F(t) = 0$ .

(iii)  $F$  ist von links stetig<sup>1)</sup>.

(iv) Es gibt ein  $t_0 > 0$  mit  $F(t_0) > 0$ .

Falls dann  $F(s+t) = F(s)F(t)$  für alle  $s, t \geq 0$  gilt, so gibt es ein  $\lambda > 0$  mit der Eigenschaft: Es ist  $F(t) = e^{-\lambda t}$  für alle  $t$ .

**Beweis:** Wir werden gleich ausnutzen, dass die Bedingung  $F(s+t) = F(s)F(t)$  weitreichende Konsequenzen hat. Es ist dann nämlich auch  $F(s_1 + \dots + s_n) = F(s_1) \cdots F(s_n)$ . (Das ist durch einen Induktionsbeweis schnell einzusehen.)

Zur Abkürzung schreiben wir  $\alpha := F(1)$ , diese Zahl wird gleich eine wichtige Rolle spielen. Angenommen, es wäre  $\alpha = 0$ . Für jedes  $n$  wäre dann auch  $F(1/n) = 0$ , denn aus der Voraussetzung würde  $0 = F(1) = F(1/n + \dots + 1/n) = (F(1/n))^n$  folgen. Da  $F$  monoton fällt, wäre  $F(t)$  für alle positiven  $t$  gleich Null im Widerspruch zur Voraussetzung (iv).

$\alpha \geq 1$  ist auch nicht möglich, denn dann wäre  $F(n) = F(1+\dots+1) = \alpha^n \geq 1$ , und das widerspricht Voraussetzung (ii). Also liegt  $\alpha$  in  $]0, 1[$ .

Nun zeigen wir, dass  $F$  bei allen  $t$  auch von rechts stetig ist. Zunächst behandeln wir den Fall  $t = 0$ . Dazu müssen wir beweisen, dass  $\lim_{t \rightarrow 0} F(t) = 1$  gilt. Da  $F$  monoton fällt, ist nur zu zeigen, dass es für jedes  $\beta < 1$  ein  $\varepsilon > 0$  gibt, so dass  $F(s) \geq \beta$  für  $s \leq \varepsilon$ .

Angenommen, das wäre nicht der Fall. Dann gäbe es ein  $\beta < 1$ , so dass wir für jedes  $n$  ein  $s_n$  mit  $0 < s_n < 1/n$  und  $F(s_n) < \beta$  finden könnten.

Dann würde  $F(n \cdot s_n) = F(s_n + \dots + s_n) \leq \beta^n$  gelten. Für genügend große  $n$  wäre dann  $\beta^n < \alpha$ , und das kann nicht sein, da  $n \cdot c_n$  kleiner als 1 ist und  $F$  monoton fallend sein soll.

Es folgt nun leicht die Stetigkeit von rechts bei allen  $t$ . Ist nämlich  $(s_n)$  monoton fallend mit  $\lim s_n = t$ , so ist auch  $s_n - t$  monoton fallend, und diese Folge konvergiert gegen 0. Nach dem vorigen Beweisteil gilt also  $F(s_n - t) \rightarrow 1$ , und daraus schließen wir

$$F(s_n) = F(t + (s_n - t)) = F(t)F(s_n - t) \rightarrow F(t).$$

Nun sind wir gleich fertig. Wir setzen  $\lambda := -\log \alpha$ , das ist unser Kandidat. Um zu zeigen, dass stets  $F(t) = e^{-\lambda t}$  gilt, argumentieren wir so:

1. Die Gleichheit gilt nach Definition von  $\lambda$  für  $t = 1$ .
2. Da sowohl  $F$  als auch  $t \mapsto e^{-\lambda t}$  Summen in Produkte überführen, folgt aus  $F(t) = e^{-\lambda t}$  auch  $F(n \cdot t) = e^{-\lambda n t}$  für jedes  $n \in \mathbb{N}$ .

---

<sup>1)</sup>D.h. für jedes  $t > 0$  und jede monoton steigende Folge  $(s_n)$  mit  $\lim_n s_n = t$  ist  $\lim_n F(s_n) = F(t)$ .

3. Die Funktionen stimmen aber dann auch bei  $t/m$  überein. Es ist nämlich

$$(F(t/m))^m = F(t/m + \dots + t/m) = F(t) = e^{-\lambda t} = (e^{-\lambda t/m})^m,$$

und da  $F(t(m))$  und  $e^{-\lambda t/m}$  in  $[0, \infty[$  liegen, muss  $F(t/m) = e^{-\lambda t/m}$  gelten.

4. Damit haben  $F$  und  $t \mapsto e^{-\lambda t}$  auf den positiven rationalen Zahlen die gleichen Werte, und da diese Menge dicht in  $[0, +\infty[$  liegt und beide Funktionen stetig sind, gilt die Gleichheit überall.

Wir bemerken noch, dass alle Funktionen des Typs  $e^{-\lambda t}$  mit  $\lambda > 0$  die Eigenschaften (i) – (iv) haben. Sie sind also dadurch charakterisiert.  $\square$

Mit dieser Vorbereitung ist der Charakterisierungssatz leicht beweisbar:

**Satz 6.1.3.** (i) Es sei  $T$  eine Wartezeit, für die  $\mathbb{P}_T$  exponentialverteilt zu irgendeinem Parameter  $\lambda > 0$  ist. Dann ist  $T$  gedächtnislos.

(ii) Sei umgekehrt  $T$  eine gedächtnislose Wartezeit. Wir setzen voraus, dass  $T$  im folgenden Sinn nichttrivial ist: Es gibt ein  $t_0 > 0$  mit  $\mathbb{P}(\{T \geq t_0\}) > 0$ . ( $T$  ist also nicht die Wartezeit, bei der  $\{T = 0\}$  Wahrscheinlichkeit Eins hat.) Dann gibt es ein  $\lambda > 0$ , so dass  $\mathbb{P}_T$  exponentialverteilt zum Parameter  $\lambda$  ist: Es ist

$$\mathbb{P}(\{a \leq T \leq b\}) = \lambda \int_a^b e^{-\lambda x} dx$$

für alle Intervalle  $[a, b]$ .

**Beweis:** (i) ist leicht einzusehen: Für exponentialverteilte  $T$  ist

$$\mathbb{P}(\{T \geq t\}) = \lambda \int_t^\infty e^{-\lambda x} dx = e^{-\lambda t},$$

und deswegen gilt offensichtlich

$$\mathbb{P}(\{T \geq s+t\}) = \mathbb{P}(\{T \geq s\}) \mathbb{P}(\{T \geq t\}).$$

Zum Beweis von (ii) definieren wir eine Funktion  $F : [0, +\infty[ \rightarrow [0, +\infty[$  durch  $F(t) := \mathbb{P}(\{T \geq t\})$  für  $t \geq 0$ . Dann erfüllt  $F$  die vier Bedingungen des vorigen Lemmas: (i) ist klar. Für (ii) nutzen wir Satz 1.3.2 aus: Geht  $(t_n)$  monoton gegen  $+\infty$ , so folgt

$$\begin{aligned} 0 &= \mathbb{P}(\emptyset) \\ &= \mathbb{P}\left(\bigcap_n \{T \geq t_n\}\right) \\ &= \lim_n \mathbb{P}(\{T \geq t_n\}) \\ &= \lim_n F(t_n). \end{aligned}$$

Ganz analog zeigt man die Stetigkeit von links bei allen  $t$ , sie folgt sofort aus  $[t, +\infty[ = \bigcap_n [s_n, +\infty[$  für Folgen  $(s_n)$ , die von unten monoton gegen  $t$  konvergieren. Eigenschaft (iv) ist vorausgesetzt, und  $F(s+t) = F(s)F(t)$  gilt wegen der Gedächtnislosigkeit.

Es gibt also ein positives  $\lambda$ , so dass

$$\mathbb{P}(\{T \geq t\}) = \lambda \int_t^{+\infty} e^{-\lambda x} dx$$

für alle  $t$  ist. Da  $F$  stetig ist, haben Mengen der Form  $\{T = t\}$  das Maß Null. So folgt

$$\begin{aligned}\mathbb{P}(\{T \in [a, b]\}) &= \mathbb{P}(\{T \in [a, b[\}) \\ &= \mathbb{P}(\{T \geq a\} \setminus \{T \geq b\}) \\ &= \mathbb{P}(\{T \geq a\}) - \mathbb{P}(\{T \geq b\}) \\ &= \lambda \int_a^{+\infty} e^{-\lambda x} dx - \lambda \int_b^{+\infty} e^{-\lambda x} dx \\ &= \lambda \int_a^b e^{-\lambda x} dx.\end{aligned}$$

Damit ist der Satz vollständig bewiesen.  $\square$

Der Erwartungswert einer Exponentialverteilung zum Parameter  $\lambda$  ist gleich  $1/\lambda$  (vgl. Seite 83). Das führt zu der

#### **Faustregel zum Arbeiten mit gedächtnislosen Wartezeiten**

Angenommen, es geht um eine Wartezeit  $T$ , und aus irgendwelchen Gründen ist es plausibel anzunehmen, dass sie gedächtnislos ist (Beratung, besetztes Telefon usw.)

Wenn dann  $T$  im Mittel den Wert  $\mu$  hat, so kann  $T$  durch eine Exponentialverteilung mit Parameter  $\lambda := 1/\mu$  modelliert werden.

Hier ein *Beispiel*: Wenn Sie bei Ihrer besten Freundin anrufen, ist es oft besetzt, und im Mittel dauert es 20 Minuten, bis der Anschluss wieder frei ist. Solche Situationen, bei denen es „ungewiss lange“ dauert, können gut durch gedächtnislose Wartezeiten modelliert werden. Aufgrund der vorstehenden Faustregel kann man zur Modellierung eine Exponentialverteilung mit dem Parameter  $\lambda = 1/20$  wählen.

Und nun können alle interessierenden Wahrscheinlichkeiten berechnet werden. Wie wahrscheinlich ist es zum Beispiel, dass die Leitung schon in 10 Minuten wieder frei ist? Die Lösung:

$$\mathbb{P}(\{T \leq 10\}) = \frac{1}{20} \int_0^{10} e^{-t/20} dt = -e^{-t/20} \Big|_0^{10} = 1 - e^{-0.5} \approx 0.39.$$

Und wie lange sollte man sich gedulden, damit man eine 50-prozentige Chance dafür hat, dass die Leitung frei ist? Bezeichnet man die fragliche Zeit mit  $t_0$ , so muss  $(1/20) \int_0^{t_0} e^{-t/20} dt = 1 - e^{-t_0/20} = 0.5$  gelten.

Folglich ist  $t_0 = -20 \log(0.5) \approx 13.86$ . Für eine 90-prozentige Sicherheit ergibt sich analog  $t_0 = -20 \log(0.1) \approx 46.05$ .

## 6.2 Kombinationen gedächtnisloser Wartezeiten

Es gibt verschiedene Situationen, in denen mehrere Wartezeiten kombiniert werden müssen:

- *Summen:* Als Beispiel denke man an das Warten im Wartezimmer des Hausarztes. Wenn dort  $n$  Patienten mit Behandlungszeiten  $T_1, \dots, T_n$  vor einem selbst ins Sprechzimmer wollen, wird die eigene Wartezeit durch die Zufallsvariable  $T_1 + \dots + T_n$  bestimmt.

Oder: Sie haben für Ihre Stehlampe fünf Glühbirnen auf Vorrat gekauft. (Kurz danach wurde der Verkauf durch eine EU-Richtlinie verboten.) Wenn die Zufallsvariable  $T_i$  die Lebensdauer der  $i$ -ten Glühbirne beschreibt, kann die Verteilung von  $T_1 + T_2 + T_3 + T_4 + T_5$  interessant sein. Wie wahrscheinlich ist es zum Beispiel, dass man die Lampe auch in sieben Jahren noch nutzen kann?

- *Maxima:* Wenn sich drei Freunde  $F_1, F_2, F_3$  auf eine Klausur vorbereiten und die Vorbereitungszeit von  $F_i$  als Wartezeit  $T_i$  aufgefasst wird, so ist die Zufallsvariable  $\max\{T_1, T_2, T_3\}$  die Wartezeit, bis sich alle drei fit für die Klausur fühlen.

Oder: Wenn  $T_1$  bzw.  $T_2$  die Lebensdauer Ihrer Fahrrad-Lampe bzw. Ihres Rücklichts misst, so kann man an  $\max\{T_1, T_2\}$  ablesen, ab wann Sie ganz ohne Beleuchtung fahren müssen.

- *Minima:* Angenommen, alle Parkplätze vor dem Mathe-Gebäude sind besetzt. Es gibt  $n$  Parkplätze, und die Wartezeit, bis der  $i$ -te frei wird, sei  $T_i$ . Es ist dann interessant zu wissen, wie sich die Zufallsvariable  $\min\{T_1, \dots, T_n\}$  verhält, denn sie ist dafür verantwortlich, auf welche Wartezeit man sich einstellen muss, um hier parken zu können.

Oder: Wenn ein elektronisches Gerät aus  $n$  wichtigen Bauteilen mit den Lebensdauern  $T_1, \dots, T_n$  besteht, so beschreibt  $\min\{T_1, \dots, T_n\}$  die Wartezeit, bis es nicht mehr funktionieren wird.

Wir wollen nun zeigen, dass es im Fall gedächtnisloser Wartezeiten für diese drei Kombinationen explizite Beschreibungen gibt.

Summen

**Satz 6.2.1.**  $T_1, \dots, T_n$  seien Zufallsvariable auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{E}, \mathbb{P})$ . Wir setzen voraus, dass sie unabhängig und exponentialverteilt zu den Parametern  $\lambda_1, \dots, \lambda_n$  sind.

(i) Ist  $\lambda_1 \neq \lambda_2$ , so hat  $\mathbb{P}_{T_1+T_2}$  auf  $[0, +\infty[$  die Dichtefunktion

$$h(x) = \frac{\lambda_1 \lambda_2}{\lambda_1 - \lambda_2} (e^{-\lambda_2 x} - e^{-\lambda_1 x}).$$

(ii) Gilt  $\lambda_1 = \dots = \lambda_n =: \lambda$ , so hat  $\mathbb{P}_{T_1+\dots+T_n}$  die Dichtefunktion

$$h_n(x) = \frac{\lambda^n x^{n-1}}{(n-1)!} e^{-\lambda x}.$$

**Beweis:** (i) Wir wenden Satz 4.6.3 an, danach ist die Dichtefunktion  $h$  von  $\mathbb{P}_{T_1+T_2}$  die Faltung der Dichtefunktionen von  $\mathbb{P}_{T_1}$  und  $\mathbb{P}_{T_2}$ , also von  $f_1(x) := \lambda_1 e^{-\lambda_1 x}$  und  $f_2(x) := \lambda_2 e^{-\lambda_2 x}$ :

$$\begin{aligned} h(x) &= (f_1 * f_2)(x) \\ &= \int_0^x f_1(y) f_2(x-y) dy \\ &= \lambda_1 \lambda_2 \int_0^x e^{-\lambda_1 y} e^{-\lambda_2(x-y)} dy \\ &= \lambda_1 \lambda_2 e^{-\lambda_2 x} \int_0^x e^{-(\lambda_1 - \lambda_2)y} dy \\ &= \lambda_1 \lambda_2 e^{-\lambda_2 x} \left( -\frac{e^{-(\lambda_1 - \lambda_2)y}}{\lambda_1 - \lambda_2} \Big|_0^x \right) \\ &= \lambda_1 \lambda_2 e^{-\lambda_2 x} \left( -\frac{e^{-(\lambda_1 - \lambda_2)x}}{\lambda_1 - \lambda_2} + \frac{1}{\lambda_1 - \lambda_2} \right) \\ &= \frac{\lambda_1 \lambda_2}{\lambda_1 - \lambda_2} (e^{-\lambda_2 x} - e^{-\lambda_1 x}). \end{aligned}$$

(ii) Es sei  $f(x) := \lambda e^{-\lambda x}$ . Wir müssen zeigen, dass die  $n$ -fache Faltung von  $f$  mit sich gleich  $h_n$  ist. Das wird durch Induktion bewiesen: Für  $n = 1$  ist die Aussage klar, und für den Induktionsschluss ist nachzurechnen, dass  $h_n * f = h_{n+1}$  gilt. Wirklich ist

$$\begin{aligned} (h_n * f)(x) &= \int_0^x h_n(y) f(x-y) dy \\ &= \frac{\lambda^{n+1}}{(n-1)!} \int_0^x y^{n-1} e^{-\lambda y} e^{-\lambda(x-y)} dy \\ &= \frac{\lambda^{n+1} e^{-\lambda x}}{(n-1)!} \int_0^x y^{n-1} dy \\ &= \frac{\lambda^{n+1} x^n}{n!} e^{-\lambda x}. \end{aligned}$$

□

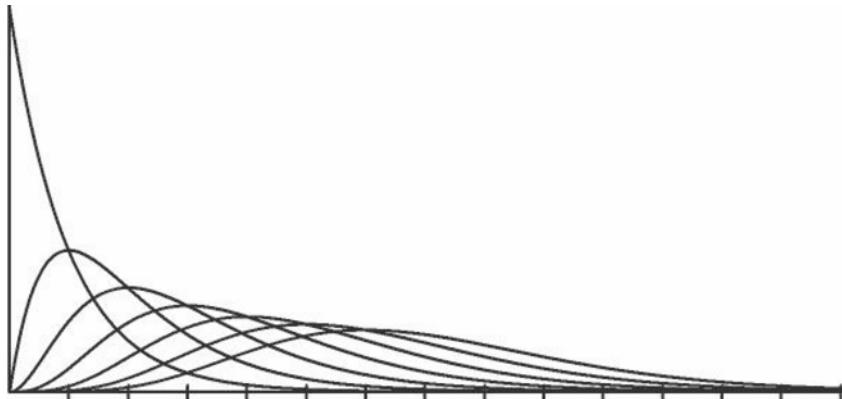


Bild 6.2.1: Dichten der Summen von  $n$  unabhängigen Exponentialverteilungen,  $\lambda = 1$ :  
 $n = 1, \dots, 7$ .

#### Bemerkungen:

1. Wie zu erwarten war, werden im vorstehenden Bild die Erwartungswerte und Streuungen immer größer: Für die Summe aus  $n$  unabhängigen Exponentialverteilungen ist der Erwartungswert  $n/\lambda$  und die Streuung gleich  $\sqrt{n}/\lambda$ .
2. Auch für beliebige  $\lambda_1, \dots, \lambda_n$  lässt sich die Dichte von  $\mathbb{P}_{T_1+\dots+T_n}$  explizit beschreiben. Die Formeln werden jedoch viel umübersichtlicher als in den hier behandelten Spezialfällen.
3. Der Satz zeigt, dass Summen exponentialverteilter Zufallsvariablen nicht exponentialverteilt sind. Das ist auch plausibel:

- Für exponentialverteilte Zufallsvariable  $T$  gibt es eine positive Konstante  $c$  (nämlich  $c = \lambda$ ), so dass  $\mathbb{P}(\{T \leq \varepsilon\}) \approx c\varepsilon$  für „kleine“  $\varepsilon$  gilt.
- Damit eine Summe  $T_1 + T_2$  klein ist, müssen  $T_1$  und  $T_2$  klein. Dafür sind die Wahrscheinlichkeiten durch  $c_1\varepsilon$  und  $c_2\varepsilon$  abschätzbar, und wenn man Unabhängigkeit voraussetzt, tritt beides gleichzeitig nur mit Wahrscheinlichkeit  $c_1c_2\varepsilon^2$  ein.

Und das ist, wie davor bemerkt, bei exponentialverteilten Wartezeiten nicht der Fall.

Als Folgerung aus dem Satz ergibt sich ein interessanter Zusammenhang zur Poissonverteilung:

**Korollar 6.2.2.** Es seien  $T_0, T_1, T_2, \dots$  unabhängige Wartezeiten, die alle exponentialverteilt zum Parameter 1 sind. Weiter sei  $\lambda > 0$ . Für jedes  $n \in \mathbb{N}_0$  gilt dann

$$\mathbb{P}(\{T_0 + \dots + T_n > \lambda\}) - \mathbb{P}(\{T_0 + \dots + T_{n-1} > \lambda\}) = \frac{\lambda^n}{n!} e^{-\lambda} (= p(n; \lambda)).$$

Das liefert ein Verfahren, die Poissonverteilung zu simulieren: Erzeuge unabhängige exponentialverteilte Zufallsvariable  $T_0, T_1, \dots$ , und zwar so lange, bis erstmals  $T_0 + \dots + T_n > \lambda$  gilt. Dieses  $n$  wird ausgegeben, die Wahrscheinlichkeit dafür ist  $p(n; \lambda)$ .

Konkret bedeutet das: Erzeuge so lange gleichverteilte  $x_i$  in  $[0, 1]$ , bis erstmals

$$-\log(x_0) - \log(x_1) - \dots - \log(x_n) > \lambda$$

gilt. Dieses  $n$  ist auszugeben.

Es geht sogar noch schneller, wenn man beachtet, dass die vorstehende Ungleichung gleichbedeutend mit  $x_0 \cdot x_1 \cdots x_n < e^{-\lambda}$  ist.

**Beweis:** Setze  $p_n := \mathbb{P}(\{T_0 + \dots + T_n > \lambda\})$ . Da  $T_0 + \dots + T_n$  eine Summe aus  $n+1$  unabhängigen exponentialverteilten Summanden ist, gilt aufgrund des vorstehenden Satzes

$$p_n = \frac{1}{n!} \int_{\lambda}^{+\infty} x^n e^{-x} dx.$$

Daraus folgt

$$\begin{aligned} p_n - p_{n-1} &= \frac{1}{n!} \int_{\lambda}^{+\infty} (x^n - nx^{n-1}) e^{-x} dx \\ &= \frac{1}{n!} \left( x^n e^{-x} \Big|_{\lambda}^{+\infty} \right) \\ &= \frac{\lambda^n}{n!} e^{-\lambda}. \end{aligned}$$

Für die Rechtfertigung der Simulationsvorschrift ist nur zu beachten, dass „Es gilt erstmals  $T_0 + \dots + T_n > \lambda$ “ gleichwertig zu „Es ist  $T_0 + \dots + T_n > \lambda$ , aber  $T_0 + \dots + T_{n-1} \leq \lambda$ “ ist und dass für gleichverteilte  $x$  in  $[0, 1]$  die Zahlen  $-\log(x)$  exponentialverteilt zum Parameter 1 sind (vgl. Seite 60).

Bei der Simulation der Poissonverteilung auf Seite 56 ist dieses Verfahren angewandt worden.  $\square$

### Maxima

Es sei  $(\Omega, \mathcal{E}, \mathbb{P})$  ein Wahrscheinlichkeitsraum, auf dem unabhängige  $[0, +\infty [-$ -wertige Zufallsvariable  $T_1, \dots, T_n$  definiert sind. Dabei soll  $T_i$  zum Parameter  $\lambda_i$  exponentialverteilt sein. Wir definieren  $T := \max\{T_1, \dots, T_n\}$ , und wir wollen wissen, ob  $\mathbb{P}_T$  eine Dichtefunktion  $h$  hat.

Für das gesuchte  $h : [0, +\infty[ \rightarrow [0, +\infty[$  muss doch gelten:

$$\mathbb{P}(\{T \in [a, b]\}) = \int_a^b h(x) dx$$

(alle  $[a, b]$ ). Wegen  $\int_a^b h(x) dx = \int_0^b h(x) dx - \int_0^a h(x) dx$  reicht es, ein  $h$  so zu bestimmen, dass stets  $\mathbb{P}(\{T \leq x\}) = \int_0^x h(t) dt$  gilt. Es folgt: Wenn wir  $\mathbb{P}(\{T \leq x\})$  als bekannte Funktion  $\phi(x)$  darstellen können, muss  $h$  nach dem Hauptsatz der Differential- und Integralrechnung gleich der Ableitung  $\phi'$  von  $\phi$  sein.

So eine Formel lässt sich herleiten, wenn man erstens bemerkt, dass  $T \leq x$  genau dann gilt, wenn  $T_i \leq x$  für alle  $i$  richtig ist, und wenn man sich zweitens daran erinnert, dass Wahrscheinlichkeiten für das gleichzeitige Eintreten unabhängiger Zufallsvariabler Produkte der Einzelwahrscheinlichkeiten sind.

Im vorliegenden Fall geht es um die Wahrscheinlichkeit von  $\{T_i \leq x\}$ , also um  $\lambda_i \int_0^x e^{-\lambda_i t} dt = 1 - e^{-\lambda_i x}$ . Es ist also

$$\phi(x) = (1 - e^{-\lambda_1 x}) \cdots (1 - e^{-\lambda_n x}),$$

und die Dichte für  $\mathbb{P}_T$  ist gleich  $\phi'$ . Das ist ein etwas unübersichtlicher Ausdruck. Für den Spezialfall, dass alle  $\lambda_i$  gleich einer Zahl  $\lambda$  sind, ist alles etwas einfacher. Wir formulieren das Ergebnis als

**Satz 6.2.3.** *Sind die  $T_i$  unabhängig und exponentialverteilt zum Parameter  $\lambda$  und setzt man  $T := \max\{T_1, \dots, T_n\}$ , so hat  $\mathbb{P}_T$  die Dichtefunktion*

$$h(x) = n\lambda e^{-\lambda x} (1 - e^{-\lambda x})^{n-1}.$$

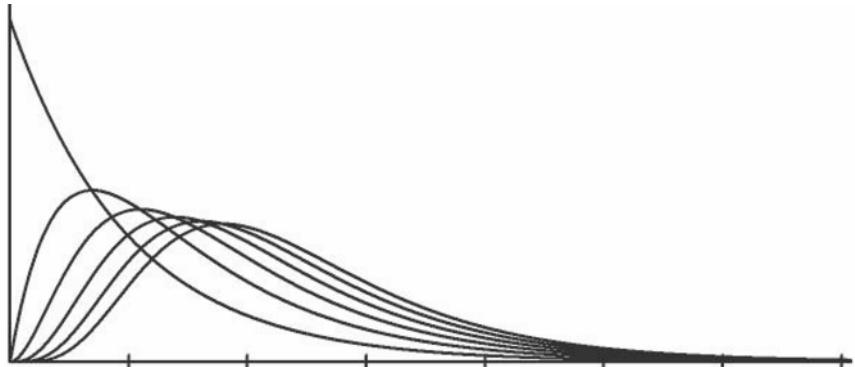


Bild 6.2.2: Dichten der Maxima von  $n$  unabhängigen Exponentialverteilungen,  $\lambda = 1$ :  
 $n = 1, \dots, 6$ .

Insbesondere folgt, dass das Maximum *nicht* wieder exponentialverteilt ist<sup>2)</sup>.

---

<sup>2)</sup>Das ist auch – wie im Fall der Summen – plausibel. Beachte nur, dass  $T$  genau dann klein ist, wenn alle  $T_i$  klein sind.

Was kann man denn über den Erwartungswert von  $T = \max\{T_1, \dots, T_n\}$  im Fall  $\lambda_1 = \dots = \lambda_n =: \lambda$  aussagen? Wegen  $T \geq T_1$  ist sicher  $\mathbb{E}(T) \geq \mathbb{E}(T_1) = 1/\lambda$ , doch wie groß kann er werden, wenn  $n$  gegen Unendlich geht? Wir behaupten, dass er *beliebig groß* werden kann. Dazu kombinieren wir die folgenden Tatsachen:

- Die vorstehend eingeführte Funktion  $\phi(x) = (1 - e^{-\lambda x})^n$  ist von der Form  $1 - e^{-\lambda x} g(x)$  mit einer beschränkten Funktion  $g$ , und deswegen konvergiert  $x\phi(x) - x$  für  $x \rightarrow +\infty$  gegen Null.
- Es ist  $h(x) = \phi'(x)$ , bei der Berechnung des Erwartungswerts der Zufallsvariablen  $\max\{T_1, \dots, T_n\}$  ist also  $\int_0^{+\infty} x\phi'(x) dx$  auszuwerten. Mit partieller Integration und der vorstehenden Bemerkung folgt

$$\begin{aligned}\mathbb{E}(\max\{T_1, \dots, T_n\}) &= \lim_{b \rightarrow +\infty} \int_0^b x\phi'(x) dx \\ &= \lim_{b \rightarrow +\infty} (x\phi(x)|_0^b - \int_0^b \phi(x) dx) \\ &= \lim_{b \rightarrow +\infty} (b\phi(b) - \int_0^b \phi(x) dx) \\ &= \lim_{b \rightarrow +\infty} (b - \int_0^b \phi(x) dx) \\ &= \lim_{b \rightarrow +\infty} \int_0^b (1 - \phi(x)) dx \\ &= \int_0^{+\infty} (1 - \phi(x)) dx.\end{aligned}$$

- Und dieses Integral kann mit der Substitution  $u := 1 - e^{-\lambda x}$  auf ein explizit lösbares Integral zurückgeführt werden:

$$\begin{aligned}\int_0^1 \frac{1}{\lambda} \frac{1 - u^n}{1 - u} du &= \int_0^1 (1 + u + \dots + u^{n-1}) du \\ &= \frac{1}{\lambda} \left(1 + \frac{1}{2} + \dots + \frac{1}{n}\right)\end{aligned}$$

Der Erwartungswert ist also das Produkt aus  $1/\lambda$  und einer Partialsumme der harmonischen Reihe. Daraus kann man folgern, dass die Erwartungswerte der  $\max\{T_1, \dots, T_n\}$  für  $n \rightarrow \infty$  gegen Unendlich konvergieren.

Durch das Ergebnis versteht man besser, warum es bei großem  $n$  manchmal ziemlich lange dauern kann, bis die längste der Wartezeiten  $T_1, \dots, T_n$  vorbei ist: Man denke an die Zeit, bis eine Touristengruppe wieder vollzählig am Bus ist. Oder daran, wie viel Zeit vergeht, bis sich alle in der Klasse für den Sportunterricht umgezogen haben.

Minima

**Satz 6.2.4.** Es sei  $(\Omega, \mathcal{E}, \mathbb{P})$  ein Wahrscheinlichkeitsraum, und  $T_1, \dots, T_n$  seien auf  $\Omega$  definierte unabhängige reellwertige Zufallsvariable. Ist dann  $T_i$  exponentialverteilt zum Parameter  $\lambda_i$  für  $i = 1, \dots, n$ , so ist  $T := \min\{T_1, \dots, T_n\}$  ebenfalls exponentialverteilt. Der zugehörige Parameter ist  $\lambda_1 + \dots + \lambda_n$ .

Der Erwartungswert von  $T$  ist folglich  $1/(\lambda_1 + \dots + \lambda_n)$ . Wenn alle  $T_i$  die gleiche Verteilung haben, wenn also  $\lambda_1 = \dots = \lambda_n =: \lambda$  gilt, fällt beim Übergang von  $T_1$  zu  $\min\{T_1, \dots, T_n\}$  der Erwartungswert von  $1/\lambda$  auf  $1/(n\lambda)$ .

**Beweis:** Die Beweistechnik ist ähnlich wie im Fall des Maximums. Diesmal spielt die Tatsache eine entscheidende Rolle, dass  $T$  genau dann größer oder gleich  $x$  ist, wenn das für alle  $T_i$  gilt.

Nun ist  $\mathbb{P}(\{T_i \geq x\}) = \lambda_i \int_x^{+\infty} e^{-\lambda_i t} dt = e^{-\lambda_i x}$ . Folglich ist wegen der Unabhängigkeit der  $T_i$

$$\mathbb{P}(\{T \geq x\}) = e^{-(\lambda_1 + \dots + \lambda_n)x} = (\lambda_1 + \dots + \lambda_n) \int_x^{+\infty} e^{-(\lambda_1 + \dots + \lambda_n)t} dt.$$

Das beweist, dass  $(\lambda_1 + \dots + \lambda_n)e^{-(\lambda_1 + \dots + \lambda_n)x}$  Dichtefunktion zu  $\mathbb{P}_T$  ist<sup>3)</sup>, d.h.  $T$  ist exponentialverteilt zum Parameter  $\lambda_1 + \dots + \lambda_n$ .  $\square$

### 6.3 Diskrete gedächtnislose Wartezeiten

In den vorigen Abschnitten haben wir Wartezeiten für  $[0, +\infty[$ -wertige Zufallsvariable behandelt. Was in diesem Fall „gedächtnislos“ heißen soll, wurde in Definition 6.1.1 festgesetzt.

Dieser Ansatz ist sinnvoll, wenn es um Zeiten oder Längen geht. Wenn die möglichen Werte aber diskret sind, wenn also etwa nur Werte in  $\mathbb{N}$  vorkommen wie beim Warten auf die erste Sechs beim Würfeln, muss die Definition sicher modifiziert werden. Aus dem Charakterisierungssatz 6.1.3 folgt nämlich, dass für gedächtnislose Wartezeiten  $T$  eine Dichte für  $\mathbb{P}_T$  existiert, und das ist im diskreten Fall nicht zu erwarten.

Man kann aber auch dann präzisieren, was „Warten verändert die zukünftige Wartesituation nicht“ bedeuten soll:

**Definition 6.3.1.** Es sei  $(\Omega, \mathcal{E}, \mathbb{P})$  ein Wahrscheinlichkeitsraum, und  $T : \Omega \rightarrow \mathbb{N}$  sei eine Zufallsvariable<sup>4)</sup>,  $T$  heißt gedächtnislos, wenn

$$\mathbb{P}(\{T > m\}) = \mathbb{P}(\{T > n + m\} \mid \{T > n\})$$

für alle  $n, m \in \mathbb{N}$  gilt.

---

<sup>3)</sup>Beachte, dass  $\int_a^b f(x) dx = \int_a^{+\infty} f(x) dx - \int_b^{+\infty} f(x) dx$  gilt: Deswegen reicht es, die Wahrscheinlichkeiten nur für die Intervalle  $[x, +\infty[$  zu bestimmen.

<sup>4)</sup>Dass die Bilder natürliche Zahlen sein sollen, ist nur eine Frage des Maßstabs. Analog geht man vor, wenn sie zum Beispiel in  $3\mathbb{N}$  oder in  $0.5\mathbb{N}$  liegen.

Eine Charakterisierung ist einfach:

**Satz 6.3.2.**  *$T : \Omega \rightarrow \mathbb{N}$  ist genau dann gedächtnislos, wenn  $T$  geometrisch verteilt ist: Es gibt ein  $q \in [0, 1]$  so,  $\mathbb{P}(\{T = n\}) = (1 - q)q^{n-1}$  für alle  $n \in \mathbb{N}$  gilt.*

**Beweis:** Ist  $T$  geometrisch verteilt, so ist

$$\mathbb{P}(\{T > n\}) = (q - 1)(q^n + q^{n+1} + \dots) = q^n.$$

Gedächtnislosigkeit ist aber äquivalent zu

$$\mathbb{P}(\{T > n + m\}) = \mathbb{P}(\{T > n\})\mathbb{P}(\{T > m\})$$

(alle  $n, m$ ), und das ist im vorliegenden Fall dann sicher erfüllt<sup>5)</sup>.

Ist umgekehrt  $T$  gedächtnislos, so setze  $\phi(n) := \mathbb{P}(\{T > n\})$  für  $n \in \mathbb{N}$ . Nach Voraussetzung liegen alle  $\phi(n)$  in  $[0, 1]$  und es gilt stets  $\phi(n+m) = \phi(n)\phi(m)$ . Mit  $q := \phi(1) \in [0, 1]$  ist dann  $\phi(n) = q^n$  für alle  $n$ . Damit sind wir auch schon fertig:

$$\mathbb{P}(\{T = n\}) = \phi(n-1) - \phi(n) = (1 - q)q^{n-1}$$

für alle  $n$ . □

Wir haben schon auf Seite 167 bemerkt, dass das Warten auf den ersten Erfolg bei einer unabhängigen Folge von Bernoulli-Experimenten mit Erfolgswahrscheinlichkeit  $p$  geometrisch verteilt mit Parameter  $q = 1 - p$  ist. Es ist nicht wirklich überraschend, dass wir es hier mit einer gedächtnislosen Wartezeit zu tun haben, denn die Unabhängigkeit der Experimente sollte doch garantieren, dass auch nach  $n$  Versuchen die Situation exakt so ist wie am Anfang.

Als Ergänzung kann noch einmal daran erinnert werden, dass die geometrische Verteilung Erwartungswert  $1/(1 - q)$  hat<sup>6)</sup>. Man muss also im Mittel  $1/p$  Versuche bis zum ersten Erfolg durchführen, wenn die Erfolgswahrscheinlichkeit  $p$  ist: Sechs Mal bis zur ersten Sechs beim Würfeln, 13.983.816 Mal bis zum ersten Sechser im Lotto usw.

Das kann man auch umgekehrt zum *Schätzen von  $p$*  heranziehen. Wie wahrscheinlich ist es zum Beispiel, dass ein zufällig ausgewählter Mensch in London rothaarig ist? Sie gehen dort durch die Straßen und stellen fest, dass so ungefähr jeder Zwanzigste, der Ihnen entgegenkommt, rote Haare hat. Dann schätzen Sie natürlich das fragliche  $p$  als  $1/20$ .

Und was passiert, wenn man diskrete gedächtnislose Wartezeiten kombiniert? Ähnlich wie im kontinuierlichen Fall kann man in allen Fällen die auftretenden Wahrscheinlichkeiten ermitteln, aber nur bei Minima ergeben sich wieder gedächtnislose Wartezeiten:

<sup>5)</sup>Genau genommen gilt der vorstehende Beweis nur für  $q < 1$ . Die Aussage ist aber auch – wie leicht zu sehen – im Fall  $q = 1$  richtig.

<sup>6)</sup>Vgl. Seite 82.

**Satz 6.3.3.**  $T_1, \dots, T_k : \Omega \rightarrow \mathbb{N}$  seien unabhängige geometrisch verteilte Zufallsvariable zu den Parametern  $q_1, \dots, q_k$ .

(i) Die Verteilung von  $T_1 + \dots + T_k$  kann durch Berechnung diskreter Faltungen leicht berechnet werden.

Sind alle  $q_i$  gleich einer Zahl  $q$ , so liegt die negative Binomialverteilung aus Abschnitt 5.1 vor.

(ii) Sei  $T' := \max\{T_1, \dots, T_k\}$ . Es ist

$$\mathbb{P}(\{T' = n\}) = (1 - q_1^n) \cdots (1 - q_k^n) - (1 - q_1^{n-1}) \cdots (1 - q_k^{n-1}).$$

Gilt  $q_1 = \dots = q_k =: q$  und ist  $q > 0$ , so wird der Erwartungswert von  $T$  für  $k$  gegen Unendlich beliebig groß.

(iii)  $T'' := \min\{T_1, \dots, T_k\}$  ist geometrisch verteilt zum Parameter  $q_1 \cdots q_k$ . Der Erwartungswert des Minimums ist folglich  $1/(1 - (q_1 \cdots q_k))$ .

**Beweis:** Für den Beweis von (i) ist nur an Satz 4.6.3 zu erinnern, und es ist zu beachten, dass  $T_1 + \dots + T_k$  gerade die Wartezeit bis zum  $k$ -ten Erfolg ist.

Der Beweis von (ii) beginnt wie im kontinuierlichen Fall mit der Bemerkung, dass  $T'$  genau dann kleiner oder gleich  $n$  ist, wenn  $T_i \leq n$  für alle  $i$  gilt. Es ist aber  $\mathbb{P}(\{T_i \leq n\}) = 1 - \mathbb{P}(\{T_i \geq n+1\}) = 1 - q_i^n$ . Deswegen ist  $\mathbb{P}(\{T' \leq n\}) = (1 - q_1^n) \cdots (1 - q_k^n)$ , und nun ist nur noch zu beachten, dass  $\mathbb{P}(\{T' = n\}) = \mathbb{P}(\{T' \leq n\}) - \mathbb{P}(\{T' \leq n-1\})$ .

Sind alle  $q_i = q$ , so vereinfacht sich die Formel zu

$$\mathbb{P}(\{T' = n\}) = (1 - q^n)^k - (1 - q^{n-1})^k,$$

der Erwartungswert hat also den Wert

$$\sum_{n=1}^{\infty} n \left[ (1 - q^n)^k - (1 - q^{n-1})^k \right].$$

Um zu zeigen, dass diese Zahl mit wachsendem  $k$  beliebig groß werden kann, setzen wir zur Abkürzung  $c_n^k := (1 - q^n)^k$ . Für festes  $k$  wachsen die  $c_n^k$ , und es ist  $c_0^k = 0$ . Wir summieren in der  $m$ -ten Partialsumme der Reihe für den Erwartungswert um:

$$\begin{aligned} \sum_{n=1}^m n(c_n^k - c_{n-1}^k) &= (c_1^k - c_0^k) + 2(c_2^k - c_1^k) + \dots + m(c_m^k - c_{m-1}^k) \\ &= -c_0^k - c_1^k - \dots - c_{m-1}^k + m c_m^k \\ &= (c_m^k - c_0^k) + (c_m^k - c_1^k) + \dots + (c_m^k - c_{m-1}^k). \end{aligned}$$

Nun sei  $r \in \mathbb{N}$  beliebig. Wähle  $k$  so groß, dass  $c_r^k \leq 0.25$ : Das ist wegen  $q < 1$  möglich, und es gilt dann auch  $c_n^k \leq 0.25$  für  $n = 1, \dots, r$ . Wähle weiter  $m$ ,

so dass für dieses  $k$  die Ungleichung  $c_m^k \geq 0.75$  gilt. Wir können die obige Abschätzung dann mit

$$\begin{aligned} \dots &\geq (c_m^k - c_0^k) + (c_m^k - c_1^k) + \dots + (c_m^k - c_r^k) \\ &\geq r(c_m^k - 0.25) \\ &\geq r/2 \end{aligned}$$

fortsetzen. Das zeigt, dass die Partialsummen für wachsende  $k$  beliebig groß werden können.

Der noch fehlende Beweis für (iii) ist sehr einfach. Es gilt  $T'' > n$  genau dann, wenn  $T_i > n$  für alle  $T_i$  gilt. Es ist aber  $\mathbb{P}(\{T_i > n\}) = q_i^n$ , aus der Unabhängigkeit folgt damit  $\mathbb{P}(\{T'' > n\}) = q_1^n \cdots q_k^n$ . Mit  $q := q_1 \cdots q_k$  ist damit

$$\mathbb{P}(\{T'' = n\}) = \mathbb{P}(\{T'' > n - 1\}) - \mathbb{P}(\{T'' > n\}) = q^{n-1}(1 - q)$$

wie behauptet.  $\square$

Am Ende dieses Abschnitts soll noch gezeigt werden, wie man gedächtnislose Wartezeiten im kontinuierlichen Fall als Grenzwerte diskreter gedächtnisloser Wartezeiten auffassen kann:

#### **Erster Erfolg bei infinitesimalen Chancen**

Es sei  $\lambda > 0$  und  $n$  „sehr groß“. Wir betrachten ein Bernoulli-Experiment mit der winzigen Erfolgswahrscheinlichkeit  $p := \lambda/n$ . Pro Zeiteinheit fragen wir  $n$  Mal ab, in  $t$  Zeiteinheiten also  $t \cdot n$  Mal. Die Wahrscheinlichkeit, dass wir bis zur Zeit  $t$  noch keinen Erfolg hatten, ist dann gleich  $(1-p)^{t \cdot n}$ . Da  $(1+x/n)^n$  durch  $e^x$  approximiert werden kann, ist diese Zahl in guter Näherung gleich  $e^{-\lambda t}$ , also gleich der Wahrscheinlichkeit, dass  $\mathbb{P}(\{T \geq t\})$  für eine zum Parameter  $\lambda$  exponentialverteilte Wartezeit  $T$ . Zusammengefasst heißt das:

*Gedächtnisloses Warten in kontinuierlicher Zeit kann dadurch interpretiert werden, dass man auf den ersten Erfolg bei den Abfragen eines Bernoulli-Experiment mit winziger Erfolgswahrscheinlichkeit wartet.*

## 6.4 Verständnisfragen

### Zu Abschnitt 6.1

#### *Sachfragen*

**S1:** Was ist eine Wartezeit?

**S2:** Wann heißt eine reellwertige Wartezeit gedächtnislos?

**S3:** Wie kann man gedächtnislose Wartezeiten charakterisieren?

**S4:** Ein Lemma aus der Analysis spielt eine wichtige Rolle: Es sei  $F$  von  $[0, +\infty[$  nach  $\mathbb{R}$  eine stetige Funktion, und für alle  $s, t$  gelte  $F(s+t) = F(s) + F(t)$ . Was weiß man dann über  $F$ ?

*Methodenfragen*

**M1:** Nachprüfen können, ob eine konkret gegebene Wartezeit gedächtnislos ist.

### Zu Abschnitt 6.2

*Sachfragen*

**S1:** Nennen Sie Beispiele aus dem „täglichen Leben“, bei denen Summen von Wartezeiten auftreten.

**S2:** Welche Verteilung haben Summen gedächtnisloser Wartezeiten?

**S3:** Finden Sie konkrete Beispiele, bei denen das Maximum von Wartezeiten auftritt.

**S4:** Welche Verteilung hat das Maximum von endlich vielen unabhängigen gedächtnisloser Wartezeiten?

**S5:** Bei welchen Beispielen aus dem „täglichen Leben“ tritt das Minimum von Wartezeiten auf?

**S6:** Welche Verteilung hat das Minimum von endlich vielen unabhängigen gedächtnisloser Wartezeiten?

*Methodenfragen*

**M1:** Verteilungen für Kombinationen von gedächtnislosen Wartezeichen berechnen können.

### Zu Abschnitt 6.3

*Sachfragen*

**S1:** Was versteht man unter einer diskreten gedächtnislosen Wartezeit?

**S2:** Wie kann man solche Wartezeiten charakterisieren?

**S3:** Angenommen, die Erfolgswahrscheinlichkeit bei einem Bernoulliexperiment ist  $p$ . Wie lange muss man dann im Mittel auf den ersten Erfolg warten? Wie kann man das für eine Schätzung von  $p$  ausnutzen?

*Methodenfragen*

**M1:** Nachprüfen können, ob eine diskrete Wahrscheinlichkeit gedächtnislos ist.

**M2:** Die Erfolgswahrscheinlichkeit  $p$  aus dem Erwartungswert der Zufallsvariablen „erster Erfolg“ ermitteln können.

## 6.5 Übungsaufgaben

### Zu Abschnitt 6.1

**Ü6.1.1** Sei  $X$  eine zum Parameter  $\lambda > 0$  exponentialverteilte Zufallsvariable. Für welche Werte von  $\lambda$  ist  $P(X \in [2, 4]) = 5/36$ ?

**Ü6.1.2** Sei  $X$  eine zum Parameter  $\lambda > 0$  exponentialverteilte Zufallsvariable. Für welchen Wert von  $\lambda$  ist  $P(X \in [2, 4])$  maximal?

**Ü6.1.3** Beweisen Sie, dass positive Vielfache gedächtnisloser Wartezeiten wieder gedächtnislos sind, und zwar

- direkt unter Verwendung der Definition „gedächtnislos“;
- unter Verwendung von Satz 6.1.3.

**Ü6.1.4** Die Zeit, die Sie benötigen, um eine stark befahrene Vorfahrtsstraße mit Ihrem PKW zu überqueren, soll durch eine gedächtnislose Wartezeit modelliert werden. Im Mittel warten Sie 2 Minuten.

- Wie wahrscheinlich ist es, dass es heute länger als 3 Minuten dauert?
- Wie groß ist die bedingte Wahrscheinlichkeit

$$\mathbb{P}(\text{Es dauert länger als drei Minuten} \mid \text{Es dauert weniger als fünf Minuten}) ?$$

**Ü6.1.5**  $T : \Omega \rightarrow \mathbb{R}^+$  sei eine gedächtnislose Wartezeit. Finden Sie alle Zahlen  $a > 0$ ,  $b \geq 0$ , so dass auch  $aT + b$  gedächtnislos ist.

**Ü6.1.6** Das Intervall  $[0, +\infty]$  sei durch die stetige Dichtefunktion  $f$  zu einem Wahrscheinlichkeitsraum gemacht worden. Es gelte für alle  $t \geq 0$

$$\frac{\int_t^\infty (s-t)f(s) ds}{\int_t^\infty f(s) ds} = \text{const.} =: \frac{1}{\lambda}.$$

Zeigen Sie, dass  $f$  die Dichte der Exponentialverteilung zum Parameter  $\lambda$  ist.

Der Quotient der Integrale ist gerade der Erwartungswert der Wartezeit unter der Bedingung, dass schon  $t$  Zeiteinheiten gewartet wurde. Die Bedingung sagt dann, dass dieser Wert unabhängig von  $t$  ist: Es nutzt nichts, schon lange gewartet zu haben, der Erwartungswert der restlichen Wartezeit ist immer gleich.

## Zu Abschnitt 6.2

**Ü6.2.1** Sie stehen vor drei besetzten Telefonzellen, vor Ihnen steht sonst niemand. Die jeweiligen Wartezeiten sind exponentialverteilt zu den Parametern 2, 3 und 5 (in Minuten). Mit welcher Wahrscheinlichkeit wird eine der Telefonzellen innerhalb der nächsten 60 Sekunden frei?

**Ü6.2.2**  $X_1, X_2, X_3, X_4$  seien unabängig und exponentialverteilt zu den Parametern  $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ . Denken Sie sich eine Situation aus dem „wirklichen Leben“ aus, in dem die Zufallsvariable  $X_1 + \min\{X_2, X_3, X_4\}$  von Interesse sein könnte und berechnen Sie die Dichtefunktion der zugehörigen Verteilung.

**Ü6.2.3** Sei  $X$  eine zum Parameter  $\lambda > 0$  exponentialverteilte Zufallsvariable. Wir definieren eine weitere Zufallsvariable  $Y$  durch  $Y(\omega) :=$  „die kleinste ganze Zahl, die größer oder gleich  $\omega$  ist“. (Anders ausgedrückt: Für  $\omega \in ]n-1, n]$  ist  $Y(\omega) = n$ ,  $n = 1, 2, \dots$ )

- Beweisen Sie, dass  $Y$  eine Zufallsvariable ist.
- Bestimmen Sie  $\mathbb{P}_Y$ . Zu welcher bekannten Klasse von Verteilungen gehört diese induzierte Wahrscheinlichkeit?

**Zu Abschnitt 6.3**

**Ü6.3.1**  $X$  sei geometrisch verteilt zum Parameter  $\lambda$ . Was ist wahrscheinlicher: „ $X$  ist gerade“ oder „ $X$  ist ungerade“?

**Ü6.3.2** Die Zufallsvariable  $X$  sei geometrisch verteilt zum Parameter  $0 < q < 1$ . Bestimmen Sie  $\mathbb{P}(\{X \in \{1, 2, 3\}\} | X \text{ ist gerade})$ .

**Ü6.3.3** Es sei  $T : \Omega \rightarrow \mathbb{N}$  eine gedächtnislose diskrete Wartezeit. Es gilt also  $P(T > s + t | T > s) = P(T > t)$  für alle  $s, t \in \mathbb{N}_0$ . Dann ist  $T$  *nicht gedächtnislos*, wenn man  $T$  als  $\mathbb{R}^+$ -wertig interpretiert: Es ist nicht richtig, dass  $P(T \geq s + t | T \geq s) = P(T \geq t)$  für alle  $s, t \geq 0$  gilt.

**Ü6.3.4** Jemand wirft zwei Würfel so lange, bis er das erste Mal „Augensumme 11“ erhält.

- Wie oft muss er im Mittel würfeln?
- Nun hat er schon 20 Mal gewürfelt, ohne dass die Augensumme einmal 11 gewesen ist. Mit welcher Wahrscheinlichkeit wird er auch in den nächsten beiden Versuchen keinen Erfolg haben?

## Teil IV

# Der Zufall verschwindet im Unendlichen

# Kapitel 7

## Konvergenz von Zufallsvariablen

Grenzwerte sind in vielen mathematischen Teilgebieten wichtig. In fast allen Fällen werden sie auf den Konvergenzbegriff für Zahlen zurückgeführt. Hier zwei Beispiele aus der Analysis, die Funktionen betreffen:

- Es seien  $f, f_1, f_2, \dots$  reellwertige Funktionen auf  $[0, 1]$ . Die Folge  $(f_n)_n$  heißt *punktwise konvergent* gegen  $f$ , wenn für alle  $x \in [0, 1]$  die Zahl  $f(x)$  Grenzwert der Folge  $(f_n(x))_n$  ist.
- Man spricht von *gleichmäßiger Konvergenz*, wenn die Folge  $(\|f_n - f\|)_n$  gegen Null konvergiert; dabei bezeichnet  $\|\cdot\|$  die Supremumsnorm<sup>1)</sup>.

Man kann Konvergenz auf sehr unterschiedliche Weise definieren. Neben punktweiser und gleichmäßiger Konvergenz gibt es noch viele weitere Möglichkeiten, der jeweils passende Konvergenzbegriff hängt von der gerade interessierenden Problemstellung ab. Wenn man etwa garantieren möchte, dass mit den  $f_n$  auch  $f$  stetig ist, so muss man gleichmäßige Konvergenz verlangen, denn der punktweise Limes einer Folge stetiger Funktionen muss nicht stetig sein.

Auch hier in der Wahrscheinlichkeitstheorie wird es um die Konvergenz von Funktionen – also von Zufallsvariablen – gehen. Dabei spielen *drei Konvergenzbegriffe* eine Rolle, die maßgeschneidert zur Beschreibung des Verhaltens des Zufalls im Unendlichen sind: Konvergenz in Wahrscheinlichkeit, Konvergenz punktweise fast sicher und Konvergenz in Verteilung.

Sie werden *in den Abschnitten 7.1, 7.2 und 7.3* eingeführt und – nur so ausführlich, wie für spätere Zwecke erforderlich – diskutiert. *In den Abschnitten 7.4 und 7.5* findet man dann Verständnisfragen und Übungsaufgaben.

---

<sup>1)</sup>Sie ist durch  $\|g\| := \sup_{x \in M} |g(x)|$  für beschränkte Funktionen  $g : M \rightarrow \mathbb{R}$  definiert.

## 7.1 Konvergenz in Wahrscheinlichkeit

In diesem Abschnitt diskutieren wir die erste Möglichkeit, durch die man ausdrücken kann, dass eine Folge von Zufallsvariablen  $(X_n)_{n \in \mathbb{N}}$  für  $n$  gegen Unendlich eine Zufallsvariable  $X$  „besser und besser“ beschreibt. Beim hier zu besprechenden Ansatz besteht die Grundidee darin, dass man verlangt, dass die Zufallsvariable  $X_n$  für große  $n$  mit hoher Wahrscheinlichkeit nur wenig von  $X$  abweicht. Präzisiert wird das so:

**Definition 7.1.1.** Es sei  $(\Omega, \mathcal{E}, \mathbb{P})$  ein Wahrscheinlichkeitsraum, und  $X$  sowie  $X_1, X_2, \dots$  seien auf  $\Omega$  definierte reellwertige Zufallsvariable. Wir sagen, dass die  $X_n$  in Wahrscheinlichkeit gegen  $X$  konvergieren, wenn für jedes  $\varepsilon > 0$  gilt:

$$\lim_{n \rightarrow \infty} \mathbb{P}(\{\omega \mid |X_n(\omega) - X(\omega)| > \varepsilon\}) = 0.$$

Anders ausgedrückt: Wenn man beliebige  $\varepsilon, \delta > 0$  vorschreibt, so kann man ein  $n_0$  finden, so dass  $\mathbb{P}(\{\omega \mid |X_n(\omega) - X(\omega)| > \varepsilon\}) \leq \delta$  für  $n \geq n_0$ .

Wenn diese Bedingung erfüllt ist, schreiben wir  $\lim_{i.W.n \rightarrow \infty} X_n = X$ .

### Ein erstes Beispiel und eine Warnung:

1. Wenn die  $X_n$  gleichmäßig gegen  $X$  gehen, folgt natürlich  $\lim_{i.W.n \rightarrow \infty} X_n = X$ , denn dann ist  $\{|X_n - X| > \varepsilon\}$  für genügend große  $n$  die leere Menge. Im nächsten Satz werden wir sehen, dass sogar punktweise Konvergenz ausreicht, um Konvergenz in Wahrscheinlichkeit garantieren zu können.

2. Für jedes  $n$  sei  $I_n \subset [0, 1]$  ein Intervall, so dass die Längen der  $I_n$  gegen Null gehen. Ist dann  $X_n$  die Indikatorfunktion  $\chi_{I_n}$  von  $I_n$ , so konvergieren die  $X_n$  offensichtlich in Wahrscheinlichkeit gegen die Nullfunktion.

Die Warnung: Das gilt auch dann, wenn man  $X_n$  durch  $X_n := c_n \chi_{I_n}$  mit beliebigen gigantisch großen Zahlen  $c_n$  definiert. Auch wenn also die Folge der  $X_n$  in Wahrscheinlichkeit konvergiert, können mit geringer Wahrscheinlichkeit riesengroße Abweichungen auftreten.

3. Wählt man im vorstehenden Beispiel die  $I_n$  so, dass jedes  $x \in [0, 1]$  in unendlich vielen  $I_n$  liegt<sup>2)</sup>, so ist die zugehörige Folge  $(X_n)$  an keiner Stelle punktweise konvergent.

Für spätere Zwecke stellen wir einige Eigenschaften dieser Konvergenzbegriffs zusammen:

**Satz 7.1.2.**  $X, Y, X_1, X_2, \dots, Y_1, Y_2, \dots$  seien reellwertige Zufallsvariable auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{E}, \mathbb{P})$ .

(i) Aus  $\lim_{i.W.n \rightarrow \infty} X_n = X$  und  $\lim_{i.W.n \rightarrow \infty} Y_n = Y$  folgt, dass die Wahrscheinlichkeit des Ereignisses  $\mathbb{P}\{X = Y\}$  gleich Eins ist. Da Ereignisse  $E$  mit  $\mathbb{P}(E) = 1$  fast sicher genannt werden, heißt das: Der Limes in Wahrscheinlichkeit ist, falls er existiert, fast sicher eindeutig bestimmt.

---

<sup>2)</sup>Man kann z.B. die Folge  $(I_n)$  als  $[0, 1], [0, 1/2], [1/2, 1], [0, 1/3], [1/3, 2/3], \dots$  wählen.

- (ii)  $\lim_{i.W.n \rightarrow \infty} X_n = X$  impliziert  $\lim_{i.W.n \rightarrow \infty} cX_n = cX$  für jedes  $c \in \mathbb{R}$ .
- (iii) Gilt  $\lim_{i.W.n \rightarrow \infty} X_n = X$  und  $\lim_{i.W.n \rightarrow \infty} Y_n = Y$ , so folgt  $\lim_{i.W.n \rightarrow \infty} X_n + Y_n = X + Y$ .
- (iv) Gibt es ein Ereignis  $N$  mit  $\mathbb{P}(N) = 0$ , so dass  $\lim_n X_n(\omega) = X(\omega)$  für  $\omega \in \Omega \setminus N$  gilt, wenn also die  $X_n$  bis auf eine Nullmenge punktweise gegen  $X$  konvergent sind, so folgt  $\lim_{i.W.n \rightarrow \infty} X_n = X$ .

**Beweis:** (i) Angenommen, die  $X_n$  konvergieren in Wahrscheinlichkeit gegen  $X$  und gegen  $Y$ . Fixiert man ein  $\varepsilon > 0$ , so muss dann  $\mathbb{P}(\{|X - Y| > \varepsilon\}) = 0$  sein. Es ist nämlich für jedes  $n$

$$\{|X - Y| > \varepsilon\} \subset \{|X - X_n| > \varepsilon/2\} \cup \{|X_n - Y| > \varepsilon/2\},$$

und deswegen gilt

$$\mathbb{P}(\{|X - Y| > \varepsilon\}) \leq \mathbb{P}(\{|X - X_n| > \varepsilon/2\}) + \mathbb{P}(\{|X_n - Y| > \varepsilon/2\}).$$

Die Summanden auf der rechten Seite gehen mit  $n$  gegen Unendlich nach Voraussetzung gegen Null, und deswegen ist wirklich  $\mathbb{P}(\{|X - Y| > \varepsilon\}) = 0$ .

Dann ist aber auch  $\mathbb{P}(\{X \neq Y\}) = 0$ , denn da  $\{X \neq Y\}$  der absteigende Durchschnitt der Mengen  $\{|X - Y| > 1/n\}$  ist, gilt wegen der Stetigkeit von Wahrscheinlichkeitsmaßen<sup>3)</sup>  $\mathbb{P}(\{X \neq Y\}) = \lim_n \mathbb{P}(\{|X - Y| > 1/n\}) = 0$ .

(ii) Das gilt offensichtlich für  $c = 0$ , und für  $c \neq 0$  folgt die Aussage sofort aus

$$\{|cX_n - cX| > \varepsilon\} = \{|X - X_n| > \varepsilon/|c|\}.$$

(iii) Die Beweistechnik ist ähnlich wie im Beweis von (i): Für  $\varepsilon > 0$  ist das Ereignis  $\{|(X_n + Y_n) - (X + Y)| > \varepsilon\}$  in der Vereinigung von  $\{|X_n - X| > \varepsilon/2\}$  und  $\{|Y_n - Y| > \varepsilon/2\}$  enthalten, und deswegen gehen die Wahrscheinlichkeiten  $\mathbb{P}(\{|(X_n + Y_n) - (X + Y)| > \varepsilon\})$  gegen Null.

(iv) Sei  $\varepsilon > 0$ , wir müssen  $\mathbb{P}(\{|X_n - X| > \varepsilon\}) \rightarrow 0$  beweisen. Für  $n \in \mathbb{N}$  definieren wir  $E_n$  als die Menge der  $\omega$ , für die  $|X_m(\omega) - X(\omega)| \leq \varepsilon$  für alle  $m \geq n$  ist. Es gilt  $E_1 \subset E_2 \subset \dots$ , und nach Voraussetzung enthält  $\bigcup_n E_n$  alle Elemente von  $\Omega \setminus N$ . Damit folgt  $\mathbb{P}(\bigcup_n E_n) = 1$ . Es ist aber wieder wegen Satz 1.3.2 –  $\mathbb{P}(\bigcup_n E_n) = \lim \mathbb{P}(E_n)$ , und damit gilt  $\mathbb{P}(\Omega \setminus E_n) = 1 - \mathbb{P}(E_n) \rightarrow 0$ .

Nun ist nur noch zu beachten, dass  $\{|X_n - X| > \varepsilon\} \subset \Omega \setminus E_n$ . So folgt wirklich  $\mathbb{P}(\{|X_n - X| > \varepsilon\}) \rightarrow 0$ .  $\square$

## 7.2 Fast sicher punktweise Konvergenz

Dieser Konvergenzbegriff – er spielte schon in Satz 7.1.2 eine Rolle – beschreibt „im Wesentlichen“ die punktweise Konvergenz:

---

<sup>3)</sup>Vgl. Satz 1.3.2(iv).

**Definition 7.2.1.** Sind  $X, X_1, X_2, \dots$  auf  $(\Omega, \mathcal{E}, \mathbb{P})$  definierte reellwertige Zufallsvariable, so sagen wir, dass die  $X_n$  punktweise fast sicher (oder auch kürzer: fast sicher) gegen  $X$  konvergieren, wenn es eine Nullmenge<sup>4)</sup>  $N$  so gibt, dass  $\lim_n X_n(\omega) = X(\omega)$  für alle  $\omega \notin N$  gilt.

In diesem Fall schreiben wir  $\lim_{f.s.} X_n = X$ .

Alle Beispiele müssen nach Definition dadurch entstehen, dass man aus  $\Omega$  eine Nullmenge  $N$  entfernt, auf  $\Omega \setminus N$  die  $X_n$  so wählt, dass sie punktweise gegen  $X$  konvergieren, und dann die Definition von  $X$  und  $X_1, X_2, \dots$  auf  $N$  im Wesentlichen beliebig ergänzt: Man muss bei diesen Schritten immer nur darauf achten, dass die Bedingung für Zufallsvariable erfüllt ist (Urbilder von Borelmengen sind Ereignisse). Ist zum Beispiel  $\Omega = [0, 1]$  mit der Gleichverteilung versehen und definiert man  $X_n(x) := x^n$  für  $x < 1$  und  $X_n(1) := (-1)^n n!$ , so konvergiert die Folge  $(X_n)$  fast sicher gegen die Nullfunktion.

Wie bei der Konvergenz in Wahrscheinlichkeit lassen sich auch für die fast sichere Konvergenz einige Permanenzeigenschaften beweisen:

**Satz 7.2.2.**  $X, Y, X_1, \dots, Y_1, \dots$  seien reellwertige Zufallsvariable auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{E}, \mathbb{P})$ .

- (i) Aus  $\lim_{f.s. n \rightarrow \infty} X_n = X$  und  $\lim_{f.s. n \rightarrow \infty} Y_n = Y$  folgt, dass  $X$  und  $Y$  fast sicher übereinstimmen: Der fast sichere Limes ist, falls er existiert, fast sicher eindeutig bestimmt.
- (ii)  $\lim_{f.s. n \rightarrow \infty} X_n = X$  impliziert  $\lim_{f.s. n \rightarrow \infty} cX_n = cX$  für jedes  $c \in \mathbb{R}$ .
- (iii) Gilt  $\lim_{f.s. n \rightarrow \infty} X_n = X$  und  $\lim_{f.s. n \rightarrow \infty} Y_n = Y$ , so folgt  $\lim_{f.s. n \rightarrow \infty} X_n + Y_n = X + Y$ .

**Beweis:** (i) Nach Voraussetzung gibt es Nullmengen  $N_X, N_Y$ , so dass gilt:

- Für  $\omega \in \Omega \setminus N_X$  ist  $\lim_n X_n(\omega) = X(\omega)$ .
- Für  $\omega \in \Omega \setminus N_Y$  ist  $\lim_n Y_n(\omega) = Y(\omega)$ .

Für die  $\omega$ , die nicht in  $N_X \cup N_Y$  liegen, gilt folglich  $X(\omega) = Y(\omega)$ , denn in  $\mathbb{R}$  sind Limites eindeutig bestimmt.

Und da  $N_X \cup N_Y$  eine Nullmenge ist, ist damit  $X = Y$  fast sicher.

(ii) Das folgt sofort daraus, dass  $a_n \rightarrow a$  für jedes  $c$  die Aussage  $ca_n \rightarrow ca$  impliziert.

(iii) Bis auf eine Nullmenge  $N_X$  (bzw.  $N_Y$ ) geht nach Voraussetzung  $(X_n(\omega))_n$  gegen  $X(\omega)$  (bzw.  $(Y_n(\omega))_n$  gegen  $Y(\omega)$ ). Außerhalb der Nullmenge  $N_X \cup N_Y$  kann damit  $(X_n + Y_n)(\omega) \rightarrow (X + Y)(\omega)$  garantiert werden, und das beweist die Behauptung.  $\square$

---

<sup>4)</sup>Zur Erinnerung: Das sind Ereignisse mit Wahrscheinlichkeit Null.

### 7.3 Konvergenz in Verteilung

Dieser Konvergenzbegriff ist sicher unter den drei Ansätzen, die in diesem Kapitel besprochen werden, am schwierigsten zugänglich.

Die Idee: Man möchte für Zufallsvariable  $X, Y$  sagen, dass sie „nahe beieinander“ liegen, wenn die Werte, die sie bei Abfrage produzieren, „kaum voneinander zu unterscheiden“ sind.  $X, Y$  können dabei auf unterschiedlichen Wahrscheinlichkeitsräumen definiert sein. Das ist noch recht vage, wir betrachten einige Beispiele:

- Es sei  $\Omega_X = \Omega_Y = \{0, 1\}$  und sowohl  $X$  als auch  $Y$  bilden 0 auf 0 und 1 auf 1 ab. Wenn dann  $\mathbb{P}_X(\{1\}) = 0.6$  und  $\mathbb{P}_Y(\{1\}) = 0.600001$  ist, wenn es also um Bernoulliverteilungen mit Erfolgswahrscheinlichkeiten 0.6 und 0.600001 geht, so sollte es doch keine wesentliche Rolle spielen, ob man  $X$  oder  $Y$  betrachtet.
- Ähnliches gilt für die folgende Situation:
  - $X$  beschreibe einen fairen Würfel, d.h.  $\Omega_X = \{1, 2, \dots, 6\}$  und es ist  $\mathbb{P}_X(\{i\}) = 1/6$  für  $i = 1, \dots, 6$ .
  - $Y$  dagegen ist die auf  $\Omega_Y = \{1.01, 2.01, \dots, 6.01\}$  definierte Abbildung  $x \mapsto x$ , und

$$\mathbb{P}_Y(\{1.01\}) := \frac{1}{6} - 0.05, \mathbb{P}_Y(\{2.01\}) = \dots = \mathbb{P}_Y(\{6.01\}) := \frac{1}{6} + 0.01.$$

Auch in diesem Fall sollte  $X \approx Y$  gelten.

- $X$  und  $Y$  seien beide die Abbildung  $x \mapsto x$  auf  $[0, 1]$ , und bei  $X$  werden die Wahrscheinlichkeiten durch die Dichtefunktion  $2x$ , bei  $Y$  aber durch die Dichtefunktion  $2.0001x^{1.001}$  erzeugt. Sicher sind auch hier  $X$  und  $Y$  praktisch ununterscheidbar.

Doch wie kann man das präzisieren? Die Idee, die sich bewährt hat, besteht darin, „Nähe“ für Zufallsvariable dadurch zu definieren, dass man verlangt, dass damit zusammenhängende Erwartungswerte nahe beieinander liegen. So gelangt man zu der nachstehenden Definition. Dass dabei nur gewisse „gutartige“ Funktionen  $g$  auftreten, liegt daran, dass man die Existenz der auftretenden Erwartungswerte garantieren möchte.

**Definition 7.3.1.** Es seien  $(\Omega, \mathcal{E}, \mathbb{P})$  und  $(\Omega_n, \mathcal{E}_n, \mathbb{P}_n)$  Wahrscheinlichkeitsräume und  $X : \Omega \rightarrow \mathbb{R}$  sowie  $X_n : \Omega_n \rightarrow \mathbb{R}$  Zufallsvariable ( $n = 1, 2, \dots$ ). Wir sagen, dass die  $X_n$  in Verteilung gegen  $X$  konvergieren, wenn gilt: Wenn  $g : \mathbb{R} \rightarrow \mathbb{R}$  eine beliebige beschränkte stetige Funktion ist<sup>5)</sup>, so ist

$$\lim_{n \rightarrow \infty} \mathbb{E}(g \circ X_n) = \mathbb{E}(g \circ X).$$

---

<sup>5)</sup>Wenn es also eine Zahl  $R$  so gibt, dass  $|g(x)| \leq R$  für alle  $x$ . Unter dieser Voraussetzung sind die  $g \circ X$ ,  $g \circ X_n$  beschränkte Zufallsvariable, und deswegen existieren die Erwartungswerte.

Wir schreiben dann  $\lim_{i.V.n \rightarrow \infty} X_n = X$

Wenn man in den obigen Beispielen  $Y$  durch eine Folge  $X_n$  ersetzt, für die die Approximation immer besser wird<sup>6)</sup>, so gilt wirklich, dass die  $X_n$  in Verteilung gegen  $X$  konvergieren.

Es hat sich gezeigt, dass die vorstehende Definition maßgeschneidert ist um auszudrücken, dass statt  $X$  auch die  $X_n$  mit großem  $n$  verwendet werden können, wenn es nur um induzierte Wahrscheinlichkeiten geht.

Hier die wichtigsten Eigenschaften:

**Satz 7.3.2.** (*Die Bezeichnungen seien wie in der vorstehenden Definition.*)

- (i)  $\lim_{i.V.n \rightarrow \infty} X_n = X$  impliziert  $\lim_{i.V.n \rightarrow \infty} cX_n = cX$  für jedes  $c \in \mathbb{R}$ .
- (ii) Ist  $(\Omega, \mathcal{E}, \mathbb{P})$  ein Wahrscheinlichkeitsraum und sind darauf reellwertige Zufallsvariable  $X, X_1, \dots$  definiert, so gilt:  
Aus  $\lim_{i.W.n \rightarrow \infty} X_n = X$  folgt  $\lim_{i.V.n \rightarrow \infty} X_n = X$ .
- (iii) Es gelte  $\lim_{i.V.n \rightarrow \infty} X_n = X$ , und es sei  $[a, b]$  ein abgeschlossenes Intervall, für das  $\mathbb{P}(\{X = a\}) = \mathbb{P}(\{X = b\}) = 0$  gilt. (Das ist zum Beispiel dann erfüllt, wenn  $\mathbb{P}_X$  eine Dichtefunktion hat.) Dann folgt

$$\lim_{n \rightarrow \infty} \mathbb{P}_n(\{X_n \in [a, b]\}) = \mathbb{P}(\{X \in [a, b]\}).$$

**Beweis:** (i) Das ist erfreulich einfach einzusehen. Ist nämlich  $g$  eine beschränkte stetige Funktion, so hat auch  $g_c$ , definiert durch  $x \mapsto g(cx)$ , diese Eigenschaft. Deswegen ist

$$\begin{aligned} \mathbb{E}(g \circ (cX)) &= \mathbb{E}(g_c \circ X) \\ &= \lim_{n \rightarrow \infty} \mathbb{E}(g_c \circ X_n) \\ &= \lim_{n \rightarrow \infty} \mathbb{E}(g \circ (cX_n)). \end{aligned}$$

(ii) Es gelte  $\lim_{i.W.n \rightarrow \infty} X_n = X$ . Wir geben eine stetige beschränkte Funktion  $g : \mathbb{R} \rightarrow \mathbb{R}$  vor, und es ist zu zeigen, dass

$$\lim_{n \rightarrow \infty} \mathbb{E}(g \circ X_n) = \mathbb{E}(g \circ X).$$

Für den Nachweis sind mehrere Tatsachen zu kombinieren. Dabei sind zunächst  $\eta$  und  $\varepsilon'$  beliebige positive Zahlen.

- Erstens ist doch  $X$  bis auf eine Menge beliebig kleinen Maßes beschränkt.  
Genauer:  $\Omega$  ist aufsteigende Vereinigung der Ereignisse  $E_k := \{|X| \leq k\}$ ,  $k = 1, 2, \dots$ , deswegen gilt  $1 = \mathbb{P}(\Omega) = \lim_k \mathbb{P}(E_k)$  (Satz 1.3.2). Folglich gibt es ein  $k_\eta$ , so dass  $\mathbb{P}(E_{k_\eta}) > 1 - \eta$ . Kurz: Bis auf ein Ereignis mit Wahrscheinlichkeit höchstens  $\eta$  gilt  $-k_\eta \leq X(\omega) \leq k_\eta$ .

---

<sup>6)</sup>Etwas in Beispiel 1 zu Beginn dieses Abschnitts: Die Erfolgswahrscheinlichkeiten für die  $X_n$  konvergieren gegen die Erfolgswahrscheinlichkeiten für  $X$ .

- Zweitens sind stetige Funktionen auf kompakten Teilmengen des Definitionsbereiches gleichmäßig stetig. Insbesondere kann man ein  $\delta \in ]0, 1]$  so finden, dass gilt: Für  $x, y \in [-k_\eta - 1, k_\eta + 1]$  mit  $|x - y| \leq \delta$  ist  $|g(x) - g(y)| \leq \varepsilon'$ .
- Die  $X_n$  konvergieren in Wahrscheinlichkeit gegen  $X$ . Deswegen konvergieren die  $\mathbb{P}(\{|X - X_n| > \delta\})$  gegen Null.

Nun sei  $\omega \in \Omega$ . Falls  $\omega$  zu  $E_{k_\eta} \cap \{|X - X_n| \leq \delta\}$  gehört, liegen  $X(\omega)$  und  $X_n(\omega)$  in  $[-k_\eta - 1, k_\eta + 1]$ , und der Abstand dieser Zahlen ist höchstens  $\delta$ . Folglich ist  $|(g \circ X(\omega)) - (g \circ X_n(\omega))| \leq \varepsilon'$ . Andernfalls liegt  $\omega$  in  $(\Omega \setminus E_{k_\eta}) \cup \{|X - X_n| > \delta\}$ . Die Wahrscheinlichkeit dieses Ereignisses ist höchstens  $\eta + \mathbb{P}(\{|X - X_n| > \delta\})$ , und der Wert von  $|(g \circ X(\omega)) - (g \circ X_n(\omega))|$  ist höchstens  $2R$ , wobei  $R$  eine obere Schranke der  $|g(x)|$  ist.

Nun fehlt nur noch eine elementare Beobachtung: Ist  $Y : \Omega \rightarrow \mathbb{R}$  eine Zufallsvariable und gilt  $|Y(\omega)| \leq a$  für  $\omega \in A$  und  $|Y(\omega)| \leq b$  für  $\omega \in B$ , wobei  $A \cup B$  eine disjunkte Zerlegung von  $\Omega$  ist<sup>7)</sup>, so ist  $\mathbb{E}(|Y|) \leq a\mathbb{P}(A) + b\mathbb{P}(B)$ . (Zur Begründung ist nur zu beachten, dass  $|Y| \leq a\chi_A + b\chi_B$  gilt und dass der Übergang zum Erwartungswert nach den Ergebnissen aus Abschnitt 3.3 linear und monoton ist.)

Hier heißt das:  $\mathbb{E}(|X - X_n|) \leq 2R(\eta + \mathbb{P}(\{|X - X_n| > \delta\})) + \varepsilon'$ . Es bleibt nur noch dafür zu sorgen, dass das für große  $n$  kleiner als ein vorgegebenes  $\varepsilon$  wird.

So kommen wir zum Finale. Es sei  $\varepsilon > 0$ . Definiere  $\eta := \varepsilon/(6R)$  und  $\varepsilon' := \varepsilon/3$ . Finde zu diesen Werten das  $\delta$  wie vorstehend. Nach Voraussetzung gibt es ein  $n_0$ , so dass für  $n \geq n_0$  die Ungleichung  $\mathbb{P}(\{|X - X_n| > \delta\}) \leq \varepsilon/(6R)$  gilt. Und das heißt

$$\mathbb{E}(|X - X_n|) \leq 2R(\eta + \mathbb{P}(\{|X - X_n| > \delta\})) + \varepsilon' \leq \varepsilon.$$

Dann ist aber auch  $|\mathbb{E}(X) - \mathbb{E}(X_n)| \leq \varepsilon$ , und damit ist alles gezeigt<sup>8)</sup>.

(iii) Wir geben ein Intervall  $[a, b]$  mit  $\mathbb{P}(\{X = a\}) = \mathbb{P}(\{X = b\}) = 0$  und ein  $\varepsilon > 0$  vor. Das Ziel: Für genügend große  $n$  ist

$$|\mathbb{P}_n(\{X_n \in [a, b]\}) - \mathbb{P}(\{X \in [a, b]\})| \leq \varepsilon.$$

Die Idee besteht darin, Wahrscheinlichkeiten durch Erwartungswerte zu approximieren. Das Ereignis  $\{X = a\}$  hat Wahrscheinlichkeit Null, und es ist der absteigende Durchschnitt der Ereignisse  $\{X \in [a - 1/k, a + 1/k]\}$ ,  $k \in \mathbb{N}$ . Deswegen gibt es nach Satz 1.3.2 ein  $k'$ , so dass  $\mathbb{P}(\{X \in [a - 1/k', a + 1/k']\}) \leq \varepsilon/4$ . Ganz analog findet man ein  $k''$  mit  $\mathbb{P}(\{X \in [b - 1/k'', b + 1/k'']\}) \leq \varepsilon/4$ , und wir setzen  $k := \max\{k', k''\}$ .

Nun definieren wir zwei Funktionen  $g_1, g_2 : \mathbb{R} \rightarrow \mathbb{R}$  die „sehr nahe“ bei  $\chi_{[a, b]}$  liegen:

<sup>7)</sup>  $A, B$  sind also disjunkt, und es gilt  $A \cup B = \Omega$ .

<sup>8)</sup> Beachte für den letzten Schritt: Ist  $Y$  Zufallsvariable, so ist  $-|Y| \leq Y \leq |Y|$ , also gilt (wegen der Monotonie der Zuordnung  $Y \mapsto \mathbb{E}(Y) - \mathbb{E}(|Y|) \leq \mathbb{E}(Y) \leq \mathbb{E}(|Y|)$ ). So folgt  $|\mathbb{E}(Y)| \leq \mathbb{E}(|Y|)$ , und das ist hier für  $Y = X - X_n$  anzuwenden.

- $g_1$ : Diese Funktion ist Null auf  $]-\infty, a - 1/k] \cup [b + 1/k, +\infty[$  und 1 auf  $[a, b]$ . Und auf  $[a - 1/k, a]$  und  $[b, b + 1/k]$  wird der Graph durch eine Strecke stetig ergänzt.
- $g_2$  ist Null auf  $]-\infty, a] \cup [b, +\infty[$  und 1 auf  $[a + 1/k, b - 1/k]$ . Auf den Intervallen  $[a, a + 1/k]$  und  $[b - 1/k, b]$  wird sie ebenfalls linear und stetig ergänzt.



Bild 7.3.1: Skizze der Funktionen  $g_1$  und  $g_2$ .

Dann sind  $g_1, g_2$  beschränkte stetige Funktionen mit  $g_2 \leq \chi_{[a, b]} \leq g_1$ , und  $g_1 - g_2$  liegt zwischen 0 und der Indikatorfunktion von

$$\Delta := [a - 1/k, a + 1/k] \cup [b - 1/k, b + 1/k].$$

Damit gilt

$$g_2 \circ X \leq \chi_{X \in [a, b]} \leq g_1 \circ X \text{ sowie } 0 \leq (g_1 - g_2) \circ X \leq \chi_{X \in \Delta},$$

und das impliziert

$$\mathbb{E}(g_2 \circ X) \leq \mathbb{P}(\{X \in [a, b]\}) \leq \mathbb{E}(g_1 \circ X),$$

$$0 \leq \mathbb{E}(g_1 \circ X) - \mathbb{E}(g_2 \circ X) \leq \mathbb{P}(\{X \in \Delta\}) \leq \varepsilon/2.$$

Anders ausgedrückt heißt das, dass  $\mathbb{P}(\{X \in [a, b]\})$  von oben bzw. von unten durch  $\mathbb{E}(g_1 \circ X)$  bzw.  $\mathbb{E}(g_2 \circ X)$  so approximiert wird, dass der Fehler jeweils höchstens  $\varepsilon/2$  ist.

Nach Voraussetzung gilt  $\mathbb{E}(g_1 \circ X_n) \rightarrow \mathbb{E}(g_1 \circ X)$  und  $\mathbb{E}(g_2 \circ X_n) \rightarrow \mathbb{E}(g_2 \circ X)$ . Außerdem ist – mit gleicher Begründung wie vor wenigen Zeilen –

$$\mathbb{E}(g_2 \circ X_n) \leq \mathbb{P}(\{X_n \in [a, b]\}) \leq \mathbb{E}(g_1 \circ X_n).$$

Nach diesen Vorbereitungen können wir den Beweis abschließen:

- Es gibt ein  $n_0$ , so dass für  $n \geq n_0$  die Zahlen  $\mathbb{E}(g_2 \circ X_n)$  und  $\mathbb{E}(g_1 \circ X_n)$  in  $I := [\mathbb{E}(g_2 \circ X) - \varepsilon/4, \mathbb{E}(g_1 \circ X) + \varepsilon/4]$  liegen.
- Auch  $\mathbb{P}_n(\{X_n \in [a, b]\})$  und  $\mathbb{P}(\{X \in [a, b]\})$  gehören zu  $I$ .
- Die Länge von  $I$  ist höchstens  $\varepsilon$ .

Deswegen ist

$$|\mathbb{P}_n(\{X_n \in [a, b]\}) - \mathbb{P}(\{X \in [a, b]\})| \leq \varepsilon.$$

für  $n \geq n_0$ , und das beweist (iii).  $\square$

Zu diesem Satz sind einige *Kommentare* angebracht.

**1.** Wegen Teil (ii) und Satz 7.1.2(iv) können wir die gegenseitigen Beziehungen zwischen den verschiedenen Konvergenzbegriffen so zusammenfassen:

- Aus  $\lim_{f.s. n \rightarrow \infty} X_n = X$  folgt  $\lim_{i.W.n \rightarrow \infty} X_n = X$ .
- Aus  $\lim_{i.W.n \rightarrow \infty} X_n = X$  folgt  $\lim_{i.V.n \rightarrow \infty} X_n = X$ .

Die Umkehrungen gelten in beiden Fällen nicht.

Für die zweite Implikation ist es nicht einmal sinnvoll, nach der Umkehrung zu fragen, da die Definitionsbereiche der  $X_n$  bei der Konvergenz in Verteilung alle unterschiedlich sein können, doch bei der Konvergenz in Wahrscheinlichkeit müssen die  $X_n$  auf dem gleichen Raum definiert sein. Aber selbst wenn das erfüllt ist, muss die Umkehrung nicht stimmen: Betrachte einfach  $\Omega = \{0, 1\}$  mit der Gleichverteilung und definiere  $X_n(0) := 0$ ,  $X_n(1) := 1$  für gerade  $n$  sowie  $X_n(0) := 1$ ,  $X_n(1) := 0$  für ungerade  $n$ . Fast sichere Konvergenz liegt nicht vor, wohl aber gilt  $\lim_{i.V.n \rightarrow \infty} X_n = X_1$ .

**2.** Manche Leser werden ein Ergebnis vermisst haben: Wie bei allen sonst gebräuchlichen Konvergenzbegriffen sollte doch aus  $\lim_{i.V.n \rightarrow \infty} X_n = X$  und  $\lim_{i.V.n \rightarrow \infty} Y_n = Y$  stets  $\lim_{i.V.n \rightarrow \infty} X_n + Y_n = X + Y$  folgen. Aber:

- Erstens wäre es nicht sinnvoll, so etwas zu formulieren, da  $X_n + Y_n$  nur dann erklärt ist, wenn beide Zufallsvariable auf dem gleichen Wahrscheinlichkeitsraum definiert sind. Das muss aber nicht der Fall sein.
- Aber selbst wenn das erfüllt ist, muss es nicht stimmen. Für ein Gegenbeispiel betrachten wir noch einmal die Zufallsvariablen aus dem vorstehenden Punkt „1.“. Setzt man  $Y_n := X_{n+1}$ , so gilt  $\lim_{i.V.n \rightarrow \infty} Y_n = X_0$ , aber alle  $X_n + Y_n$  sind die konstante Funktion 1, die sicher nicht in Verteilung gegen  $2X_0$  konvergiert.

**3.** Warum muss die Definition so kompliziert sein? Warum kann man sie nicht durch die Bedingung ersetzen, dass  $\lim_n \mathbb{P}_{X_n}(B) = \mathbb{P}_X(B)$  für alle Borelmengen  $B$  gilt, dass also die Wahrscheinlichkeit,  $X_n$  in  $B$  zu finden, gegen die Wahrscheinlichkeit konvergiert,  $X$  in  $B$  zu finden?

Der Grund: Dann würden zum Beispiel die konstanten Funktionen  $X_n = 1/n$  nicht gegen die Nullfunktion  $X = 0$  konvergieren, denn  $\mathbb{P}(\{X_n \in [0, 0]\}) = 0$ , aber  $\mathbb{P}(\{X \in [0, 0]\}) = 1$ . Konvergenz in Verteilung liegt aber vor.

**4.** In Teil (iii) des Satzes ist die Zusatzvoraussetzung (Wahrscheinlichkeit 0 für  $X = a$  und  $X = b$ ) wesentlich. Das zeigt das vorstehende Beispiel unter „2.“: Konvergenz in Verteilung liegt vor, aber die  $\mathbb{P}(\{X_n \in [0, 0]\})$  konvergieren nicht gegen  $\mathbb{P}(\{X \in [0, 0]\})$ .

Wir werden hauptsächlich – beim Beweis des zentralen Grenzwertsatzes in Abschnitt 8.4 – an Fällen interessiert sein, in denen  $X$  standard-normalverteilt

ist, und da ist Teil (iii) des vorstehenden Satzes dann wirklich für alle  $[a, b]$  anwendbar.

Der Abschnitt schließt mit einem eher technischen Ergebnis, das in Abschnitt 8.4 eine fundamentale Rolle spielen wird. Es besagt, dass man Konvergenz in Verteilung mit vergleichsweise einfachen Funktionen testen kann:

**Lemma 7.3.3.** *Die Zufallsvariable  $X$  und die  $X_n$  seien wie in Definition 7.3.1. Für jede Funktion  $h : \mathbb{R} \rightarrow \mathbb{R}$ , die beliebig oft differenzierbar ist und außerhalb eines beschränkten Intervalls gleich Null ist, gelte  $\mathbb{E}(h \circ X_n) \rightarrow \mathbb{E}(h \circ X)$ . Dann gilt  $\lim_{i.V.n \rightarrow \infty} X_n = X$ .*

**Beweis:** Es sei  $g : \mathbb{R} \rightarrow \mathbb{R}$  eine beliebige stetige beschränkte Funktion. Wir müssen  $\mathbb{E}(g \circ X_n) \rightarrow \mathbb{E}(g \circ X)$  zeigen. Dazu geben wir  $\varepsilon > 0$  vor, gesucht ist ein  $n_0$ , so dass  $|\mathbb{E}(g \circ X_n) - \mathbb{E}(g \circ X)| \leq \varepsilon$  für  $n \geq n_0$ .

Für unsere Analyse wird es praktisch sein, mit einem beliebigen  $\eta > 0$  anzufangen, das wir erst später geeignet festlegen.

1. *Schritt:* Wie im Beweis des vorigen Satzes beginnen wir mit der Beobachtung, dass  $\mathbb{R}$  die aufsteigende Vereinigung der  $\{|X| \leq k\}$  ist und dass es deswegen ein  $k_\eta$  geben muss, so dass  $\mathbb{P}(\{|X| \leq k_\eta\}) \geq 1 - \eta$  gilt.

Nun übersetzen wir wieder Wahrscheinlichkeiten in Erwartungswerte. Wir wählen eine beliebig oft differenzierbare Funktion  $h_1$  mit

$$\chi_{[-k_\eta, k_\eta]} \leq h_1 \leq \chi_{[-k_\eta - 1, k_\eta + 1]}.$$

Für beliebige reellwertige Zufallsvariable  $Y$  gilt dann

$$\mathbb{P}(\{|Y| \leq k_\eta\}) \leq \mathbb{E}(h_1 \circ Y) \leq \mathbb{P}(\{|Y| \leq k_\eta + 1\}).$$

Hier heißt das, dass  $\mathbb{E}(h_1 \circ X) \geq 1 - \eta$  gilt, und da die  $\mathbb{E}(h_1 \circ X_n)$  nach Voraussetzung gegen  $\mathbb{E}(h_1 \circ X)$  konvergieren, gibt es ein  $n_1$ , so dass  $\mathbb{E}(h_1 \circ X_n) \geq 1 - 2\eta$  für  $n \geq n_1$ . Aufgrund der vorstehenden Ungleichung können wir folgern, dass  $\mathbb{P}_n(|X_n| \geq k_{\eta+1}) \geq 1 - 2\eta$  und damit auch  $\mathbb{P}_n(\{|X_n| > k_\eta + 1\}) \leq 2\eta$  gilt.

Kurz: Für große  $n$  ist  $\mathbb{P}_{X_n}$  im Wesentlichen auf  $[-k_\eta - 1, k_\eta + 1]$  konzentriert.

2. *Schritt:* Nun approximieren wir die zu Beginn des Beweises vorgelegte Funktion  $g$  durch eine „einfache Funktion“  $h_2 : \mathbb{R} \rightarrow \mathbb{R}$ . Genauer soll gelten:

1.  $h_2$  ist beliebig oft differenzierbar und außerhalb eines beschränkten Intervalls gleich Null.
2. Für alle  $x$  ist  $h_2(x) \leq M + 1$ ; dabei ist  $M$  eine obere Schranke der  $|g(x)|$ ,  $x \in \mathbb{R}$ .
3. Für alle  $x \in [-k_\eta - 1, k_\eta + 1]$  ist  $|g(x) - h_2(x)| \leq \eta$ .

Die Tatsache, dass so ein  $h_2$  wirklich gefunden werden kann, übernehmen wir aus der Analysis.

Nun nutzen wir ein zweites Mal die Voraussetzung aus: Mit einem geeigneten  $n_2$  ist  $|\mathbb{E}(h_2 \circ X) - \mathbb{E}(h_2 \circ X_n)| \leq \eta$  für  $n \geq n_2$ .

Nach diesen Vorbereitungen können wir den Beweis beenden. Uns interessiert doch der Abstand zwischen  $\mathbb{E}(g \circ X)$  und  $\mathbb{E}(g \circ X_n)$ , wir betrachten nur noch Indizes  $n$  mit  $n \geq n_0 := \max\{n_1, n_2\}$ :

*Abstand von  $\mathbb{E}(g \circ X)$  und  $\mathbb{E}(h_2 \circ X)$ :* Auf  $[-k_\eta - 1, k_\eta + 1]$  sind  $h_2 \circ X$  und  $g \circ X$  höchstens um  $\eta$  voneinander entfernt und auf den anderen  $x \in \mathbb{R}$  höchstens um  $2(M + 2)$ . Die Wahrscheinlichkeit von  $\{|X| > k_\eta + 1\}$  ist aber höchstens  $\eta$ , und damit folgt

$$|\mathbb{E}(g \circ X) - \mathbb{E}(h_2 \circ X)| \leq \eta + (2M + 2)\eta.$$

*Abstand von  $\mathbb{E}(g \circ X_n)$  und  $\mathbb{E}(h_2 \circ X_n)$ :* Ganz analog kommt man zu der Abschätzung

$$|\mathbb{E}(g \circ X_n) - \mathbb{E}(h_2 \circ X_n)| \leq \eta + (2M + 2)2\eta.$$

Und da der Abstand zwischen  $\mathbb{E}(h_2 \circ X)$  und  $\mathbb{E}(h_2 \circ X_n)$  höchstens  $\eta$  ist, erhalten wir unter Verwendung der Dreiecksungleichung insgesamt

$$|\mathbb{E}(g \circ X) - \mathbb{E}(g \circ X_n)| \leq \eta(6M + 6).$$

Wenn wir also am Anfang  $\eta$  als  $\varepsilon/(6M + 6)$  definieren, können wir wirklich garantieren, dass  $|\mathbb{E}(g \circ X) - \mathbb{E}(g \circ X_n)| \leq \varepsilon$  für genügend große  $n$  gilt.  $\square$

## 7.4 Verständnisfragen

### Zu Abschnitt 7.1

#### *Sachfragen*

**S1:** Was versteht man unter „Konvergenz in Wahrscheinlichkeit“?

**S2:** Wie ist der Zusammenhang zwischen „Konvergenz in Wahrscheinlichkeit“ und „Konvergenz fast sicher“?

**S3:** Welche Permanenzaussagen zum Thema „Konvergenz in Wahrscheinlichkeit“ kennen Sie?

#### *Methodenfragen*

**M1:** Nachprüfen können, ob Konvergenz in Wahrscheinlichkeit vorliegt.

**M2:** Einfache Eigenschaften zu diesem Konvergenzbegriff nachweisen können.

### Zu Abschnitt 7.2

#### *Sachfragen*

**S1:** Was versteht man unter „Konvergenz (punktweise) fast sicher“?

**S2:** Muss eine Folge von Zufallsvariablen, die in Wahrscheinlichkeit konvergiert, auch punktweise fast sicher konvergieren?

**S3:** Welche Permanenzaussagen zum Thema „Konvergenz fast sicher“ kennen Sie?

*Methodenfragen*

**M1:** Nachprüfen können, ob fast sichere Konvergenz vorliegt.

**M2:** Einfache Eigenschaften zu diesem Konvergenzbegriff nachweisen können.

### Zu Abschnitt 7.3

*Sachfragen*

**S1:** Was versteht man unter „Konvergenz in Verteilung“?

**S2:** Angenommen, die  $X_n$  konvergieren in Verteilung gegen  $X$ . Unter welchen Voraussetzungen kann dann  $\mathbb{P}_X([a, b])$  für genügend große  $n$  durch  $\mathbb{P}_{X_n}([a, b])$  approximiert werden?

*Methodenfragen*

**M1:** Nachprüfen können, ob Konvergenz in Verteilung vorliegt.

**M2:** Einfache Eigenschaften zu diesem Konvergenzbegriff nachweisen können.

## 7.5 Übungsaufgaben

### Zu Abschnitt 7.1

**Ü7.1.1**  $\Omega$  sei höchstens abzählbar, und  $X, X_1, \dots$  seien darauf definierte reellwertige Zufallsvariable. Zeigen Sie, dass die  $X_n$  fast sicher gegen  $X$  gehen, falls Konvergenz in Wahrscheinlichkeit vorliegt.

Anders ausgedrückt: „Konvergenz in Wahrscheinlichkeit“ und „Konvergenz fast sicher“ sind in diesem Fall äquivalent.

**Ü7.1.2**  $X_1, X_2, \dots : \Omega \rightarrow \mathbb{R}$  seien Zufallsvariable, die nur die Werte 0 und 1 annehmen. Weiter sei  $p_n := \mathbb{P}(\{X_n = 1\})$ . Man zeige, dass die  $X_n$  genau dann in Wahrscheinlichkeit gegen die Nullfunktion gehen, wenn  $p_n \rightarrow 0$  gilt.

**Ü7.1.3**  $X, X_1, X_2, \dots : \Omega \rightarrow ]0, +\infty[$  seien Zufallsvariable, und es gelte  $X = \lim_{i.W.n \rightarrow \infty} X_n$ . Gilt dann auch  $1/X = \lim_{i.W.n \rightarrow \infty} 1/X_n$ ?

**Ü7.1.4** Gibt es eine zu Satz 7.1.2(iii) analoge Aussage, wenn man „+“ durch „-“ ersetzt?

**Ü7.1.5** Es seien  $A, A_1, A_2, \dots$  Ereignisse in einem Wahrscheinlichkeitsraum. Zeigen Sie, dass die Indikatorfunktionen  $\chi_{A_n}$  genau dann gegen  $\chi_A$  gehen, wenn  $\mathbb{P}(A_n \Delta A) \rightarrow 0$  gilt<sup>9)</sup>

### Zu Abschnitt 7.2

**Ü7.2.1** Es gelte  $\lim_{f.s.} X_n = X$ . Wir fixieren nun ein  $r > 0$  und „beschneiden“  $X$  und die  $X_n$ : Neue Zufallsvariable werden durch

$$Y := \min\{\max\{X, -r\}, r\}, \quad Y_n := \min\{\max\{X_n, -r\}, r\}$$

---

<sup>9)</sup>Hierbei bezeichnet  $\Delta$  die symmetrische Differenz:  $A \Delta B := (A \setminus B) \cup (B \setminus A)$ .

deiniert. Zeigen Sie, dass dann  $\lim_{f.s.} Y_n = Y$  gilt.

**Ü7.2.2** Gibt es eine zu Satz 7.2.2(iii) analoge Aussage, wenn man „+“ durch „·“ ersetzt?

**Ü7.2.3** Es seien  $A_1, A_2 \dots$  Ereignisse. Finden Sie Bedingungen dafür, dass  $\lim_{f.s.} \chi_{A_n} = 0$  gilt.

### Zu Abschnitt 7.3

**Ü7.3.1**  $\Omega = [0, 1]$  sei mit der Gleichverteilung versehen. Die Zufallvariablen  $X_n$  seien definiert durch

$$X_n(\omega) := \begin{cases} \omega, & \text{wenn } n \text{ ungerade} \\ 1 - \omega, & \text{wenn } n \text{ gerade.} \end{cases}$$

Konvergiert die Folge  $(X_n)_{n \in \mathbb{N}}$

- fast sicher,
- in Wahrscheinlichkeit,
- in Verteilung?

**Ü7.3.2**  $\lim_{i.V.} X_n = X$  impliziert  $\lim_{i.V.} (X_n + c) = X + c$  für alle  $c \in \mathbb{R}$ .

**Ü7.3.3** Man sagt, dass eine für Elementarereignisse sinnvoll formulierbare Eigenschaft *fast sicher* gilt, wenn es ein Ereignis  $N$  mit Wahrscheinlichkeit 0 so gibt, dass alle nicht in  $N$  gelegenen Elementarereignisse diese Eigenschaft haben.

Nun seien  $X, X_1, X_2, \dots$  reellwertige Zufallsvariable, die alle auf dem gleichen Wahrscheinlichkeitsraum definiert sind. Wir setzen voraus, dass die  $X_n$  für  $n = 1, \dots$  fast sicher nichtnegativ sind.

Die  $X_n$  sollen gegen  $X$  konvergent sein. Welche Art der Konvergenz impliziert, dass auch  $X$  fast sicher nichtnegativ ist: fast sicher, in Wahrscheinlichkeit, in Verteilung?

**Ü7.3.4** Es seien  $X, X_1, X_2, \dots$  Zufallsvariable, und  $F_X, F_{X_1}, F_{X_2}, \dots$  seien die zugehörigen Verteilungsfunktionen. Zeigen Sie: Wenn die  $X_n$  in Verteilung gegen  $X$  gehen, so ist  $\lim_n F_{X_n}(x) = F_X(x)$  für jedes  $x$ , bei dem  $F_X$  stetig ist<sup>10)</sup>.

---

<sup>10)</sup>Die Umkehrung gilt übrigens auch, das ist aber etwas schwieriger einzusehen.

# Kapitel 8

## Die Gesetze der großen Zahlen

Dieses Kapitel ist unter verschiedenen Aspekten von Bedeutung. Es beginnt in *Abschnitt 8.1* mit einer überraschenden Tatsache: Wenn auch im Allgemeinen Wahrscheinlichkeiten beliebige Werte in  $[0, 1]$  annehmen können, so gibt es doch Situationen, bei denen man garantieren kann, dass nur einer der extremen Werte – also Null oder Eins – vorkommen wird. Wichtigster Vertreter dieser so genannten *Null-Eins-Gesetze* sind die Lemmata von *Borel-Cantelli*, die hier bewiesen werden. Sie werden in den weiteren Abschnitten eine wichtige Rolle spielen.

In *Abschnitt 8.2* lernen wir das *schwache Gesetz der großen Zahlen* kennen: Mittelwerte von unabhängigen Zufallsvariablen tendieren dazu, mit hoher Wahrscheinlichkeit in der Nähe des Erwartungswertes zu liegen. Es folgt aus der *Tschebyscheff-Ungleichung*, ein Ergebnis, das auch eine große praktische Bedeutung hat. Wie viele Leute muss man zum Beispiel danach fragen, was sie am nächsten Sonntag wählen würden, um eine zuverlässige Wahlprognose erstellen zu können?

*Abschnitt 8.3* enthält so etwas wie eine theoretische Absicherung unseres axiomatischen Zugangs zur Wahrscheinlichkeitsrechnung: das *starke Gesetz der großen Zahlen*. Unser heuristisches Konzept, dass „Wahrscheinlichkeit von  $E$ “ so etwas ist wie der prozentuale Anteil der Fälle, in denen wir bei „vielen“ Versuchen das Ergebnis „ $\omega \in E$ “ erhalten, wird nun unter Verwendung des richtigen Konvergenzbegriffs beweisbar.

Inwieweit die *Normalverteilung* eine zentrale Rolle in der Wahrscheinlichkeitsrechnung spielt, wird in *Abschnitt 8.4* präzisiert, in dem der *zentrale Grenzwertsatz* bewiesen werden wird.

In den ersten vier Abschnitten wurde untersucht, auf welch unterschiedliche Weise der Zufall mehr und mehr abgeschwächt wird, wenn sich viele unabhängige Zufallseinflüsse überlagern. In *Abschnitt 8.5* geht es um das Gegenteil: Wieviel Zufallseinfluss bleibt auch im Unendlichen erhalten?

Das Kapitel schließt – in den Abschnitten 8.6, 8.7 und 8.8 – mit Ergänzungen, Verständnisfragen und Übungsaufgaben.

## 8.1 Die Lemmata von Borel-Cantelli

Es sei  $M$  eine Menge, und  $E_1, E_2, \dots$  seien Teilmengen. Uns interessieren diejenigen  $x \in M$ , die in unendlich vielen der  $E_n$  liegen. Die Gesamtheit dieser  $x$  soll  $\limsup E_n$  heißen<sup>1)</sup>:

$$\limsup E_n := \{x \in M \mid \text{es gibt unendlich viele } n \text{ mit } x \in E_n\}.$$

Man kann, wie nicht schwer zu sehen,  $\limsup E_n$  auch anders darstellen, nämlich als

$$E = \bigcap_{m \in \mathbb{N}} \bigcup_{n \geq m} E_n.$$

Damit gibt es eine gewisse formale Ähnlichkeit zur Definition des Limes superior der Analysis:

„ $\leq$ “ ist eine Ordnungsrelation auf  $\hat{\mathbb{R}} := [-\infty, +\infty]$ , und alle Teilmengen von  $\hat{\mathbb{R}}$  haben ein Supremum und ein Infimum. Insbesondere existieren für jede Folge  $(x_n)$  in  $\hat{\mathbb{R}}$  die Zahlen

$$\limsup x_n := \inf_m \sup_{n \geq m} x_n, \quad \liminf x_n := \sup_m \inf_{n \geq m} x_n;$$

dabei sind die Werte  $+\infty$  und  $-\infty$  zugelassen.

Und wenn  $M$  eine Menge ist, definiert „ $\subset$ “ eine Ordnungsrelation auf der Potenzmenge von  $M$ . Beliebige Teilmengen der Potenzmenge haben dann ein Supremum und ein Infimum, nämlich die Vereinigung bzw. den Durchschnitt über die Elemente dieser Teilmenge<sup>2)</sup>. Und folglich ist  $\bigcap_{m \in \mathbb{N}} \bigcup_{n \geq m} E_n$  so etwas wie der Limes superior der Mengenfolge  $E_1, E_2, \dots$

Zum besseren Kennenlernen der Definition folgen einige einfache *Bemerkungen und Beispiele*:

1. Es sei  $M = \mathbb{R}$ .

- Für  $E_n := [0, 1/n]$  ist  $\limsup E_n = \{0\}$ .
- Sind die  $E_n$  durch  $E_n := [-n, n]$  definiert, so ist  $\limsup E_n = \mathbb{R}$ .

Zur Begründung ist in beiden Fällen an das Archimedescxiom zu erinnern: Es ist beim ersten Beispiel wichtig zu wissen, dass für  $x > 0$  ein  $n \in \mathbb{N}$  mit  $1/n < x$  existiert, und für das zweite Beispiel ist wesentlich, dass es für jedes  $x \in \mathbb{R}$  ein  $n \in \mathbb{N}$  mit  $x \in [-n, n]$  gibt.

<sup>1)</sup>Gesprochen: „Limes superior der  $E_n$ “.

<sup>2)</sup>Beachte: Der Durchschnitt (bzw. die Vereinigung) über das leere Mengensystem muss als  $M$  (bzw.  $\emptyset$ ) definiert werden.

**2.** Ist  $M$  eine beliebige Menge,  $A$  eine Teilmenge und sind alle  $E_n$  gleich  $A$ , so ist  $\limsup E_n = A$ .

**3.**  $\limsup E_n$  besteht gerade aus denjenigen  $x$ , bei denen die Funktion  $\sum_n \chi_{E_n}$  den Wert Unendlich annimmt.

**4.** Es seien alle  $E_n$  Ereignisse in einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{E}, \mathbb{P})$ . Dann ist auch  $\limsup E_n$  ein Ereignis, denn an der Konstruktion dieser Teilmenge sind nur abzählbar viele Ereignisse beteiligt.

Es gibt in der Wahrscheinlichkeitstheorie zwei Ergebnisse über den Limes Superior der  $E_n$ , die in sehr vielen Beweisen eine unverzichtbare Rolle spielen. Es sind die Lemmata von Borel und Cantelli, die wir nun behandeln werden.

**Lemma 8.1.1.** (*Lemma von Borel<sup>3)</sup>-Cantelli, Teil 1*)

Es sei  $(\Omega, \mathcal{E}, \mathbb{P})$  ein Wahrscheinlichkeitsraum, und  $E_1, E_2, \dots$  seien Ereignisse.

Gilt dann  $\sum_n \mathbb{P}(E_n) < \infty$ , so ist  $\mathbb{P}(\limsup E_n) = 0$ .

**Beweis:** Es sei  $F_m := E_m \cup E_{m+1} \cup \dots$  für  $m \in \mathbb{N}$ . Dann gilt  $F_1 \supset F_2 \supset F_3 \supset \dots$ , und  $\limsup E_n = \bigcap_m F_m$ . Deswegen ist aufgrund der Stetigkeit von Maßen  $\mathbb{P}(\limsup E_n) = \lim_m \mathbb{P}(F_m)$  (vgl. Satz 1.3.2(vi)). Da Wahrscheinlichkeitsmaße subadditiv sind (Satz 1.3.2(iv)), ist  $\mathbb{P}(F_m) \leq \sum_{n=m}^{\infty} \mathbb{P}(E_n)$ . Die Zahlen auf der rechten Seite gehen gegen Null, da  $\sum_n \mathbb{P}(E_n)$  eine konvergente Reihe ist. Damit ist alles gezeigt.  $\square$

**Lemma 8.1.2.** (*Lemma von Borel-Cantelli<sup>4)</sup>, Teil 2)*

$E_1, E_2, \dots$  seien unabhängige Ereignisse in einem Wahrscheinlichkeitsraum.

Gilt dann  $\sum_m \mathbb{P}(E_n) = \infty$ , so ist  $\mathbb{P}(\limsup E_n) = 1$ .

**Beweis:** Wir werden zeigen, dass  $F := \Omega \setminus \limsup E_n$  Wahrscheinlichkeit Null hat. Nun ist „ $x \in F$ “ gleichbedeutend zu „ $x$  liegt nur in endlich vielen  $E_n$ “ und folglich zu „für ein geeignetes  $m$  ist  $x \in \bigcap_{n \geq m} (\Omega \setminus E_n)$ “. Das bedeutet  $F = \bigcup_m \bigcap_{n \geq m} F_n$ , wobei wir  $F_n$  als  $\Omega \setminus E_n$  definiert haben<sup>5)</sup>.

Um  $\mathbb{P}(F) = 0$  zu zeigen, reicht es folglich,  $\mathbb{P}(\bigcap_{n \geq m} F_n) = 0$  für  $m = 1, 2, \dots$  zu beweisen, denn wegen Satz 1.3.2(iv) ist die Vereinigung von abzählbar vielen Nullmengen wieder eine Nullmenge.

Für diesen noch fehlenden Beweisschritt kombinieren wir die Voraussetzung der Unabhängigkeit mit einer Ungleichung der Analysis. Wir fixieren ein  $m$  und ein  $k$  und betrachten die Ereignisse  $F_m, \dots, F_{m+k}$ . Sie sind unabhängig, und folglich ist

$$\begin{aligned} \mathbb{P}(F_m \cap \dots \cap F_{m+k}) &= \mathbb{P}(F_m) \cdots \mathbb{P}(F_{m+k}) \\ &= (1 - \mathbb{P}(E_m)) \cdots (1 - \mathbb{P}(E_{m+k})). \end{aligned}$$

<sup>3)</sup>Emile Borel, 1871 bis 1956. Er war Professor an der Sorbonne in Paris, viele mathematische Bereiche wurden durch seine Ergebnisse wesentlich vorangebracht. Insbesondere trug er viel zur präzisen Begründung der Maßtheorie bei.

<sup>4)</sup>Francesco Cantelli, 1875 bis 1966. Er war Professor für Versicherungsmathematik in Catania und Rom. Die Lemmata von Borel-Cantelli wurden von Borel und Cantelli unabhängig voneinander gefunden.

<sup>5)</sup>Diese Tatsache kann auch als  $\Omega \setminus \limsup E_n = \liminf_n (\Omega \setminus E_n)$  interpretiert werden.



Borel



Cantelli

Aus der Analysis übernehmen wir die Ungleichung  $1 - x \leq e^{-x}$ . (Wir benötigen sie nur für  $x \in [0, 1]$ , und dafür folgt sie sofort daraus, dass die Reihenglieder in der Entwicklung von  $e^{-x} = 1 - x + x^2/2! - x^3/3! + \dots$  betragsmäßig fallen und alternierende Vorzeichen haben.) Daraus schließen wir, dass

$$\mathbb{P}(F_m \cap \dots \cap F_{m+k}) \leq e^{-(\mathbb{P}(E_m) + \dots + \mathbb{P}(E_{m+k}))}$$

gilt. Und da die Reihe  $\mathbb{P}(E_n)$  divergiert, wird die rechte Seite der Ungleichung beliebig klein. Anders ausgedrückt (und hier nutzen wir noch einmal die Stetigkeit von  $\mathbb{P}$  aus):

$$\mathbb{P}\left(\bigcap_{n \geq m} F_n\right) = \lim_k \mathbb{P}\left(\bigcap_{n=m}^{n=m+k} F_n\right) = 0.$$

Das beweist die Behauptung.  $\square$

Es folgen zwei *Bemerkungen*:

**1.** Im ersten Lemma von Borel-Cantelli konnten die  $E_n$  beliebige Ereignisse sein, im zweiten Lemma wurde die *Unabhängigkeit* vorausgesetzt. Für ganz allgemeine Situationen ist ein entsprechendes Ergebnis auch nicht zu erwarten, denn ist  $A$  ein Ereignis mit  $\mathbb{P}(A) \in ]0, 1[$  und sind alle  $E_n$  gleich  $A$ , so ist  $\sum \mathbb{P}(E_n) = +\infty$ , aber  $\mathbb{P}(\limsup_n E_n) = \mathbb{P}(A) < 1$ .

Ein Beweis dafür, dass es beim zweiten Lemma überraschenderweise ausreicht, die paarweise Unabhängigkeit zu fordern, findet man in den Ergänzungen zu diesem Kapitel auf Seite 258.

**2.** Angenommen, die Ereignisse  $E_1, E_2, \dots$  sind unabhängig. Dann gibt es für die Wahrscheinlichkeit von  $\limsup_n E_n$  nur zwei extreme Möglichkeiten: Sie ist entweder gleich Null oder gleich Eins, je nachdem, ob  $\sum_n \mathbb{P}(E_n)$  konvergiert oder divergiert. Das ist also ein typisches *Null-Eins-Gesetz*.

Zum Abschluss dieses Abschnitts werden wir *zwei Anwendungen* der Lemmata von Borel-Cantelli studieren.

Anwendung von Lemma 1: Konvergenz

Mehrere der später in diesem Kapitel zu beweisenden Resultate beruhen auf einer geschickten Anwendung des folgenden Satzes:

**Satz 8.1.3.**  *$X_1, X_2, \dots$  seien reellwertige Zufallsvariable auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{E}, \mathbb{P})$ . Wir nehmen an, dass für jedes  $\varepsilon > 0$*

$$\sum_n \mathbb{P}(|X_n| > \varepsilon) < +\infty$$

*gilt. Dann konvergieren die  $X_n$  punktweise fast sicher gegen die Nullfunktion.*

**Beweis:** Fixiere  $\varepsilon > 0$ , setze  $E_n^\varepsilon := \{|X_n| > \varepsilon\}$  und bezeichne mit  $N_\varepsilon$  den Limes superior der  $E_n^\varepsilon$ . Aufgrund des ersten Lemmas von Borel-Cantelli ist  $\mathbb{P}(N_\varepsilon) = 0$ .

Das heißt doch: Abgesehen von den  $\omega$  in der Nullmenge  $N_\varepsilon$  ist es richtig, dass  $\omega$  in nur endlich vielen  $E_n^\varepsilon$  liegt. Oder anders ausgedrückt: Für die  $\omega \notin N_\varepsilon$  gibt es ein  $n_0$ , so dass  $|X_n(\omega)| \leq \varepsilon$  für  $n \geq n_0$ .

Das ist schon beinahe die Behauptung, wir brauchen die entsprechende Aussage aber für *alle* positiven  $\varepsilon$ . Dazu betrachten wir nacheinander spezielle  $\varepsilon$ , nämlich  $\varepsilon = 1, 1/2, 1/3, \dots$  und definieren  $N$  als Vereinigung der dadurch erzeugten Nullmengen  $N_1, N_{1/2}, N_{1/3}, \dots$  Das ist eine Nullmenge, und für  $\omega \notin N$  kann die vorstehende Überlegung für  $\varepsilon = 1, 1/2, \dots$  angewendet werden. Es gilt folglich  $X_n(\omega) \rightarrow 0$ .  $\square$

Unter den Bedingungen des Satzes liegt also Konvergenz vor, und das wird, wie wir sehen werden, interessante Folgerungen gestatten. Es soll aber nicht verschwiegen werden, dass so gut wie immer noch wichtige Fragen offen bleiben. Das Problem kann schon im Rahmen der Analysis erläutert werden:

Mal angenommen, wir haben mit irgendwelchen abstrakten Methoden bewiesen, dass  $\lim x_n = x$  gilt. Dann weiß man doch (zum Beispiel), dass es ein  $n_0$  gibt, dass  $|x_n - x_0| \leq 1/1000$  für  $n \geq n_0$ .

*Doch wie groß muss man  $n_0$  wählen?*

Wie groß muss etwa  $n$  sein, dass die Eulersche Zahl  $e$  bis auf drei Stellen nach dem Komma mit  $1 + 1/1! + \dots + 1/n!$  übereinstimmt?

Die Antworten auf die entsprechenden Fragen bei der punktweisen Konvergenz fast sicher können leider aus dem vorstehenden Satz (und dem zugehörigen Beweis) nicht abgelesen werden. Schlimmer noch: Wenn ein konkretes  $\omega \in \Omega$  vorliegt so weiß man erstens nicht, ob  $X_n(\omega) \rightarrow 0$  gilt (denn  $\omega$  könnte ja in der Ausnahme-Nullmenge liegen), und selbst wenn Konvergenz gegen Null vorliegt, kann nicht gesagt werden, wie groß bei konkret vorgegebenem  $\varepsilon$  das  $n_0$  sein muss, damit  $|X_n(\omega)| \leq \varepsilon$  für  $n \geq n_0$  gilt. Beweise, die die Lemmata von Borel-Cantelli verwenden, geben das nicht her.

#### Anwendung von Lemma 2: Der Affe an der Schreibmaschine

Der nächste Satz könnte auch so heißen: „Alles, was eine positive Wahrscheinlichkeit hat, passiert auch irgendwann einmal“:

**Satz 8.1.4.** *Es seien  $X_1, X_2, \dots$  unabhängige und identisch verteilte reellwertige Zufallsvariable auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{E}, \mathbb{P})$ . Weiter seien  $B_1, \dots, B_k$  Borelmengen in  $\mathbb{R}$ , so dass  $\mathbb{P}(\{X_n \in B_i\})$  für alle  $n$  und  $i = 1, \dots, k$  positiv ist<sup>6)</sup>.*

*Dann hat die Menge der  $\omega$ , für die es ein  $n$  mit der Eigenschaft*

$$X_{n+1}(\omega) \in B_1, \dots, X_{n+k}(\omega) \in B_k$$

<sup>6)</sup>Da die  $X_n$  identisch verteilt sind, hätte es gereicht zu fordern, dass alle  $\mathbb{P}(\{X_1 \in B_i\})$ ,  $i = 1, \dots, k$ , positiv sind.

gibt, Wahrscheinlichkeit Eins.

Fasst man also  $X_1, X_2, \dots$  als Zufallsspaziergang in  $\mathbb{R}$  auf, so kann man sich (mit Wahrscheinlichkeit Eins) darauf verlassen, dass der Spaziergang „irgendwann einmal“ die  $B_1, \dots, B_k$  in der richtigen Reihenfolge besucht.

**Beweis:** Für  $r = 0, 1, 2, \dots$  sei  $E_r$  das Ereignis

$$E_r := \{\omega \mid X_{k \cdot r+1}(\omega) \in B_1, \dots, X_{k \cdot r+k}(\omega) \in B_k\}.$$

Da die  $X_n$  unabhängig und identisch verteilt sind, haben alle  $E_r$  die Wahrscheinlichkeit  $\alpha := \mathbb{P}(\{X_1 \in B_1\}) \cdots \mathbb{P}(\{X_k \in B_k\})$ ; diese Zahl ist nach Voraussetzung positiv. Außerdem sind die  $E_1, E_2, \dots$  unabhängig, denn diese Ereignisse sind aus den  $X_n$  konstruiert worden, wobei die auftretenden Indexmengen paarweise disjunkt sind (vgl. Satz 4.4.8).

Damit kann das zweite Lemma von Borel-Cantelli angewendet werden: Mit Wahrscheinlichkeit Eins liegt ein  $\omega$  in unendlich vielen  $E_r$ , und das entspricht der Behauptung<sup>7)</sup>.  $\square$

Die Konsequenzen dieses Satzes sind erstaunlich:

1. Für alle  $n$  sei  $\mathbb{P}_{X_n}$  die Gleichverteilung auf  $\{1, 2, 3, 4, 5, 6\}$ . (Das kann man sich als nicht abbrechende Folge von Würfelwürfen vorstellen.) Wenn man dann eine beliebig große aus den Ziffern 1, 2, 3, 4, 5, 6 gebildete Zahl vorgibt, so werden die Ziffern dieser Zahl in der Folge dieser Würfelwürfe mit Wahrscheinlichkeit Eins irgendwann einmal in der richtigen Reihenfolge auftauchen. Sogar unendlich oft.
2. Wenn ein Zufallsgenerator Elemente aus  $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$  erzeugt, so dass jede Ziffer eine positive Wahrscheinlichkeit hat, so kann man bei genügend häufiger Wiederholung (fast) sicher sein, dass in der Folge der Abfragen die eigene Telefonnummer auftaucht. Oder die eigene Geheimzahl am Bankautomaten. Oder ...
3. Sei  $E$  ein Ereignis mit positiver Wahrscheinlichkeit in einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{E}, \mathbb{P})$ . Dabei kann  $\mathbb{P}(E)$  beliebig klein sein.

Wenn man dann immer wieder Zufallsabfragen  $\omega$  aus  $\Omega$  erzeugt, wird mit Wahrscheinlichkeit Eins unendlich oft das  $\omega$  in  $E$  liegen: Alles, was eine positive Wahrscheinlichkeit hat, passiert auch. Sogar unendlich oft. Jedenfalls mit Wahrscheinlichkeit Eins.

(Um das auf die vorstehenden Ergebnisse zurückzuführen, kann man unabhängige Kopien der Zufallsvariablen  $\chi_E$  betrachten.)

Am bekanntesten ist aber die folgende Veranschaulichung des Ergebnisses.

---

<sup>7)</sup>Es folgt sogar mehr: Erstens weiß man, dass das  $n$  aus der Aussage des Satzes von der Form  $r \cdot k$  gewählt werden kann, und zweitens gibt es mit Wahrscheinlichkeit Eins sogar unendlich viele  $n$  mit der fraglichen Eigenschaft.

### Der Affe an der Schreibmaschine

Wir setzen einen Affen an eine Schreibmaschine und lassen ihn darauf herumtippen. (Das ist die übliche Verkleidung der Geschichte. Man könnte statt der Schreibmaschine genauso gut eine Computer-tastatur nehmen.) Wird etwas Vernünftiges herauskommen?

Aufgrund der hier besprochenen Ergebnisse ist die Antwort ein klares „Ja!“. Jedes erdenkliche Werk wird „irgenwann“ einmal einge-tipt werden: Das Berliner Telefonbuch, das Manuskript dieses Bu-ches, usw.

Und wozu, bitte, braucht man da noch so etwas wie Kreativität, wenn auch jeder Affe – zum Beispiel – jedes Gedicht von Baudelaire produzieren kann, wenn man ihm nur genug Zeit lässt?

Es ist wirklich etwas verwirrend, dass der Zufall der Produzent auch der anspruchsvollsten literarischen Produkte sein kann. Es besteht trotzdem kein Grund zur Beunruhigung. Es folgen zwei beruhigende Argumente:

- Tatsächlich müsste man ziemlich lange warten, um etwas Sinnvolles zu erhalten. Als Beispiel wollen wir einmal überschlagen, wie lange man warten muss, bis das Wort „STOCHASTIK“ getippt wird. Dazu setzen wir den Affen an eine Tastatur, mit der man der Einfachheit halber nur die 26 Großbuchstaben des Alphabets erzeugen kann. Nun werden – immer wieder – zehn Buchstaben getippt. Wie oft muss das versucht werden, damit das Ergebnis „STOCHASTIK“ ist?

Wir wollen annehmen, dass der Affe keine Buchstaben bevorzugt und die einzelnen Buchstaben unabhängig voneinander produziert. Dann können wir das Experiment mit der Gleichverteilung auf den 26 Buchstaben erzeugen, und die Wahrscheinlichkeit, dass ein Versuch zum Ergebnis „STO-CHASTIK“ führt, ist gleich

$$p := 1/26^{10} = 1/141.167.095.653.376,$$

also etwa Eins zu 140 Billionen.

Wir haben es also mit einem Bernoulliexperiment mit winziger Erfolgs-wahrscheinlichkeit  $p$  zu tun. Aus Abschnitt 6.3 wissen wir, dass der Er-wartungswert der Anzahl der Versuche bis zum ersten Erfolg gleich  $1/p$  ist: So viele Experimente sollten wir also einkalkulieren.

Und wie lange dauert das? Wir können ja immer neue Affen beschäftigen und rund um die Uhr tippen lassen. Wenn dann ein Versuch 30 Sekunden dauert, schafft man 120 pro Stunde,  $24 \cdot 120 = 2880$  am Tag und folglich  $365 \cdot 24 \cdot 120 = 1.051.200$  pro Jahr. Wir sollten also

$$\frac{141.167.095.653.376}{1.051.200} \approx 0.134 \cdot 10^9$$

Jahre, also rund 134 Millionen Jahre einkalkulieren. Das ist ziemlich viel für das Warten auf auf so etwas Einfaches wie „STOCHASTIK“ ...

→  
Programm!

Selbst wenn wir Affen durch Computer ersetzen, würde das Warten noch unvertretbar lange dauern. Sogar der wesentlich bescheidenere Wunsch, ein „STOCH“ zu erzeugen, hat eine Wahrscheinlichkeit von etwa 1 zu 12 Millionen, des entspricht beinahe der Wahrscheinlichkeit für einen Sechser im Lotto: Mit 12 Millionen Schritten ist also zu rechnen.

Es ist auch zu beachten, dass die Streuung der einzukalkulierenden Versuchsanzahl erheblich ist: Wenn  $p$  die Erfolgswahrscheinlichkeit ist, ist der Wert gleich  $\sqrt{1-p}/p$ , also für kleine  $p$  in etwa  $1/p$ . (Vgl. die Tabelle in Abschnitt 3.3., dort ist bei der geometrischen Verteilung  $q := 1 - p$  zu setzen.)

- Zweitens ist es naiv zu glauben, dass ein intelligentes Produkt „einfach so“ plötzlich da ist. Ein intelligents Wesen ist doch erforderlich, um inmitten des Datenmülls etwas Sinnvolles zu erkennen. Egal, ob es das Wort „STOCHASTIK“ oder etwas wirklich Anspruchsvolles ist.



Bild 8.1.1: Ein Gedicht? Ein mathematisches Theorem? (Fotomontage: Elke Behrends)

Das Fazit: Es besteht kein Grund zur Beunruhigung: Die Mozarts, Einsteins, Gaußs sind weiterhin unverzichtbar.

## 8.2 Das schwache Gesetz der großen Zahlen

In diesem und den folgenden Abschnitten geht es darum, genauer zu untersuchen, inwieweit „der Zufall im Unendlichen verschwindet“. Allen Resultaten ist gemeinsam, dass wir es mit einer Zufallsvariablen  $X$  zu tun haben. Was ist dann bei vielen Abfragen im Mittel zu erwarten? Der Rahmen, so etwas mathematisch exakt behandeln zu können, ist in den Abschnitten 4.4 und 4.5 bereitgestellt worden. Wir wollen zur Untersuchung von  $X$  eine Folge  $(X_n)_{n=1,2,\dots}$  betrachten, so dass erstens die  $X_n$  unabhängig sind und zweitens  $\mathbb{P}_X = \mathbb{P}_{X_n}$  für alle  $n$  gilt. Um das nicht immer wieder neu präzisieren zu müssen, vereinbaren wir:

**Definition 8.2.1.** Es sei  $(X_n)$  eine auf einem Wahrscheinlichkeitsraum definierte Folge reellwertiger Zufallsvariablen. Wir sagen, dass sie unabhängig und identisch verteilt mit Verteilung  $\mathbb{P}_X$  ist, wenn gilt:

- (i)  $X$  ist eine reellwertige Zufallsvariable, die evtl. auf einem anderen Raum definiert ist.
- (ii) Die  $X_n$  sind unabhängig.
- (iii) Für alle  $n$  ist  $\mathbb{P}_{X_n} = \mathbb{P}_X$ .

Wir wissen dann schon, dass man für beliebige  $X$  solche  $X_1, X_2, \dots$  finden kann. Das ist der „Klonsatz“ aus Abschnitt 4.5. Auch haben wir in Abschnitt 4.3 gezeigt: Wenn für  $X$  der Erwartungswert (bzw. die Varianz) existiert, so gilt das auch für  $X_n$ . In diesem Fall ist  $\mathbb{E}(X_n) = \mathbb{E}(X)$  (bzw.  $V(X_n) = V(X)$ ) für alle  $n$ , denn Erwartungswert und Varianz hängen nur von der Verteilung ab.

Die nächsten Schritte werden die folgenden sein:

- *Schwaches Gesetz* (dieser Abschnitt): Die Mittelwerte  $(X_1 + \dots + X_n)/n$  konvergieren (unter gewissen Bedingungen) in Wahrscheinlichkeit gegen die konstante Funktion  $\mathbb{E}(X)$ .
- *Starkes Gesetz* (Abschnitt 8.3): Die Mittelwerte  $(X_1 + \dots + X_n)/n$  konvergieren (unter gewissen Bedingungen) punktweise fast sicher gegen die konstante Funktion  $\mathbb{E}(X)$ .
- *Zentraler Grenzwertsatz* (Abschnitt 8.4): Nach richtiger Skalierung (im  $n$ -ten Schritt wird durch  $\sqrt{n}\sigma$  geteilt) konvergiert  $X_1 + \dots + X_n - n\mathbb{E}(X)$  mit  $n \rightarrow \infty$  in Verteilung gegen die Standard-Normalverteilung.
- *Satz vom iterierten Logarithmus* (Abschnitt 8.5): Für große  $n$  ist, wie in den Abschnitten 8.2 und 8.3 gezeigt,  $(X_1 + \dots + X_n)/n \approx \mathbb{E}(X)$ , d.h.  $X_1 + \dots + X_n$  sollte dann nicht allzuweit von  $n\mathbb{E}(X)$  entfernt sein. Es kommt aber dabei immer wieder zu Schwankungen, die sich mit dem Satz vom iterierten Logarithmus recht genau beschreiben lassen.

Die Tschebyscheff-Ungleichung und die Markov-Ungleichung

In Abschnitt 4.3 hatten wir die Varianz einer Zufallsvariable als Maß für die mittlere quadratische Abweichung vom Erwartungswert eingeführt. Tatsächlich lässt sich damit auch etwas über die Abweichung selber aussagen. Genauer:

**Satz 8.2.2. (Tschebyscheff-Ungleichung<sup>8)</sup>** *Es sei  $X$  eine reellwertige Zufallsvariable auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{E}, \mathbb{P})$ . Wir nehmen an, dass Erwartungswert  $\mathbb{E}(X)$  und Varianz  $\sigma^2(X)$  existieren<sup>9)</sup>. Ist dann  $\varepsilon > 0$ , so gilt*

$$\mathbb{P}(\{|X - \mathbb{E}(X)| \geq \varepsilon\}) \leq \frac{\sigma^2(X)}{\varepsilon^2}.$$

**Beweis:** Ein  $\varepsilon > 0$  sei vorgegeben. Wir definieren ein Ereignis  $E_\varepsilon$  als die Menge  $\{|X - \mathbb{E}(X)| \geq \varepsilon\}$ . Dann ist für jedes  $\omega$

$$|X(\omega) - \mathbb{E}(X)|^2 \geq \varepsilon^2 \chi_{E_\varepsilon}(\omega),$$

<sup>8)</sup>Pafnuty Tschebyscheff (im Englischen: Chebyshev), 1821 bis 1894. Die wichtigsten Beiträge dieses russischen Mathematikers betreffen die Zahlentheorie und die mathematische Physik. Eine der am häufigsten verwendeten wahrscheinlichkeitstheoretischen Ungleichungen ist nach ihm benannt.

<sup>9)</sup>Eigentlich müsste es  $(\sigma(X))^2$  heißen, da  $\sigma(X) = \sqrt{V(X)}$ . Im Interesse der besseren Lesbarkeit schreiben wir einfacher  $\sigma^2(X)$ .



Tschebyscheff

denn für  $\omega \notin E_\varepsilon$  ist die rechte Seite Null, und für  $\omega \in E_\varepsilon$  hat die linke Seite nach Definition mindestens den Wert  $\varepsilon^2$ . Und da der Übergang zu Erwartungswerten monoton ist, folgt

$$\begin{aligned}\sigma^2(X) &= \mathbb{E}(|X - \mathbb{E}(X)|^2) \\ &\geq \mathbb{E}(\varepsilon^2 \chi_{E_\varepsilon}) \\ &= \varepsilon^2 \mathbb{P}(E_\varepsilon).\end{aligned}$$

Dabei haben wir auch ausgenutzt, dass  $\mathbb{E}(cY) = c\mathbb{E}(Y)$  und  $\mathbb{E}(\chi_E) = \mathbb{P}(E)$  gilt (vgl. Satz 3.3.4.).

Nun muss nur noch durch  $\varepsilon^2$  geteilt werden.  $\square$

Hier einige *Kommentare* zu dieser Ungleichung:

**1.** Durch die Tschebyscheff-Ungleichung wird die Wahrscheinlichkeit, dass die Zahl  $|X(\omega) - \mathbb{E}(X)|$  groß ist, *nach oben* abgeschätzt: Sie ist höchstens soundso groß. Durch Übergang zur Komplementärmenge von  $\{|X - \mathbb{E}(X)| \geq \varepsilon\}$  kann man auch Abschätzungen *nach unten* für die Wahrscheinlichkeit finden, dass  $|X(\omega) - \mathbb{E}(X)|$  klein ist. Aus  $\mathbb{P}(\Omega \setminus E) = 1 - \mathbb{P}(E)$  folgt nämlich sofort

$$\mathbb{P}(\{|X - \mathbb{E}(X)| < \varepsilon\}) \geq 1 - \frac{\sigma^2(X)}{\varepsilon^2}.$$

**2.** In manchen Fällen wird die Tschebyscheff-Ungleichung zu einer recht nutzlosen Information führen. Zum Beispiel liefert sie im Fall  $\sigma^2(X) = 1$  und  $\varepsilon = 0.1$  die Abschätzung  $\mathbb{P}(\{|X - \mathbb{E}(X)| \geq 0.1\}) \leq 100$ . Das ist wirklich nicht spannend, denn Wahrscheinlichkeiten sind ja immer durch Eins beschränkt.

**3.** Im Beweis haben wir sehr grob abgeschätzt, als wir von  $|X - \mathbb{E}(X)|^2$  zu  $\varepsilon^2 \chi_{E_\varepsilon}$  übergegangen sind. Deswegen wird in den meisten konkreten Fällen die Wahrscheinlichkeit von  $\{|X - \mathbb{E}(X)| \geq \varepsilon\}$  viel kleiner sein als  $\sigma^2(X)/\varepsilon^2$ .

Die Tschebyscheff-Ungleichung kann als Spezialfall eines viel allgemeineren Ergebnisses aufgefasst werden, das wir etwas später in diesem Abschnitt benötigen werden:

**Satz 8.2.3. (Markov-Ungleichung<sup>10)</sup>)** Ist  $g : [0, \infty[ \rightarrow [0, \infty[$  eine monoton wachsende Funktion und  $X$  eine Zufallsvariable, so gilt für jedes  $\varepsilon > 0$ :

$$\mathbb{P}(|X| \geq \varepsilon) \leq \frac{\mathbb{E}(g \circ |X|)}{g(\varepsilon)}.$$

Dabei wird vorausgesetzt, dass der Erwartungswert von  $g \circ |X|$  existiert.

---

<sup>10)</sup> Andrej Markov, 1856 bis 1922. Er ist einer der wichtigsten Vertreter der russischen wahrscheinlichkeitstheoretischen Schule. Seit 1886 war er Professor in St. Petersburg. Die von ihm systematisch untersuchten (heute so genannten) Markovprozesse spielen eine wichtige Rolle in der Wahrscheinlichkeitsrechnung.



Markov

**Beweis:** Der *Beweis* ist ganz ähnlich wie beim Beweis der Tschebyscheff-Ungleichung: Setze diesmal  $E_\varepsilon := \{|g(X)| \geq \varepsilon\}$  und beachte, dass (wegen der Monotonie von  $g$ )  $|g \circ X| \geq g(\varepsilon) \chi_{E_\varepsilon}$  gilt.  $\square$

Die Tschebyscheff-Ungleichung ergibt sich, wenn man die Markovungleichung auf die durch  $\tilde{X}(\omega) := X(\omega) - \mathbb{E}(X)$  definierte Zufallsvariable  $\tilde{X}$  und die monotone Funktion  $g(x) := x^2$  (für  $x \geq 0$ ) anwendet.

Direkte Folgerungen aus der Tschebyscheff-Ungleichung

Geht man von einer Zufallsvariablen zu Mittelwerten unabhängiger Kopien über, so werden die Varianzen kleiner<sup>11)</sup>. Mit diesem Ergebnis lässt sich die folgende Variante der Tschebyscheff-Ungleichung beweisen:

**Satz 8.2.4.** *Die  $X_1, X_2, \dots, X_n$  seien unabhängig und identisch verteilt mit Verteilung  $\mathbb{P}_X$  (vgl. Definition 8.2.1). Außerdem nehmen wir an, dass Erwartungswert  $\mathbb{E}(X)$  und Varianz  $\sigma^2(X)$  von  $X$  existieren. Dann ist*

$$\mathbb{P}\left(\left\{\left|\frac{X_1 + \dots + X_n}{n} - \mathbb{E}(X)\right| \geq \varepsilon\right\}\right) \leq \frac{\sigma^2(X)}{n\varepsilon^2}.$$

**Beweis:** Sei  $\tilde{X}$  die Zufallsvariable  $(X_1 + \dots + X_n)/n$ . Da alle  $X_i$  wie  $X$  verteilt sind, hat  $\tilde{X}$  Erwartungswert  $\mathbb{E}(X)$ , und wegen Satz 4.6.2 ist  $\sigma^2(\tilde{X}) = \sigma^2(X)/n$ . Nun ist nur noch die Tschebyscheff-Ungleichung auf  $\tilde{X}$  anzuwenden.  $\square$

Wir behandeln einige *typische Anwendungen* dieser Ungleichung.

**1.**  $X$  nehme die Werte 1 und 0 jeweils mit Wahrscheinlichkeit  $p$  und  $1 - p$  an. (Es geht also um ein Bernoulliexperiment mit Erfolgswahrscheinlichkeit  $p$ .) Dann ist  $\mathbb{E}(X) = p$  und  $\sigma^2(X) = p(1 - p)$ . Der Ausdruck  $(X_1 + \dots + X_n)/n$  ist folglich die relative Erfolgsanzahl bei diesem Experiment, wenn  $n$  Versuche durchgeführt werden. Der Satz besagt dann:  $(X_1 + \dots + X_n)/n$  wird höchstens mit Wahrscheinlichkeit  $p(1 - p)/(n\varepsilon^2)$  um mehr als  $\varepsilon$  von  $p$  abweichen.

**2.** Mit einer Zusatzüberlegung kann man die vorstehende Abschätzung verwenden, um Informationen über das eventuell noch unbekannte  $p$  zu erhalten. Man muss nur bemerken, dass  $p(1 - p)$  für die  $p \in [0, 1]$  durch  $1/4$  nach oben abgeschätzt werden kann.

Hier ein Beispiel. Wir wollen wissen, mit welcher Wahrscheinlichkeit  $p$  eine in die Luft geworfene Reißzwecke nach dem Fallen so liegt, dass die Spitze nach oben zeigt. Wir machen 1000 Versuche und stellen fest, dass das in 653 Fällen eingetreten ist. Satz 8.2.4 (zusammen mit der vorstehenden Zusatzüberlegung) besagt, dass das Ergebnis  $|653/1000 - p| \geq \varepsilon$  höchstens mit Wahrscheinlichkeit  $(1/4)/(1000\varepsilon^2)$  zu erwarten war. Oder anders ausgedrückt: Mit Wahrscheinlichkeit von mindestens  $1 - (1/4)/(1000\varepsilon^2)$  ist der Abstand von  $p$  zu 0.635 höchstens  $\varepsilon$ . Für  $\varepsilon = 0.1$  etwa heißt das: Mit Wahrscheinlichkeit von mindestens  $1 - 0.25/10 = 97.5$  liegt  $p$  in  $[0.635 - 0.1, 0.635 + 0.1]$ .

<sup>11)</sup>Das haben wir als Wurzel- $n$ -Gesetz in Satz 4.6.2 bewiesen.

Das ist eine recht schwache Aussage angesichts der 1000 Versuche, aber eine größere Genauigkeit ist bei der experimentellen Bestimmung von Wahrscheinlichkeiten nur mit sehr vielen Versuchen zu erreichen.

**3.** Jetzt wollen wir es genauer wissen:  $p$  soll auf 0.01 genau bestimmt werden, und zwar mit 98 Prozent Sicherheit. Wie groß muss dann die Anzahl  $n$  der Versuche sein?

Wir müssen garantieren, dass  $1 - (1/4)/(n \cdot 0.01^2) \geq 0.98$  gilt. Wenn man das nach  $n$  auflöst, erhält man  $n \geq (1/4)/(0.02 \cdot 0.01^2) = 125.000$ .

**4.** In der Formel steht das  $\varepsilon$  auf der rechten Seite quadriert im Nenner. Wenn man es halbiert, muss man also von  $n$  Versuchen zu  $4n$  Versuchen übergehen, um die gleiche Abschätzung für die Wahrscheinlichkeit zu erhalten.

### Das schwache Gesetz der großen Zahlen

Aus der Tschebyscheff-Ungleichung lässt sich leicht ablesen, dass Mittelwerte aus unabhängigen Abfragen immer besser den Erwartungswert approximieren. Genauer gilt:

**Satz 8.2.5.** (*Schwaches Gesetz der großen Zahlen*) Die  $X_1, X_2, \dots$  seien unabhängig und identisch verteilt mit Verteilung  $\mathbb{P}_X$  (vgl. Definition 8.2.1). Auch diesmal nehmen wir an, dass Erwartungswert und Varianz von  $X$  existieren.

Unter diesen Voraussetzungen geht die Folge der Mittelwerte

$$\left( \frac{X_1 + \dots + X_n}{n} \right)_{n \in \mathbb{N}}$$

mit  $n \rightarrow \infty$  in Wahrscheinlichkeit gegen die konstante Funktion  $\mathbb{E}(X)$ .

**Beweis:** Es ist doch zu zeigen, dass

$$\mathbb{P} \left( \left\{ \left| \frac{X_1 + \dots + X_n}{n} - \mathbb{E}(X) \right| \geq \varepsilon \right\} \right)$$

für jedes  $\varepsilon > 0$  gegen Null konvergiert. Das folgt aber sofort aus Satz 8.2.4, denn danach sind diese Zahlen durch  $\sigma^2(X)/(n\varepsilon^2)$  abschätzbar.  $\square$

Für unseren Beweis war wesentlich, dass die Varianz von  $X$  existiert. Die kommt in der Aussage des Satzes allerdings nicht vor, und man kann sich fragen, ob die Aussage auch dann gilt, wenn über die Existenz von  $\sigma^2(X)$  nichts bekannt ist. Die Antwort ist „Ja!“. Der Beweis ist allerdings wesentlich schwieriger als der vorstehende. Wer möchte, kann gleich zu Abschnitt 8.3 weiterblättern.

**Satz 8.2.6.** (*Schwaches Gesetz der großen Zahlen, allgemeine Version*) Die Aussage ist exakt so wie in Satz 8.2.5. Wir setzen allerdings nicht voraus, dass  $\sigma^2(X)$  existiert.

**Beweis:** Wir zeigen das Ergebnis in mehreren Schritten.

1. Eine Definition: Es sei  $(X_n)$  eine Folge von Zufallsvariablen, die auf dem gleichen Wahrscheinlichkeitsraum definiert sind und für die der Erwartungswert existiert. Wir sagen, dass die  $(X_n)$  dem schwachen Gesetz der großen Zahlen genügen, wenn

$$\frac{X_1 + \dots + X_n - (\mathbb{E}(X_1) + \dots + \mathbb{E}(X_n))}{n}$$

in Wahrscheinlichkeit gegen die Nullfunktion geht. Das ist nach der schon bewiesenen Variante des schwachen Gesetzes zum Beispiel dann der Fall, wenn die  $X_n$  unabhängig und identisch verteilt sind und die Varianz existiert.

2. Einige vorbereitende Ergebnisse: Wir beweisen als Vorbereitung für den Hauptbeweis einige Hilfsaussagen.

2a) Genügen  $(X_n)$  und  $(Y_n)$  dem schwachen Gesetz (beide Folgen sollen auf dem gleichen Wahrscheinlichkeitsraum definiert sein), so auch  $(X_n + Y_n)$ .

Diese Aussage folgt unmittelbar aus Satz 7.1.2(iii) und der Linearität des Erwartungswerts.

2b) Die  $(X_n)$  sollen alle existierende Erwartungswerte und Varianzen haben, und sie seien unabhängig. Weiter gebe es eine Konstante  $C$ , so dass alle Varianzen  $\sigma^2(X_n)$  durch  $C$  beschränkt sind. Dann genügt  $(X_n)$  dem schwachen Gesetz der großen Zahlen.

*Beweis dazu:* Setze  $Y_n := (X_1 + \dots + X_n)/n$ . Wegen Satz 4.6.2(i) ist

$$\sigma^2(Y_n) = \frac{\sigma^2(X_1) + \dots + \sigma^2(X_n)}{n^2} \leq \frac{C}{n},$$

und mit der Tschebyscheffungleichung für  $Y_n$  folgt

$$\begin{aligned} \mathbb{P}\left(\left|\frac{X_1 + \dots + X_n - (\mathbb{E}(X_1) + \dots + \mathbb{E}(X_n))}{n}\right| \geq \varepsilon\right) &= \mathbb{P}(|Y_n - \mathbb{E}(Y_n)| \geq \varepsilon) \\ &\leq \frac{C}{n \cdot \varepsilon^2}; \end{aligned}$$

das geht mit  $n \rightarrow \infty$  gegen Null.

Der Beweis zeigt sogar, dass es gereicht hätte zu fordern, dass die Varianz von  $(X_1 + \dots + X_n)/n$  gegen Null geht. Das wird gleich für den Hauptbeweis wichtig werden.

2c) Ist  $(X_n)$  eine Folge mit  $\mathbb{E}(|X_n|) \rightarrow 0$ , so genügt  $(X_n)$  dem schwachen Gesetz der großen Zahlen.

*Beweis dazu:* Zunächst muss man sich klar machen, dass eine Folge  $(Y_n)$  von Zufallsvariablen in Wahrscheinlichkeit gegen Null geht, wenn  $\mathbb{E}(|Y_n|)$  gegen Null konvergiert. Das folgt sofort aus der Markovungleichung mit  $g(x) := x$ :

$$\mathbb{P}(|Y_n| \geq \varepsilon) \leq \frac{\mathbb{E}(|Y_n|)}{\varepsilon} \rightarrow 0.$$

Und das kann auf  $Y_n = (X_1 + \dots + X_n - (\mathbb{E}(X_1) + \dots + \mathbb{E}(X_n))) / n$  angewendet werden:

$$\mathbb{E}(|Y_n|) \leq 2 \frac{\mathbb{E}(|X_1|) + \dots + \mathbb{E}(|X_n|)}{n} \rightarrow 0.$$

(Hier spielte ein klassisches Ergebnis der Analysis eine Rolle: Aus  $x_n \rightarrow 0$  folgt  $(x_1 + \dots + x_n)/n \rightarrow 0$ .)

*3. Der Hauptbeweis:* Wir werden nun die wie im Satz vorgegebene Folge als Summe von zwei Folgen zu schreiben, die beide dem schwachen Gesetz genügen. Der Trick besteht darin, die  $X_n$  „geschickt“ abzuschneiden.

Genauer: Definiere für jedes  $n$  eine Zufallsvariable  $Z_n$  durch  $Z_n(\omega) := X_n(\omega)$ , falls  $|X_n(\omega)| \leq n^{1/4}$  und  $Z_n(\omega) := 0$  sonst: Durch diese ziemlich technische Definition wird von  $X_n$  alles abgeschnitten, was „zu groß“ ist. (Dass  $n^{1/4}$  eine gute Wahl ist, wird sich gleich zeigen.)

Setzt man noch  $W_n := X_n - Z_n$ , so ist sicher  $X_n = Z_n + W_n$ . Wir zeigen, dass sowohl  $(Z_n)$  als auch  $(W_n)$  dem schwachen Gesetz der großen Zahl genügen. Wegen der Vorbereitungen wären wir dann fertig.

Die  $(Z_n)$ . Alle  $Z_n$  liegen zwischen  $-n^{1/4}$  und  $n^{1/4}$ , und wir erhalten mit Satz

$$3.3.9(\text{vi}): \sigma^2(Z_n) = \mathbb{E}(Z_n^2) - (\mathbb{E}(Z_n))^2 \leq \mathbb{E}(Z_n^2) \leq n^{1/2}.$$

Das impliziert

$$\begin{aligned} \sigma^2\left(\frac{Z_1 + \dots + Z_n}{n}\right) &= \frac{1}{n^2}(\sigma^2(Z_1) + \dots + \sigma^2(Z_n)) \\ &\leq \frac{n \cdot n^{1/2}}{n^2} \\ &= n^{-1/2}. \end{aligned}$$

(Hier sieht man, dass  $n^{1/4}$  eine gute Wahl für das Abschneiden war.) Aus dem (Zusatz zum) vorletzten Beweis-Vorbereitungsschritt 2b folgt, dass die  $(Z_n)$  dem schwachen Gesetz genügen.

Die  $(W_n)$ . Hier müssen wir eine Anleihe an ein Ergebnis der Maßtheorie machen: Ist  $Y$  eine Zufallsvariable mit existierendem Erwartungswert und  $(a_n)$  eine monoton steigende reelle Folge mit  $a_n \rightarrow \infty$ , so gilt

$$E(|Y^{[a_n]}|) \rightarrow 0;$$

dabei ist  $Y^{[a_n]}$  die unterhalb von  $a_n$  gekappte Funktion  $Y$ , also  $Y^{[a_n]}(\omega) := Y(\omega)$  für  $|Y^{[a_n]}(\omega)| > a_n$  und  $Y^{[a_n]}(\omega) := 0$  sonst<sup>12)</sup>.

Das ist hier so anzuwenden:  $W_n$  ist nach Definition die Funktion  $X_n^{[n^{1/4}]}$ . Sie hat, da  $X_n$  wie  $X$  verteilt ist, die gleiche Verteilung wie  $X^{[n^{1/4}]}$ . Und weil

<sup>12)</sup>Das Ergebnis folgt sofort aus dem Satz von der monotonen Konvergenz (vgl. Seite 358). Er ist auf die Funktionen  $Y - Y^{[a_n]}$  anzuwenden, die monoton gegen  $Y$  steigen. Deswegen konvergiert die Folge der Erwartungswerte der  $Y - Y^{[a_n]}$  gegen  $E(Y)$ , d. h. die  $E(Y^{[a_n]})$  gehen gegen Null.

für die Berechnung von Erwartungswerten nur die Verteilung eine Rolle spielt, gehen aufgrund des Ergebnisses aus dem vorigen Absatz die  $\mathbb{E}(|W_n|)$  gegen Null. Damit der Beweis vollständig geführt.  $\square$

### 8.3 Das starke Gesetz der großen Zahlen

Um den Unterschied zwischen dem schwachen Gesetz der großen Zahlen aus dem vorigen Abschnitt und dem starken Gesetz besser erläutern zu können, betrachten wir ein spezielles Beispiel.

Wir stellen uns einen fairen Würfel vor und würfeln ihn immer wieder. Nach dem schwachen Gesetz gilt: Ist  $n$  „groß“, so ist die Wahrscheinlichkeit, dass der Mittelwert der Augenzahlen um mehr als  $\varepsilon$  vom Erwartungswert (also von 3.5) abweicht, klein. Denn diese Wahrscheinlichkeiten gehen gegen Null.

Das ist noch nicht ganz das, was eigentlich plausibel wäre. Man erwartet doch, dass „im Regelfall“ (also mit Wahrscheinlichkeit Eins) die Mittelwerte der Augenzahlen gegen 3.5 konvergent sind. Diese Aussage – sie entspricht der Konvergenz fast sicher von  $(X_1 + \dots + X_n)/n$  gegen die konstante Funktion  $\mathbb{E}(X)$  – folgt aber nicht aus dem schwachen Gesetz, denn Konvergenz in Wahrscheinlichkeit impliziert *nicht* die Konvergenz fast sicher.

In diesem Abschnitt werden wir dieses weitergehende Ergebnis, das *starke Gesetz der großen Zahlen*, beweisen<sup>13)</sup>.

Es wird zwei Versionen des Beweises geben, nämlich

- Einen Beweis unter einer ziemlich starken Annahme an die  $X_n$ , die allerdings in so gut wie allen praktisch interessierenden Fällen erfüllt ist.
- Einen weiteren Beweis, bei dem wir nur voraussetzen, dass die Varianzen der  $X_n$  existieren.

Damit ist allerdings noch nicht alles gesagt, was man zum starken Gesetz weiß. Es gilt nämlich ohne irgendwelche Voraussetzungen an die beteiligten Zufallsvariablen. (Die Erwartungswerte müssen natürlich existieren, denn sonst lässt sich die Aussage nicht sinnvoll formulieren.). Der sehr technische Beweis würde jedoch den Rahmen dieses Buches sprengen. Interessierte finden das ganz allgemeine Ergebnis (den *Satz von Entemadi*) zum Beispiel im Buch von Klenke.

Ein (vergleichsweise) einfacher Beweis unter einer Zusatzannahme

Ausnahmsweise soll nicht ein fertiger und möglichst perfekter Beweis vorgeführt werden. Wir versetzen uns in die Situation eines Mathematikers, der das starke Gesetz beweisen möchte. Und dabei sehen wir ihm sozusagen über

<sup>13)</sup>Da Konvergenz fast sicher die Konvergenz in Wahrscheinlichkeit impliziert, sind viele Ergebnisse aus dem vorigen Abschnitt Korollare zu den hier bewiesenen. Es ist trotzdem sinnvoll, sie extra zu beweisen, da die beim Beweis des schwachen Gesetzes verwendeten Methoden wesentlich elementarer sind.

die Schulter, wie er nach und nach einen Beweis findet. Das Ergebnis wird am Ende der Überlegungen als Satz 8.3.1 formuliert.

*Was ist denn zu zeigen?* Es soll doch bewiesen werden, dass unter geeigneten Voraussetzungen  $(X_1 + \dots + X_n)/n$  punktweise fast sicher gegen die konstante Funktion  $\mathbb{E}(X)$  konvergiert. Das ist gleichbedeutend damit, dass die Folge der Funktionen  $(Y_1 + \dots + Y_n)/n$  punktweise fast sicher gegen die Nullfunktion geht, wenn wir  $Y_i := X_i - \mathbb{E}(X)$  setzen. Die  $Y_i$  haben aber Erwartungswert Null, und das bedeutet: Ohne Beschränkung der Allgemeinheit dürfen wir annehmen, dass  $\mathbb{E}(X) = 0$  gilt.

*Eine Erinnerung.* Als Anwendung des ersten Lemmas von Borel-Cantelli haben wir doch in Satz 8.1.3 bewiesen, dass  $Z_n \rightarrow 0$  fast sicher gilt, wenn man weiß, dass  $\sum_n \mathbb{P}\{|Z_n| < \varepsilon\} < +\infty$  für jedes positive  $\varepsilon$  gilt. (Das ist hier für  $Z_n := (X_1 + \dots + X_n)/n$  anzuwenden.)

Doch wie kann man garantieren, dass die Summe der  $\mathbb{P}(\{|Z_n| \geq \varepsilon\})$  endlich ist? Man könnte vielleicht die Markovungleichung (Satz 8.2.3) anwenden: Danach ist  $\mathbb{P}(\{|Z_n| \geq \varepsilon\}) \leq \mathbb{E}(Z_n)/\varepsilon$ , wenn man in dieser Ungleichung  $g(x) := x$  setzt.

Nutzt das hier etwas? Wir wären fertig, wenn

$$\sum_n \mathbb{E}\left(\left|\frac{X_1 + \dots + X_n}{n}\right|\right) < \infty$$

wäre. Doch das ist recht unhandlich, da es keine Formeln gibt, mit denen man den Erwartungswert eines Betrages ausrechnen kann.

*Die erste Idee.*  $|Z| \geq \varepsilon$  ist doch gleichwertig zu  $Z^2 \geq \varepsilon^2$ . Die Markovungleichung impliziert daher (diesmal mit  $g(x) := x^2$ )

$$\mathbb{P}(\{|Z| \geq \varepsilon\}) = \mathbb{P}(\{Z^2 \geq \varepsilon^2\}) \leq \frac{\mathbb{E}(Z^2)}{\varepsilon^2}.$$

Das interessiert uns hier für die  $Z = Z_n$ . Das sieht vielversprechend aus, denn

$$\begin{aligned} \mathbb{E}(Z_n^2) &= \frac{1}{n^2} \sum_{i,j=1}^n \mathbb{E}(X_i X_j) \\ &= \frac{1}{n^2} \left( \sum_{i,j, i \neq j} \mathbb{E}(X_i) \mathbb{E}(X_j) + \sum_{i=1}^n \mathbb{E}(X_i^2) \right) \\ &= \frac{1}{n^2} \sum_{i=1}^n V(X_i) \\ &= \frac{1}{n^2} \cdot n \cdot V(X) \\ &= \frac{V(X)}{n}. \end{aligned}$$

Man beachte: Bei diesen Umformungen haben wir es eigentlich mit den  $n^2$  Summanden  $\mathbb{E}(X_i X_j)$  zu tun. Wegen der Unabhängigkeit der  $X_i$  verschwinden aber die allermeisten, denn für  $i \neq j$  kann  $\mathbb{E}(X_i X_j)$  zu  $\mathbb{E}(X_i) \mathbb{E}(X_j)$  umgeformt werden, und beide Faktoren sind nach Voraussetzung Null.

Auf diese Weise gelangt man zu der Abschätzung

$$\mathbb{P}(\{|Z_n| \geq \varepsilon\}) = \mathbb{P}(\{Z_n^2 \geq \varepsilon^2\}) \leq V(X)/n.$$

Schade, das hat noch nicht gereicht: Auch wenn wir die Existenz der Varianz von  $X$  voraussetzen, können wir mit dieser Abschätzung nicht die Endlichkeit von  $\sum_n \mathbb{P}(\{|Y_n| \geq \varepsilon\})$  garantieren, denn die harmonische Reihe  $\sum_n 1/n$  divergiert.

*Die zweite Idee.* Eigentlich könnte man es nun mit der dritten Potenz in der Markov-Ungleichung versuchen:  $|Z| \geq \varepsilon$  ist gleichwertig zu  $|Z|^3 \geq \varepsilon^3$ . Das soll hier für  $Z := Z_n := (X_1 + \dots + X_n)/n$  ausgewertet werden, doch das führt nicht weiter, weil man über die Erwartungswerte der  $|X_1 + \dots + X_n|^3$  im Allgemeinen nichts weiß.

Deswegen sind die vierten Potenzen der nächste Kandidat, mit dem wir es versuchen. Statt  $\mathbb{P}(\{|Z_n| \geq \varepsilon\})$  (diese Zahlen interessieren hier) kann man auch  $\mathbb{P}(\{Z_n^4 \geq \varepsilon^4\})$  berechnen, und diese Wahrscheinlichkeiten sind durch  $\mathbb{E}(Y_n^4)/\varepsilon^4$  abschätzbar.

Für die Auswertung von  $\mathbb{E}(Z_n^4)$  muss man sich um den Erwartungswert von  $(X_1 + \dots + X_n)^4$  kümmern. Es gibt  $n^4$  Summanden des Typs  $\mathbb{E}(X_i X_j X_k X_l)$ . Folgende Fälle sind dabei möglich:

- Es ist  $i = j = k = l$ , davon gibt es  $n$  Summanden. In diesem Fall ist  $\mathbb{E}(X_i X_j X_k X_l) = \mathbb{E}(X^4)$ , denn die  $X_i$  sind wie  $X$  verteilt. Um weiterzukommen, müssen wir also sicher voraussetzen, dass  $\gamma := \mathbb{E}(X^4)$  eine endliche Zahl ist.
- Unter den Indizes  $i, j, k, l$  gibt es zwei Pärchen, wir sind aber nicht im vorigen Fall. Es ist zum Beispiel  $i = j \neq k = l$ . Dann ist der Erwartungswert von  $X_i X_j X_k X_l$  gleich  $\mathbb{E}(X_i^2) \mathbb{E}(X_k^2)$ , denn  $X_i^2, X_k^2$  sind unabhängig. Folglich ist  $\mathbb{E}(X_i X_j X_k X_l) = V^2$ , wobei  $V$  die Varianz von  $X$  bezeichnet. Unter den  $n^4$  möglichen  $i, j, k, l$  treten  $3n(n-1)$  solche Situationen auf (nämlich  $i = j \neq k = l, i = k \neq j = l, i = l \neq j = k$  jeweils  $n(n-1)$  Mal).
- In allen anderen Fällen ist  $X_i X_j X_k X_l$  von der Form  $X_i$  mal eine von  $X_i$  unabhängige Zufallsvariable  $Z$ , und deswegen ist der Erwartungswert gleich  $\mathbb{E}(X_i) \mathbb{E}(Z) = 0$ .

Zusammen: Von den  $n^4$  Zahlen  $\mathbb{E}(X_i X_j X_k X_l)$  sind die meisten Null. Man muss nur  $n$  Summanden der Größe  $\gamma$  und  $3n(n-1)$  Summanden der Größe  $V^2$  berück-

sichtigen. Es folgt

$$\begin{aligned}
 \mathbb{P}(\{|Z_n| \geq \varepsilon\}) &= \mathbb{P}(\{Z_n^4 \geq \varepsilon^4\}) \\
 &\leq \frac{\mathbb{E}(Z_n^4)}{\varepsilon^4} \\
 &= \frac{\mathbb{E}((X_1 + \dots + X_n)^4)}{n^4 \varepsilon^4} \\
 &= \frac{n\gamma + 3n(n-1)V^2}{n^4 \varepsilon^4} \\
 &\leq \frac{c}{n^2}.
 \end{aligned}$$

Dabei haben wir  $c$  durch  $(\gamma + 3V^2)/\varepsilon^4$  definiert. Und das sieht sehr gut aus, denn  $\sum_n 1/n^2 < \infty$ .

Wir fassen zusammen: Wenn man die fast sichere Konvergenz der Mittelwerte  $(X_1 + \dots + X_n)/n$  mit Hilfe des ersten Lemmas von Borel-Cantelli, einer Umformung der Wahrscheinlichkeiten  $\mathbb{P}(\{|Y_n| \geq \varepsilon\})$  zu  $\mathbb{P}(\{|Y|^k \geq \varepsilon^k\})$ , Anwendung der Markovungleichung mit  $g(x) := x^k$  und Auswertung der  $\mathbb{E}(X_{i_1}X_{i_2}\dots X_{i_k})$  beweisen möchte, so stellt sich heraus, dass  $k = 4$  eine gute Wahl ist. Wir formulieren das als

**Satz 8.3.1.** (*Starkes Gesetz der großen Zahlen, Variante 1*) *Es sei  $X$  eine Zufallsvariable, und die  $X_1, X_2, \dots$  seien unabhängige Kopien wie in Definition 8.2.1. Wenn man voraussetzt, dass  $\mathbb{E}(X)$ , die Varianz von  $X$  und  $\mathbb{E}(X^4)$  existieren, so gilt: Die Mittelwerte  $(X_1 + \dots + X_n)/n$  konvergieren punktweise fast sicher gegen die konstante Funktion  $\mathbb{E}(X)$ .*

Eine allgemeinere Version des starken Gesetzes

Tatsächlich kann man das gleiche Ergebnis auch zeigen, wenn man für  $X$  nur die Existenz der Varianz fordert. Auch die Voraussetzung bezüglich der Unabhängigkeit kann abgeschwächt werden: Im vorstehenden Beweis mussten je vier der  $X_i$  unabhängig sein, im folgenden wird es reichen, paarweise Unabhängigkeit vorauszusetzen. Der Beweis ist allerdings etwas komplizierter.

**Satz 8.3.2.** (*Starkes Gesetz der großen Zahlen*): *Es sei  $X$  eine reellwertige Zufallsvariable mit existierendem Erwartungswert  $\mathbb{E}(X)$  und existierender Streuung  $\sigma := \sigma(X)$ . Weiter seien  $X_1, X_2, \dots$  paarweise unabhängige Kopien von  $X$ , die auf einem geeigneten Wahrscheinlichkeitsraum definiert sind. Dann gilt punktweise fast sicher*

$$\frac{X_1 + \dots + X_n}{n} \rightarrow \mathbb{E}(X).$$

**Beweis:** Wir nehmen zunächst an, dass  $X \geq 0$  gilt und setzen  $S_n := X_1 + \dots + X_n$  für  $n \in \mathbb{N}$ . Es ist zu zeigen, dass  $S_n/n$  punktweise fast sicher gegen  $\mathbb{E}(X)$  konvergiert. Der Beweis besteht aus zwei Schritten:

- Wir zeigen, dass eine geeignete Teilfolge der Folge  $(S_n/n)$  fast sicher gegen  $\mathbb{E}(X)$  geht.
- Danach wird bewiesen, dass sogar für die ganze Folge fast sichere Konvergenz vorliegt.

*1. Schritt:* Sei  $\varepsilon > 0$  beliebig. Wir definieren (für alle  $n \in \mathbb{N}$ ) Zahlen  $k_n$  als größte natürliche Zahl  $\leq (1+\varepsilon)^n$ . Die folgenden einfach nachzuweisenden Eigenschaften der  $k_n$  werden gleich wichtig werden:

- $k_n \leq k_{n+1}$ .
- $k_n \geq (1 + \varepsilon)^n/2$  für genügend große  $n$ . (Das gilt sicher dann, wenn  $n$  so groß ist, dass  $(1 + \varepsilon)^n/2 \leq (1 + \varepsilon)^n - 1$ .)
- $k_{n+1} \leq (1 + 2\varepsilon)k_n$  für genügend große  $n$ . (Es ist  $k_{n+1} = (1 + \varepsilon)^{n+1} + x$  und  $k_n = (1 + \varepsilon)^n + y$  mit geeigneten  $x, y \in [0, 1[$ . Die Behauptung gilt also, unabhängig von  $x, y$ , wenn  $n$  die Bedingung  $\varepsilon(1 + \varepsilon)^n \geq 1$  erfüllt.)

Sei  $A_n$  die Menge

$$\left\{ \left| \frac{S_{k_n}}{k_n} - \mathbb{E}(X) \right| \geq \frac{1}{(1 + \varepsilon)^{n/4}} \right\}.$$

Da  $\mathbb{E}(X)$  der Erwartungswert von  $S_{k_n}/k_n$  ist, gilt aufgrund der Tschebyscheff-Ungleichung

$$\begin{aligned} \mathbb{P}(A_n) &\leq (1 + \varepsilon)^{n/2} V(S_{k_n}/k_n) \\ &= (1 + \varepsilon)^{n/2} V(X)/k_n \\ &\leq 2(1 + \varepsilon)^{n/2} (1 + \varepsilon)^{-n} V(X) \\ &= 2(1 + \varepsilon)^{-n/2} V(X); \end{aligned}$$

hier haben wir ausgenutzt, dass die  $X_n$  (paarweise) unabhängig sind und das deswegen die Varianz der Summe gleich der Summe der Varianzen ist. Diese Abschätzung impliziert, dass

$$\sum_n \mathbb{P}(A_n) < +\infty,$$

denn aufgrund der vorstehenden Rechnungen kann diese Reihe durch ein Vielfaches der geometrischen Reihe  $1 + q + q^2 + \dots$  mit  $q := (1 + \varepsilon)^{-0.5} (< 1)$  abgeschätzt werden.

Mit Satz 8.1.3 folgt, dass fast sicher  $S_{k_n}(\omega)/k_n \rightarrow \mathbb{E}(X)$  gilt.

*2. Schritt:* In diesem Schritt wird wichtig, dass alle  $X_n$  nichtnegativ sind. Sei ein  $l$  aus  $\{k_n, k_n + 1, \dots, k_{n+1}\}$  vorgegeben. Dann gilt (für genügend große  $n$ )

die folgende Ungleichungskette:

$$\begin{aligned} \frac{1}{(1+2\varepsilon)} \frac{1}{k_n} S_{k_n} &\leq \frac{1}{k_{n+1}} S_{k_n} \\ &\leq \frac{S_l}{l} \\ &\leq \frac{S_{k_{n+1}}}{k_n} \\ &\leq (1+2\varepsilon) \frac{S_{k_{n+1}}}{k_{n+1}}. \end{aligned}$$

Da sich  $1+2\varepsilon$  beliebig wenig von  $1/(1+2\varepsilon)$  unterscheidet und da die Folge  $S_{k_n}/k_n$  fast sicher gegen  $\mathbb{E}(X)$  konvergiert, ist  $S_l/l$  quasi zwischen zwei konvergenten Folgen „eingesperrt“, wobei beide gegen  $\mathbb{E}(X)$  gehen<sup>14)</sup>. Und deswegen muss auch  $\lim_l S_l/l = \mathbb{E}(X)$  fast sicher gelten.

Es bleibt noch zu zeigen, dass der Satz auch dann richtig ist, wenn die Bedingung  $X \geq 0$  nicht erfüllt ist. Ist  $X$  beliebig vorgegeben, so zerlegen wir jedes  $X_n$  in  $X_n^+ - X_n^-$  mit nichtnegativen  $X_n^+, X_n^-$ . Dazu setzen wir  $X_n^+ := \max\{X_n, 0\}$  sowie  $X_n^- := \max\{-X_n, 0\}$ . Dann sind die  $X_n^+$  (paarweise) unabhängig und so verteilt wie  $X^+$ , und wegen  $|X^+| \leq |X|$  existieren Erwartungswert und Varianz von  $X_n^+$ . Aufgrund des ersten Beweisteils gilt dann

$$\frac{X_1^+ + \cdots + X_n^+}{n} \rightarrow \mathbb{E}(X^+)$$

fast sicher, und analog folgt die fast sichere Konvergenz von

$$\frac{X_1^- + \cdots + X_n^-}{n} \rightarrow \mathbb{E}(X^-).$$

Zieht man noch beide konvergenten Folgen voneinander ab, so ergibt sich mit

$$\mathbb{E}(X) = \mathbb{E}(X^+ - X^-) = \mathbb{E}(X^+) - \mathbb{E}(X^-)$$

die Behauptung. □

---

<sup>14)</sup>Etwas genauer: Sei  $\varepsilon_0 > 0$ . Bestimme  $n_0$ , so dass

$$\mathbb{E}(X) - \frac{\varepsilon_0}{2} \leq S_{k_n}/k_n \leq \mathbb{E}(X) + \frac{\varepsilon_0}{2}$$

für  $n \geq n_0$ . Wähle  $\varepsilon > 0$  so klein, dass

$$\mathbb{E}(X) - \varepsilon_0 \leq \frac{1}{1+2\varepsilon} \left( \mathbb{E}(X) - \frac{\varepsilon_0}{2} \right), \quad (1+2\varepsilon) \left( \mathbb{E}(X) + \frac{\varepsilon_0}{2} \right) \leq \mathbb{E}(X) + \varepsilon_0.$$

Dann ist  $\mathbb{E}(X) - \varepsilon_0 \leq S_l/l \leq \mathbb{E}(X) + \varepsilon_0$  für genügend große  $l$ .

### Hat der Würfel ein schlechtes Gewissen?

Im Zusammenhang mit den Gesetzen der großen Zahlen ergibt sich scheinbar ein Problem, denn es sollten doch die beiden folgenden Tatsachen gelten:

- Der Zufall hat kein Gedächtnis: Bei unabhängigen Zufallsexperimenten ist das jeweils nächste Ergebnis völlig unbeeinflusst von den Ergebnissen der Vergangenheit.
- Die Mittelwerte von unabhängig gezogenen Ausgaben einer Zufallsvariablen konvergieren gegen den Erwartungswert.

Das bedeutet zum Beispiel beim Würfeln, dass das nächste Ergebnis unbeeinflusst von den vorhergehenden gefunden wird, dass aber andererseits bei vielen Versuchen zu erwarten ist, dass – zum Beispiel – ein Sechstel der Würfelwürfe zu einer Sechs führt.

Das kann aber doch wohl nicht gleichzeitig wahr sein. Wenn mit einem Würfel lange keine Sechs gewürfelt wurde, muss er sich doch jetzt wohl ein bisschen anstrengen, um auf die richtige Quote zu kommen, dass also in einer langen Versuchsreihe ein Sechstel der Würfe die Sechs zeigt.

Anders ausgedrückt: Wenn lange keine Sechs kam, sollte die Wahrscheinlichkeit für eine Sechs steigen. Viele glauben, dass das tatsächlich so ist. Es stimmt aber nicht, doch wie kann man diesen scheinbaren Widerspruch auflösen?



Bild 8.3.1: Immer nur Dreien? Kann das sein?

Um zu begründen, dass alles mit rechten Dingen zugeht, machen wir ein Gedankenexperiment:

In einem großen Saal befinden sich 1024 Personen. Jede wirft eine faire Münze. Wir nehmen der Einfachheit halber an, dass genau die Hälfte, also 512 Personen, „Kopf“ würfelt. Sie sollen noch einmal würfeln: Davon die Hälfte, also (etwa) 256 Personen würfeln noch einmal „Kopf“. So wird das fortgesetzt, und es ist dann alles andere als überraschend, dass es nach 10 Durchgängen eine Person geben wird, bei der die Münze *immer „Kopf“* zeigte.

Anders ausgedrückt heißt das, dass überraschende Ergebnisse durchaus auftreten können, allerdings – und das ist hier wesentlich – nur mit einer minimalen Wahrscheinlichkeit.

Wenn also der Würfel lange keine Sechs zeigt, so widerspricht das keineswegs der Theorie. Im Gegenteil, es ist zu erwarten, dass das bei einer kleinen Anzahl von Versuchsreihen wirklich auftritt.

Zum Abschluss soll noch einmal die *grundätzliche Bedeutung des starken Gesetzes* betont werden. Es garantiert, dass man sich darauf verlassen kann, dass bei einer Folge von Abfragen einer Zufallsvariable die Mittelwerte gegen den Erwartungswert konvergieren. Dass diese Mittelwerte im Allgemeinen nicht konvergieren müssen, war aufgrund der Simulationsbeispiele zur Cauchy-Verteilung zu erwarten (vgl. Seite 106).

Bemerkenswerter Weise ist das sogar eine Charakterisierung. Das besagt das folgende Ergebnis von *Kolmogoroff*:

**Satz 8.3.3.** *Es sei  $X$  eine Zufallsvariable, und  $X_1, X_2, \dots$  seien unabhängige Kopien von  $X$ . Dann sind äquivalent:*

- (i) *Der Erwartungswert von  $X$  existiert.*
- (ii) *Die Folge  $((X_1 + \dots + X_n)/n)$  ist mit einer positiven Wahrscheinlichkeit konvergent.*
- (iii) *Das Ereignis  $\{\omega \mid X_n(\omega)/n \rightarrow 0\}$  hat eine positive Wahrscheinlichkeit.*

*Wenn eine dieser Bedingungen erfüllt ist, so sind die Wahrscheinlichkeiten in (ii) und (iii) gleich Eins, und in (ii) ist der Limes die konstante Funktion  $\mathbb{E}(X)$ .*

**Beweis:** Dass „(ii)  $\rightarrow$  (i)“ gilt, ist gerade das Gesetz der großen Zahl in der allgemeinen Version von Entemadi, den wir schon auf Seite 235 erwähnt, in diesem Buch allerdings nicht bewiesen haben.

(ii)  $\Rightarrow$  (iii): Wir setzen  $S_n := X_1 + \dots + X_n$ . Ist  $((S_n(\omega)/n))$  konvergent, so muss  $X_n(\omega)/n \rightarrow 0$  gelten, denn

$$\frac{X_n(\omega)}{n} = \frac{S_n(\omega) - S_{n-1}(\omega)}{n} = \frac{S_n(\omega)}{n} - \frac{n-1}{n} \frac{S_{n-1}(\omega)}{n-1},$$

und der Limes von  $(S_n(\omega)/n)S_n$  ist – falls existent – gleich dem Limes von  $(S_{n-1}(\omega)/(n-1))$ .

(iii)  $\Rightarrow$  (i): Sei  $A_n$  das Ereignis  $\{|X_n| \geq n\}$ . Die  $A_n$  sind unabhängig, und  $\limsup A_n$  hat nach Voraussetzung eine Wahrscheinlichkeit, die echt kleiner als Eins ist: Denn wenn  $X_n(\omega)/n \rightarrow 0$  gilt, liegt  $\omega$  nur in endlich vielen  $A_n$ . Damit ist nach dem zweiten Lemma von Borel-Cantelli ausgeschlossen, dass  $\sum_n \mathbb{P}(A_n) = +\infty$  gilt, denn in diesem Fall wäre  $\mathbb{P}(\limsup A_n) = 1$ . Es folgt  $\sum_n \mathbb{P}(A_n) < +\infty$ , und das impliziert die Existenz von  $\mathbb{E}(X)$  (vgl. Übungsaufgabe Ü3.3.5).

Der Zusatz wurde schon mitbewiesen. □

## 8.4 Der zentrale Grenzwertsatz

In diesem Abschnitt geht es um eines der überraschendsten Phänomene der Wahrscheinlichkeitsrechnung: Die Normalverteilung ist allgegenwärtig. Indizien dafür haben wir schon an mehreren Stellen angetroffen, man blättere etwa zu den Bildern auf den Seiten 151 und 165 zurück. Doch erst jetzt haben wir die Möglichkeit, alles präzise zu formulieren.

Wir beginnen mit einer Erinnerung: Unter der  $N(a, \sigma^2)$ -Verteilung verstehen wir das auf  $\mathbb{R}$  durch die Dichtefunktion  $e^{-(x-a)^2/2\sigma^2}/\sqrt{2\pi}\sigma$  definierte Wahrscheinlichkeitsmaß.

### Vorbereitungen

Wir beweisen im nächsten Satz, dass man Erwartungswert und Varianz solcher Verteilungen leicht ermitteln kann, und danach zeigen wir, dass die Familie der Normalverteilungen unter Multiplikation, Translation und unabhängigen Summen abgeschlossen ist

**Satz 8.4.1.** *Es sei  $a \in \mathbb{R}$  und  $\sigma > 0$ . Der Erwartungswert einer  $N(a, \sigma^2)$ -verteilten Zufallsvariablen ist  $a$ , und die Varianz ist  $\sigma^2$ .*

**Beweis:** Hat  $\mathbb{P}_X$  eine Dichtefunktion  $f$ , so ist  $\mathbb{E}(X) = \int_{-\infty}^{+\infty} xf(x) dx$  und  $V(X) = \int_{-\infty}^{+\infty} (x - \mathbb{E}(X))^2 f(x) dx$  (vgl. Satz 3.3.5). Im vorliegenden Fall ist zur Berechnung von  $\mathbb{E}(x)$  das Integral

$$\frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} xe^{-\frac{(x-a)^2}{2\sigma^2}} dx$$

auszuwerten. Dieses Integral transformieren wir mit der Substitution  $u := x - a$  in

$$\frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} (u + a)e^{-u^2/2\sigma^2} du$$

um. Das führt auf die Summe der Integrale

$$\frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} ue^{-u^2/2\sigma^2} du + a \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} e^{-u^2/2\sigma^2} du.$$

Das erste ist Null, denn es wird die schiefsymmetrische Funktion  $ue^{-u^2/2\sigma^2}$  über ein zu 0 symmetrisches Intervall integriert. Das zweite ist das  $a$ -fache des Integrals über eine Dichtefunktion, das Ergebnis ist also  $a$ . Und damit ist  $\mathbb{E}(X) = a$  bewiesen.

Nun zur Varianz, dazu müssen wir

$$\frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} (x - a)^2 e^{-\frac{(x-a)^2}{2\sigma^2}} dx$$

bestimmen. Diesmal substituieren wir  $u := (x - a)/\sigma$ , so kommen wir zu

$$\sigma^3 \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} u^2 e^{-u^2/2} du.$$

Zur Auswertung dieses Integrals hilft ein Trick weiter: Wir schreiben  $u^2 e^{-u^2/2}$  als  $u(ue^{-u^2/2})$  und wenden auf diese Zerlegung partielle Integration an:

$$\begin{aligned} \int_{-\infty}^{+\infty} u^2 e^{-u^2} du &= \lim_{a \rightarrow \infty} \int_{-a}^a u^2 e^{-u^2/2} du \\ &= \lim_{a \rightarrow \infty} \int_{-a}^a u(ue^{-u^2/2}), du \\ &= \lim_{a \rightarrow \infty} \left( -ue^{-u^2/2} \Big|_{-a}^a + \int_{-a}^a e^{-u^2} du \right) \\ &= \int_{-\infty}^{+\infty} e^{-u^2} du \\ &= \sqrt{2\pi}. \end{aligned}$$

Da haben wir ausgenutzt, dass  $\pm ue^{-u^2/2}$  mit  $u \rightarrow \infty$  gegen Null geht und dass  $e^{-u^2/2}/\sqrt{2\pi}$  eine Dichtefunktion ist.

Insgesamt ergibt sich so der Wert  $\sigma^2$  für die Varianz von  $X$ , und damit ist der Satz vollständig bewiesen.  $\square$

**Satz 8.4.2.** Es seien  $X, X_1, X_2 : \Omega \rightarrow \mathbb{R}$  normalverteilte Zufallsvariable:  $X$  bzw.  $X_1$  bzw.  $X_2$  sei  $N(a, \sigma^2)$ - bzw.  $N(a_1, \sigma_1^2)$ - bzw.  $N(a_2, \sigma_2^2)$ -verteilt. Außerdem seien  $X_1, X_2$  unabhängig.

- (i) Für  $\alpha \in \mathbb{R}$  und  $\beta \neq 0$  gilt:  $\beta X + \alpha$  ist  $N(a\beta + \alpha, \beta^2 \sigma^2)$ -verteilt.
- (ii)  $X_1 + X_2$  ist  $N(a_1 + a_2, \sigma_1^2 + \sigma_2^2)$ -verteilt.

**Beweis:** (i) Man könnte hier Satz 3.2.3 anwenden, wir beweisen die Aussage aber direkt. Sei zunächst  $\beta > 0$ . Mit  $Y := \beta X + \alpha$  ist dann für beliebige Intervalle  $[c, d]$

$$\begin{aligned} \mathbb{P}(\{Y \in [c, d]\}) &= \mathbb{P}(\{X \in [(c - \alpha)/\beta, (d - \alpha)/\beta]\}) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{(c-\alpha)/\beta}^{(d-\alpha)/\beta} e^{-(x-a)^2/2\sigma^2} dx \\ &= \frac{1}{\sqrt{2\pi}\sigma\beta} \int_c^d e^{-\left(u-(\beta a + \alpha)\right)^2/2\sigma^2\beta^2}. \end{aligned}$$

Dabei haben wir im letzten Schritt  $u := \beta x + \alpha$  gesetzt und die Substitutionsregel angewendet. Folglich ist  $\mathbb{P}(\{Y \in [c, d]\})$  das Integral über  $[c, d]$  und die zu  $N(a\beta + \alpha, \beta^2 \sigma^2)$  gehörige Dichtefunktion, d.h.  $Y$  ist  $N(a\beta + \alpha, \beta^2 \sigma^2)$ -verteilt.

Der Fall  $\beta < 0$  wird analog behandelt. Es ist nur zu beachten, dass diesmal

$$\mathbb{P}(\{Y \in [c, d]\}) = \mathbb{P}(\{X \in [(d - \alpha)/\beta, (c - \alpha)/\beta]\})$$

gilt. Aber da bei der Substitution  $u = \beta x + \alpha$  die Ableitung  $du/dx$  negativ ist, erhalten wir das gleiche Endergebnis.

(ii) Wenn wir wüssten, dass  $X_1 + X_2$  normalverteilt ist, wären wir schon fertig: Der Erwartungswert muss  $\mathbb{E}(X_1) + \mathbb{E}(X_2) = a_1 + a_2$  sein, und die Varianzen müssen sich wegen Satz 4.6.2 addieren. Leider wissen wir das noch nicht, und deswegen müssen wir etwas aufwändiger argumentieren.

Dichtefunktionen für Summen unabhängiger Zufallsvariablen ergeben sich nach Satz 4.6.5 durch Faltung. Mit den Bezeichnungen der Seite 51 ist also zu zeigen, dass

$$f_{a_1, \sigma_1^2} * f_{a_2, \sigma_2^2} = f_{a_1+a_2, \sigma_1^2+\sigma_2^2}$$

gilt. Wir berechnen dazu  $f_{a_1, \sigma_1^2} * f_{a_2, \sigma_2^2}$  bei irgendeinem  $y \in \mathbb{R}$ :

$$\begin{aligned} f_{a_1, \sigma_1^2} * f_{a_2, \sigma_2^2}(y) &= \int_{-\infty}^{+\infty} f_{a_1, \sigma_1^2}(y-x) f_{a_2, \sigma_2^2}(x) dx \\ &= \frac{1}{2\pi\sigma_1\sigma_2} \int_{-\infty}^{+\infty} e^{-(y-x-a_1)^2/2\sigma_1^2} e^{-(x-a_2)^2/2\sigma_2^2} dx. \end{aligned}$$

Und es ist zu zeigen, dass diese Zahl gleich

$$f_{a_1+a_2, \sigma_1^2+\sigma_2^2}(y) = \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}} e^{-\frac{(y-(a_1+a_2))^2}{2(\sigma_1^2+\sigma_2^2)}}$$

ist. Das beweisen wir durch die Kombination einiger elementarer Überlegungen:

- Jede quadratische Funktion  $\alpha x^2 + \beta x + \gamma$  lässt sich als  $\alpha(x - \beta')^2 + \gamma'$  schreiben. Diese elementare Tatsache spielt auch bei der  $p-q$ -Formel für quadratische Gleichungen eine Rolle.
- Insbesondere gilt das für den im Integral auftretenden Exponenten

$$-(y - x - a_1)^2/2\sigma_1^2 - (x - a_2)^2/2\sigma_2^2.$$

Wenn man da die entsprechende Umformung vornimmt, so ergibt sich, dass

$$-(y - x - a_1)^2/2\sigma_1^2 - (x - a_2)^2/2\sigma_2^2 = -(x - \alpha)^2/2\beta - \gamma$$

gilt, wobei

$$\alpha = \frac{(y - a_1)\sigma_2^2 + a_2\sigma_1^2}{\sigma_1^2 + \sigma_2^2}, \quad \beta = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}, \quad \gamma = \frac{(y - a_1 - a_2)^2}{2(\sigma_1^2 + \sigma_2^2)}.$$

- Ein Integral der Form  $\int_{-\infty}^{+\infty} e^{-(x-\alpha)^2/2\beta} dx$  hat den Wert  $\sqrt{2\pi\beta}$ . Denn  $e^{-(x-\alpha)^2/2\beta}$  ist ein Vielfaches von  $f_{\alpha, \beta}$ : Das ist eine Dichtefunktion, und folglich ist das Integral über  $\mathbb{R}$  gleich Eins.

Nach diesen Vorbereitungen setzen wir die oben begonnene Berechnung von  $f_{a_1, \sigma_1^2} * f_{a_2, \sigma_2^2}(y)$  fort:

$$\begin{aligned}\dots &= \frac{1}{2\pi\sigma_1\sigma_2} \int_{-\infty}^{+\infty} e^{-(x-\alpha)^2/2\beta} e^{-\gamma} dx \\ &= \frac{e^{-\gamma}}{2\pi\sigma_1\sigma_2} \int_{-\infty}^{+\infty} e^{-(x-\alpha)^2/2\beta} dx \\ &= \frac{e^{-\gamma}}{2\pi\sigma_1\sigma_2} \sqrt{2\pi\beta} \\ &= \frac{\sqrt{2\pi\sigma_1\sigma_2}/\sqrt{\sigma_1^2 + \sigma_2^2}}{2\pi\sigma_1\sigma_2} e^{-(y-a_1-a_2)^2/2(\sigma_1^2 + \sigma_2^2)} \\ &= f_{a_1+a_2, \sigma_1^2 + \sigma_2^2}(y).\end{aligned}$$

Damit ist der Satz bewiesen.  $\square$

**Eine wichtige Konsequenz:** Teil (i) des vorstehenden Satzes impliziert insbesondere, dass man jede Normalverteilung mit Hilfe der Standardnormalverteilung simulieren kann: Ist  $X \sim N(0, 1)$ -verteilt, so ist  $(\sigma X + a) \sim N(a, \sigma^2)$ -verteilt.

Wenn man also  $N(a, \sigma^2)$ -verteilte Zufallszahlen benötigt, so kann man  $N(0, 1)$ -verteilte Zahlen  $z$  erzeugen und dann zu  $\sigma z + a$  übergehen.

Die Simulation von standard-normalverteilten Zahlen wurde auf Seite 61 beschrieben.

### Der zentrale Grenzwertsatz

Wir kommen nun zu einem der wichtigsten Sätze der Wahrscheinlichkeitstheorie. Er besagt, dass die Normalverteilung allgegenwärtig ist:

**Satz 8.4.3.** (Zentraler Grenzwertsatz) Es seien  $X_1, X_2, \dots$  wie in Definition 8.2.1 unabhängige Kopien einer Zufallsvariablen  $X$ , für die Erwartungswert  $\mathbb{E}(X)$  und Varianz  $\sigma^2$  existieren; die Varianz soll nicht Null sein. Dann konvergieren die Zufallsvariablen  $(X_1 + \dots + X_n - n\mathbb{E}(X))/\sqrt{n}\sigma$  in Verteilung gegen die Standardnormalverteilung  $N(0, 1)$ . Insbesondere gilt also wegen Satz 7.3.2(iii):

$$\mathbb{P}\left(\left\{ \frac{X_1 + \dots + X_n - n\mathbb{E}(X)}{\sqrt{n}\sigma} \in [a, b] \right\}\right) \rightarrow \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2} dx$$

für  $n \rightarrow \infty$  und beliebige Intervalle  $[a, b]$ .



Bild 8.4.1: Die Gaußsche Glockenkurve: Bis zur Einführung des Euro war sie auf dem 10-DM-Schein abgebildet.

**Beweis:** Für diesen Satz gibt es mehrere Beweismöglichkeiten. Hier soll ein „elementarer“ Beweis geführt werden. Dabei bedeutet „elementar“, dass die bisher in diesem Buch behandelten Methoden und Ergebnisse für das Verständnis ausreichen. Ein wesentlich eleganterer Weg, der allerdings tieferliegende Techniken der Wahrscheinlichkeitstheorie ausnutzt, wird in Abschnitt 8.6 beschrieben werden.

Unser Beweis besteht aus zwei Teilen:

- *Teil 1:* Es gibt Beispiele, in denen der Satz richtig ist.
- *Teil 2:* Wenn der Satz wenigstens einmal richtig ist, so ist er immer gültig.

*Zu Teil 1:* Das ist leicht, wir haben sogar zwei Möglichkeiten, das einzusehen. Man könnte auf den Satz von de Moivre-Laplace verweisen<sup>15)</sup>, der besagt, dass der zentrale Grenzwertsatz richtig ist, wenn  $X$  binomialverteilt ist. Einfacher ist es aber, die zu Beginn dieses Abschnitts bewiesenen Ergebnisse über Normalverteilungen auszunutzen. Aus Satz 8.4.2 folgt doch, dass  $(Y_1 + \dots + Y_n)/\sqrt{n}$   $N(0, 1)$ -verteilt ist, wenn die  $Y_n$  unabhängige Kopien einer  $N(0, 1)$ -verteilten Zufallsvariablen sind.

Die Verteilungen sind also konstant, es liegt daher sicher Konvergenz in Verteilung vor.

*Zu Teil 2:* Gegeben seien die  $X, X_1, X_2, \dots$  wie im Satz. Wir dürfen annehmen, dass  $\mathbb{E}(X) = 0$  und  $V(X) = 1$  gilt, denn sobald der Satz dafür bewiesen ist, ist er durch Übergang von den  $X_n$  zu den  $(X_n - \mathbb{E}(X))/\sigma$  auch für die allgemeinere Situation gezeigt.

Mit  $Y, Y_1, Y_2, \dots$  bezeichnen wir eine Situation, für die der Satz schon gezeigt ist: So etwas gibt es nach Teil 1 des Beweises. Wie bei  $X$  können wir auch bei  $Y$  annehmen, dass der Erwartungswert Null und die Varianz Eins ist. Und wegen der allgemeineren Version des „Klonsatzes“ dürfen wir davon ausgehen, dass alle Zufallsvariablen  $X_1, X_2, \dots, Y_1, Y_2, \dots$  auf dem gleichen Wahrscheinlichkeitsraum definiert und unabhängig sind.

<sup>15)</sup>Vgl. Abschnitt 5.4.

Wir müssen beweisen, dass  $\mathbb{E}\left(h((X_1 + \dots + X_n)/\sqrt{n})\right)$  für jede stetige beschränkte Funktion  $h : \mathbb{R} \rightarrow \mathbb{R}$  gegen  $\mathbb{E}(h \circ Z)$  konvergiert, wobei  $Z$  eine  $N(0, 1)$ -verteilte Zufallsvariable ist. Wir wissen dabei schon:

- Das stimmt für die  $\mathbb{E}\left(h((Y_1 + \dots + Y_n)/\sqrt{n})\right)$ , denn für die  $Y_n$  ist der Satz ja nach Voraussetzung richtig.
- Wegen Lemma 7.3.3 müssen wir die Behauptung nur für Funktionen  $h$  zeigen, die beliebig oft differenzierbar sind und außerhalb eines beschränkten Intervalls verschwinden.

Wenn man diese Vorüberlegungen kombiniert, so hat sich der Beweis des Satzes auf das folgende Problem reduziert:

Gegeben seien ein Wahrscheinlichkeitsraum  $(\Omega, \mathcal{E}, \mathbb{P})$  und unabhängige Zufallsvariable

$$X_1, Y_1, X_2, Y_2, \dots : \Omega \rightarrow \mathbb{R}.$$

Die  $X_n$  (bzw. die  $Y_n$ ) seien verteilt wie eine Zufallsvariable  $X$  (bzw.  $Y$ ), für die  $\mathbb{E}(X) = 0$  und  $V(X) = 1$  (bzw.  $\mathbb{E}(Y) = 0$  und  $V(Y) = 1$ ) gilt.

Man zeige: Ist  $h : \mathbb{R} \rightarrow \mathbb{R}$  beliebig oft differenzierbar und verschwindet  $h$  außerhalb eines beschränkten Intervalls, so gilt

$$\mathbb{E}\left(h\left(\frac{X_1 + \dots + X_n}{\sqrt{n}}\right)\right) - \mathbb{E}\left(h\left(\frac{Y_1 + \dots + Y_n}{\sqrt{n}}\right)\right) \rightarrow 0.$$

Wir fixieren so eine Funktion  $h$ . Mit  $R$  bezeichnen wir eine Zahl, dass  $h(x) = 0$  für  $|x| > R$  gilt. Zunächst beweisen wir zwei vorbereitende Ergebnisse.

*Behauptung 1:* Es sei  $(\Omega, \mathcal{E}, \mathbb{P})$  ein Wahrscheinlichkeitsraum, und  $U, Z, W$  seien auf  $\Omega$  definierte reellwertige Zufallsvariable. Für eine geeignete Zahl  $c > 0$  soll gelten:

$$|\mathbb{E}(h \circ (U + r)) - \mathbb{E}(h \circ (Z + r))| \leq c$$

für alle  $r \in \mathbb{R}$ .

Sind dann  $U, V$  und  $U, W$  unabhängig, so gilt auch

$$|\mathbb{E}(h(U + W)) - \mathbb{E}(h(V + W))| \leq c.$$

*Beweis dazu:* Zunächst nehmen wir an, dass  $W$  von der Form  $\sum_{i=1}^{\infty} r_i \chi_{A_i}$  mit paarweise verschiedenen  $r_i$  und paarweise disjunkten  $A_i$  ist. Wegen der Unabhängigkeit von  $U$  und  $W$  sind dann die  $\chi_{A_i}$  von  $U$  und damit auch von

$h \circ (U + r_i)$  unabhängig<sup>16)</sup>, und da man  $h(U + W)$  als  $\sum_i \chi_{A_i} h(U + r_i)$  schreiben kann, folgt aus der Multiplikativität des Erwartungswerts bei unabhängigen Zufallsvariablen (Satz 4.6.1), dass

$$\begin{aligned}\mathbb{E}(h(U + W)) &= \mathbb{E}\left(\sum_i \chi_{A_i} h(U + r_i)\right) \\ &= \sum_i \mathbb{E}(\chi_{A_i}) \mathbb{E}(h(U + r_i)) \\ &= \sum_i \mathbb{P}(A_i) \mathbb{E}(h(U + r_i)).\end{aligned}$$

Eine analoge Rechnung lässt sich für  $\mathbb{E}(h(V + W))$  durchführen, und so folgt

$$\begin{aligned}|\mathbb{E}(h(U + W)) - \mathbb{E}(h(V + W))| &= \left| \sum_i \mathbb{P}(A_i) (\mathbb{E}(h(U + r_i)) - \mathbb{E}(h(V + r_i))) \right| \\ &\leq \sum_i \mathbb{P}(A_i) |\mathbb{E}(h(U + r_i)) - \mathbb{E}(h(V + r_i))| \\ &\leq \sum_i \mathbb{P}(A_i) c \\ &\leq c.\end{aligned}$$

Für allgemeinere  $W$  beweisen wir das Ergebnis durch Approximation. Ein  $\varepsilon > 0$  wird vorgegeben, und dazu wählen wir  $\delta > 0$ , so dass  $|x - y| \leq \delta$  stets  $|h(x) - h(y)| \leq \varepsilon$  impliziert: Das ist möglich, denn  $h$  ist – als Funktion, die außerhalb eines kompakten Intervalls verschwindet – gleichmäßig stetig. Nun approximieren wir  $W$  bis auf  $\delta$  durch eine Funktion des Typs  $W' = \sum_i r_i \chi_{A_i}$  mit paarweise verschiedenen  $r_i$  und paarweise disjunkten  $A_i$ , so dass alle  $\chi_{A_i}$  von  $U$  und von  $V$  unabhängig sind. Dazu kann man die Intervalle  $[\delta m, \delta(m+1)]$  mit  $m \in \mathbb{Z}$  auf beliebige Weise als  $B_1, B_2, \dots$  abzählen, die  $A_i$  als  $\{W \in B_i\}$  definieren und  $r_i$  als linken Endpunkt von  $B_i$  wählen. Da alle  $W'(\omega)$   $\delta$ -nahe bei  $U(\omega)$  liegen, wird  $h \circ (U + W)$  gleichmäßig bis auf einen Fehler  $\varepsilon$  durch  $h \circ (U + W')$  approximiert, und deswegen ist  $|\mathbb{E}(h \circ (U + W)) - \mathbb{E}(h \circ (U + W'))| \leq \varepsilon$ .

Eine entsprechende Abschätzung gilt, wenn man  $U$  durch  $V$  ersetzt. Zusammen mit der schon bewiesenen Ungleichung

$$|\mathbb{E}(h \circ (U + W')) - \mathbb{E}(h \circ (V + W'))| \leq c$$

folgt daraus, dass  $|\mathbb{E}(h \circ (U + W)) - \mathbb{E}(h \circ (V + W))| \leq c + 2\varepsilon$ , und da  $\varepsilon$  beliebig war, ist die Behauptung 1 bewiesen.

*Behauptung 2:* Es sei  $\varepsilon_0 > 0$ . Dann gibt es ein  $n_0$ , so dass für alle  $n \geq n_0$  gilt:

$$|\mathbb{E}(h(r + X_1/\sqrt{n})) - \mathbb{E}(h(r + Y_1/\sqrt{n}))| \leq \frac{\varepsilon_0}{n}; \quad \text{alle } r \in \mathbb{R}.$$

(Da Erwartungswerte nur von der Verteilung abhängen, hätten wir statt  $X_1$  auch ein beliebiges  $X_i$  und für  $Y_1$  ein beliebiges  $Y_j$  wählen können.)

<sup>16)</sup>Vgl. Satz 4.4.8.

*Beweis dazu:* Dass die links stehende Zahl für große  $n$  klein wird, ist plausibel: Dann sind  $h(r + X_1/\sqrt{n})$  und  $h(r + Y_1/\sqrt{n})$  in guter Näherung die konstante Funktion  $h(r)$ , und deswegen wird die Differenz der Erwartungswerte gegen Null gehen.

Dass der Abstand aber sogar durch  $\varepsilon_0/n$  abgeschätzt werden kann, hat analytische Gründe. Um die Idee zu verstehen, nehmen wir einmal an,  $h$  wäre die Funktion  $h(x) = a + bx + cx^2$ . (Dass sie in Wirklichkeit gar nicht vorkommen kann, da sie nicht außerhalb eines Intervalls verschwindet, soll uns im Augenblick nicht stören.) Dann ist doch  $h \circ (r + X_1/\sqrt{n})$  von der Form  $a' + b'X_1 + c'X_1^2$ , der Erwartungswert ist also  $a' + c'$ , da wir  $\mathbb{E}(X) = 0$  und  $\sigma^2(X) = 1$  vorausgesetzt haben. Und genau so ergibt sich, dass der Erwartungswert von  $h \circ (r + Y_1/\sqrt{n})$  den Wert  $a' + c'$  hat, die Differenz ist also exakt Null.

Die vorstehenden Ideen sollen nun kombiniert werden: Als beliebig oft differenzierbare Funktion ist  $h$  lokal von der Form  $a + bx + cx^2$ , und für große  $n$  kann der Erwartungswert von  $h \circ (r + X_1/\sqrt{n})$  unter Ausnutzung des Verhaltens von  $h$  in der Nähe von  $r$  ermittelt werden. Das gleiche gilt für  $h \circ (r + Y_1/\sqrt{n})$ , und *deswegen* ist der Abstand der Erwartungswerte stets klein.

Technisch ist das leider ein bisschen verwickelt. Die Hauptschwierigkeit liegt darin, dass wir  $X_1$  und  $Y_1$  nicht als beschränkt annehmen dürfen. Um das zu umgehen, beginnen wir mit der Wahl eines  $\tau > 0$ , das wir erst später festsetzen. Wir setzen  $E_N := \{|X_1| \leq N\}$  für  $N \in \mathbb{N}$  und beachten dass die Funktionen  $\chi_{E_N} X_1^2$  punktweise und monoton gegen  $X_1^2$  konvergieren. Aus dem Satz von der monotonen Konvergenz (Anhang, Seite 358) folgt, dass die  $\mathbb{E}(\chi_{E_N} X_1^2)$  gegen  $\mathbb{E}(X_1^2) = 1$  konvergieren, und deswegen muss es ein  $N_0$  geben, für das  $\mathbb{E}(\chi_{E_{N_0}} X_1^2) \geq 1 - \tau$  gilt.

Nun betrachten wir die Taylorentwicklung von  $h$  bei den  $r \in \mathbb{R}$ . Wir bezeichnen mit  $\phi_r(s)$  die quadratische Funktion  $h(r) + h'(r)s + h''(r)s^2/2$ . Setzt man  $\psi_r(s) = h(r + s)$ , so kann man den Unterschied zwischen  $\phi$  und  $\psi$  gut beschreiben: Wegen der Restgliedformel ist die Differenz gleich  $(h''(r) - h''(r + \theta s))s^2/2$ , wobei  $\theta$  eine Zahl in  $]0, 1[$  ist.

Das hat eine wichtige Konsequenz: Da  $h''$  gleichmäßig stetig ist, findet man ein  $\delta > 0$ , so dass aus  $|x - y| \leq \delta$  stets  $|h''(x) - h''(y)| \leq \varepsilon_0/2$  folgt. Und damit ist  $|\phi_r(s) - \psi_r(s)| \leq \varepsilon_0 s^2/4$ , falls  $|s| \leq \delta$ . Für beliebige  $s$  lässt sich nur garantieren, dass  $|\phi_r(s) - \psi_r(s)| \leq M s^2$  gilt, wenn man  $M$  als das Maximum der Zahlen  $|h''(x)|$ ,  $x \in \mathbb{R}$ , definiert.

Wir wählen nun  $n_0$  so, dass  $N_0/\sqrt{n_0} \leq \delta$ . Dann gilt für beliebige  $r \in \mathbb{R}$  und  $n \geq n_0$ :

- Für die  $\omega \in E_0 := \{|X_1| \leq N_0\}$  ist

$$|\phi_r(X_1(\omega)/\sqrt{n}) - \psi_r(X_1(\omega)/\sqrt{n})| \leq \varepsilon_0 \frac{X_1^2}{4n}.$$

- Für die  $\omega \notin E_0$  ist

$$|\phi_r(X_1(\omega)/\sqrt{n}) - \psi_r(X_1(\omega)/\sqrt{n})| \leq M \frac{X_1^2}{n}.$$

Da  $\Omega \setminus E_0$  höchstens Wahrscheinlichkeit  $\tau$  hat, folgt für die Erwartungswerte:

$$\begin{aligned} |\mathbb{E}(\phi_r(X_1/\sqrt{n})) - \mathbb{E}(\psi_r(X_1/\sqrt{n}))| &\leq \mathbb{E}(|\phi_r(X_1/\sqrt{n}) - \psi_r(X_1/\sqrt{n})|) \\ &= \int_{\Omega} |\phi_r(X_1/\sqrt{n}) - \psi_r(X_1/\sqrt{n})| d\mathbb{P} \\ &= \int_{E_0} |\phi_r(X_1/\sqrt{n}) - \psi_r(X_1/\sqrt{n})| dWP + \\ &\quad + \int_{\Omega \setminus E_0} |\phi_r(X_1/\sqrt{n}) - \psi_r(X_1/\sqrt{n})| d\mathbb{P} \\ &\leq \frac{\varepsilon_0}{4} \int_{E_0} \frac{X_1^2}{n} d\mathbb{P} + M \int_{\Omega \setminus E_0} \frac{X_1^2}{n} d\mathbb{P} \\ &\leq \frac{1}{4n} (\varepsilon_0 + 4M\tau). \end{aligned}$$

Es folgt: Wenn man  $\tau > 0$  so wählt, dass  $4M\tau \leq \varepsilon_0$  gilt, so ist

$$|\mathbb{E}(\phi_r(X_1/\sqrt{n})) - \mathbb{E}(\psi_r(X_1/\sqrt{n}))| \leq \varepsilon_0/2n.$$

Für  $Y_1$  kann man ähnlich argumentieren; wir wollen das  $n_0$  so wählen, dass es sowohl für  $X_1$  als auch für  $Y_1$  geeignet ist. Dann wird der Beweis von Behauptung 2 schnell zu Ende geführt. Denn für  $n \geq n_0$  gilt aufgrund der vorstehenden Vorbereitungen:

- $|\mathbb{E}(h(r + X_1/\sqrt{n})) - \mathbb{E}(\phi_r(X_1/\sqrt{n}))| \leq \varepsilon_0/2n.$
- $|\mathbb{E}(h(r + Y_1/\sqrt{n})) - \mathbb{E}(\phi_r(Y_1/\sqrt{n}))| \leq \varepsilon_0/2n.$
- $\mathbb{E}(\phi_r(X_1/\sqrt{n})) = \mathbb{E}(\phi_r(Y_1/\sqrt{n}))$ ; das ergibt sich daraus, dass  $\phi_r$  ein quadratisches Polynom ist und  $X_1$  und  $Y_1$  Erwartungswert Null und Varianz Eins haben.

Mit Hilfe der Dreiecksungleichung folgt, dass

$$|\mathbb{E}(h(r + X_1/\sqrt{n})) - \mathbb{E}(h(r + Y_1/\sqrt{n}))| \leq \varepsilon_0/n$$

für  $n \geq n_0$ .

*Schluss des Beweises.* Sei  $\varepsilon_0 > 0$ . Es fehlt noch der Nachweis, dass man ein  $n_0$  so wählen kann, dass

$$\left| \mathbb{E}\left(h\left(\frac{X_1 + \dots + X_n}{\sqrt{n}}\right)\right) - \mathbb{E}\left(h\left(\frac{Y_1 + \dots + Y_n}{\sqrt{n}}\right)\right) \right| \leq \varepsilon_0$$

für  $n \geq n_0$ . Wir wählen  $n_0$  wie vorstehend und betrachten ein  $n \geq n_0$ .

Aus schreibtechnischen Gründen wird es günstig sein, für die nächsten Zeilen eine Abkürzung einzuführen: Für  $k = 0, \dots, n$  soll  $H_k$  durch

$$H_k := \mathbb{E}\left(h((Y_1 + \dots + Y_k + X_{k+1} + \dots + X_n)/\sqrt{n})\right)$$

definiert sein. (Im Fall  $k = 0$  treten nur die  $X_i$ , im Fall  $k = n$  nur die  $Y_i$  auf.)

*Schritt 1:* Es ist  $|\mathbb{E}(h(r + X_1/\sqrt{n}) - \mathbb{E}(h(r + Y_1/\sqrt{n}))| \leq \varepsilon_0/n$  für alle  $r \in \mathbb{R}$ , und da sowohl  $X_1/\sqrt{n}$  als auch  $Y_1/\sqrt{n}$  von  $(X_2 + \dots + X_n)/\sqrt{n}$  unabhängig sind, folgt aus Behauptung 1, dass  $|H_0 - H_1| \leq \varepsilon_0/n$ .

*Schritt 2:* Es ist  $|\mathbb{E}(h(r + X_2/\sqrt{n}) - \mathbb{E}(h(r + Y_2/\sqrt{n}))| \leq \varepsilon_0/n$  für alle  $r \in \mathbb{R}$ , und da sowohl  $X_2/\sqrt{n}$  als auch  $Y_2/\sqrt{n}$  von  $(Y_1 + X_3 + \dots + X_n)/\sqrt{n}$  unabhängig sind, folgt  $|H_1 - H_2| \leq \varepsilon_0/n$ .

*Schritt 3:* Es ist  $|\mathbb{E}(h(r + X_3/\sqrt{n}) - \mathbb{E}(h(r + Y_3/\sqrt{n}))| \leq \varepsilon_0/n$  für alle  $r \in \mathbb{R}$ , und da sowohl  $X_3/\sqrt{n}$  als auch  $Y_3/\sqrt{n}$  von  $(Y_1 + Y_2 + X_4 + \dots + X_n)/\sqrt{n}$  unabhängig sind, folgt  $|H_2 - H_3| \leq \varepsilon_0/n$ .

Und so weiter: Wir tauschen nach und nach die  $X_i$  gegen die  $Y_i$  aus. Im letzten Schritt erhalten wir die Abschätzung  $|H_{n-1} - H_n| \leq \varepsilon_0/n$ .

Zum Abschluss des Beweises nutzen wir einen *Teleskopsummen-Trick* aus:

$$\begin{aligned} \left| \mathbb{E}\left(h\left(\frac{X_1 + \dots + X_n}{\sqrt{n}}\right)\right) - \mathbb{E}\left(h\left(\frac{Y_1 + \dots + Y_n}{\sqrt{n}}\right)\right) \right| &= |H_0 - H_n| \\ &= |(H_0 - H_1) + (H_1 - H_2) + \dots + (H_{n-1} - H_n)| \\ &\leq |H_0 - H_1| + |H_1 - H_2| + \dots + |H_{n-1} - H_n| \\ &\leq n \frac{\varepsilon_0}{n} \\ &= \varepsilon_0. \end{aligned}$$

Damit ist der zentrale Grenzwertsatz vollständig bewiesen.  $\square$

Es folgen einige *Beispiele zur Illustration des zentralen Grenzwertsatzes*:

1. Zunächst soll darauf hingewiesen werden, dass der Satz von de Moivre-Laplace aus Abschnitt 3.4 wirklich ein Spezialfall des zentralen Grenzwertsatzes ist. Zur Begründung betrachten wir unabhängige Kopien  $X_1, \dots, X_n$  einer Bernoulli-variablen  $X$  (Erfolgswahrscheinlichkeit  $p$ ). Im Satz von de Moivre-Laplace ist man an der Wahrscheinlichkeit des Ereignisses  $E := \{\alpha \leq X_1 + \dots + X_n \leq \beta\}$  interessiert. Setzt man  $a := (\alpha - np)/\sqrt{np(1-p)}$  und  $b := (\beta - np)/\sqrt{np(1-p)}$ , so kann man  $E$  als

$$\left\{ a \leq \frac{(X_1 + \dots + X_n) - np}{\sqrt{np(1-p)}} \leq b \right\}$$

schreiben. Beachtet man, dass  $\mathbb{E}(X) = p$  und  $\sigma(X) = \sqrt{p(1-p)}$ , so folgt aus dem zentralen Grenzwertsatz, dass die Wahrscheinlichkeit von  $E$  durch  $\int_a^b e^{-x^2/2} dx / \sqrt{2\pi}$  approximiert werden kann.

Auf das gleiche Integral kommt man auch mit dem Satz von de Moivre-Laplace, denn es ist  $x(\alpha) = a$  und  $x(\beta) = b$ , wenn man die Bezeichnungen wie in diesem Satzes verwendet<sup>17)</sup>.

<sup>17)</sup>Genau genommen sind die Integrationsgrenzen dort  $x(\alpha - 0.5)$  und  $x(\beta + 0.5)$ , aber für nicht zu kleine  $n$  ist der Unterschied vernachlässigbar.

2. Die  $X_1, \dots, X_n$  seien wie im zentralen Grenzwertsatz. Dann ist, für beliebige vorgegebene  $\alpha, \beta$ , die Aussage  $\alpha \leq X_1 + \dots + X_n \leq \beta$  gleichwertig zu

$$\frac{\alpha - n\mathbb{E}(X)}{\sqrt{n}\sigma} \leq \frac{(X_1 + \dots + X_n) - n\mathbb{E}(X)}{\sqrt{n}\sigma} \leq \frac{\beta - n\mathbb{E}(X)}{\sqrt{n}\sigma}.$$

Die Wahrscheinlichkeit dieses Ereignisses kann also durch  $\int_a^b e^{-x^2/2} dx / \sqrt{2\pi}$  approximiert werden, wenn wir

$$a := (\alpha - n\mathbb{E}(X)) / (\sqrt{n}\sigma) \text{ und } b := (\beta - n\mathbb{E}(X)) / (\sqrt{n}\sigma)$$

setzen. Wie bei den Beispielen zum Satz von de Moivre-Laplace folgt daraus die überraschende Tatsache, dass  $X_1 + \dots + X_n$  mit einer bemerkenswert hohen Wahrscheinlichkeit „sehr nahe“ bei  $n\mathbb{E}(X)$  liegt.

Wie wahrscheinlich ist es zum Beispiel, dass die Augensumme aus 10.000 Würfelwürfen zwischen 34.700 und 35.300 liegt, die Abweichung vom Erwartungswert 35.000 also höchstens 300 ist? Hier ist  $X$  gleichverteilt auf  $\{1, 2, 3, 4, 5, 6\}$ , es ist  $\mathbb{E}(X) = 3.5$  und  $\sigma^2(X) = 35/12$ . Die gesuchte Wahrscheinlichkeit ist also (approximativ)  $\int_a^b e^{-x^2/2} dx / \sqrt{2\pi}$ , wobei die Werte von  $a$  und  $b$  durch  $a = -300/\sqrt{35 \cdot 10.000/12} \approx -1.757$ ,  $b = 300/\sqrt{35 \cdot 10.000/12} \approx 1.757$  gegeben sind. Mit Tabellenhilfe erhalten wir den Wert  $0.9599 - 0.0401 = 0.9198$ . Eine derart geringe Abweichung (weniger als ein Prozent) vom Erwartungswert ist also mit fast 92 Prozent Wahrscheinlichkeit zu erwarten.

3. Nun kann man endlich auch verstehen, warum man ausgerechnet 12 in  $[0, 1]$  gleichverteilte Zufallsvariable addiert und dann 6 abzieht, um die Standard-Normalverteilung zu simulieren. Der Erwartungswert der Gleichverteilung auf  $[0, 1]$  ist 0.5 und die Varianz ist  $1/12$ . Wählt man also  $n = 12$ , so wird der Ausdruck  $\sqrt{n}\sigma$  gleich Eins. Und deswegen ist wirklich die Wahrscheinlichkeit von  $a \leq X_1 + \dots + X_{12} - 6 \leq b$  nahe bei  $\int_a^b e^{-x^2/2} dx / \sqrt{2\pi}$ , d.h.  $X_1 + \dots + X_{12} - 6$  ist approximativ standard-normalverteilt.

## 8.5 Der Satz vom iterierten Logarithmus

In den vorigen Abschnitten haben wir nachgewiesen, dass „der Zufall im Unendlichen verschwindet“: Mittelwerte  $(X_1 + \dots + X_n)/n$  unabhängiger Experimente sind immer weniger Zufallsabhängig. In diesem Abschnitt wird beschrieben, wieviel Zufälliges erhalten bleibt. Das bestmögliche Ergebnis in diesem Zusammenhang ist der *Satz vom iterierten Logarithmus*, der hier erläutert werden wird. Die zugehörigen Beweise müssen leider entfallen, sie würden den Rahmen dieses Buches sprengen.

Eine Erinnerung: Limes superior und limes inferior

Ist  $(x_n)$  eine Folge reeller Zahlen, so kann man den *Limes superior*  $\limsup x_n$  dieser Folge definieren. Man sollte wissen:

- $\limsup_n x_n$  ist der größte Häufungspunkt der Folge  $(x_n)$ . Alternativ: Suche konvergente Teilfolgen, der größtmögliche Limes ist  $\limsup x_n$ . Dieser Wert existiert stets in  $[-\infty, +\infty]$ .

Zum Beispiel ist  $\limsup(-1)^n = 1$ ,  $\limsup(-n)^n = +\infty$ ,  $\limsup x_n = x$  im Fall  $\lim x_n = x$  usw.

- Ganz analog definiert man  $\liminf x_n$  (*Limes inferior*) als den kleinsten Häufungspunkt der Folge  $(x_n)$ . Auch dieser Wert existiert stets.
- Die Zahl  $a = \limsup x_n$  ist durch folgende Eigenschaften charakterisiert: Ist  $b > a$ , so gibt es nur endlich viele  $n$  mit  $x_n \geq b$ , und ist  $b < a$ , so gibt es unendlich viele  $n$  mit  $x_n > b$ .
- Der Limes der Folge  $(x_n)$  existiert in  $[-\infty, +\infty]$  genau dann, wenn  $\limsup x_n$  und  $\liminf x_n$  übereinstimmen.

Kurz:  $\limsup x_n$  ist so etwas wie die „wesentliche Größe“ von  $(x_n)$ .

### Eine alternative Interpretation der Gesetze der großen Zahlen

Wie in den Abschnitten 8.2 und 8.3 seien die  $X_1, X_2, \dots$  auf einem Wahrscheinlichkeitsraum  $\Omega$  definierte unabhängige Kopien einer reellwertigen Zufallsvariablen  $X$ , für die Erwartungswert  $\mathbb{E}(X)$  und Streuung  $\sigma := \sigma(X)$  existieren (vgl. Definition 8.2.1). Beim schwachen und beim starken Gesetz haben wir  $(X_1 + \dots + X_n)/n$  mit  $\mathbb{E}(X)$  verglichen: Die Abstände werden kleiner und kleiner.

Man kann aber auch  $X_1 + \dots + X_n$  mit  $n \mathbb{E}(X)$  vergleichen, also zum Beispiel die Anzahl der Erfolge in  $n$  Versuchen mit  $np$  bei unabhängigen Bernoulli-Experimenten mit Erfolgswahrscheinlichkeit  $p$ . Zur Veranschaulichung verbinden wird die Punkte  $(X_1(\omega) + \dots + X_n(\omega), n)$  durch Strecken und zeichnen gleichzeitig die Gerade  $n \mapsto np$  ein:

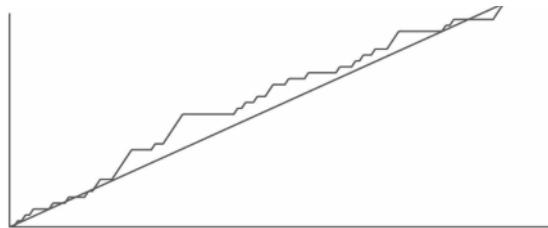


Bild 8.5.1: Ein zufälliger Pfad  $n \mapsto X_1 + \dots + X_n$ .

Das *schwache Gesetz* besagt dann: Für jedes  $\varepsilon > 0$  geht die Wahrscheinlichkeit dafür, dass  $X_1 + \dots + X_n$  um mehr als  $n\varepsilon$  von  $np$  abweicht, gegen Null.

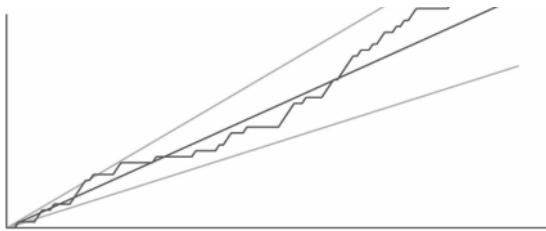


Bild 8.5.2: Liegt  $n \mapsto X_1 + \dots + X_n$  zwischen den Geraden  $n \mapsto (p \pm \varepsilon)n$ ?

Anders ausgedrückt bedeutet das: Fixiert man ein genügend großes  $n$ , so liegt der Pfad an der Stelle  $n$  mit hoher Wahrscheinlichkeit zwischen den Geraden  $(p - \varepsilon)n$  und  $(p + \varepsilon)n$ . Zur Illustration sind im nachstehenden Bild mehrere Pfade eingezeichnet:

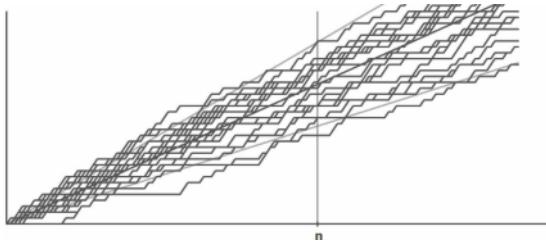


Bild 8.5.3: Mit hoher Wahrscheinlichkeit liegt  $X_1 + \dots + X_n$  in  $[pn - \varepsilon n, pn + \varepsilon n]$ .

Und das *starke Gesetz* lässt sich so umformulieren: Für jedes  $\varepsilon > 0$  liegt die Folge der  $(n, X_1 + \dots + X_n)$  von einer Stelle  $n_0$  an zwischen den Geraden  $n \mapsto (p - \varepsilon)n$  und  $n \mapsto (p + \varepsilon)n$ . Es ist zu betonen, dass das *nicht* aus dem schwachen Gesetz gefolgert werden kann.

#### Der Satz vom iterierten Logarithmus

Für große  $n$  ist viel Platz zwischen  $\mathbb{E}(X) - n\varepsilon$  und  $\mathbb{E}(X) + n\varepsilon$ , und man kann fragen, ob man das Verhalten des Zufalls nicht etwas genauer beschreiben kann.

Genau das leistet der Satz vom iterierten Logarithmus, der von dem russischen Mathematiker Alexander Khinchin<sup>18)</sup> 1924 bewiesen wurde<sup>19)</sup>. Seinen Namen hat der Satz von der Funktion  $n \mapsto \log \log n$ , die in der Formulierung des Satzes auftaucht. Ausführlich müsste man sie als  $n \mapsto \log(\log n)$  schreiben,



Khinchin

<sup>18)</sup>Aleksandr Khinchin, 1894 bis 1959. Professor in Moskau. Von ihm stammen fundamentale Beiträge in verschiedenen Gebieten: Zahlentheorie, Stochastik, mathematische Physik, theoretische Informatik.

<sup>19)</sup>Bei Khinchin geht es nur um unabhängige Bernoulli variable, der allgemeine Fall wurde erst in den vierziger Jahren des vorigen Jahrhunderts gezeigt.

es wird also der Logarithmus des Logarithmus gebildet. Man sollte sich klar machen, dass das eine extrem langsam wachsende Funktion ist<sup>20)</sup>: So ist etwa  $a := e^{(e^b)} \approx 2.85 \cdot 10^{64}$ , aber  $\log \log a$  hat nur den gegen  $a$  winzigen Wert 5.

Den Satz vom iterierten Logarithmus kann man ausführlich so formulieren:

- Sei  $\varepsilon > 0$ . Mit Wahrscheinlichkeit Eins ist die Folge  $X_1 + \dots + X_n$  nur endlich oft größer als  $n\mathbb{E}(X) + (1 + \varepsilon)\sigma\sqrt{2n \log \log n}$ .
- Für  $\varepsilon > 0$  gilt: Mit Wahrscheinlichkeit Eins ist die Folge  $X_1 + \dots + X_n$  unendlich oft größer als  $n\mathbb{E}(X) + (1 - \varepsilon)\sigma\sqrt{2n \log \log n}$ .

Anders ausgedrückt: Die Abweichungen von  $X_1 + \dots + X_n$  gegen  $n\mathbb{E}(X)$  nach oben sind – abgesehen von endlich vielen Ausnahmen – nicht größer als

$$(1 + \varepsilon)\sigma\sqrt{2n \log \log n},$$

sie sind aber immer einmal wieder mindestens

$$(1 - \varepsilon)\sigma\sqrt{2n \log \log n}.$$

Eine entsprechende Aussage gilt für Abweichungen nach unten.

Unter Verwendung der Begriffe Limes superior/inferior lässt sich das kürzer so formulieren:

**Satz 8.5.1. (Satz vom iterierten Logarithmus)**

Mit den vorstehenden Bezeichnungen gilt:

(i) Mit Wahrscheinlichkeit Eins ist

$$\limsup_{n \rightarrow \infty} \frac{X_1 + \dots + X_n - n\mathbb{E}(X)}{\sigma\sqrt{2n \log \log n}} = 1.$$

(ii) Mit Wahrscheinlichkeit Eins ist

$$\liminf_{n \rightarrow \infty} \frac{X_1 + \dots + X_n - n\mathbb{E}(X)}{\sigma\sqrt{2n \log \log n}} = -1.$$

Wie schon angekündigt, würde ein Beweis den Rahmen dieses Buches sprengen. Die Beweisidee ist allerdings leicht zu erklären: Bei Aussagen des Typs „Mit Wahrscheinlichkeit Eins tritt dieses oder jenes nur endlich oft ein“ liegt es nahe, es mit dem ersten Lemma von Borel-Cantelli zu versuchen. Und wenn „Mit Wahrscheinlichkeit Eins tritt dieses oder jenes unendlich oft ein“ zu zeigen ist, so ist möglicher Weise das zweite Lemma von Borel-Cantelli anwendbar.

So macht man es auch wirklich:

---

<sup>20)</sup>Die meisten Leser dieses Buches werden solche Funktionen noch nie angetroffen haben. In der Zahlentheorie, etwa bei der Beschreibung des Primzahlwachstums, spielen sie eine wichtige Rolle.

- Man zeigt, dass es für  $\varepsilon > 0$  Ereignisse  $B_1, B_2, \dots$  mit  $\sum_n \mathbb{P}(B_n) < +\infty$  gibt, so dass die Menge der  $\omega$ , für die

$$X_1(\omega) + \dots + X_n(\omega) > n \mathbb{E}(X) + (1 + \varepsilon)\sigma\sqrt{2n \log \log n}$$

unendlich oft eintritt, in  $\limsup B_n$  enthalten ist und folglich Wahrscheinlichkeit Null hat.

- Und man beweist, dass es für  $\varepsilon > 0$  unabhängige Ereignisse  $C_n$  mit der folgenden Eigenschaft gibt: Es ist  $\sum_n \mathbb{P}(C_n) = +\infty$ , und für  $\omega \in \limsup C_n$  – also mit Wahrscheinlichkeit Eins – tritt

$$X_1(\omega) + \dots + X_n(\omega) \geq n \mathbb{E}(X) + (1 - \varepsilon)\sigma\sqrt{2n \log \log n}$$

für unendlich viele  $n$  ein.

Beweise des Satzes vom iterierten Logarithmus findet man zum Beispiel in den Büchern von Feller und Klenke.

Für den Spezialfall  $\mathbb{E}(X) = 0$  – also zum Beispiel für ein faires Spiel – kann das Ergebnis so formuliert werden<sup>21)</sup>:

- Egal, wie klein  $\varepsilon$  ist, ein Gewinn von mehr als  $(1 + \varepsilon)\sigma\sqrt{2n \log \log n}$  in der  $n$ -ten Runde ist nur höchstens endlich oft zu erwarten. Das gleiche gilt für eine Pechsträhne: Nur endlich oft wird der Gesamtgewinn im  $n$ -ten Spiel weniger als  $-(1 + \varepsilon)\sigma\sqrt{2n \log \log n}$  sein.
- Man kann sich aber darauf verlassen, dass man unendlich oft in der  $n$ -ten Runde  $(1 - \varepsilon)\sigma\sqrt{2n \log \log n}$  gewonnen haben wird. Allerdings wird man auch unendlich oft bei weniger als  $-(1 - \varepsilon)\sigma\sqrt{2n \log \log n}$  angelangt sein.



Bild 8.5.4: Hohe Gewinne und Verluste sind mit Sicherheit zu erwarten.

Für Berufsspieler sind diese Ergebnisse nicht wirklich relevant. Erstens ist die Aussage „Dies und das passiert unendlich oft“ ziemlich nutzlos, wenn man nicht weiß, wann es eintritt. Und selbst wenn das der Fall ist, ist durchaus nicht klar, ob man so viele Runde spielen darf und ob man zwischendurch nicht irgendwann pleite ist.

---

<sup>21)</sup>Man muss vorher eine Einheit – zum Beispiel einen Euro – festgelegt haben.

## 8.6 Ergänzungen

Borel-Cantelli 2: schwächere Voraussetzungen

Wie viel Unabhängigkeit muss man für den zweiten Teil des Lemmas von Borel-Cantelli voraussetzen? Ist es wirklich notwendig zu fordern, dass  $A_1, A_2, \dots$  unabhängig sind?

Wir haben schon gesehen, dass das Lemma falsch wird, wenn man gar keine Bedingungen stellt<sup>22)</sup>. Bemerkenswerter Weise reicht es, die schwächste Form der Unabhängigkeit vorauszusetzen:

**Satz 8.6.1.** *Es seien  $A_1, A_2, \dots$  paarweise unabhängige Ereignisse in einem Wahrscheinlichkeitstraum mit  $\sum_n \mathbb{P}(A_n) = +\infty$ . Dann ist  $\mathbb{P}(\limsup_n A_n) = 1$ .*

**Beweis:** Sei  $\phi : \Omega \rightarrow [0, +\infty]$  die Abbildung  $\omega \mapsto \sum_n \chi_{A_n}(\omega)$ . Damit ist  $\phi(\omega)$  die Anzahl der  $A_n$ , in denen  $\omega$  liegt. Die Behauptung läuft darauf hinaus zu zeigen, dass  $\phi$  fast sicher den Wert  $+\infty$  annimmt.

Als Vorbereitung reduzieren wir das Problem. Für  $k \in \mathbb{N}$  sei  $B_k$  das Ereignis  $\{\phi \geq k\}$ ; das ist die Menge der  $\omega$ , die in mindestens  $k$  der  $A_n$  liegen. Mal angenommen, wir könnten zeigen, dass  $\mathbb{P}(B_k) = 1$  für jedes  $k$  gilt. Wegen  $B_1 \supset B_2 \supset \dots$  hätte dann auch der Durchschnitt  $\bigcap_n B_k$  aufgrund der Stetigkeit des Wahrscheinlichkeitsmaßes Wahrscheinlichkeit Eins. Auf diesem Durchschnitt ist aber  $\phi$  gleich Unendlich, und damit wäre die Behauptung bewiesen.

Sei  $k \in \mathbb{N}$  vorgegeben und  $\varepsilon > 0$  beliebig. Wir werden beweisen, dass  $\mathbb{P}(B_k) \geq 1 - \varepsilon$  gilt: Dann muss  $\mathbb{P}(B_k) = 1$  sein und der Beweis wäre vollständig geführt.

Für  $r \in \mathbb{N}$  definieren wir  $\phi_r$  durch  $\phi_r := \sum_{i=1}^r \chi_{A_i}$ . Der Erwartungswert dieser Zufallsvariablen ist  $E_r := \sum_{i=1}^r \mathbb{P}(A_i)$ , er wird nach Voraussetzung beliebig groß. Wir teilen durch  $E_r$ : Die Zufallsvariable  $\psi_r := \phi_r/E_r$  hat dann den Erwartungswert 1.

Die Varianz berechnen wir mit Hilfe von Satz 4.6.2. Nach diesem Satz ist die Varianz von  $\phi_r$  gleich der Summe der Varianzen der  $\chi_{A_i}$ , denn diese Zufallsvariablen sind nach Voraussetzung paarweise unabhängig. Es ist also

$$\sigma^2(\phi_r) = \sum_{i=1}^r \sigma^2(\chi_{A_i}) = \sum_{i=1}^r \mathbb{P}(A_i) - (\mathbb{P}(A_i))^2 \leq \sum_{i=1}^r \mathbb{P}(A_i).$$

Für die Varianz von  $\psi_r$  impliziert das

$$\sigma^2(\psi_r) = \frac{\sigma^2(\phi_r)}{E_r^2} \leq \frac{\sum_{i=1}^r \mathbb{P}(A_i)}{E_r^2} = \frac{1}{E_r}.$$

Da die  $E_r$  beliebig groß werden, heißt das, dass beliebig kleine Varianzen für die  $\psi_r$  zu erwarten sind.

Wähle  $r$  so groß, dass einerseits  $E_r \geq 2k$  und gleichzeitig  $\sigma^2(\psi_r) \leq \varepsilon/4$  gilt. Die Tschebyscheff-Ungleichung garantiert dann, dass  $\mathbb{P}(|\psi_r - 1| \geq 1/2) \leq$

---

<sup>22)</sup>Setze alle  $A_n$  gleich einem festen Ereignis  $A$ .

$\sigma^2(\psi_r)/(1/4) \leq \varepsilon$  gilt, und folglich ist  $\mathbb{P}(\{|\psi_r - 1| < 1/2\}) \geq 1 - \varepsilon$ . Für ein  $\omega \in \{|\psi_r - 1| < 1/2\}$  ist insbesondere  $\psi_r(\omega) \geq 1/2$ , d.h. es gilt  $\phi_r(\omega) \geq E_r/2 \geq k$ . Das impliziert, dass  $\omega$  zu  $B_k$  gehört, und damit ist wirklich  $\mathbb{P}(B_k) \geq 1 - \varepsilon$  wie behauptet.  $\square$

Diese Version des Lemmas ist – nach Kenntnis des Autors – in der Lehrbuchliteratur nicht zu finden. Die Erklärung liegt wohl darin, dass man in allen interessierenden Anwendungen Unabhängigkeit für die gesamte Folge  $A_1, A_2, \dots$  garantieren kann, schwächere Voraussetzungen also keine neuen Anwendungsmöglichkeiten eröffnen.

#### Ein alternativer Beweis des zentralen Grenzwertsatzes

Der zentrale Grenzwertsatz wurde in Kapitel 8.3 „elementar“ bewiesen: Alle Beweisschritte konnten auf dem Niveau einer elementaren Stochastikvorlesung verstanden werden.

Es gibt eine wesentlich elegantere Möglichkeit, das Ergebnis zu erhalten. Man kann es nach richtiger Übersetzung aus der elementaren Tatsache folgern, dass die Folge  $(1 + x/n)^n$  für alle  $x$  gegen  $e^x$  konvergiert.

Wenn man allerdings die Zwischenstufen alle ausführen würde, wäre der Beweisumfang erheblich, und außerdem müsste man sich dazu noch wesentlich ausführlicher um die maßtheoretischen Grundlagen der Wahrscheinlichkeitstheorie kümmern, als es hier möglich ist. Daher wird es nachstehend nur eine Skizze geben.

Die Idee besteht darin, das Problem zu transformieren: Eine Frage über Wahrscheinlichkeitsmaße wird in eine Frage über Funktionen übersetzt. Solche „Übersetzungen“ sind in der Mathematik nichts Ungewöhnliches. Schon in der Schule lernt man, dass man mit der Logarithmenrechnung multiplikative Probleme in additive verwandeln kann, mit Hilfe der Fouriertransformation werden aus gewissen Differentialgleichungsproblemen Probleme der linearen Algebra usw.

Hier ist die für unsere Zwecke relevante Definition:

**Definition 8.6.2.** Es sei  $X$  eine reellwertige Zufallsvariable. Unter der charakteristischen Funktion  $\phi_X$  von  $X$  verstehen wir dann die Abbildung

$$\phi_X : \mathbb{R} \rightarrow \mathbb{C}, \quad \phi_X(s) := \mathbb{E}(e^{isX}).$$

Genau genommen, haben wir Erwartungswerte bisher nur für reellwertige Zufallsvariable definiert, doch hier tritt die komplexwertige Funktion  $e^{isX}$  auf. Das ist kein schwerwiegendes Problem, denn wenn  $Y$  komplexwertig ist, kann man  $Y$  als  $Y_1 + iY_2$  mit reellwertigen  $Y_1, Y_2$  schreiben und dann  $\mathbb{E}(Y)$  durch die Formel  $\mathbb{E}(Y_1) + i\mathbb{E}(Y_2)$  definieren.

Bemerkenswerter ist die Tatsache, dass die Funktion  $e^{isX}$  durch Eins betragsmäßig beschränkt ist. Deswegen ist  $\phi_X(s)$  für alle  $s \in \mathbb{R}$  definiert.

Um Beispiele zu berechnen, ist es günstig, sich an Satz 3.3.5 zu erinnern: Der Erwartungswert von  $\psi(X)$  ist das Integral von  $\psi$  über  $\mathbb{R}$  bezüglich  $\mathbb{P}_X$ .

*Beispiel 1: charakteristische Funktion des Würfels:*

Es ist  $\mathbb{P}_{X(\{k\})} = 1/6$  für  $k = 1, \dots, 6$ , folglich hat  $\phi_X$  die Form

$$\phi_X(s) = \sum_{k=1}^6 \frac{e^{isk}}{6}.$$

Immer dann, wenn  $X$  nur endlich viele Werte annimmt, ist  $\phi_X$  eine endliche Konvexitätskombination von Funktionen des Typs  $s \mapsto e^{isa}$  mit  $a \in \mathbb{R}$ .

*Beispiel 2: charakteristische Funktion der Gleichverteilung auf  $[0, 1]$ :*

Aufgrund der Vorbemerkung ist  $\phi_X(s) = \int_0^1 e^{ist} dt = (e^{is} - 1)/(is)$  für  $x \neq 0$  und  $\phi_X(0) = 1$ .

*Beispiel 3: charakteristische Funktion der Poissonverteilung zum Parameter  $\lambda$ :* Zur Berechnung von  $\phi_X(s)$  müssen wir  $\sum_{k=0}^{\infty} \lambda^k e^{iks} e^{-\lambda} / k!$  auswerten. Das ergibt  $e^{\lambda(e^{is}-1)}$ .

*Beispiel 4: charakteristische Funktion der Standard-Normalverteilung:*

Für  $s \in \mathbb{R}$  ist  $\phi_X(s) = \int_{\mathbb{R}} e^{ist} e^{-t^2/2} dt / \sqrt{2\pi}$ . Das kann man (durch quadratische Ergänzung) als  $e^{-s^2/2} \int_{\mathbb{R}} e^{-(t-is)^2/2} ds / \sqrt{2\pi}$  schreiben, und das Integral  $\int_{\mathbb{R}} e^{-(t-is)^2/2} ds / \sqrt{2\pi}$  wird nach der Substitution  $u := t-is$  zu  $\int_{\mathbb{R}} e^{-u^2/2} du / \sqrt{2\pi}$ , es hat also den Wert Eins<sup>23)</sup>. Und das bedeutet:  $\phi_{N(0,1)}(s) = e^{-s^2/2}$ .

Man muss dann einige Tatsachen beweisen, die den Übergang von  $X$  zu  $\phi_X$  betreffen:

1.  $\phi_X$  hängt nur von  $\mathbb{P}_X$  ab: Das gilt offensichtlich.

2.  $\phi_{\alpha X}(s) = \phi_X(\alpha s)$  für  $\alpha \in \mathbb{R}$ . Das ist leicht einzusehen.

3. Wenn  $s$  „klein“ ist, so gilt

$$\phi_X(s) = \mathbb{E}(e^{isX}) = \mathbb{E}(1 + isX - X^2 s^2/2 + \dots) \approx 1 + is\mathbb{E}(X) - \mathbb{E}(X^2)s^2/2.$$

4. Sind  $X$  und  $Y$  unabhängig, so ist  $\phi_{X+Y} = \phi_X \phi_Y$ . Insbesondere gilt: Sind  $X_1, \dots, X_n$  unabhängige Kopien einer Zufallsvariablen  $X$ , so ist

$$\phi(X_1 + \dots + X_n)(s) = (\phi_X(s))^n.$$

5. Sind  $X, X_1, X_2, \dots$  Zufallsvariable, so konvergiert  $X_n$  in Verteilung gegen  $X$  genau dann, wenn die charakteristischen Funktionen  $\phi_{X_n}$  punktweise gegen  $\phi_X$  konvergieren.

Dabei sind die Beweise zu den Aussagen „4.“ und „5.“ die schwierigsten.

Mal angenommen, wir haben alle diese Vorbereitungen erarbeitet. Dann ist der zentrale Grenzwertsatz recht elegant beweisbar:

---

<sup>23)</sup>Hier wäre natürlich zu begründen, dass die aus dem Reellen gewohnten Substitutionsregeln auch im Komplexen angewandt werden können.

- Erstens überlegt man sich, dass es wie üblich reicht, sich auf den Fall  $\mathbb{E}(X) = 0$  und  $\sigma^2(X) = 1$  zu beschränken.
- Zweitens wüssten wir dann, dass der zentrale Grenzwertsatz äquivalent zu der Tatsache ist, dass die Zahlen

$$\phi_{(X_1+\dots+X_n)/\sqrt{n}}(s) = (\phi_X(s/\sqrt{n}))^n$$

für festes  $s$  gegen  $\phi_{N(0,1)}(s) = e^{-s^2/2}$  konvergieren.

Und das zeigt man so: Wenn  $n$  groß ist, ist  $s/\sqrt{n}$  klein, und deswegen darf  $\phi_X(s/\sqrt{n})$  durch  $1 + is\mathbb{E}(X) - \mathbb{E}(X^2)s^2/(2n) = 1 - s^2/(2n)$  approximiert werden. Deswegen ist wirklich

$$\phi_{(X_1+\dots+X_n)/\sqrt{n}} = (\phi_X(s/\sqrt{n}))^n \approx (1 - s^2/(2n))^n \approx e^{-s^2/2}.$$

Es ist klar, dass noch viele technische Feinheiten zu berücksichtigen wären, aber es sollte doch deutlich geworden sein, dass es am Ende nur an der Approximation  $(1 - s^2/(2n))^n \approx e^{-s^2/2}$  liegt, dass der zentrale Grenzwertsatz gilt.

## 8.7 Verständnisfragen

### Zu Abschnitt 8.1

#### *Sachfragen*

**S1:** Was versteht man unter dem Limes superior einer Mengenfolge?

**S2:** Was weiß man über den Limes Superior einer Ereignisfolge  $(A_n)$ , wenn  $\sum_n \mathbb{P}(A_n) < \infty$  gilt? Spielt für dieses Ergebnis die Unabhängigkeit der  $A_n$  eine Rolle?

**S3:** Was weiß man über den Limes Superior einer Ereignisfolge  $(A_n)$ , wenn  $\sum_n \mathbb{P}(A_n) = \infty$  gilt? Spielt für dieses Ergebnis die Unabhängigkeit der  $A_n$  eine Rolle? Falls ja, reicht paarweise Unabhängigkeit?

**S4:** Was weiß man über eine Folge  $X_n$  von Zufallsvariablen, wenn die Summe  $\sum_{n=1}^{\infty} \mathbb{P}(|X_n| > \varepsilon)$  für jedes  $\varepsilon$  endlich ist?

**S5:** Woran denkt ein Stochastiker, wenn vom „Affen an der Schreibmaschine“ die Rede ist? Inwieweit wirft dieses Gedankenexperiment Fragen nach der Bedeutung von Kreativität auf?

#### *Methodenfragen*

**M1:** Den Limes superior einer Mengenfolge bestimmen können.

**M2:** Die Lemmata von Borel-Cantelli anwenden können. (Zum Beispiel Konvergenzfragen mit Hilfe dieser Lemmata behandeln können).

**Zu Abschnitt 8.2***Sachfragen*

**S1:** Was besagt die Tschebyscheff-Ungleichung, was die Markov-Ungleichung?

**S2:** Was versteht man unter dem schwachen Gesetz der großen Zahlen?

**S3:** Welche Konvergenz spielt beim schwachen Gesetz eine Rolle: in Wahrscheinlichkeit, punktweise fast sicher, in Verteilung?

*Methodenfragen*

**M1:** Konkrete Rechnungen rund um die Tschebyscheff-Ungleichung (in der Version für Mittelwerte von  $n$  unabhängigen Zufallsvariablen) durchführen können.

**Zu Abschnitt 8.3***Sachfragen*

**S1:** Was besagt das starke Gesetz der großen Zahlen?

**S2:** Warum kann es als Rechtfertigung unseres axiomatischen Ansatzes interpretiert werden?

**S3:** Welche Konvergenz spielt beim starken Gesetz eine Rolle: in Wahrscheinlichkeit, punktweise fast sicher, in Verteilung?

**Zu Abschnitt 8.4***Sachfragen*

**S1:** Was versteht man unter dem zentralen Grenzwertsatz?

**S2:** Wie kommt es, dass man ausgerechnet 12 in  $[0, 1]$  gleichverteilte Zufallsvariable addiert (und dann 6 abzieht), um die Standard-Normalverteilung zu simulieren?

**S3:** Welche Konvergenz spielt beim zentralen Grenzwertsatz eine Rolle: in Wahrscheinlichkeit, punktweise fast sicher, in Verteilung?

*Methodenfragen*

**M1:** Den zentralen Grenzwertsatz zur Approximation von Wahrscheinlichkeiten anwenden können.

**Zu Abschnitt 8.5***Sachfragen*

**S1:** Was besagt der Satz vom iterierten Logarithmus?

*Methodenfragen*

**M1:** Folgerungen aus dem Satz vom iterierten Logarithmus ziehen können.

## 8.8 Übungsaufgaben

### Zu Abschnitt 8.1

**Ü8.1.1** Für  $n \in \mathbb{N}$  sei  $E_n \subset \mathbb{R}^2$  die Kreisscheibe mit dem Radius  $n^3$ . Bestimmen Sie  $\limsup E_n$  und  $\liminf E_n$ .

**Ü8.1.2** Es sei  $(X_n)$  eine Folge von unabhängigen Zufallsvariablen, die alle geometrisch verteilt zum Parameter  $0 < q < 1$  sind. Wie groß ist die Wahrscheinlichkeit, dass unendlich oft  $X_n \geq n + 1$  gilt?

**Ü8.1.3** Es sei  $(X_n)$  eine Folge von unabhängigen Zufallsvariablen, die alle exponentialverteilt zum Parameter  $\lambda > 0$  sind. Wie groß ist die Wahrscheinlichkeit, dass unendlich oft  $X_n \leq n$  gilt?

**Ü8.1.4** Es seien  $X_n : \Omega \rightarrow \mathbb{R}$  Zufallsvariable für  $n \in \mathbb{N}$ , und  $\Delta$  sei die Menge derjenigen  $\omega \in \Omega$ , für die  $\sum_n X_n(\omega)$  existiert. (Es darf ohne Beweis verwendet werden, dass  $\Delta$  ein Ereignis ist.)

Beweisen Sie: Ist  $\mathbb{P}(\{|X_n| \geq q^n\}) \leq q^n$  für eine geeignete Zahl  $q \in ]0, 1[$  und alle  $n$ , so hat  $\Delta$  Wahrscheinlichkeit 1: Die Reihe  $\sum_n X_n$  ist also fast sicher konvergent.

**Ü8.1.5** Ein Affe versucht, den Text „ELEMENTARE STOCHASTIK“ zu schreiben. Er ist unermüdlich (und unsterblich), ein Versuch dauert zehn Sekunden. Bestimmen Sie den Erwartungswert der Wartezeit, bis das zum ersten Mal klappt. (Auf seiner Schreibmaschine sind nur die Großbuchstaben und das Leerzeichen, und alle Tasten haben die gleiche Wahrscheinlichkeit, angeschlagen zu werden.)

### Zu Abschnitt 8.2

**Ü8.2.1** Zeigen Sie, dass für die Gültigkeit des schwachen Gesetzes die Unabhängigkeit der  $X_1, X_2, \dots$  wesentlich ist.

**Ü8.2.2** Nutzen Sie die Tschebyscheff-Ungleichung aus, um das folgende Problem zu lösen: Wie viele Versuche muss man machen, um eine unbekannte Wahrscheinlichkeit  $p$  bis auf einen Fehler von 0.02 so zu bestimmen, dass das Ergebnis mit 95 Prozent Wahrscheinlichkeit verlässlich ist

**Ü8.2.3** Das Integral  $\int_0^1 X(x) dx$  soll mit einem Monte-Carlo-Verfahren bestimmt werden: Als Ergebnis wird  $(X(x_1) + \dots + X(x_n))/n$  vorgeschlagen, wobei die  $x_1, \dots, x_n$  unabhängig und gleichverteilt in  $[0, 1]$  sind. Wie muss man  $n$  wählen, um das Integral bis auf einen Fehler von höchstens 0.01 und mit einer Sicherheit von 0.99 dadurch bestimmen zu können. (Achtung: Das Ergebnis hängt von  $X$  ab.)

### Zu Abschnitt 8.3

**Ü8.3.1** Zeigen Sie, dass für die Gültigkeit des starken Gesetzes die Unabhängigkeit der  $X_1, X_2, \dots$  wesentlich ist.

**Ü8.3.2** Angenommen, es existiert der Erwartungswert von  $X^6$ . Was lässt sich dann über  $\mathbb{P}(|X_1 + \dots + X_n|/n \geq \varepsilon)$  aussagen? (Vgl. den Beweis von Satz 8.3.1.)

**Ü8.3.3** Folgern Sie aus dem starken Gesetz: Ist  $E$  ein Ereignis im Wahrscheinlichkeitsraum  $(\Omega, \mathcal{E}, \mathbb{P})$  und erzeugt man unabhängige Abfragen aus  $\Omega$  gemäß  $\mathbb{P}$ , so konvergiert der Anteil der  $\omega$ , die in  $E$  liegen, fast sicher gegen  $\mathbb{P}(E)$ .

### Zu Abschnitt 8.4

**Ü8.4.1** Beweisen Sie, dass die Voraussetzung der Unabhängigkeit im zentralen Grenzwertsatz wesentlich ist.

**Ü8.4.2** Es soll gezeigt werden, dass gilt:

$$\lim_{n \rightarrow \infty} e^{-n} \sum_{k=0}^n \frac{n^k}{k!} = \frac{1}{2}.$$

Tipp: Betrachten Sie eine Folge  $X_1, X_2, \dots$  von unabhängigen Zufallsvariablen, die alle poissonverteilt zum Parameter 1 sind. Bestimmen Sie nun den Grenzwert  $\lim_{n \rightarrow \infty} \mathbb{P}(\{X_1 + \dots + X_n \leq 0\})$  auf zwei verschiedene Arten: einmal durch Anwenden des zentralen Grenzwertsatzes und einmal durch exaktes Berechnen von  $P(X_1 + \dots + X_n \leq 0)$ . Dabei muss man sich daran erinnern, was herauskommt, wenn man zwei Poisson-Verteilungen miteinander faltet.

Ein analytischer Beweis der Behauptung scheint schwieriger zu sein.

**Ü8.4.3** Das schwache Gesetz besagt doch: Wenn Erwartungswert und Streuung für  $X$  existieren, so gehen die Mittelwerte unabhängiger Kopien in Wahrscheinlichkeit gegen eine konstante Zufallsvariable (konstanter Wert =  $\mathbb{E}(X)$ ).

Zeigen Sie, dass dieses Ergebnis auch eine Folgerung aus dem zentralen Grenzwertsatz ist. (Vorbereitend ist dazu zu beweisen: Gehen Zufallsvariable in Verteilung gegen eine Konstante, so liegt sogar Konvergenz in Wahrscheinlichkeit vor.)

**Ü8.4.4** Eine Zufallsvariable  $X$  sei poissonverteilt zum Parameter 4, und die Zufallsvariablen  $X_1, X_2, \dots$  seien unabhängige Kopien. Wie groß ist (approximativ) die Wahrscheinlichkeit, dass  $X_1 + \dots + X_{1000}$  zwischen 3.9 und 4.1 liegt?

### Zu Abschnitt 8.5

**Ü8.5.1** Wir betrachten einen Zufallsspaziergänger auf  $\mathbb{Z}$ . Er startet bei Null, und im jeweils nächsten Schritt geht er mit gleicher Wahrscheinlichkeit einen Schritt nach links oder rechts. Folgern Sie aus dem Satz vom iterierten Logarithmus, dass er mit Wahrscheinlichkeit Eins jedes  $z \in \mathbb{Z}$  unendlich oft besuchen wird.

**Ü8.5.2** Stimmt die Aussage in der vorigen Aufgabe auch dann noch, wenn die Wahrscheinlichkeit für „links“ gleich  $p$  und für „rechts“ gleich  $1 - p$  ist und  $p$  von 0.5 verschieden ist?

**Ü8.5.3** Zeigen Sie, dass das starke Gesetz der großen Zahlen eine Folgerung des Satzes vom iterierten Logarithmus ist.

## Teil V

# Grundlagen der Statistik

Im letzten Teil dieses Buches werden einige Verfahren aus der Statistik vorgestellt. Stark vereinfacht kann man den Unterschied zwischen Wahrscheinlichkeitsrechnung und Statistik so beschreiben:

- In der *Wahrscheinlichkeitstheorie* sind ein Wahrscheinlichkeitsraum und eine oder mehrere Zufallsvariable gegeben. Die auftretenden Wahrscheinlichkeiten sind bekannt, und das Ziel ist es, Folgerungen zu ziehen.
- Die Ausgangssituation in der *Statistik* ist ganz anders: Auch hier spielen ein Wahrscheinlichkeitsraum ( $\Omega, \mathcal{E}, \mathbb{P}$ ) und eine darauf definierte Zufallsvariable  $X$  eine Rolle, doch die sind nicht direkt zugänglich. Alles, was man kennt, sind die Ergebnisse von endlich vielen (hoffentlich) unabhängigen Abfragen.

Ein Beispiel: Wir betrachten ein Bernoulliexperiment mit unbekannter Erfolgswahrscheinlichkeit  $p$ . Wir können dieses Experiment  $n$  Mal durchführen und protokollieren, wie viele Erfolge es dabei gibt. Was lässt sich dann über  $p$  aussagen?

Statistik ist wesentlich präsenter im öffentlichen Bewusstsein als Wahrscheinlichkeitsrechnung<sup>24)</sup>. Ein Teil der entsprechenden Meldungen ist sicher nicht besonders ernst zu nehmen. Sie beginnen mit „Wissenschaftler in X haben festgestellt, dass“, und dann folgen Aussagen, die manchmal skurril sind und oft auch wenig überraschen („Kinder älterer Väter sind klüger“. „Jungen fällt das Lernen schwerer als Mädchen“. „Gesunde Menschen sind glücklicher“. „Vegetarier leben länger“. „Ein Glas Rotwein am Abend vermindert das Infarktrisiko“. ...) Es überrascht nicht besonders, dass oft einige Wochen später das Gegenteil als neue Erkenntnis verkündet wird.

Viel schwerwiegender ist, dass Statistik sehr oft eine wesentliche Rolle spielt, wenn es um Einschätzungen und Entscheidungen geht, die für die Gesamtgesellschaft relevant sind:

- Ist die neue rein-biologische Kopfschmerztablette genauso wirkungsvoll wie die aus der Chemiefabrik?
- Verführen brutale Computerspiele zur Gewalttätigkeit?
- Ist das Krebsrisiko in der Nähe von Kohlekraftwerken erhöht?

Oft geht es um Konsequenzen mit großen Folgekosten, z.B. dann, wenn mehrere Tonnen Impfstoff hergestellt werden, um für das Übergreifen einer gefährlichen Krankheit vorbereitet zu sein. Oder dann, wenn man sicherheitshalber einige tausend Rinder (oder Schweine oder Hühner) notschlachtet und verbrennt, um ganz sicherzugehen, dass sich eine Seuche nicht ausbreitet.

Kurz: Die Einschätzung der Verlässlichkeit statistischer Methoden hat eine kaum zu überschätzende Bedeutung.

---

<sup>24)</sup>Die schafft es nur dann in die Medien, wenn es wieder einmal um die Wahrscheinlichkeit geht, den Jackpot zu gewinnen oder wenn „absolut sichere“ Gewinnstrategien – zum Beispiel beim Roulette – diskutiert werden.

Einige der wichtigsten Ansätze der *mathematischen Statistik*<sup>25)</sup> sollen in den nächsten Kapiteln vorgestellt werden.

In *Kapitel 9* ist alles noch recht elementar: Wie unterscheiden sich die verschiedenen Klassen von Daten, die man erheben kann? Durch welche Größen kann man sich einen ersten Überblick über die erhobene „Stichprobe“ machen? Welche Möglichkeiten gibt es, die Ergebnisse zu visualisieren? Mathematisch anspruchsvoller wird es dann in *Kapitel 10*. Es wird erklärt, was ein statistisches Modell ist, und es werden verschiedene Ansätze beschrieben, mit denen man einen unbekannten Parameter schätzen kann.

Oft sollen statistische Tests dazu dienen, Entscheidungen zu treffen: Dieses Medikament kann zugelassen werden, denn gefährliche Nebenwirkungen sind voraussichtlich nicht zu befürchten; diese Lieferung ist zurückzuweisen, denn mehr als fünf Prozent der gelieferten 10.000 Schaltkreise sind defekt; . . . In *Kapitel 11* wird beschrieben, wie eine „optimale“ Lösung solcher Entscheidungsprobleme aussehen könnte und wie man sie in vielen Fällen auch konkret angeben kann.

Ausgangspunkt der Untersuchungen in den Kapiteln 9, 10 und 11 waren immer irgendwelche Parameter, die man näher kennen lernen wollte oder die als Grundlage für Entscheidungen eine Rolle spielen. Es gibt aber eine Reihe von Problemen, die sich hier nicht einordnen lassen, etwa die Frage, ob zwei Zufallsvariable unabhängig sind. Man fasst die zugehörigen Verfahren unter dem Namen „Nichtparametrische Statistik“ zusammen, einige Beispiele werden in *Kapitel 12* vorgestellt.

---

<sup>25)</sup>Darunter versteht man diejenigen Teile der Statistik, bei denen die Ergebnisse mathematisch streng begründet werden können.

# Kapitel 9

## Beschreibende Statistik

Die ersten beiden Abschnitte dieses Kapitels haben mit Mathematik noch sehr wenig zu tun. In *Abschnitt 9.1* wird darauf hingewiesen, dass die von einem Statistiker erhobenen Daten einen sehr unterschiedlichen Charakter haben können, und in *Abschnitt 9.2* werden einige Möglichkeiten vorgestellt, Daten zu visualisieren.

Wenn die Daten gesammelt sind, so kann man sich fragen, durch welche damit zusammenhängenden Zahlen es möglich ist, sich einen ersten Überblick zu verschaffen. Dazu werden Stichprobenmittel, Median und Stichprobenstreuung in *Abschnitt 9.3* eingeführt, und es wird auch gezeigt, dass diese Größen alternativ als Lösung von gewissen Approximationsproblemen beschrieben werden können.

In *Abschnitt 9.4* geht es dann um Situationen, bei denen bei einer Abfrage zwei Größen gemessen wurden: Größe und Gewicht einer Person, Gehalt und Krankheitstage, usw. Der Korrelationskoeffizient hat sich als grobes Maß für die Quantifizierung von Abhängigkeiten zwischen diesen verschiedenen Aspekten bewährt. Er tritt auch auf, wenn man den zweiten Aspekt möglichst gut durch eine lineare Funktion des ersten Aspekts, also durch eine Regressionsgerade, approximieren möchte.

Das Ende des Kapitels bilden wieder Verständnisfragen und Übungsaufgaben (*in den Abschnitten 9.5 und 9.6*).

### 9.1 Statistische Daten

Statistiker sammeln Daten. Sie befragen Konsumenten, stellen fest, ob ein Medikament wirkt oder auch nicht, usw. Im einfachsten Fall wird pro „Abfrage“ nur ein einziges Ergebnis ermittelt („Gefällt Ihnen das Design?“; Antwortmöglichkeit Ja/Nein). Meist werden jedoch gleich mehrere Daten erhoben. Wenn es zum Beispiel um das Rauchverhalten geht, kann es interessant sein, nicht nur die Frage „Rauchen Sie?“ zu stellen, sondern auch nach dem Geschlecht und dem Alter zu fragen, um eventuell Zusammenhänge ermitteln zu können.

Man stellt sich dabei vor, dass die „wirkliche“ Situation durch einen Wahrscheinlichkeitsraum  $(\Omega, \mathcal{E}, \mathbb{P})$  und eine Zufallsvariable  $X : \Omega \rightarrow B$  beschrieben wird. Dabei kann – wenn viele Aspekte interessant sind –  $B$  ein mehr oder weniger komplizierter Produktraum sein, etwa

$$B = \{\text{ja,nein}\} \times \{\text{männlich,weiblich}\} \times \mathbb{N}$$

im vorstehenden Beispiel. Und der Statistiker hat nur Zugriff zu den  $X(\omega) \in B$ , daraus sollen die Informationen über  $\Omega$  und  $\mathbb{P}$  ermittelt und möglicherweise weitreichende Entscheidungen getroffen werden.

Wonach kann denn gefragt werden? Man unterscheidet:

- *Quantitativen Merkmale*. Dabei ist das Ergebnis eine Zahl, es wird etwas gemessen oder gezählt. Man unterscheidet *diskrete* und *kontinuierliche* quantitative Merkmale, je nachdem, ob man als Messwerte Zahlen in  $\mathbb{N}$  (evtl. nach Skalierung) oder in  $\mathbb{R}$  erwartet. Die mathematische Statistik beschäftigt sich überwiegend mit diesen Merkmalen.

Typische Beispiele sind Gewichte, Temperaturen, Längen, Wartezeiten usw. Nur in diesen Fällen ist es sinnvoll, Erwartungswerte und Varianzen bestimmten<sup>1)</sup>.

- *Rangmerkmale*. Manchmal kann man die möglichen Antworten in eine natürliche Reihenfolge bringen: „Wie hat Ihnen das Hotel gefallen: sehr gut / gut / mittel / schlecht / sehr schlecht ?“
- *Qualitative Merkmale*: Da geht es um Fragen, bei denen die Antwort in einer Menge ohne irgendeine Struktur ist. Beispiele sind: „Welche Automarke fahren Sie?“; „In welchem Land haben Sie zuletzt Urlaub gemacht?“; „Rauchen Sie?“.

## 9.2 Visualisierung von statistischen Daten

Die „Beschreibende Statistik“ ist ein wichtiges Teilgebiet der Statistik. Da der mathematische Anteil allerdings nicht allzu hoch ist, werden wir nur sehr kurz darauf eingehen.

Immer geht es um die Frage, wie man statistische Daten so visualisieren kann, dass interessante Kenngrößen und Zusammenhänge besser erkennbar werden. Fast täglich ist so etwas in der Zeitung zu sehen: Wahlergebnisse, die wirtschaftliche Entwicklung usw.

---

<sup>1)</sup>Doch Achtung: Das stimmt nicht in allen Fällen, wenn die Ergebnisse der Abfragen Zahlen sind. So ist zum Beispiel eine Aussage des Typs „Der Mittelwert der erfragten Hausnummern war 22.4“ völlig belanglos.

*Histogramme und Tortendiagramme*

Angenommen, bei der Umfrage geht es um ein qualitatives Merkmal, das  $n$  Ausprägungen haben kann<sup>2)</sup>. Dann gibt es zwei naheliegende Möglichkeiten, das Ergebnis zu visualisieren:

1. Die Anteile können durch Rechtecke dargestellt werden, wobei der Flächeninhalt dem Anteil entspricht. Damit das für die einzelnen Merkmale besser vergleichbar ist, sollte die Grundseite für alle Rechtecke die gleiche Länge haben.

Hier ein Beispiel, bei dem ein Phantasie-Wahlergebnis visualisiert wird:

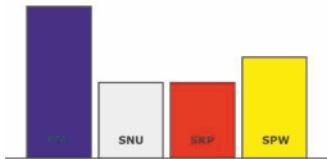


Bild 9.2.1: Visualisierung eines Wahlergebnisses.

Man spricht dann von einem *Histogramm*.

2. Wenn die Anzahl der Aspekte nicht zu groß ist, kann man die einzelnen Anteile auch durch die Größe der „Tortenstücke“ einer „Torte“ darstellen. So etwas wird dann ein *Tortendiagramm* genannt. Als *Beispiel* ist hier noch einmal das Phantasie-Wahlergebnis dargestellt:

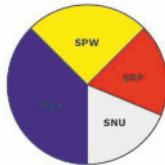


Bild 9.2.2: Ein Tortendiagramm.

Statt mit Rechtecken kann man auch mit Säulen arbeiten, die alle die gleiche quadratische Grundfläche haben sollten. Das bietet sich insbesondere dann an, wenn zwei qualitative Aspekte gleichzeitig dargestellt werden sollen. Hat der erste  $k$  und der zweite  $l$  Ausprägungen, so startet man mit einem  $k \times l$ -Schachbrett und platziert darauf Säulen so, dass die jeweiligen Höhen dem jeweiligen Anteil entsprechen. Tritt zum Beispiel bei 7 Prozent der Daten die Ausprägung  $i$  von Aspekt 1 und Ausprägung  $j$  von Anteil 2 auf, so sollte auf dem  $j$ -Feld der  $i$ -ten Reihe eine Säule der Höhe 7 stehen. Alles ist dann nur noch zweidimensional perspektivisch abzubilden<sup>3)</sup>.

<sup>2)</sup>Es könnte zum Beispiel um Automarken gehen, und dabei werden 20 Fabrikate berücksichtigt.

<sup>3)</sup>Oft ist es sinnvoll, die Grundfläche der Säulen kleiner zu wählen als die des Schachbrettfelds, damit keine der weiter hinten liegenden ganz verdeckt sind.

Als Beispiel betrachten wir eine Umfrage über das Rauchverhalten, bei der zusätzlich das Geschlecht verzeichnet wurde. Das Ergebnis lässt sich übersichtlich in einer so genannten *Kontingenztafel* darstellen:

|          | Raucher | Nichtraucher | gesamt |
|----------|---------|--------------|--------|
| männlich | 310     | 188          | 498    |
| weiblich | 210     | 292          | 502    |
| gesamt   | 520     | 480          | 1000   |

Ein dreidimensionales Histogramm könnte in diesem Fall so aussehen:

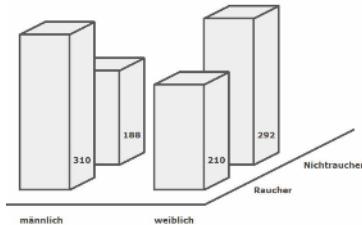


Bild 9.2.3: Dreidimensionale Darstellung von zwei Aspekten.

### Histogramme für quantitative Daten

Falls quantitative Merkmale vorliegen, kann man eine ähnliche Visualisierung verwenden. Wenn die Ergebnisse etwa zwischen  $a$  und  $b$  liegen, so unterteile man  $[a, b]$  in  $n$  Teilintervalle  $I_1, \dots, I_n$  gleicher Länge. Über  $I_j$  wird dann ein Rechteck errichtet, dessen Höhe proportional zur Anzahl der Messwerte in  $I_j$  ist.

Es wird von der Situation abhängen, wie man die Zahl  $n$  wählen sollte. In der Regel wird man einen Kompromiss finden müssen, denn ist  $n$  zu klein oder zu groß, so ist das Histogramm nur wenig aussagekräftig.

### Boxplots

Das ist eine weitere Möglichkeit, sich einen ersten Eindruck über die wichtigsten Charakteristika einer Stichprobe zu verschaffen. Wir gehen davon aus, dass wir Daten  $x_1, \dots, x_n \in \mathbb{R}$  gemessen haben, die einem quantitativen Merkmal entsprechen. Dann sucht man

- eine Zahl  $m$ , so dass  $n/2$  der  $x_i$  kleiner oder gleich  $m$  und  $n/2$  der  $x_i$  größer oder gleich  $m$  sind<sup>4)</sup>;
- Zahlen  $q_1$  und  $q_2$ , so dass jeweils ein Viertel der  $x_i$  kleiner oder gleich  $q_1$  bzw. größer oder gleich  $q_2$  ist: das sind die *Quartile*.

<sup>4)</sup>  $m$  ist ein Median, mehr dazu findet man im nächsten Abschnitt.

Dann kann man den zu den Daten gehörigen *Boxplot* zeichnen:

- Zeichne über einer Skala ein Rechteck (die „Box“), dessen Schmalseiten von  $q_1$  bis  $q_2$  reichen.
- Füge eine zu den Seitenlinien parallele Linie bei der Koordinate  $m$  ein.
- Verlängere die Box nach links bzw. rechts durch so genannte „Antennen“ bis zum kleinsten bzw. bis zum größten der  $x_i$ . (Für die Einzeichnung der Antennen kann auch anders verfahren werden, indem man zum Beispiel „Ausreißer“ ignoriert. Weitere Einzelheiten findet man auf der Internetseite <http://de.wikipedia.org/wiki/Boxplot>.)

Hier ein Beispiel, es sollen die folgenden Daten dargestellt werden:

39.41 82.04 118.56 71.36 76.91 91.28 114.80 109.31 83.25 69.47 56.04 56.21  
 68.14 79.75 106.04 46.53 56.92 88.82 73.44 73.72 43.72 76.56 78.40 117.46  
 96.32 103.91 110.79 128.37 48.57 118.72 78.40 62.46 41.32 104.37 68.39  
 82.47 97.88 100.60 70.63 68.26

Hier ist der Median gleich 78.40, das 0.25-Quartil ist gleich 62.46, und das 0.75-Quartil ist gleich 100.60.

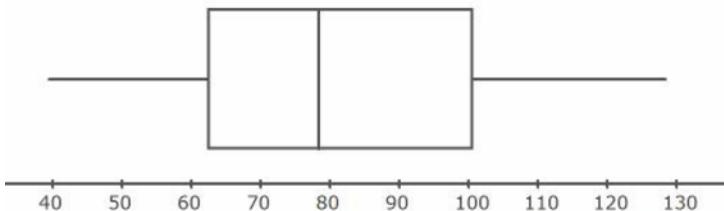


Bild 9.2.4: Boxplot-Darstellung der vorstehenden Daten.

### 9.3 Stichprobenmittel und Stichprobenvarianz

Nun wird es etwas mathematischer. Gegeben seien  $n$  Zahlen, die den Messungen  $x_1, \dots, x_n$  eines quantitativen Merkmals entsprechen. Man spricht von einer *Stichprobe*. In diesem Abschnitt geht es um einige Maßzahlen, die sich zur großen Beschreibung von Stichproben bewährt haben. Sie dienen zur Einschätzung gewisser typischer Aspekte der Situation.

#### Das Stichprobenmittel

Das *Stichprobenmittel* ist einfach der Mittelwert der Messwerte, es wird mit  $\bar{x}$  (gesprochen „ $x$  quer“) bezeichnet:

$$\bar{x} := \frac{1}{n} \sum_{i=1}^n x_i.$$

Die Idee dabei: Das Stichprobenmittel soll eine Einschätzung davon geben, wie groß die  $x_i$  „im Mittel“ sind. Das wird in den meisten Fällen auch wirklich erreicht, manchmal ist es aber kein geeignetes Maß:

- Wenn 10 Mitarbeiter einer Firma befragt werden, davon 9 jeweils 2000 Euro verdienen und ein einziger (Aufsichtsrat!) 22.000 Euro bekommt, so ist der Mittelwert 4000 Euro. Diese Zahl dient sicher nicht dazu, über die Gehaltsstruktur etwas Sinnvolles auszusagen.
- Genau so könnte man in einem Dritte-Welt-Land nach einer Umfrage (viele Arme, einige Multimillionäre) zu dem beruhigenden Ergebnis kommen, dass es den Leuten doch gar nicht so schlecht geht.
- In einem physikalischen Labor soll die Erdbeschleunigung durch Experimente ermittelt werden. Wenn dann einige Ausreißer dabei sind (grobe Ablesefehler, die U-Bahn fährt vorbei, ...), so wird der Mittelwert sicher kein guter Ausgangspunkt für eine Präzisionsmessung sein.

### Die Stichprobenvarianz

Das Stichprobenmittel ist natürlich die statistische Variante des Erwartungswertes aus der Wahrscheinlichkeitsrechnung. Nun kommen wir zum *Analogon der Varianz*. Wieder geht es darum, ein Maß dafür zu finden, wie stark die Werte um den Mittelwert streuen. Dafür gibt es viele Möglichkeiten, man könnte die Größe des Vektors

$$(x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x})$$

je nach Problemstellung in verschiedenen Normen des  $\mathbb{R}^n$  berechnen. Als wichtigste Maßzahl hat sich dabei eine geeignete Skalierung der euklidischen Norm herausgestellt. Bei dieser Norm werden Messwerte, für die der Abstand zu  $\bar{x}$  kleiner als Eins ist, wenig gewichtet, aber Abstände, die größer als Eins sind, gehen besonders stark in die Berechnung ein. Wie in der Wahrscheinlichkeitsrechnung ist die Bevorzugung der quadratischen Wichtung der Abweichung eher pragmatisch als logisch zu begründen<sup>5)</sup>.

In der nun folgenden Definition wird bei der Mittelwertbildung durch  $n - 1$  geteilt, man hätte hier eigentlich eher die Zahl  $n$  erwartet<sup>6)</sup>:

**Definition 9.3.1.** Unter der Stichprobenvarianz (auch: empirische Varianz) der Stichprobe  $x_1, \dots, x_n$  versteht man die Zahl

$$V_x := \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

---

<sup>5)</sup>Ein wichtiger Grund ist sicher die Tatsache, dass man bei Verwendung der euklidischen Norm viele Konzepte zur Verfügung hat, die in allgemeinen normierten Räumen nicht sinnvoll betrachtet werden können: Winkel zwischen Vektoren, Orthogonalität, eindeutige beste Approximationen an Unterräume, usw.

<sup>6)</sup>Der Grund für die Wahl von  $n - 1$  wird in Kapitel 10, Satz 10.2.2, klar werden.

Die Stichproben-Streuung – sie wird mit  $s_x$  bezeichnet – ist die positive Wurzel aus der Stichprobenvarianz.

### Der Median

Um die größten Probleme auszugleichen, die sich bei der Betrachtung des Stichprobenmittels als Maß für das mittlere Verhalten ergeben können, wird eine weitere Maßzahl studiert, der *Median*. Sie spielte schon im vorigen Abschnitt bei der Beschreibung der Darstellung durch Boxplots eine Rolle.

Hier die Definition:

**Definition 9.3.2.** Eine Zahl  $m$  heißt ein Median der Stichprobe, wenn mindestens die Hälfte der  $x_i$  größer gleich  $m$  und gleichzeitig mindestens die Hälfte der  $x_i$  kleiner gleich  $m$  ist.

Sind etwa alle  $x_i$  verschieden und ist  $n$  ungerade, so gibt es einen eindeutig bestimmten Median, nämlich „das mittlere“  $x_i$ ; ist  $n$  gerade, so bilden die Mediane ein Intervall.

Man kann das Stichprobenmittel und Mediane durch eine Approximationsbedingung charakterisieren, diese Zahlen minimieren die quadratischen bzw. die absoluten Abstände zu den  $x_i$ :

**Satz 9.3.3.** Es seien  $x_1, \dots, x_n \in \mathbb{R}$ .

- (i) Für  $x \in \mathbb{R}$  ist  $x$  genau dann gleich dem Stichprobenmittel  $\bar{x}$ , wenn die Zahl  $\sum_{i=1}^n (x_i - x)^2$  minimal ist.
- (ii) Ist  $x$  eine reelle Zahl, so ist  $x$  genau dann ein Median der  $x_1, \dots, x_n$ , wenn  $\sum_{i=1}^n |x_i - x|$  minimal ist.

**Beweis:** (i) Man kann diese Tatsache sehr elementar mit Hilfe der Differentialrechnung zeigen (Ableitung Null setzen usw.). Zur Übung der entsprechenden Methoden führen wir den Beweis aber im Rahmen der Theorie der euklidischen Räume. Wir arbeiten im  $\mathbb{R}^n$  mit der euklidischen Norm, die wichtigsten Fakten dazu sind im Anhang auf Seite 359 zusammengestellt. Es sei nun  $x \in \mathbb{R}$ . Betrachtet man die Vektoren  $A := (\bar{x}, \bar{x}, \dots, \bar{x})$ ,  $B := (x, x, \dots, x)$  und  $C := (x_1, \dots, x_n)$ , so steht  $B - A$  senkrecht auf  $A - C$ . Das folgt aus der Definition von  $\bar{x}$ :

$$\begin{aligned} \langle B - A, A - C \rangle &= \sum_{i=1}^n (x - \bar{x})(\bar{x} - x_i) \\ &= (x - \bar{x}) \sum_{i=1}^n (\bar{x} - x_i) \\ &= (x - \bar{x})(n\bar{x} - \sum_{i=1}^n x_i) \\ &= 0. \end{aligned}$$

Nach dem Satz von Pythagoras für euklidische Räume heißt das

$$\|B - C\|^2 = \|(B - A) + (A - C)\|^2 = \|B - A\|^2 + \|A - C\|^2.$$

Dieser Wert wird, wenn  $B$  alle möglichen Vektoren durchläuft, folglich am kleinsten, wenn  $B = A$  gilt.

Geometrisch kann man sich das so vorstellen: Ist  $W \subset \mathbb{R}^n$  der eindimensionale Unterraum der Vektoren der Form  $(x, \dots, x)$  (alle Einträge sind also gleich), so ist der Vektor  $A$  das Element bester Approximation in  $W$  an  $C$ . Daher die Orthogonalität.

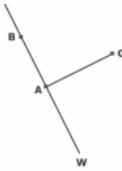


Bild 9.3.1:  $B - A$  steht senkrecht auf  $A - C$ .

(ii) Wir betrachten die Funktion

$$\phi : x \mapsto \sum_{i=1}^n |x_i - x|.$$

$\phi$  ist stetig und geht für  $|x| \rightarrow \infty$  gegen Unendlich, folglich wird das Minimum angenommen.

Sei  $x$  eine Minimalstelle. Es gebe  $a$  bzw.  $b$  bzw.  $c$  Indizes  $i$  mit  $x_i = x$  bzw.  $x_i < x$  bzw.  $x_i > x$ . (Es ist also  $a + b + c = n$ .) Wenn man dann (mit einem positiven kleinen  $\varepsilon$ ) von  $x$  zu  $x - \varepsilon$  übergeht, so verändert sich  $\phi$  um den Wert  $\varepsilon a - \varepsilon b + \varepsilon c$ . Da  $x$  Minimalstelle war, heißt das  $\varepsilon(a - b + c) \geq 0$ . Entsprechend folgt beim Betrachten von  $\phi(x + \varepsilon)$ , dass  $\varepsilon(a + b - c) \geq 0$ . Aus den beiden Ungleichungen  $a + b - c, a - b + c \geq 0$  folgt dann, dass  $a + b \geq c = n - (a + b)$  und dass  $a + c \geq b = n - (a + c)$ , d.h.  $a + b \geq n/2$  und  $a + c \geq n/2$  gilt. Das heißt, dass mindestens die Hälfte der  $x_i$  links und ebenfalls mindestens die Hälfte rechts von  $x$  liegt.

Damit ist gezeigt: Jeder minimierende Wert ist ein Median, und da es Minimalwerte gibt und  $\phi$  auf der Menge der Mediane konstant ist, ist auch die Umkehrung bewiesen.  $\square$

Mediane sind wesentlich stabiler gegen „Ausreißer“ als das Stichprobenmittel. In Situationen, in denen derartige Verfälschungen zu befürchten sind, ist eine Bewertung durch den Median daher realistischer.

Als kleine Anekdote aus einem mathematischen Fachbereich ist in diesem Zusammenhang zu berichten, dass die mittlere *Studiendauer* dort früher durch den Mittelwert der Studienzeiten der Absolventen gemessen wurde. Da es einige Super-Langzeit-Studenten gab, führte das zu beschämend schlechten Werten. Irgendwann konnte dann die Universitätsspitze davon überzeugt werden, dass der Median ein realistischeres Maß ist. Prompt wurde die mittlere Studiendauer um zwei Semester reduziert, die Studienzeiten waren nun irgendwo im bundesweiten Vergleich im Mittelfeld.

## 9.4 Korrelation und Regression

In diesem Abschnitt geht es um den Versuch, Zusammenhänge zwischen zwei quantitativen Merkmalen zu messen. Steigt der Ernteertrag mit der Düngemittelzugabe? Nimmt die Reisefreudigkeit mit dem Alter ab? Der *Korrelationskoeffizient* ist ein sehr grobes Hilfsmittel, um dazu Informationen zu bekommen.

Die Idee ist einfach. Die Stichproben  $x_i$  und  $y_i$  ( $i = 1, \dots, n$ ) für zwei quantitative Merkmale seien vorgelegt. Es könnte zum Beispiel sein, dass man bei  $n$  Personen das Gewicht und die Größe ermittelt hat. Wie üblich bezeichnen wir die Mittelwerte mit  $\bar{x}$  und  $\bar{y}$ . Nun betrachten wir die Produkte  $p_i := (x_i - \bar{x})(y_i - \bar{y})$ . Dann gilt doch:

- Wenn das Verhalten von  $x_i$  mit dem von  $y_i$  nichts zu tun hat, dann wird  $p_i$  positive und negative Werte annehmen, es wird keine bevorzugte Tendenz geben.
- Ist dagegen  $x_i$  in der Regel dann groß (bzw. klein), wenn das auch für  $y_i$  gilt, so sind die  $p_i$  von der Tendenz her eher positiv.
- Haben die  $x_i$  und die  $y_i$  eine eher gegensätzliche Tendenz ( $x_i$  ist groß wenn  $y_i$  klein ist und umgekehrt), so sind die  $p_i$  meist negativ.

Zusammengefasst heißt das, dass große Werte von  $\sum p_i$  dafür sprechen, dass eine Situation wie im ersten Fall vorlag. Kleine Werte sind typisch für den dritten Fall, und Werte in der Nähe von Null können in vielen Fällen so gedeutet werden, dass der durch die  $x_i$  gemessene Aspekt mit dem durch die  $y_i$  gemessenen nichts zu tun hat.

Da man eine Größe haben möchte, die maßstabsunabhängig ist, wird noch entsprechend geteilt. Die genaue Definition steht in

**Definition 9.4.1.**  $x_1, \dots, x_n$  und  $y_1, \dots, y_n$  seien quantitative Merkmale. Unter dem Korrelationskoeffizienten versteht man dann die Zahl

$$r_{xy} := \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}.$$

Dabei wird angenommen, dass nicht alle  $x_i$  gleich  $\bar{x}$  und nicht alle  $y_i$  gleich  $\bar{y}$  sind<sup>7)</sup>.

<sup>7)</sup>Denn nur dann ist der Nenner von Null verschieden.

**Bemerkungen:**

1.  $r_{xy}$  hat eine *geometrische Interpretation*. Wir betrachten wieder das am Ende des vorigen Abschnitts auf dem  $\mathbb{R}^n$  eingeführte Skalarprodukt. Zunächst gehen wir von  $(x_i)$  zu  $(x_i - \bar{x})$  und entsprechend von  $(y_i)$  zu  $(y_i - \bar{y})$  über, nehmen also o.B.d.A. an, dass die jeweiligen Stichprobenmittel verschwinden. Dann ist  $r_{xy}$  gerade der Quotient

$$\frac{\langle (x_i), (y_i) \rangle}{\| (x_i) \| \| (y_i) \|},$$

denn die Norm eines Vektors  $(z_i) \in \mathbb{R}^n$  ist durch  $\sqrt{\sum z_i^2}$  definiert. Dieser Ausdruck ist ein guter alter Bekannter. Man weiß aus der linearen Algebra<sup>8)</sup>:

- Er liegt zwischen  $-1$  und  $+1$ , das folgt aus der Cauchy-Schwarzschen Ungleichung. Der zugehörige Arcuscosinus wird als Winkel zwischen  $(x_i)$  und  $(y_i)$  interpretiert.
- Der Wert  $+1$  (bzw.  $-1$ ) wird genau dann angenommen, wenn  $y_i = ax_i$  für alle  $i$  und ein geeignetes  $a > 0$  (bzw.  $a < 0$ ) gilt. Das ist genau dann der Fall, wenn die Tupel  $(x_i, y_i)$  auf einer Geraden durch den Nullpunkt mit positiver bzw. negativer Steigung liegen<sup>9)</sup>.

2. Nach der Vorrede sollte klar sein: Ist  $r_{xy}$  in der Nähe von  $1$ , so haben die  $x_i$  die gleiche „Tendenz“ wie die  $y_i$  (ist  $x_i$  groß, so in der Regel auch  $y_i$ , und ist  $x_i$  klein, so auch  $y_i$ ), für  $r_{xy} \approx -1$  liegt eine gegensätzliche Tendenz vor (für große  $x_i$  ist  $y_i$  klein und umgekehrt), und Unabhängigkeit sollte zu  $r_{xy} \approx 0$  führen. Im Fall  $r_{xy} > 0$  bzw.  $r_{xy} < 0$  spricht man von einer *positiven* bzw. *negativen Korrelation*. Vorstellen kann man sich das so:

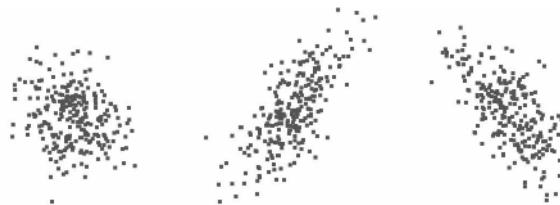


Bild 9.4.1: Korrelation nahe bei Null (links), positiv (Mitte), negativ (rechts).

<sup>8)</sup>Vgl. auch Seite 359 im Anhang.

<sup>9)</sup>Dabei ist „durch den Nullpunkt“ zu streichen, wenn man beliebige Situationen – also nicht notwendig  $\bar{x} = \bar{y} = 0$  – betrachtet: Die Approximationsmöglichkeit durch eine Gerade wird durch Verschieben des Koordinatensystems ja nicht beeinflusst.

### Eine Warnung

Es ist dringend davor zu warnen, eine positive Korrelation so zu interpretieren, dass der eine Aspekt den anderen sozusagen steuert. Manchmal ist das ja richtig: Zum Beispiel ist die Auslenkung einer Feder positiv korreliert (mit einem Korrelationskoeffizienten sehr nahe bei Eins) zu der Kraft, mit der sie auseinandergezogen wird.

Im Allgemeinen dürfen aus positiver Korrelation solche Schlüsse nicht gezogen werden. Das ist schon deswegen klar, weil bei der Aussage „Die Korrelation liegt nahe bei Eins“ die  $x_i$  und die  $y_i$  eine symmetrische Rolle spielen. Und es kann ja wohl nicht sein, dass die  $x_i$  die  $y_i$  steuern und das Umgekehrte gleichzeitig auch gilt.

Als konkreteres Beispiel betrachten wir die an verschiedenen Tagen durchgeföhrten Erhebungen „Helligkeitsdauer“ und „Öffnungszeiten der Eisdiele“. Niemand käme auf die Idee, die eine Größe als Ursache der anderen anzusehen: Sonst könnte man ja die Eisdiele länger offen halten, um einen späteren Sonnenuntergang zu erzwingen.

Leider muss diese Vorsicht bei der Interpretation statistischer Daten immer wieder angemahnt werden. So gab es kürzlich eine Studie, nach der die Friedfertigkeit männlicher Jugendlicher mit der Anzahl der Stunden, die sie in der Oper verbringen, stark positiv korreliert war. Und prompt kam der Vorschlag auf, allen Jugendlichen Freikarten zu schenken ...

Liegt  $r_{xy}$  in der Nähe von  $+1$  oder  $-1$ , kann man versuchen, den linearen Zusammenhang zwischen den  $x_i$  und den  $y_i$  etwas genauer zu untersuchen. Das Problem stellt sich so:

Gegeben seien  $(x_i, y_i)$  für  $i = 1, \dots, n$ , man kann sich diese Menge als „Punktwolke“ in der Ebene vorstellen. Finde eine Gerade, also eine Funktion der Form  $x \mapsto a + bx$ , die sich dieser Punktwolke „möglichst gut anpasst“.

Die Anführungszeichen deuten schon an, dass noch Präzisierungsbedarf besteht: Was soll *möglichst gut* heißen? Mit dem folgenden Ansatz lässt es sich am besten arbeiten:

**Definition 9.4.2.** Eine Gerade  $x \mapsto a + bx$  heißt eine Regressionsgerade, wenn der quadratische Abstand zur Punktwolke<sup>10)</sup> so klein wie möglich wird, wenn also

$$\sum_i (y_i - (a + bx_i))^2$$

unter allen möglichen Wahlen von  $a, b$  minimal ist.

---

<sup>10)</sup>Genauer: die Summe der Abstandsquadrate zwischen den Punkten der Punktwolke und den Punkten der Geraden.

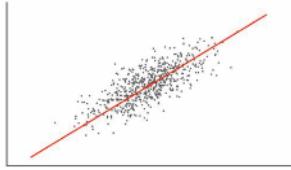


Bild 9.4.2: Eine Regressionsgerade.

Es gibt Regressionsgeraden, und sie sind eindeutig bestimmt:

**Satz 9.4.3.** *Die  $(x_i)$  und die  $(y_i)$  seien gegeben, es gelte  $s_x > 0^{11})$ .*

*(i) Ist  $\bar{x} = \bar{y} = 0$ , so ist die (eindeutig bestimmte) Regressionsgerade durch  $x \mapsto a + bx$  mit  $a = 0$  und*

$$b = \frac{r_{xy} s_y}{s_x} = \frac{\sum_{i=1}^n x_i y_i}{\sum_i^n x_i^2}$$

*gegeben.*

*(ii) Im allgemeinen Fall sind die Koeffizienten  $a, b$  der (eindeutig bestimmten) Regressionsgeraden  $x \mapsto a + bx$  durch*

$$b = \frac{r_{xy} s_y}{s_x}, \quad a = \bar{y} - b\bar{x}$$

*definiert.*

**Beweis:** (i) Es gelte  $\bar{x} = \bar{y} = 0$ . Zunächst sollte man sich daran erinnern, wie man Extremwertaufgaben in mehreren Veränderlichen löst. Wir definieren

$$\phi(a, b) := \sum_i (y_i - (a + bx_i))^2,$$

gesucht ist ein Minimum von  $\phi$  auf dem  $\mathbb{R}^2$ . Nun geht  $\phi$  für  $a, b \rightarrow \infty$  gegen Unendlich<sup>12)</sup>. Folglich wird das Minimum aus Stetigkeitsgründen angenommen. Wir können es dadurch finden, dass wir Punkte suchen, bei denen die partiellen Ableitungen von  $\phi$  nach  $a$  und nach  $b$  gleichzeitig verschwinden. Man rechnet leicht aus:  $\partial\phi/\partial a = 0$  ist gleichwertig zu  $\bar{y} = a + b\bar{x}$ , und  $\partial\phi/\partial b = 0$  lässt sich zu

$$\sum_i x_i y_i = a \sum_i x_i + b \sum_i x_i^2$$

umformen. Diese beiden Gleichungen haben eine eindeutig bestimmte Lösung, nämlich die, die im Satz angegeben ist. Aus der Problemstellung ist klar, dass es sich um ein Minimum handelt<sup>13)</sup>, die Eindeutigkeit wurde mitbewiesen.

<sup>11)</sup>  $s_x$  und  $s_y$  sind die Stichproben-Streuungen für die  $x$ - und die  $y$ -Werte; vgl. Seite 275.

<sup>12)</sup> Hier wird gebraucht, dass es mindestens zwei verschiedene  $x_i$ -Werte gibt; wir wollen das voraussetzen, sonst ist die Suche nach einer Geraden ja auch nicht sehr sinnvoll.

<sup>13)</sup> Man kann es auch streng einsehen: Die Hessematrix ist nämlich diagonal mit Einträgen  $2n$  und  $2 \sum_i x_i^2$ , ist also positiv definit.

- (ii) Dieser Teil kann durch Übergang von den  $x_i$  bzw. den  $y_i$  zu den  $x_i - \bar{x}$  bzw.  $y_i - \bar{y}$  auf den vorstehenden Spezialfall zurückgeführt werden.  $\square$

### Achtung: Paradoxien

Hat eine Regressionsgerade eine positive Steigung, so wird das oft so interpretiert, dass eine Zunahme des  $x$ -Merkmals „in der Regel“ eine Zunahme des  $y$ -Merkmals impliziert. (Es wurde oben schon betont, dass das nicht immer gerechtfertigt ist.) Hier gibt es viele Fallen, berühmt ist das *Simpson-Paradoxon*.

Zur Illustration betrachten wir die folgende Punktfolke. Sie könnte entstanden sein als Menge der Tupel

(Studiendauer, Anfangsgehalt)

bei einer Umfrage unter Universitätsabsolventen der Mathematik:

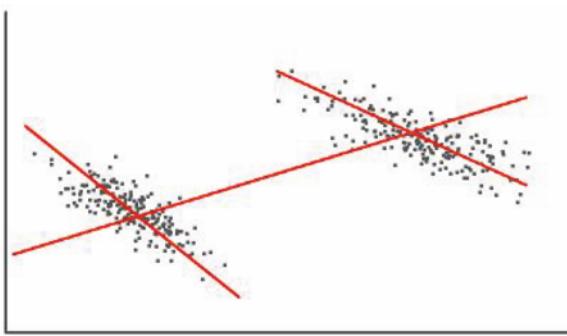


Bild 9.4.3: Das Simpson-Paradoxon.

Dabei betrifft die linke „Wolke“ Bachelorabsolventen und die rechte Diplomkandidaten. Wenn man nun eine Regressionsgerade durch die Menge aller Tupel legt, so gibt es eine positive Steigung. Fazit:

Je länger man studiert, um so höher ist das Anfangsgehalt.

In Wirklichkeit ist es aber gerade umgekehrt. Wenn man nur die Bachelorkandidaten oder nur die Diplomkandidaten betrachtet, so sieht man, dass Langzeitstudenten eher ein geringeres Anfangsgehalt bekommen.

Ein ähnliches Paradoxon gibt es für bedingte Wahrscheinlichkeiten. Ausgangspunkt ist die folgende Tatsache aus der Bruchrechnung:

$$\text{Aus } a_1/b_1 < x_1/y_1 \text{ und } a_2/b_2 < x_2/y_2 \text{ folgt nicht} \\ (a_1 + a_2)/(b_1 + b_2) < (x_1 + x_2)/(y_1 + y_2).$$

Das kann für die Statistik wichtig sein. Bewerben sich etwa  $b_i$  Männer und  $y_i$  Frauen für das Studienfach  $S_i$  ( $i = 1, 2$ ) und haben  $a_i$  Männer bzw.  $x_i$  Frauen Erfolg, so kann folgende Situation eintreten:

Es ist  $a_1/b_1 < x_1/y_1$  und  $a_2/b_2 < x_2/y_2$ , d.h. in  $S_1$  und  $S_2$  ist die Quote der erfolgreichen Männer schlechter als die der Frauen. Trotzdem gilt  $(a_1+a_2)/(b_1+b_2) > (x_1+x_2)/(y_1+y_2)$ , d.h. der Anteil der Erfolgreichen ist bei den Männern höher als bei den Frauen.

### Eine Ergänzung: Ein Schritt in die Nichtlinearität

Durch die Regressionsgerade sollte doch ein *linearer* (eigentlich: affiner) Zusammenhang aufgedeckt werden: Wenn „in Wirklichkeit“  $y = a + bx$  gilt, aber nur fehlerbehaftete Messungen von  $y$  bei verschiedenen  $x$  zur Verfügung stehen, wie kann man dann sinnvolle Kandidaten für  $a$  und  $b$  finden?

Konstruktion und Beweis sind eindeutig auf den linearen Fall zugeschnitten. Trotzdem ist es nicht schwer, die gleichen Überlegungen auch auf gewisse nichtlineare Situationen zu übertragen.

#### *Exponentielles Wachstum*

Mal angenommen, es liegen Messpunkte  $(x_i, y_i)$  vor, die – wenn man sie als Punktwolke skizziert – an eine Exponentialfunktion erinnern. So etwas passiert häufig, wenn Wachstums- oder Zerfallsvorgänge beobachtet werden. Wie lassen sich dann  $a$  und  $b$  so bestimmen, dass die Kurve  $y = a \cdot e^{bx}$  eine möglichst gute Approximation darstellt?

Dazu wird die Gleichung  $y = a \cdot e^{bx}$  zu  $\log y = \log a + bx$  umgeformt. Folglich reicht es, die Punktwolke  $(x_i, \log y_i)$  mit den bekannten Methoden durch eine Regressionsgerade  $\alpha + \beta x$  zu approximieren und dann  $a := e^\alpha$  und  $b := \beta$  zu setzen<sup>14)</sup>.

#### *Wachstum wie bei einer Potenz $x^r$*

Manchmal ist es günstiger, es mit dem Ansatz  $y = a \cdot x^b$  versuchen. Auch hier findet man  $a$  und  $b$  nach einer Umformung:  $y = a \cdot x^b$  ist gleichwertig zu  $\log y = \log a + b \log x$ . Man muss also nur eine Regressionsgerade  $\alpha + \beta x$  für die Paare  $(\log x_i, \log y_i)$  finden und dann wieder  $a := e^\alpha$  und  $b := \beta$  setzen<sup>15)</sup>.

## 9.5 Verständnisfragen

### Zu Abschnitt 9.1

#### *Sachfragen*

**S1:** Was sind quantitative Merkmale?

**S2:** Was sind Rangmerkmale?

**S3:** Was sind qualitative Merkmale?

---

<sup>14)</sup>Wenn man es „von Hand“ machen möchte, empfiehlt es sich, die  $(x_i, y_i)$  in einfach logarithmisches Papier einzutragen.

<sup>15)</sup>Um eine Vorstellung darüber zu bekommen, ob so ein Modell aussichtsreich ist, empfiehlt sich, die  $(x_i, y_i)$  in doppelt-logarithmischem Papier einzutragen. Die Punkte sollten dann ungefähr auf einer Geraden liegen.

*Methodenfragen*

**M1:** Entscheiden können, zu welchem Typ die in der Stichprobe erhobenen Merkmale gehören.

**Zu Abschnitt 9.2***Sachfragen*

**S1:** Was ist ein Histogramm?

**S2:** Was ist ein Tortendiagramm?

**S3:** Welche Informationen enthält ein Boxplot?

*Methodenfragen*

**M1:** Graphische Darstellungen von Stichproben erstellen und interpretieren können.

**Zu Abschnitt 9.3***Sachfragen*

**S1:** Was ist das Stichprobenmittel? Durch welches Extremwertproblem ist es charakterisiert?

**S2:** Was ist ein Median einer Stichprobe? Durch welches Extremwertproblem sind Mediane charakterisiert?

**S3:** Was versteht man unter Stichprobenvarianz und Stichprobenstreuung?

*Methodenfragen*

**M1:** Stichprobenmittel, Mediane sowie Stichprobenvarianzen und -streuungen bestimmen können.

**M2:** Einfache Eigenschaften im Zusammenhang mit diesen Definitionen herleiten können.

**Zu Abschnitt 9.4***Sachfragen*

**S1:** Was ist der Korrelationskoeffizient, der zu einer Stichprobe von zwei quantitativen Merkmalen gehört?

**S2:** Durch welche Eigenschaft ist eine Regressionsgerade definiert?

**S3:** Was besagt das Simpsonparadoxon?

*Methodenfragen*

**M1:** Die Regressionsgerade zu einer Familie  $(x_i, y_i), i = 1, \dots, n$  bestimmen können.

**M2:** Einfache Eigenschaften im Zusammenhang mit Korrelation und Regressionsgeraden beweisen können.

## 9.6 Übungsaufgaben

### Zu Abschnitt 9.3

**Ü9.3.1** Es sei  $\psi : \mathbb{R} \rightarrow \mathbb{R}$  stetig differenzierbar.  $\psi'$  sei streng monoton steigend und bei 0 gleich Null; 0 ist also das eindeutig bestimmte Minimum von  $\psi$ . (Beispiele:  $x^{2k}$ ,  $e^{x^4}$ , ...) Für eine Stichprobe  $x_1, \dots, x_n \in \mathbb{R}$  betrachten wir

$$\phi(x) := \sum_{i=1}^n a_i \psi(x - x_i)$$

für alle  $x \in \mathbb{R}$ ; dabei seien  $a_1, \dots, a_n > 0$ .

Beweisen Sie, dass es ein eindeutig bestimmtes  $x_0$  zwischen  $\min x_i$  und  $\max_i x_i$  gibt, bei dem  $\phi$  minimiert wird.

(Neben der minimalen quadratischen Abweichung könnte man also auch allgemeiner mit einer gewichteten  $\psi$ -Abweichung arbeiten.)

**Ü9.3.2** Es sei  $x_1, \dots, x_n \in \mathbb{R}$  eine Stichprobe. Für welche  $x$  wird  $\max_i |x - x_i|$  minimal?

**Ü9.3.3** Beweisen oder widerlegen Sie die folgende Aussagen.

- a) Entfernt man aus einer Stichprobe  $x_1, \dots, x_n$  ein  $x_i$  mit  $x_i = \max_j x_j$ , so wird das Stichprobenmittel kleiner.
- b) Das Stichprobenmittel von  $x_1, \dots, x_n$  ist genau dann Null, wenn (im euklidischen Raum  $\mathbb{R}^n$ )  $(x_1, \dots, x_n)$  senkrecht auf  $(1, \dots, 1)$  steht.
- c) Das Stichprobenmittel ist immer ein Median.
- d) Sei  $x_1, \dots, x_n$  eine Stichprobe, in der jedes  $\alpha \in \mathbb{R}$  genauso oft wie  $-\alpha$  auftritt. Dann ist das Stichprobenmittel ein Median.
- e) Ist in einer Stichprobe das Stichprobenmittel enthalten, so ist das Stichprobenmittel ein Median.

### Zu Abschnitt 9.4

**Ü9.4.1** Die  $(x_i, y_i)_{i=1, \dots, n}$  seien eine zweidimensionale Stichprobe. Sind die nachstehenden Aussagen richtig oder falsch (mit Begründung)?

- a) Wenn der Korrelationskoeffizient  $r_{xy}$  verschwindet, so gibt es keine Regressionsgerade.
- b) Wenn der Korrelationskoeffizient  $r_{xy}$  positiv ist, so hat die Regressionsgerade eine positive Steigung.
- c) Sind  $\alpha, \beta > 0$ , so ist der Korrelationskoeffizient für die Stichprobe  $(\alpha x_i, \beta y_i)$  der gleiche wie für  $(x_i, y_i)_{i=1, \dots, n}$ .
- d) Ist  $\alpha > 0$ , so ist der Korrelationskoeffizient für die Stichprobe  $(\alpha x_i, \alpha y_i)$  der gleiche wie für  $(x_i, y_i)_{i=1, \dots, n}$ .

**Ü9.4.2** Es sei  $(x_i, y_i)_{i=1, \dots, n+m}$  eine zweidimensionale Stichprobe. Die Stichprobenvarianz der  $(x_i)_{i=1, \dots, n}$  und die Stichprobenvarianz der  $(x_i)_{i=n+1, \dots, n+m}$  seien von Null verschieden. Die Regressionsgeraden von  $(x_i, y_i)_{i=1, \dots, n}$  und von  $(x_i, y_i)_{i=n+1, \dots, n+m}$  seien identisch.

Beweisen oder widerlegen Sie: Dann ist diese Gerade auch die Regressionsgerade zu  $(x_i, y_i)_{i=1, \dots, n+m}$ .

(Tipp: Erinnern Sie sich an die Definition der Regressionsgeraden als Lösung eines Optimierungsproblems.)

**Ü9.4.3** Geben Sie ein Beispiel für eine Stichprobe  $(x_i, y_i)_{i=1, \dots, n}$  an, bei dem sich die Regressionsgerade ändert, wenn man die  $x_i$  mit den  $y_i$  vertauscht.

**Ü9.4.4** Die Parabel  $x \mapsto ax^2 + bx + c$  heißt Regressionsparabel der Stichprobe  $(x_i, y_i)_{i=1, \dots, n}$ , falls

$$\sum_{i=1}^n (y_i - (ax_i^2 + bx_i + c))^2$$

minimal ist. Zeigen Sie die Existenz und die Eindeutigkeit der Regressionsparabel und finden Sie ein Gleichungssystem, mit dem man  $a, b, c$  bestimmen kann.

**Ü9.4.5** Für  $k = 1, 2, \dots$  sei  $(x_i, y_i^{(k)})_{i=1, \dots, n}$  eine zweidimensionale Stichprobe. Die Stichprobenvarianz der  $x_i$  sei von Null verschieden. Die zur  $k$ -ten Stichprobe gehörige Regressionsgerade sei als  $a^{(k)} + b^{(k)}x$  geschrieben.

Es wird vorausgesetzt, dass für  $i = 1, \dots, n$  die Folge der  $(y_i^{(k)})_k$  konvergiert, der Limes werde mit  $y_i$  bezeichnet.

Man zeige: Ist  $a + bx$  die Regressionsgerade zu  $(x_i, y_i)_{i=1, \dots, n}$ , so gilt  $a^{(k)} \rightarrow a$  sowie  $b^{(k)} \rightarrow b$ .

(Anders ausgedrückt: Die Regressionsgerade hängt stetig von den  $y_i$  ab.)

# Kapitel 10

## Schätzen

Schätzprobleme lassen sich durch viele Beispiele illustrieren: Wie lange wird eine frisch gekaufte Glühlampe halten? Wie groß ist die Wahrscheinlichkeit, dass ein zufällig ausgewählter Wahlberechtigter die SPD wählen wird? ...

Allgemein geht es darum, dass man einen Wahrscheinlichkeitsraum  $(\Omega, \mathcal{E}, \mathbb{P})$  vor sich hat, von dem man nur weiß, dass er zu einer bestimmten Familie von Wahrscheinlichkeitsräumen gehört: Zu jedem Wert  $\theta$  aus einer Indexmenge  $\Theta$  gehört so ein Raum. Nun wird  $n$  Mal aus  $\Omega$  abgefragt, und anhand der so erzeugten Stichprobe  $\omega_1, \dots, \omega_n$  soll man das  $\theta$  „möglichst gut“ schätzen, das zu  $\mathbb{P}$  gehört.

Das ist, zugegeben, noch recht vage, eine Präzisierung der Problemstellung wird es gleich in *Abschnitt 10.1* geben. Danach klären wir in *Abschnitt 10.2* die Frage, was denn beim Schätzen „möglichst gut“ heißen soll. Einige Beispiele für Schätzverfahren werden dann in *Abschnitt 10.3* diskutiert; dabei lassen sich für einige spezielle Situationen sogar die beweisbar besten Lösungen angeben.

In *Abschnitt 10.4* wird es dann um einen anderen Ansatz beim Schätzen gehen. In den ersten Abschnitten waren die Ergebnisse immer von der Form: „Aufgrund der Stichprobe wird geschätzt, dass der fragliche Parameter den Wert soundso hat“. Jetzt wird es um vorsichtiger Formulierungen gehen: Eine typische Aussage könnte sein: „Aufgrund der Stichprobe liegt die unbekannte Wahrscheinlichkeit  $p$  im Intervall  $[0.2, 0.25]$ “. Diese Theorie der *Konfidenzbereiche* lässt besonders im Fall normalverteilter Zufallsvariablen sehr präzise Aussagen zu: Das ist der Gegenstand von *Abschnitt 10.5*.

Den Schluss des Kapitels bilden dann wieder (*in den Abschnitten 10.6 und 10.7*) Verständnisfragen und Übungsaufgaben.

## 10.1 Das statistische Modell, Schätzfunktionen

### *Die Fragestellung*

Bevor wir später alles mit Hilfe der in den vorigen Kapiteln eingeführten Begriffe präzisieren, illustrieren wir das Schätzproblem an einigen Beispielen.

*Beispiel 1:* Vor Ihnen liegt ein aus 20 Skatkarten bestehender Kartenstapel, der  $r$  rote und  $20 - r$  schwarze Karten enthält. Die Zahl  $r$  ist dabei unbekannt. Nun wird 10 Mal hintereinander mit Zurücklegen gezogen, und nach jedem Ziehen wird gut gemischt. Sie erfahren allerdings nur, ob die gezogene Karte rot oder schwarz war. Und danach sollen Sie die Anzahl  $r$  der roten Karten schätzen.

Auch ohne Mathematik liegt es nahe, wie eine Lösung aussehen könnte: Gab es beim Ziehen  $k$  Mal die Farbe „rot“, so sollte doch  $r$  in etwa gleich  $2k$  sein, denn es ist zu erwarten, dass der Anteil der roten Karten in der Stichprobe gleich der Anzahl der roten Karten im Kartenspiel ist. Das entspricht übrigens genau der maximum-likelihood-Schätzung, die wir in Abschnitt 3.5 behandelt haben.

Doch lässt sich irgendwie begründen, dass das die beste Möglichkeit ist,  $r$  zu schätzen?

*Beispiel 2:* In einem geschlossenen Kasten befinden sich  $n_0$  Zettel, diese Zahl ist uns aber nicht bekannt. Auf dem ersten steht „1“, auf dem zweiten „2“, usw. Nun wird mit Zurücklegen 10 Mal gezogen, und die auf dem gezogenen Zettel stehende Zahl wird notiert. Mal angenommen, man hat als Ergebnis die Werte

$$44, 3, 56, 31, 84, 11, 31, 66, 81, 29$$

erhalten. Wie groß wird dann wohl  $n_0$  gewesen sein? Sicher mindestens 84, wahrscheinlich aber größer. Allerdings wohl auch nicht viel größer ...

Wie könnte eine optimale Schätzung aussehen? Im nächsten Abschnitt wird das ausführlich diskutiert werden.

*Beispiel 3:* Wir kommen auf die *Reißzwecke* zurück, die schon in Abschnitt 8.2 auf Seite 231 auftrat: Mit welcher Wahrscheinlichkeit  $p_0$  wird die Spitze nach dem Werfen nach oben zeigen? Wir dürfen die Reißzwecke 1000 Mal werfen, und wenn sie 635 Mal auf dem Rücken gelandet ist, ist es naheliegend,  $p_0$  durch  $635/1000$  zu schätzen. Inwiefern ist das eine gute Schätzung?

*Beispiel 4:*  $[0, 1]$  sei mit der Gleichverteilung versehen, und darauf betrachten wir die durch  $X_n(x) := x^n$  definierten Zufallsvariablen ( $n \in \mathbb{N}$ ). Ein  $n_0$  wird ausgewählt, doch wir wissen nicht, welches es ist.  $X_{n_0}$  wird 5 Mal abgefragt, und wir sollen schätzen, wie groß der Erwartungswert von dieser Zufallsvariable ist. Wenn man  $n_0$  kennt, kann man den Erwartungswert leicht ausrechnen:  $\mathbb{E}(X_{n_0}) = \int_0^1 x^{n_0} dx = 1/(n_0 + 1)$ . Doch wie soll man diesen Wert aufgrund der Abfragen schätzen?

Die wesentlichen Aspekte in allen diesen Beispielen sind die folgenden:

- Man hat gewisse Vorinformationen: Der Wahrscheinlichkeitsraum, der gleich eine wesentliche Rolle spielen wird, ist zwar unbekannt, man weiß aber, dass er zu einer bekannten Familie von Wahrscheinlichkeitsräumen gehört. (Im zweiten Beispiel etwa ging es um die Gleichverteilungen auf  $\{1, \dots, n_0\}$ , im vierten um die durch die  $X_n$  induzierten Wahrscheinlichkeitsräume.)
- Zu jedem der Wahrscheinlichkeitsräume gehört eine Zahl  $z$ , an der wir interessiert sind: Die Zahl  $r$  in Beispiel 1, die Zahl  $n_0$  in Beispiel 2,  $p_0$  in Beispiel 3, der Erwartungswert von  $X_n$  in Beispiel 4.
- Es dürfen  $n$  Zufallsexperimente gemacht werden. Und aufgrund des Ergebnisses soll dasjenige  $z$  geschätzt werden, das zu dem gerade ausgewählten Raum gehört.

*Die Präzisierung*

Es wird nun etwas abstrakter. In ausreichender Allgemeinheit kann die Ausgangssituation so beschrieben werden: Es sind eine Menge  $\Omega$  und eine  $\sigma$ -Algebra  $\mathcal{E}$  vorgegeben. Weiter gibt es eine Familie  $\mathbb{P}_\theta$  von Wahrscheinlichkeitsmaßen auf  $(\Omega, \mathcal{E})$ , wobei  $\theta$  zu einer Menge  $\Theta$  gehört. Eines der möglichen  $\mathbb{P}_\theta$ , das Maß  $\mathbb{P}_{\theta_0}$ , wird ausgewählt, auf diese Weise entsteht ein Wahrscheinlichkeitsraum  $(\Omega, \mathcal{E}, \mathbb{P}_{\theta_0})$ . Wir können diesen Raum nun  $n$  Mal abfragen, als Ergebnis erhalten wir die Stichprobe  $x_1, \dots, x_n$ . Und mit Hilfe dieser Werte sollen nun Aussagen über  $\theta_0$  gefunden werden. Meist ist es das Ziel,  $\theta_0$  möglichst genau zu schätzen, in manchen Fällen interessiert aber nur  $\gamma(\theta_0)$ , wobei  $\gamma : \Theta \rightarrow \mathbb{R}$  eine Funktion ist<sup>1)</sup>. Formal besteht die Aufgabe also darin, eine „geeignete“ Abbildung  $S : \Omega^n \rightarrow \mathbb{R}$  zu finden:  $S(x_1, \dots, x_n)$  ist unsere Schätzung für  $\gamma(\theta_0)$ . Was dabei „geeignet“ heißen soll, wird bald genauer erläutert werden. Im Prinzip gibt es zunächst keine Einschränkungen für  $S$ .

Bevor wir das in der nächsten Definition präzisieren, folgen einige Beispiele und Erläuterungen:

**1.** Es könnte – wie beim vorstehenden Reißzwecken-Beispiel – darum gehen, eine unbekannte Wahrscheinlichkeit  $p_0$  zu schätzen. Das modellieren wir dadurch, dass auf  $\Omega = \{0, 1\}$  alle BernoulliVerteilungen zugelassen sind, d.h. alle  $\mathbb{P}_p$  für beliebige Erfolgswahrscheinlichkeiten  $p \in [0, 1]$ . In diesem Fall ist  $\Theta = [0, 1]$ .

$\Omega$  ist mit  $\mathbb{P}_{p_0}$  versehen, und wir fragen  $n$  Mal ab. Es wird sich eine Folge  $x_1, \dots, x_n$  aus Nullen (= Misserfolg) und Einsen (= Erfolg) ergeben, und aufgrund dieses Ergebnisses wollen wir  $p_0$  schätzen. In diesem Fall geht es also um eine Abbildung  $S$  von  $\{0, 1\}^n$  nach  $[0, 1]$ .

Ein offensichtlicher Kandidat für einen Schätzer ist die Abbildung

$$S(x_1, \dots, x_n) := \frac{x_1 + \dots + x_n}{n},$$

also die relative Anzahl der Erfolge: Wenn es zum Beispiel bei 42 von 100 Bernoulliexperimenten einen Erfolg gab, so sollte 0.42 eine gute Schätzung von  $p_0$  sein.

---

<sup>1)</sup> $\gamma$  kann auch vektorwertig sein, doch hier werden wir uns auf reellwertige  $\gamma$  konzentrieren.

**2.** Das obige Beispiel 1 mit dem Kartenspiel wird ähnlich modelliert. Es geht um Bernoulliexperimente zur Wahrscheinlichkeit  $p_\theta$ , wobei  $\theta \in \Theta := \{0, 1, \dots, 20\}$  und  $p_\theta := \theta/20$ .

**3.** Zur Modellierung des vorstehenden Beispiels 2 setzt man  $\Omega := \mathbb{N}$ , und für  $n \in \Theta := \mathbb{N}$  ist  $\mathbb{P}_n$  die Gleichverteilung auf  $\{1, \dots, n\}$ . In diesem Fall ist  $\gamma(n) = n$ , denn dieser Wert soll geschätzt werden.

**4.** Im obigen Beispiel mit den  $X_n$  trägt  $\Omega = [0, 1]$  die induzierten Wahrscheinlichkeiten  $\mathbb{P}_n := \mathbb{P}_{X_n}$ , es ist  $\Theta = \mathbb{N}$ , und  $\gamma : \Theta \rightarrow \mathbb{R}$  ist durch  $\gamma(n) := \mathbb{E}(X_n)$  definiert.

**5.** Wenn alle Poissonverteilungen auf  $\mathbb{N}_0$  zugelassen sind, ist  $\Theta = [0, +\infty[$ , und zu  $\lambda \in \Theta$  ist  $\mathbb{P}_\lambda$  die Poissonverteilung zum Parameter  $\lambda$ . Das könnte dann ein sinnvolles Modell sein, wenn es um die Anzahl der Druckfehler auf einer zufällig ausgewählten Seite eines Buches geht.

Es ist  $\Omega = \{0, 1, 2, \dots\}$ , und dieses  $\Omega$  wird mit  $\mathbb{P}_{\lambda_0}$  versehen:  $\lambda_0$  ist unbekannt. Wenn die Abfragen die Werte  $x_1, \dots, x_n \in \Omega$  ergeben, soll damit  $\lambda_0$  geschätzt werden. (Im Buch-Beispiel würde man auf  $n$  unabhängig gewählten Seiten die Anzahl der Druckfehler zählen.)

**6.** Möchte man alle Normalverteilungen  $N(a, \sigma^2)$  betrachten, wird  $\Theta$  aus der Menge der möglichen  $(a, \sigma)$  bestehen:  $\Theta = \mathbb{R} \times ]0, +\infty[$ .

Falls man nur an  $a$  interessiert ist, ist  $\gamma$  durch  $\gamma(a, \sigma) := a$  zu definieren.

**7.** Die Wahl des Modells stellt eine wichtige Entscheidung dar. Dabei gehen Vor-Informationen über die Situation ein. Außerdem ist ein Kompromiss zu finden: zwischen einer möglichst angemessenen Beschreibung (= detailgetreues Modell) und mathematischer Einfachheit (= einfaches Modell). In Beispiel 1 und Beispiel 2 ist es naheliegend, mit Bernoulliverteilungen zu arbeiten, in den Beispielen 5 und 6 wäre zu überlegen, ob die Annahme gerechtfertigt ist, dass es sich – wenigstens in guter Näherung – im konkreten Fall um Poissonverteilungen bzw. um Normalverteilungen handelt. Dabei werden die Ergebnisse aus den Abschnitten 5.3 (Poissonverteilung als „Überlagerung“ von vielen Bernoulliverteilungen mit kleiner Erfolgswahrscheinlichkeit) und 8.4 (zentraler Grenzwertsatz) eine wichtige Rolle spielen.

**8.** Am Beginn dieses Abschnitts war von „ $n$  Abfragen aus  $(\Omega, \mathcal{E}, \mathbb{P}_{\theta_0})$ “ die Rede. Gemeint sind natürlich *unabhängige* Abfragen. Mit Hilfe von Satz 4.5.3 kann man das ganz präzise formulieren: Eine aus  $n$  Elementen bestehende Stichprobe entspricht einer einzigen Abfrage aus dem Wahrscheinlichkeitsraum  $(\Omega^n, \mathcal{E}_n, (\mathbb{P}_{\theta_0})^n)$ .

Denkt man noch an die Messbarkeit der auftretenden Abbildungen (um später induzierte Wahrscheinlichkeiten, Erwartungswerte usw. ausrechnen zu können), so gelangt man zu

**Definition 10.1.1.** Ein statistisches Modell besteht aus einer Familie von Wahrscheinlichkeitsräumen  $(\Omega, \mathcal{E}, \mathbb{P}_\theta)_{\theta \in \Theta}$  und einer Abbildung  $\gamma : \Theta \rightarrow \mathbb{R}$ .

Eine Schätzfunktion (für Stichproben von  $n$  Elementen) ist eine messbare Abbildung  $S : \Omega^n \rightarrow \mathbb{R}$ , so dass  $\{S \in B\} \in \mathcal{E}_n$  für alle Borelmengen  $B$  gilt<sup>2)</sup>.

Ist bei beliebigem  $\theta_0 \in \Theta$  der Messraum  $(\Omega^n, \mathcal{E}_n)$  mit dem Wahrscheinlichkeitsmaß  $(\mathbb{P}_{\theta_0})^n$  versehen, so wird  $S$  als Schätzer für  $\gamma(\theta_0)$  interpretiert.

Wir werden diesen Ansatz in den nächsten Abschnitten ausführlich diskutieren. Hier gibt es nur ein – leider recht unrealistisches – Beispiel das zeigt, dass exaktes Schätzen theoretisch durchaus möglich ist.

Für dieses Beispiel ist  $\Omega = \{1, 2, 3, 4\}$ ,  $\Theta = \{1, 2\}$ , und  $\mathbb{P}_1$  bzw.  $\mathbb{P}_2$  sind durch

$$\mathbb{P}_1(\{1\}) := \mathbb{P}_1(\{2\}) := 0.5, \quad \mathbb{P}_1(\{3\}) := \mathbb{P}_1(\{4\}) := 0$$

$$\mathbb{P}_2(\{1\}) := \mathbb{P}_2(\{2\}) := 0, \quad \mathbb{P}_2(\{3\}) := \mathbb{P}_2(\{4\}) := 0.5$$

definiert. Wenn dann  $\Omega$  mit  $\mathbb{P}_\theta$  versehen ist, ist schon nach einer einzigen Abfrage klar, um welches  $\theta$  es sich handelt: Im Fall  $\theta = 1$  sind nur die Werte 1 und 2 als Ergebnis möglich, und im Fall  $\theta = 2$  nur die Werte 3 und 4.

Im Allgemeinen wird sich absolute Sicherheit nicht erreichen lassen. Wie man die verbleibende Ungewissheit misst und worauf man beim Schätzen Wert legen sollte, wird im nächsten Abschnitt diskutiert.

## 10.2 Güteeigenschaften für Schätzer

In Definition 10.1.1 konnte  $S$  eine beliebige Funktion sein. Wir wollen nun präzisieren, was wir unter einem „guten“ Schätzer verstehen wollen.

### *Erwartungstreue Schätzer*

Mindestens sollte es doch so sein, dass man „im Mittel“ das richtige Ergebnis schätzt:

**Definition 10.2.1.** Es sei  $S$  eine Schätzfunktion für den Aspekt  $\gamma$  wie in Definition 10.1.1.  $S$  heißt erwartungstreu, wenn gilt:

Für jedes  $\theta \in \Theta$  existiert der Erwartungswert von  $S$ , wenn man  $S$  als Zufallsvariable auf dem Raum  $(\Omega^n, \mathcal{E}_n, (\mathbb{P}_\theta)^n)$  auffasst, und dieser Erwartungswert – wir werden ihn mit  $\mathbb{E}_{\mathbb{P}_\theta^n}(S)$  bezeichnen – ist gleich  $\gamma(\theta)$ .

Wir werden es im Folgenden fast nur mit erwartungstreuen Schätzfunktionen zu tun haben. Es ist jedoch offensichtlich, dass es solche  $S$  nicht in allen Fällen geben muss:

Wenn  $\Omega$  endlich ist, ist jedes  $S : \Omega^n \rightarrow \mathbb{R}$  beschränkt, und deswegen bilden auch die Zahlen  $\mathbb{E}_{\mathbb{P}_\theta^n}(S)$ ,  $\theta \in \Theta$ , eine beschränkte Menge. Ist  $\gamma$  unbeschränkt, kann es folglich keine erwartungstreuen Schätzer geben.

---

<sup>2)</sup> $S$  ist also eine auf  $(\Omega^n, \mathcal{E}_n)$  definierte Zufallsvariable.

Als konkretes Beispiel könnte man eine Urne betrachten, in der 10 rote und  $w$  weiße Kugeln enthalten sind. Dabei ist  $w \in \mathbb{N}$  unbekannt. Es wird 5 Mal mit Zurücklegen gezogen, und  $w$  soll aufgrund dieser Stichprobe geschätzt werden. Hier ist  $\Theta = \mathbb{N}$ ,  $\Omega = \{0, \dots, 5\}$  (die Anzahl der roten Kugeln in der Stichprobe), und die  $\mathbb{P}_w$  ergeben sich mit Hilfe der hypergeometrischen Verteilung:

$$\mathbb{P}_w(\{i\}) = \frac{\binom{10}{i} \binom{w}{5-i}}{\binom{10+w}{5}}$$

für  $i = 0, \dots, 5$ . Die Funktion  $\gamma$ , definiert durch  $\gamma(w) := w$ , lässt sich sicher nicht erwartungstreue schätzen.

Im Fall endlicher  $\Omega$  kann man das Problem, erwartungstreue Schätzer zu finden, in die Frage nach der Lösbarkeit eines linearen Gleichungssystems übersetzen. Sei etwa  $\Omega = \{1, \dots, r\}$ , und die Maße  $\mathbb{P}_\theta$  seien jeweils durch die  $r$  Zahlen  $p_1^\theta, \dots, p_r^\theta$  definiert (für  $\theta \in \Theta$ ). Wir konzentrieren uns auf den Fall  $n = 1$  und kürzen  $S(i)$  für  $i \in \Omega$  durch  $S_i$  ab. Dann heißt doch Erwartungstreue, dass  $p_1^\theta S_1 + \dots + p_r^\theta S_r = \gamma(\theta)$  für alle  $\theta$  gilt. Das sind so viele lineare Gleichungen für die  $S_1, \dots, S_r$ , wie  $\Theta$  Elemente hat. Deswegen ist es klar, dass es Situationen geben wird, in denen man kein, genau ein oder auch unendlich viele erwartungstreue  $S$  finden kann.

Als Beispiel betrachten wir  $\Omega = \{1, 2, 3\}$ ,  $\Theta = \{1, 2\}$  und die durch

$$\mathbb{P}_1(\{1\}) := \mathbb{P}_1(\{2\}) := \mathbb{P}_1(\{3\}) := \frac{1}{3},$$

$$\mathbb{P}_2(\{1\}) := \frac{1}{2}, \quad \mathbb{P}_2(\{2\}) := \mathbb{P}_2(\{3\}) := \frac{1}{4}$$

definierten Wahrscheinlichkeitsmaße  $\mathbb{P}_1, \mathbb{P}_2$ . Wenn  $\gamma(i) := i$  für  $i = 1, 2$  ist, so muss das Gleichungssystem

$$\frac{1}{3}S_1 + \frac{1}{3}S_2 + \frac{1}{3}S_3 = 1, \quad \frac{1}{2}S_1 + \frac{1}{4}S_2 + \frac{1}{4}S_3 = 2$$

gelöst werden. Für beliebige  $t \in \mathbb{R}$  erfüllen  $S_1 = 5, S_2 = t, S_3 = -2 - t$  diese Gleichungen. Insbesondere heißt das (wenn  $t$  „groß“ ist), dass erwartungstreue Schätzer sehr weit daneben liegen können. Im Mittel gleichen sich die Abweichungen aber aus. Auch sollte man doch erwarten, dass wir in diesem Fall auch Schätzer finden, die nur die Werte 1 und 2 annehmen, denn nur die können als mögliche  $\gamma(\theta)$  auftreten. Es gibt aber keinen einzigen erwartungstreuen Schätzer dieses Typs!

Fazit: Es ist in vielen Fällen nicht einmal leicht, die Minimalforderung der Erwartungstreue zu realisieren. Es gibt aber ein bemerkenswertes, sehr allgemeines, positives Ergebnis:

**Satz 10.2.2.** (i) Sei  $X$  eine reellwertige Zufallsvariable, für die der Erwartungswert existiert. Das Stichprobenmittel schätzt diesen Erwartungswert erwartungstreue<sup>3)</sup>.

---

<sup>3)</sup>Da das nur etwas gekünstelt im Rahmen der statistischen Modelle präzisiert werden kann, soll die Aussage noch einmal anders direkt formuliert werden: Sind  $X_1, \dots, X_n$  unabhängige Kopien von  $X$ , so ist der Erwartungswert von  $(X_1 + \dots + X_n)/n$  gleich  $\mathbb{E}(X)$ .

(ii) Ist  $X$  eine reellwertige Zufallsvariable, für die Erwartungswert und Varianz existieren, so ist die Stichprobenvarianz  $V_x$  ein erwartungstreuer Schätzer für  $\sigma^2(X) (= V(X))$ .

**Beweis:** (i) Es wurde schon bemerkt, dass die Behauptung auf

$$\mathbb{E}\left(\frac{X_1 + \dots + X_n}{n}\right) = \mathbb{E}(X)$$

hinausläuft. Das ist aber wegen der Linearität des Erwartungswerts und  $\mathbb{E}(X_i) = \mathbb{E}(X)$  (alle  $i$ ) klar.

(ii) Wir setzen zur Abkürzung  $\sigma^2 = \sigma^2(X)$ . Die Behauptung ist eine Umformulierung der folgenden Aussage:

Sind  $X_1, \dots, X_n$  unabhängige Kopien einer Zufallsvariablen  $X$  und bezeichnet man mit  $\bar{X}$  die Zufallsvariable  $(X_1 + \dots + X_n)/n$ , so gilt

$$\mathbb{E}\left(\frac{1}{n-1} \sum_i (X_i - \bar{X})^2\right) = \sigma^2.$$

Zur Vorbereitung des Beweises bemerken wir, dass

$$\begin{aligned}\mathbb{E}[(X_i - \mathbb{E}(X))(\bar{X} - \mathbb{E}(X))] &= \sigma^2/n, \\ \mathbb{E}(\bar{X} - \mathbb{E}(X))^2 &= \sigma^2/n;\end{aligned}$$

beides liegt an der Definition von  $\sigma^2$  und der Tatsache, dass der Erwartungswert für unabhängige Zufallsvariable multiplikativ ist.

Nun können wir so rechnen:

$$\begin{aligned}\mathbb{E}\left(\frac{1}{n-1} \sum_i (X_i - \bar{X})^2\right) &= \mathbb{E}\left(\frac{1}{n-1} \sum_i [(X_i - \mathbb{E}(X)) - (\bar{X} - \mathbb{E}(X))]^2\right) \\ &= \mathbb{E}\left(\frac{1}{n-1} \sum_i [(X_i - \mathbb{E}(X))^2 + \right. \\ &\quad \left. - 2(X_i - \mathbb{E}(X))(\bar{X} - \mathbb{E}(X)) + (\bar{X} - \mathbb{E}(X))^2]\right) \\ &= \frac{1}{n-1} \sum_i [\sigma^2 - \frac{2}{n}\sigma^2 + \frac{1}{n}\sigma^2] \\ &= \sigma^2.\end{aligned}$$

□

**Warum  $1/(n-1)$  und nicht  $1/n$  bei der Stichprobenvarianz?**

Es war für manche Leser sicher überraschend, dass bei der Definition von  $V_x$  in Abschnitt 9.3 im Nenner die Zahl  $n - 1$  und nicht, wie eigentlich plausibel wäre, die Zahl  $n$  steht. Der vorstehende Satz liefert die Begründung: Nur bei der Wahl von  $n - 1$  ergibt sich ein erwartungstreuer Schätzer für  $\sigma^2(X)$ .

Mit einiger Überlegung kann man das auch plausibel finden:

- $\sigma^2(X)$  ist der Erwartungswert von  $(X - \mathbb{E}(X))^2$ , also aufgrund von Teil (i) des vorigen Satzes der Erwartungswert von

$$\sum_i (x_i - \mathbb{E}(X))^2 / n,$$

wenn die  $x_1, \dots, x_n$  unabhängige Abfragen von  $X$  sind.

- Die Zahl  $\bar{x}$  liegt nach Definition immer irgendwo zwischen den Punkten  $x_1, \dots, x_n$ . Für  $\mathbb{E}(X)$  muss das aber nicht stimmen, es könnten etwa (allerdings mit geringer Wahrscheinlichkeit) alle  $x_i$  links von  $\mathbb{E}(X)$  liegen.
- Deswegen ist die Zahl  $\sum_i (x_i - \mathbb{E}(X))^2$  tendenziell *größer* als  $\sum_i (x_i - \bar{x})^2$ .

Wenn man also  $\sum_i (x_i - \bar{x})^2$  durch  $n$  teilt, so ist ein Wert zu erwarten, der *kleiner* als  $\sigma^2(X)$  ist. Und deswegen ist es plausibel, dass der Nenner etwas kleiner als  $n$  sein sollte. (Dass  $n - 1$  die richtige Wahl ist, wird durch diese Betrachtung allerdings noch nicht klar.)

Die Beweisidee von Teil (i) des Satzes lässt sich verwenden, um erwartungstreue Schätzer in unseren vorstehenden Beispielen zu erhalten. Betrachten wir etwa das Reißzwecken-Problem. Wir behaupten: Sind  $\lambda_1, \dots, \lambda_n \geq 0$  und  $\sum_i \lambda_i = 1$ , so ist  $S(x_1, \dots, x_n) := \lambda_1 x_1 + \dots + \lambda_n x_n$  ein erwartungstreuer Schätzer. Das folgt daraus, dass  $\mathbb{E}(\lambda_1 X_1 + \dots + \lambda_n X_n) = \mathbb{E}(X) = p_0$  gilt, wenn die  $X_i$  unabhängige Kopien der Bernoullivariablen  $X$  sind, die zu einem Bernoulliraum mit Erfolgswahrscheinlichkeit  $p_0$  gehört. So sind zum Beispiel  $x_1, (2x_1 + 7x_3)/9$  und  $(x_1 + \dots + x_n)/n$  erwartungstreu.

Das dritte Beispiel scheint aber „irgendwie besser“ als die beiden ersten zu sein, da nur dort alle  $x_i$  gleichberechtigt berücksichtigt werden.

**Ein weiteres Beispiel: Schätzen des Maximalwerts**

Auch für Nichtmathematiker ist es plausibel, dass eine unbekannte Wahrscheinlichkeit durch die relative Erfolgsanzahl geschätzt wird. Im Allgemeinen ist es aber völlig unklar, wie man eine Schätzfunktion definieren sollte, und es ist offen, welches die optimale Wahl ist, wenn man verschiedene Kandidaten gefunden hat.

Ein solcher Fall soll nun ausführlich diskutiert werden<sup>4)</sup>.

Für eine unbekannte Zahl  $a > 0$  betrachten wir die Gleichverteilung auf  $[0, a]$ . Es werden  $n$  unabhängige Abfragen  $x_1, \dots, x_n$  durchgeführt, und anschließend soll  $a$  geschätzt werden.

*Vorschlag 1:* Bei diesem Vorschlag argumentieren wir mit dem Erwartungswert. Eine in  $[0, a]$  gleichverteilte Zufallsvariable hat doch den Erwartungswert  $a/2$ . Das Stichprobenmittel schätzt den Erwartungswert erwartungstreu, folglich sollte  $(x_1 + \dots + x_n)/n$  ein guter Schätzer für  $a/2$  sein. Da wir an  $a$  interessiert sind, lautet unser Vorschlag: Schätze durch

$$S_1(x_1, \dots, x_n) := \frac{2}{n}(x_1 + \dots + x_n).$$

$S_1$  ist dann ein erwartungstreuer Schätzer für  $a$ .

*Vorschlag 2:* Diesmal argumentieren wir anders, wir analysieren das Verhalten von  $\max\{X_1, \dots, X_n\}$  von unabhängigen Zufallsvariablen  $X_i$ , die alle auf  $[0, a]$  gleichverteilt sind. Die Wahrscheinlichkeit, dass für alle  $i$  die Ungleichung  $X_i \leq a - \varepsilon$  gilt, ist doch  $\left(\frac{a-\varepsilon}{a}\right)^n$ , und das geht für  $n \rightarrow \infty$  gegen Null. Folglich wird der Wert

$$\tilde{S}_2(x_1, \dots, x_n) = \max\{x_1, \dots, x_n\}$$

mit hoher Wahrscheinlichkeit nahe bei  $a$  liegen, das könnte ebenfalls ein guter Kandidat für eine Schätzfunktion sein.

Ist  $\tilde{S}_2$  erwartungstreu? Dazu erinnern wir an einige Tatsachen aus den vorigen Kapiteln:

- Sind  $X_1, \dots, X_n$  unabhängig, so ist

$$\mathbb{P}(\max\{X_1, \dots, X_n\} \leq x) = \prod_i \mathbb{P}\{X_i \leq x\}.$$

- Sind insbesondere die  $X_i$  unabhängige Abfragen der Gleichverteilung auf  $[0, a]$ , so ist die Wahrscheinlichkeit, dass das Maximum  $\leq x$  ist, durch  $(x/a)^n$  gegeben.
  - Ist  $\mu$  ein Wahrscheinlichkeitsmaß auf  $\mathbb{R}$  mit einer Dichte  $f$  und kennt man die Funktion  $x \mapsto \mu([-\infty, x])$ , so ist  $f$  die Ableitung dieser Funktion.
- Im vorliegenden Fall heißt das: Das Maximum aus  $n$  Gleichverteilungen auf  $[0, a]$  hat eine Dichtefunktion, nämlich die Funktion  $nx^{n-1}/a^n$  (definiert auf  $[0, a]$ ).
- Hat ein Wahrscheinlichkeitsmaß auf  $\mathbb{R}$  eine Dichte  $f$ , so ist der Erwartungswert der Identität durch  $\int xf(x)dx$  gegeben.

---

<sup>4)</sup>Das Beispiel – es ist eine Variante von Beispiel 2 auf Seite 288 – habe ich im Buch von Georgii gefunden. Es eignet sich sehr gut für die Illustration der mit Schätzfunktionen auftretenden Probleme.

Und so folgt: Das Maximum aus  $n$  Gleichverteilungen auf  $[0, a]$  hat den Erwartungswert

$$\int_0^a \frac{nxx^{n-1}}{a^n} dx = \frac{n}{n+1}a.$$

Damit sieht man, dass  $\tilde{S}_2$  nicht erwartungstreu schätzt<sup>5)</sup>. Das ist aber leicht zu reparieren, man muss nur mit  $(n+1)/n$  multiplizieren. Wenn wir also

$$S_2(x_1, \dots, x_n) := \frac{n+1}{n} \max\{x_1, \dots, x_n\}$$

setzen, so ist auch  $S_2$  ein erwartungstreuer Schätzer.

Mit  $S_1$  und  $S_2$  sind offensichtlich auch alle Funktionen  $\lambda S_1 + (1-\lambda)S_2$  (für  $\lambda$  in  $[0, 1]$ ) erwartungstreu, und es lassen sich auch noch viele weitere Kandidaten finden. Doch welcher dieser Schätzer ist „bestmöglich“? Wir werden gleich (in Definition 10.2.3) präzisieren, was wir darunter verstehen wollen.

### Konsistente Schätzfolgen

Ein weiterer Begriff, mit dem die Güte von Schätzverfahren beurteilt wird, ist die *Konsistenz*. Man möchte ausdrücken, dass die Schätzung „immer besser“ wird, wenn die Stichprobengröße wächst. Genauer: Für jedes  $n$  soll eine Schätzfunktion  $S_n$  definiert sein, die aus einer Stichprobe  $x_1, \dots, x_n$  eine Schätzung für  $\gamma(\theta_0)$  liefert, wenn die Stichprobe gemäß  $\mathbb{P}_{\theta_0^n}$  erzeugt wurde. Die Schätzfolge  $(S_n)$  heißt dann *konsistent*, wenn sie in Wahrscheinlichkeit gegen die konstante Funktion  $\gamma(\theta_0)$  geht. Anders ausgedrückt: Für „große“  $n$  ist  $S_n(x_1, \dots, x_n)$  mit hoher Wahrscheinlichkeit „nahe bei“  $\gamma(\theta_0)$ .

Wenn man das wirklich präzise formulieren möchte, müssen die  $S_n$  alle auf dem gleichen Raum definiert sein. Man wählt  $\Omega_\infty$ , den Raum der Folgen in  $\Omega$  und darauf das unendliche Produkt  $\mathbb{P}_{\theta_0}^\infty$ ; das entspricht einer Verallgemeinerung von Satz 4.5.3 auf unendlich viele Faktoren. Die  $S_n$  sind zwar auf  $\Omega_\infty$  definiert,  $S_n$  darf aber nur von den ersten  $n$  Folgengliedern abhängen: Ist  $x_k = y_k$  für  $k \leq n$ , so ist  $S_n(x_1, \dots) = S_n(y_1, \dots)$ . Und für eine solche Folge  $(S_n)$  kann dann sinnvoll definiert werden, was Konsistenz bedeutet.

Konsistenz ist eher von theoretischem Interesse, auch wird der Begriff im Folgenden keine wichtige Rolle spielen. Es soll nur noch bemerkt werden, dass das schwache Gesetz der großen Zahlen gerade besagt, dass die Stichprobenmittel  $(x_1 + \dots + x_n)/n$  eine konsistente Schätzfolge für den Erwartungswert einer Zufallsvariablen bilden.

### Schätzer mit gleichmäßig kleinster Varianz

Erwartungstreue kann sicher nicht das einzige Gütekriterium für Schätzfunktionen  $S$  sein. Dann ist zwar  $S(x_1, \dots, x_n)$  im Mittel gleich  $\gamma(\theta_0)$ , es ist aber

---

<sup>5)</sup>Diese Tatsache ist nicht wirklich überraschend, denn  $\tilde{S}_2$  schätzt systematisch zu niedrig.

sicher wünschenswert, dass die Schätzwerte nicht allzu stark um  $\gamma(\theta_0)$  herumstreuen.

Ein naheliegendes Maß für die Streuung ist die *Varianz* von  $S$ . Genauer:

**Definition 10.2.3.** (i) Es sei  $S$  eine erwartungstreue Schätzfunktion gemäß Definition 10.1.1. Wir setzen voraus, dass für alle  $\theta$  die Varianz von  $S$ , aufgefasst als Zufallsvariable auf dem Raum  $(\Omega^n, \mathcal{E}_n, \mathbb{P}_\theta^n)$ , existiert. Mit  $V_S : \Theta \rightarrow \mathbb{R}$  bezeichnen wir dann die Funktion, die jedem  $\theta$  die entsprechende Varianz zuordnet.

(ii) Ein erwartungstreuer Schätzer  $S^*$  heißt Schätzer mit gleichmäßig bester Varianz, wenn gilt: Es ist  $V_{S^*}(\theta) \leq V_S(\theta)$  (alle  $\theta \in \Theta$ ) für alle erwartungstreuen Schätzer  $S$ .

Wenn  $S_1$  und  $S_2$  erwartungstreue Schätzer sind, so werden sie im Allgemeinen nicht vergleichbar sein: Für manche  $\theta$  ist  $V_{S_1}(\theta)$  kleiner als  $V_{S_2}(\theta)$ , und für andere ist es umgekehrt. Wenn es nur einen einzigen erwartungstreuen Schätzer gibt, so ist das natürlich ein Schätzer mit gleichmäßig bester Varianz. Doch wie sieht es aus, wenn mehrere gefunden werden können? Im Allgemeinen wird es dann kein  $S^*$  geben. Allerdings lässt sich zeigen, dass man das Schätzproblem für viele wichtige Fälle optimal lösen kann. Diese Frage greifen wir im nächsten Abschnitt noch einmal auf.

Hier wollen wir noch einmal auf die vorstehend gefundenen *Schätzervorschläge für den Maximalwert* eingehen. Was ist besser: das verdoppelte Stichprobenmittel  $S_1$  oder der durch  $(n+1) \max x_i/n$  definierte Schätzer  $S_2$ ? Dazu müssen wir ein bisschen rechnen<sup>6)</sup>:

*Die Varianzen für  $S_1$ :* Die Varianz der Gleichverteilung auf  $[0, a]$  ist bekanntlich gleich  $a^2/12$ . Der Mittelwert aus  $n$  Abfragen führt zu einem  $n$  im Nenner, und da mit 2 multipliziert wird, erhalten wir für die Varianz noch einen Faktor 4. Fasst man alles zusammen, so ergibt sich für die Varianz von  $S_1$  der Wert  $\frac{a^2}{3n}$ .

*Die Varianzen für  $S_2$ :* Wir kümmern uns zunächst um  $\tilde{S}_2$ , also das Maximum der Stichprobe. Nun gilt für eine Zufallsvariable  $X$ :

- Es ist  $V(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$ .
- Hat  $\mathbb{P}_X$  eine Dichtefunktion  $f$ , so ist  $\mathbb{E}(X^2) = \int_{\mathbb{R}} x^2 f(x) dx$ .

In unserem Fall heißt das: Die Varianz des Maximums aus  $n$  Gleichverteilungen auf  $[0, a]$  ist gleich

$$\int_0^a n x^2 x^{n-1} a^{-n} dx - \left( \frac{na}{n+1} \right)^2 = \frac{na^2}{(n+1)^2(n+2)}.$$

$\tilde{S}_2$  hat also eine kleinere Varianz als  $S_1$ , allerdings schwanken die Werte noch um den falschen Wert. Macht man noch die  $(n+1)/n$ -Modifikation, so ist die

---

<sup>6)</sup>Die hier verwendeten Tatsachen sind in Abschnitt 3.3 bewiesen worden.

eben berechnete Zahl mit dem Quadrat von  $(n+1)/n$  zu multiplizieren, es ergibt sich die Varianz

$$V_{S_2}(a) = \frac{(n+1)^2}{n^2} \frac{na^2}{(n+1)^2(n+2)} = \frac{a^2}{n(n+2)}.$$

Es folgt, dass  $S_2$  für alle  $a$  ein besserer Schätzer ist als  $S_1$ . Doch ist  $S_2$  schon bestmöglich? Diese Frage wird im nächsten Abschnitt beantwortet werden.

### 10.3 Beispiele für Punktschätzer

Wir gehen wieder von einem statistischen Modell wie zu Beginn von Abschnitt 10.1 aus. Es wurde schon erwähnt, dass es im Allgemeinen keine erwartungstreuen Schätzer geben muss und dass, selbst wenn welche existieren, es sein kann, dass kein Schätzer mit gleichmäßig bester Varianz dabei ist.

Es gibt aber eine Reihe von Beispielen, in denen man solche optimalen Schätzfunktionen explizit angeben kann. Die wichtigsten können mit dem *Satz von Lehmann-Scheffé* gefunden werden. Die exakte Formulierung und erst recht der Beweis dieses Satzes sind Gegenstand weiterführender Vorlesungen. Die folgenden Begriffe spielen dabei eine wichtige Rolle.

**1.** Eine Schätzfunktion ist doch eine Vorschrift, aus einer Stichprobe  $x_1, \dots, x_n$  eine Zahl – manchmal auch einen Vektor – zu berechnen. Formal handelt es sich dabei um eine Abbildung von  $\Omega^n$  nach  $\mathbb{R}$ . Manchmal ist es dazu notwendig, die Daten „vorzubehandeln“: Sie können zum Beispiel sortiert oder zusammengefasst werden. Anders formuliert heißt das, dass  $S$  als  $S' \circ T$  geschrieben wird, wobei  $T(x_1, \dots, x_n)$  die „vorbehandelten“ Daten sind, auf die dann nur noch  $S'$  angewandt wird.

Ein Beispiel: Beim Schätzen von Wahrscheinlichkeiten hatten wir doch  $S(x_1, \dots, x_n) = (x_1 + \dots + x_n)/n$  betrachtet. Es ist  $S = S' \circ T$  mit  $T(x_1, \dots, x_n) = x_1 + \dots + x_n$  (das ist die Anzahl der Erfolge) und  $S' : y \mapsto y/n$ .

Solche Abbildungen  $T$  nennt man eine *Statistik*. Besonders wichtig sind Statistiken, durch die „keine Informationen über die Stichprobe verloren gehen“. So sollte es für das Schätzen von Wahrscheinlichkeiten nur wesentlich sein zu wissen, wie viele Erfolge es bei  $n$  Versuchen gab. Dagegen ist es sicher nicht sinnvoll, im Fall  $n > 2$  mit  $T(x_1, \dots, x_n) := (x_1, x_2)$  zu arbeiten, also die Ergebnisse der Abfragen 3 bis  $n$  nicht zu berücksichtigen.

Geht durch  $T$  keine wesentliche Information verloren, so heißt  $T$  *suffizient*. Im diskreten Fall wird das wie folgt präzisiert: Ist  $T : \Omega^n \rightarrow D$  eine Abbildung, so heißt  $T$  *suffizient*, wenn für jeden Bildwert  $y$  die durch die  $\mathbb{P}_\theta$  auf  $T^{-1}(\{y\})$  induzierten Wahrscheinlichkeitsmaße für alle  $\theta \in \Theta$  übereinstimmen. Es soll also zu jedem  $y \in D$  mit  $T^{-1}(\{y\}) \neq \emptyset$  ein Wahrscheinlichkeitsmaß  $Q_y$  auf  $\Omega_y := T^{-1}(\{y\})$  geben, so dass für alle  $\theta$  und alle  $E \subset \Omega_y$  gilt:  $Q_y(E) = \mathbb{P}_\theta^n(E)/\mathbb{P}_\theta^n(\Omega_y)$ . Diese recht technische Definition leistet wirklich das Gewünschte. Zum Beispiel ist  $(x_1, \dots, x_n) \mapsto x_1 + \dots + x_n$  suffizient,  $(x_1, \dots, x_n) \mapsto (x_1, x_2)$  aber nicht.

**2.** Suffizienz ist also so etwas wie eine zulässige Informationskompression: nichts Wesentliches geht verloren. Solche  $T$  gibt es sehr viele, neben  $(x_1, \dots, x_n) \mapsto x_1 + \dots + x_n$  sind zum Beispiel für das Schätzen von Wahrscheinlichkeiten auch  $(x_1, \dots, x_n) \mapsto (x_1, x_2 + \dots + x_n)$  oder  $(x_1, \dots, x_n) \mapsto (x_1, x_2, x_3 + \dots + x_n)$  suffizient. Das erste Beispiel scheint aber stärker zu komprimieren.

Eine Statistik, die die Daten „sehr stark“ komprimiert, für die also die  $\Omega_y$  sehr „groß“ sind, heißt *vollständig*. Auch in diesem Fall ist die Definition sehr technisch.

Hier ist die Präzisierung für den diskreten Fall.  $T : \Omega^n \rightarrow D$  heißt *vollständig*, wenn für jede Funktion  $f : \Omega^n \rightarrow \mathbb{R}$ , die auf allen  $\Omega_y$  konstant ist und für die der Erwartungswert  $\mathbb{E}(f)$  unter allen  $\mathbb{P}_\theta^n$  gleich Null ist, folgt, dass  $f$  die Nullfunktion sein muss. Ein extremes Beispiel sind konstante Funktionen  $T$ , aber auch die Erfolgsanzahl  $(x_1, \dots, x_n) \mapsto x_1 + \dots + x_n$  ist für das Schätzen von Wahrscheinlichkeiten vollständig.

**3.** Der *Satz von Lehmann-Scheffé* besagt dann, dass man einen erwartungstreuen Schätzer  $S^*$  mit gleichmäßig kleinster Varianz angeben kann, wenn folgende Bedingungen erfüllt sind:

- Man hat einen erwartungstreuen Schätzer  $S$  gefunden.
- Es gibt eine Statistik  $T : \Omega^n \rightarrow D$ , die suffizient und vollständig ist.

$S^*$  kann dann so definiert werden:  $S^*(x_1, \dots, x_n)$  ist der Erwartungswert der Einschränkung von  $S$  auf  $\Omega_y$ ; dabei ist  $y := T(x_1, \dots, x_n)$ , und  $\Omega_y$  ist mit dem Wahrscheinlichkeitsmaß  $Q_y$  aus der Definition von „Suffizienz“ versehen.

Der theoretische Hintergrund ist sogar für den Fall diskreter Räume recht verwickelt. Für Räume mit Dichten erfordern die Definitionen von Suffizienz und Vollständigkeit weitere umfangreiche Vorbereitungen<sup>7)</sup>. Einzelheiten findet man – zum Beispiel – im Buch von Georgii, wir sammeln hier nur einige Ergebnisse:

1. Geht es um Bernoulliräume mit  $p \in [0, 1]$  (also um das Schätzen einer unbekannten Wahrscheinlichkeit in  $[0, 1]$ ), so hat der Schätzer  $\sum_{i=1}^n x_i/n$  die gleichmäßig beste Varianz. Das entspricht dem in Beispiel 1 auf Seite 289 angegebenen naheliegenden Ansatz.
2. Besteht das statistische Modell aus allen Gleichverteilungen auf den Intervallen  $[0, a]$  mit  $a > 0$  und soll  $a$  geschätzt werden, so ist der im vorigen Abschnitt diskutierte Schätzer

$$(x_1, \dots, x_n) \mapsto \frac{n+1}{n} \max\{x_1, \dots, x_n\}$$

bestmöglich: Er ist erwartungstreu und hat gleichmäßig die kleinste Varianz.

3. Den Parameter  $\lambda$  einer Poissonverteilung schätzt man am besten durch das Stichprobenmittel.

---

<sup>7)</sup>Man muss zum Beispiel wissen, was bedingte Wahrscheinlichkeitsmaße sind.

4. Die diskrete Variante von Beispiel 2 hat eine überraschende Lösung. Wir betrachten die Gleichverteilungen auf allen  $\{1, \dots, a\}$  mit  $a \in \mathbb{N}$  und wollen  $a$  schätzen. Ein optimaler, erwartungstreuer Schätzer mit gleichmäßig kleinster Varianz ist dann durch

$$S^*(x_1, \dots, x_n) := \frac{y^{n+1} - (y-1)^{n+1}}{y^n - y^{n+1}}$$

gegeben, wobei  $y$  durch  $y := \max\{x_1, \dots, x_n\}$  definiert ist.

Anders als bei den vorigen Beispielen wäre man darauf mit dem „gesunden Menschenverstand“ sicher nicht gekommen.

Im zweiten Teil dieses Abschnitts kommen wir noch einmal auf die *maximum-likelihood-Schätzungen* zurück, von denen in Kapitel 3 schon die Rede war (vgl. Seite 102). Die Definition ist – wenigstens für diskrete Wahrscheinlichkeitsräume – leicht auf den Fall statistischer Modelle zu übertragen:

Es sei  $\Omega$  endlich oder abzählbar. Ein Schätzer  $S$  heißt *maximum-likelihood-Schätzer*, wenn  $S$  so definiert ist:  $S(x_1, \dots, x_n)$  ist das Bild unter  $\gamma$  von demjenigen  $\theta$ , für das  $\mathbb{P}_\theta^n(\{x_1, \dots, x_n\})$  maximal ist.

Offensichtlich sind Beispiele denkbar, bei denen kein solches  $\theta$  existieren muss, und selbst wenn es existiert, muss es nicht eindeutig bestimmt sein. Der Ansatz ist aber naheliegend und führt in vielen Fällen zu sinnvollen Ergebnissen.

Die Idee ist nicht so ohne Weiteres auf nicht-diskrete Räume zu übertragen, denn es kann passieren, dass alle  $\mathbb{P}_\theta^n(\{x_1, \dots, x_n\})$  gleich Null sind: Wie soll man da etwas Maximales auswählen?

Man behilft sich im Fall von Räumen mit Dichten mit einer nahe liegenden Idee. Wir denken als Beispiel an das Intervall  $[0, 1]$ , das durch eine Dichtefunktion  $f$  zu einem Wahrscheinlichkeitsraum gemacht wurde. Für alle  $x$  ist dann  $\mathbb{P}(\{x\}) = 0$ , aber man hat doch den Eindruck – auch wenn das völlig unmathematisch ist –, dass die  $x$  mit großem  $f(x)$  „irgendwie wahrscheinlicher“ sind als die mit kleinem  $f(x)$ . (Denn im ersten Fall wird für kleine  $\varepsilon$  die Wahrscheinlichkeit  $\mathbb{P}([x - \varepsilon, x + \varepsilon])$  größer sein als im zweiten.) Das führt uns zu der folgenden

**Definition 10.3.1.** Vorgelegt sei ein statistisches Modell. Wir nehmen an, dass  $\Omega$  ein Intervall ist und dass für alle  $\theta \in \Theta$  das Maß  $\mathbb{P}_\theta$  durch eine auf  $\Omega$  definierte Dichtefunktion  $f_\theta$  definiert ist; wegen Satz 4.4.4 hat  $\mathbb{P}_\theta^n$  dann die durch

$$f_\theta^n : (x_1, \dots, x_n) \mapsto f_\theta(x_1) \cdots f_\theta(x_n)$$

definierte Dichtefunktion.

Ein Schätzer  $S$  heißt ein *maximum-likelihood-Schätzer*, wenn  $S(x_1, \dots, x_n)$  als dasjenige  $\gamma(\theta)$  definiert ist, für das  $f_\theta^n(x_1, \dots, x_n)$  maximal ist.

Auch hier kann es natürlich sein, dass solche  $\theta$  nicht existieren, und sie müssen auch nicht eindeutig bestimmt sein. In solchen Fällen wird es dann keinen maximum-likelihood-Schätzer geben.

Wir diskutieren einige *Beispiele*:

- Angenommen, wir wollen eine unbekannte Wahrscheinlichkeit  $p$  schätzen. Bei  $n$  unabhängigen Bernoulliexperimenten gab es  $k$  Erfolge, welches  $p$  sollte man vorschlagen?

Die Wahrscheinlichkeit für  $k$  Erfolge ist durch die Binomialverteilung gegeben, also gleich  $\binom{n}{k} p^k (1-p)^{n-k}$ . Mit elementarer Analysis (Ableitung Null setzen) folgt schnell, dass dieser Ausdruck für  $k/n$  maximal wird. *Das* ist der maximum-likelihood-Schätzer, er stimmt mit dem optimalen Schätzer überein.

- Diesmal betrachten wir Exponentialverteilungen, das unbekannte  $\lambda$  soll geschätzt werden. Die Dichtefunktion ist durch  $\lambda e^{-\lambda x}$  gegeben, zur Modellierung von  $n$  unabhängigen Abfragen muss  $[0, +\infty[^n$  also mit der Dichtefunktion  $(x_1, \dots, x_n) \mapsto \lambda^n e^{-\lambda(x_1 + \dots + x_n)}$  versehen werden. Wieder führt elementare Analysis zum Ziel: Dieser Ausdruck wird – bei gegebenen  $x_1, \dots, x_n$  und variablem  $\lambda$  – maximal für  $\lambda = n/(x_1 + \dots + x_n)$ , also für das Inverse des Stichprobenmittels. Das ist plausibel, denn der Erwartungswert einer Exponentialverteilung ist  $1/\lambda$ .

- Es sei  $\Omega = \mathbb{R}^+$ , wir betrachten alle Gleichverteilungen auf Intervallen des Typs  $[0, a]$ . Nun werde die Stichprobe  $(x_1, \dots, x_n)$  gezogen. Für festes  $a$  führen  $n$  unabhängige Abfragen der Gleichverteilung auf  $[0, a]$  zur Gleichverteilung auf  $[0, a]^n \subset \mathbb{R}^n$ . Die zugehörige Dichtefunktion – sie wird als Funktion vom  $\mathbb{R}^n$  nach  $\mathbb{R}$  aufgefasst – ist Null, falls  $\max x_i > a$  und  $1/a^n$  sonst. Folglich ist  $\max x_i$  ein Maximum-likelihood-Schätzer.

Das zeigt, dass solche Schätzer nicht erwartungstreu sein müssen.

- Wir betrachten nun alle Normalverteilungen  $N(a, \sigma^2)$  und suchen einen maximum-likelihood-Schätzer für  $(a, \sigma^2)$ , falls die Stichprobe  $(x_1, \dots, x_n)$  vorliegt. Das läuft auf das folgende analytische Problem hinaus:

Finde zu vorgegebenem Vektor  $x := (x_1, \dots, x_n) \in \mathbb{R}^n$  dasjenige Tupel  $(a, \sigma^2)$ , für das

$$f(a, v) := \frac{1}{(2\pi v)^{n/2}} \prod_i \exp[-(x_i - a)^2 / 2v] = \frac{1}{(2\pi v)^{n/2}} \exp\left(-\sum_i (x_i - a)^2 / 2v\right)$$

maximal ist; dabei ist  $v := \sigma^2$ .

Sicher muss man dazu versuchen,  $\sum (x_i - a)^2$  zu minimieren: Die Lösung kennen wir schon, es ist das Stichprobenmittel  $\bar{x}$  (vgl. Satz 9.3.3). Und nun muss noch ein optimales  $v$  gefunden werden. Das findet man durch Nullsetzen der Ableitung des Logarithmus von  $f$ , also aus der Gleichung

$$0 = -\frac{n}{2v} + \frac{1}{2v^2} \sum (x_i - \bar{x})^2.$$

Wir erhalten

$$v = \frac{1}{n} \sum_i (x_i - \bar{x})^2,$$

also bis auf den Faktor  $n/(n - 1)$  die Stichprobenvarianz<sup>8)</sup>.

## 10.4 Konfidenzbereiche

Bisher haben wir beim Schätzen immer aufgrund von  $n$  unabhängigen Abfragen  $x_1, \dots, x_n$  aus  $(\Omega, \mathcal{E}, \mathbb{P}_{\theta_0})$  einen konkreten Wert für  $\gamma(\theta_0)$  vorgeschlagen. Sollte zum Beispiel eine unbekannte Wahrscheinlichkeit geschätzt werden, so hatten wir als Schätzung den Wert „Anzahl der Erfolge bei  $n$  Versuchen, geteilt durch  $n$ “ angegeben.

In diesen Ansatz kann leicht eine Exaktheit hineininterpretiert werden, die gar nicht beabsichtigt ist und die man aufgrund der vorliegenden Informationen auch gar nicht rechtfertigen kann. Zum Beispiel glaubt doch niemand im Ernst, dass die Wahrscheinlichkeit für „Die Reißzwecke fällt auf den Rücken“ gleich 0.631 ist, wenn sie bei 1000 Versuchen 631 Mal auf dem Rücken gelandet ist.

Um dieses Problem zu vermeiden, wird eine vorsichtigere Formulierung gewählt: Statt aufgrund des Ergebnisses der  $n$  Abfragen aus  $(\Omega, \mathcal{E}, \mathbb{P}_{\theta_0})$  zu sagen „ $\gamma(\theta_0)$  ist soundso groß“, formuliert man vorsichtiger „ $\gamma(\theta_0)$  liegt in der Menge  $\Delta_{x_1, \dots, x_n}$ “, wobei  $\Delta_{x_1, \dots, x_n}$  in Abhängigkeit von den  $x_1, \dots, x_n$  angegeben wird.

Bei der Wahl dieser Mengen gibt es zwei gegensätzliche Tendenzen. Einerseits möchte man  $\Delta_{x_1, \dots, x_n}$  möglichst klein wählen, denn bei großem  $\Delta_{x_1, \dots, x_n}$  nutzt die Aussage, dass  $\gamma(\theta_0)$  darin liegt, in der Regel nicht viel. Doch andererseits ist es verführerisch  $\Delta_{x_1, \dots, x_n}$  groß zu wählen, denn dann ist das Risiko, dass die Prognose „ $\gamma(\theta_0) \in \Delta_{x_1, \dots, x_n}$ “ falsch ist, geringer.

Der bewährte Ausweg aus diesem Dilemma besteht darin, von vornherein festzulegen, mit welcher Wahrscheinlichkeit man bereit ist, eine falsche Prognose zu stellen. Man gibt dazu eine Zahl  $\alpha \in ]0, 1[$ , die *Irrtumswahrscheinlichkeit* vor, oft verwendete Werte sind 0.1, 0.05 und 0.01. Die Zahl  $\beta := 1 - \alpha$  ist dann die Wahrscheinlichkeit, mit der man auf eine richtige Voraussage hoffen kann, sie heißt das *Konfidenzniveau*. Hier die Präzision dieses Ansatzes:

**Definition 10.4.1.** Wieder betrachten wir  $(\Omega, \mathcal{E})$  zusammen mit einer Familie  $(\mathbb{P}_\theta)_{\theta \in \Theta}$  von Wahrscheinlichkeitsmaßen und eine Abbildung  $\gamma : \Theta \rightarrow \mathbb{R}$ . Weiter sei eine Irrtumswahrscheinlichkeit  $\alpha \in ]0, 1[$  vorgegeben.

Unter einer Konfidenzbereichsschätzung zum Irrtumsniveau  $\alpha$  (bzw. zum Konfidenzniveau  $\beta = 1 - \alpha$ ) verstehen wir dann eine Abbildung, die jedem  $n$ -Tupel  $(x_1, \dots, x_n)$  eine Teilmenge  $\Delta_{x_1, \dots, x_n}$  von  $\mathbb{R}$  zuordnet, so dass die folgenden Bedingungen erfüllt sind:

(i) Für jedes  $\theta_0 \in \Theta$  liegt die Menge  $\{(x_1, \dots, x_n) \mid \gamma(\theta_0) \in \Delta_{x_1, \dots, x_n}\}$  in der  $\sigma$ -Algebra  $\mathcal{E}_n$ . Es ist also sinnvoll, von der Wahrscheinlichkeit dieser Menge unter  $\mathbb{P}_{\theta_0}^n$  zu sprechen.

(ii) Für alle  $\theta_0$  gilt  $\mathbb{P}_{\theta_0}^n(\{(x_1, \dots, x_n) \mid \gamma(\theta_0) \in \Delta_{x_1, \dots, x_n}\}) \geq \beta$ <sup>9)</sup>.

---

<sup>8)</sup>Auch in diesem Fall – schätzen von  $\sigma^2$  – liegt also kein erwartungstreuer Schätzer vor. Wegen Satz 10.2.2 ist  $nv/(n - 1)$  erwartungstreu,  $v$  selbst also nicht.

<sup>9)</sup>Gleichwertig dazu ist natürlich die Forderung  $\mathbb{P}(\{(x_1, \dots, x_n) \mid \gamma(\theta_0) \notin \Delta_{x_1, \dots, x_n}\}) \leq \alpha$ .

Kurz: Stets liegt  $\gamma(\theta_0)$  mit mindestens Wahrscheinlichkeit  $\beta$  in  $\Delta_{x_1, \dots, x_n}$ .

**... aber  $\gamma(\theta_0)$  ist doch ein fester Wert ...**

Ist die Bedingung in der vorstehenden Definition, dass  $\gamma(\theta_0)$  „mit mindestens Wahrscheinlichkeit  $\beta$  in  $\Delta_{x_1, \dots, x_n}$ “ liegt, überhaupt sinnvoll? Für eine feste Zahl – hier  $\gamma(\theta_0)$  – kann man doch eindeutig entscheiden, ob sie in einer Menge liegt oder nicht.

Im Allgemeinen stimmt das auch, aber hier ist die Menge, um die es geht, vom Zufall abhängig. Je nach  $n$ -Tupel  $x_1, \dots, x_n$  entsteht eine andere Menge  $\Delta_{x_1, \dots, x_n}$ . Manchmal wird  $\gamma(\theta_0)$  dazugehören, manchmal aber auch nicht. Und es wird gefordert, dass die Wahrscheinlichkeit dafür, dass der erste Fall eintritt, mindestens  $\beta$  ist.

Solche Verfahren haben wirklich die Eigenschaft, dass die Irrtumswahrscheinlichkeit kontrolliert werden kann, das ist in die Definition eingebaut. Formal sind auch sehr „feige“ Konfidenzbereichsschätzungen zugelassen: Man könnte zum Beispiel stets  $\Delta_{x_1, \dots, x_n} := \mathbb{R}$  setzen, dann ist die Irrtumswahrscheinlichkeit sogar Null. Die Konfidenzmengen sollen aber zu vorgelegten  $x_1, \dots, x_n$  möglichst klein sein, um möglichst präzise Aussagen zu erhalten. Der Einfachheit halber wählt man oft Intervalle (in  $\mathbb{N}$  oder  $\mathbb{R}$ ), man spricht dann auch von *Konfidenzintervallen*.

Wie findet man „kleine“ Konfidenzbereiche? Hier ist ein *allgemeines Verfahren*:

$\alpha$  sei vorgegeben. Wir wählen für jedes  $\theta$  ein „möglichst kleines“ Ereignis  $\Omega_{\theta, \alpha} \subset \Omega^n$ , so dass  $\mathbb{P}_{\theta}^n(\Omega_{\theta, \alpha}) \geq 1 - \alpha$ . (Im Idealfall sollte „=“ gelten, das ist im diskreten Fall aber nicht immer zu erreichen.)

Nun können leicht Konfidenzbereiche gefunden werden: Definiere, für  $(x_1, \dots, x_n) \in \Omega^n$ , die Konfidenzmenge  $\Delta_{x_1, \dots, x_n}$  durch

$$\Delta_{x_1, \dots, x_n} := \{\gamma(\theta) \mid (x_1, \dots, x_n) \in \Omega_{\theta, \alpha}\}.$$

Es ist dann durch die Definition sichergestellt, dass es sich wirklich um Konfidenzmengen handelt.

Begründung: Es sei  $\theta_0$  der „wirkliche“ Parameter. Die gemäß  $\mathbb{P}_{\theta_0}^n$  erzeugte Stichprobe  $(x_1, \dots, x_n)$  wird – nach Konstruktion von  $\Omega_{\theta_0, \alpha}$  – mit mindestens Wahrscheinlichkeit  $1 - \alpha$  in  $\Omega_{\theta_0, \alpha}$  liegen, was gerade nach Definition von  $\Delta_{x_1, \dots, x_n}$  bedeutet, dass  $\Delta_{x_1, \dots, x_n}$  mit einer Wahrscheinlichkeit von mindestens  $1 - \alpha$  den Wert  $\gamma(\theta_0)$  enthält.

Es folgen einige Beispiele dazu:

1. Wir suchen Konfidenzbereiche für eine unbekannte Wahrscheinlichkeit. Es sind also  $\alpha$  und  $n$  fixiert, und bei unbekanntem  $p_0$  werden  $n$  unabhängige Abfragen aus einer Bernoulli-verteilung durchgeführt, bei der die Erfolgswahrscheinlichkeit  $p_0$  ist.  $p_0$  soll zum Konfidenzniveau  $1 - \alpha$  geschätzt werden.

Nach dem vorstehend beschriebenen Verfahren verfährt man so:

- Sei  $p \in [0, 1]$ . Wähle ein möglichst kleines Intervall  $I_{p,\alpha}$  in  $\mathbb{N}$ , so dass die Erfolgsanzahl mit mindestens Wahrscheinlichkeit  $1 - \alpha$  in  $I_{p,\alpha}$  liegt.

Dazu kann man so vorgehen: Suche ein möglichst großes  $k'_p \in \mathbb{N}$  mit  $\sum_{k=0}^{k'_p-1} b(k, n; p) < \alpha/2$  sowie ein möglichst kleines  $k''_p$  mit  $\sum_{k=k''_p+1}^n b(k, n; p) < \alpha/2$ . Setze dann  $I_{p,\alpha} := \{k'_p, k'_p + 1, \dots, k''_p\}$ . Dann liegt die Erfolgsanzahl mit mindestens Wahrscheinlichkeit  $1 - \alpha$  in  $I_{p,\alpha}$ .

- Wenn die  $I_{p,\alpha}$  für alle  $p$  bestimmt sind, definiere  $\Delta_k$  als die Menge der  $p \in [0, 1]$  mit  $k \in I_{p,\alpha}$ . Das ist ein Intervall  $[p_u(k), p_o(k)]$ .

Falls dann die Wahrscheinlichkeit wirklich gleich  $p_0$  ist und  $k$  Erfolge bei  $n$  Versuchen beobachtet wurden, wird  $p_0$  mit mindestens Wahrscheinlichkeit  $1 - \alpha$  in  $\Delta_k$  liegen.

Man kann die  $\Delta_k$  für einige  $\alpha$  mit dem zum Buch gehörigen Computerprogramm berechnen lassen (vgl. den Anhang, Seite 366).

Hier eine typische Anwendung, dabei seien  $n = 50$  und  $\alpha = 0.02$  (also  $\beta = 0.98$ ). Ergeben sich dann 24 Erfolge, so führt das zu dem Konfidenzintervall  $[0.314, 0.650]$ : Die hohe Sicherheit wird durch ein recht großes Konfidenzintervall erkauft.

3. Mal angenommen, wir haben eine Normalverteilung  $N(a, \sigma^2)$  mit unbekanntem  $a$  und bekanntem  $\sigma$  vor uns. Sie wird  $n$ -mal abgefragt, daraus soll ein Konfidenzintervall für  $a$  zum Fehlerniveau  $\alpha$  gefunden werden. Wir argumentieren so:

- $x_1, \dots, x_n$  sei das Ergebnis der Abfrage, mit  $\bar{x}$  bezeichnen wir wie üblich das Stichprobenmittel. Es ist bekannt, dass  $\bar{x} \sim N(a, \sigma^2/n)$  verteilt ist. Beachte:  $(\bar{x} - a)\sqrt{n}/\sigma$  ist dann  $N(0, 1)$ -verteilt.
- Wähle ein Intervall der Form  $[-r, r]$ , so dass

$$\mathbb{P}([-r, r]) = 1 - \alpha$$

unter  $N(0, 1)$ . (Das kann mit Hilfe einer Tafel von  $N(0, 1)$  leicht gefunden werden: Bestimme  $r$  so, dass  $\mathbb{P}_{N(0,1)}(x \leq r) = 1 - \alpha/2$ .)

- Mit Wahrscheinlichkeit  $1 - \alpha$  liegt  $(\bar{x} - a)\sqrt{n}/\sigma$  in  $[-r, +r]$ . Das ist gleichwertig zu: Mit Wahrscheinlichkeit  $1 - \alpha$  liegt  $a$  in  $[\bar{x} - r\sigma/\sqrt{n}, \bar{x} + r\sigma/\sqrt{n}]$ .

Und das heißt:  $[\bar{x} - r\sigma/\sqrt{n}, \bar{x} + r\sigma/\sqrt{n}]$  ist ein Konfidenzintervall für  $a$  zum Konfidenzniveau  $1 - \alpha$ .

Ein *Beispiel* dazu: Es sei  $\alpha = 0.05$ ,  $\sigma = 10$  und  $n = 40$ . Der Wert von  $r$  aus den vorstehenden Überlegungen ist dann 1.96. Damit ist  $10 \cdot 1.96/\sqrt{40} = 3.10$  auszurechnen. Fazit: Das Konfidenzintervall für  $a$  ist  $[\bar{x} - 3.10, \bar{x} + 3.10]$ .

Wenn man die Idee erst einmal verstanden hat, lassen sich auch leicht Varianten des Problems diskutieren. Wie groß muss zum Beispiel  $n$  sein, damit das

Konfidenzintervall zum Fehlerniveau  $0.05$  die Länge  $l$  hat? Die Lösung: Wähle  $n$  so, dass  $1.96 \cdot 10/\sqrt{n} \leq l/2$ , d.h.  $n \geq 4 \cdot 19.6/l^2$ .

Oder: Die Länge  $l$  des Konfidenzintervalls und die Versuchsanzahl  $n$  werden vorgegeben. Welche Irrtumswahrscheinlichkeit  $\alpha$  ist anzusetzen? Die Lösung: Mit den obigen Bezeichnungen soll für  $r$  die Bedingung  $r\sigma/\sqrt{n} = l/2$  erfüllt sein, und damit lässt sich  $r$  berechnen.  $\alpha$  ergibt sich dann mit Hilfe einer Tabelle der Normalverteilung:  $\alpha$  ist derjenige Wert, für den  $\mathbb{P}_{N(0,1)}([-r, r]) = 1 - \alpha$  gilt.

Ein Beispiel dazu: Es sei  $l = 0.2$ ,  $\sigma = 1$  und  $n = 100$ . Dann muss für  $r$  die Gleichung  $r/\sqrt{100} = 0.1$  gelten, es folgt  $r = 1$ . Der Tabelle der Normalverteilung entnehmen wir, dass  $\mathbb{P}_{N(0,1)}([-1, 1]) = 0.6826$ . Das ergibt ein Irrtumsniveau von knapp 32 Prozent. Das deutet darauf hin, dass genaue Prognosen bei dieser Konstellation nur recht unzuverlässig zu erwarten sind<sup>10)</sup>. Wenn wir uns mit  $l = 0.4$  begnügen, führt das zu  $r = 2$  und damit – wegen  $\mathbb{P}_{N(0,1)}([-2, 2]) = 0.9544$  – zu  $\alpha = 0.0456$ : In nur knapp fünf Prozent aller Fälle wird unsere Prognose falsch sein.

## 10.5 Konfidenzintervalle bei normalverteilten Zufallsvariablen

Am Ende des letzten Abschnitts haben wir Konfidenzintervalle für  $a$  berechnet, wenn es um Normalverteilungen  $N(a, \sigma^2)$  geht und  $\sigma$  bekannt ist. Was ist zu tun, wenn  $\sigma$  unbekannt ist? Wie findet man Konfidenzintervalle für  $\sigma$ , wenn  $a$  bekannt bzw. unbekannt ist?

Die Lösung besteht darin, die Stichprobenwerte  $x_1, \dots, x_n$  so in eine Formel  $F$  einzusetzen, dass gilt:

- In  $F$  kommen die  $x_1, \dots, x_n$ , bekannte Größen und der gerade interessierende Parameter  $\theta$  vor.
- Die Verteilung von  $F$  ist von  $\theta$  unabhängig, die zugehörigen Wahrscheinlichkeiten können in einer Tabelle nachgeschlagen werden.

Folglich lässt sich zu vorgegebenem  $\alpha$  ein Intervall  $I$  finden, so dass  $F \in I$  mit Wahrscheinlichkeit  $1 - \alpha$  gilt. Löst man „ $F \in I$ “ noch nach  $\theta$  auf, formt also  $F \in I$  in  $\theta \in J$  um, hat man einen Konfidenzbereich  $J$  zum Konfidenzniveau  $1 - \alpha$  für  $\theta$  gefunden.

Im Beispiel am Ende des vorigen Abschnitts hatten wir mit  $F = (\bar{x} - a)\sqrt{n}/\sigma$  gearbeitet. (Der unbekannte Parameter  $\theta$  ist hier der Erwartungswert  $a$  der Verteilung  $N(a, \sigma^2)$ .)  $F$  ist  $N(0, 1)$ -verteilt, und zwar unabhängig von  $a$ . So fanden wir  $I = [-r, r]$ , und nachdem wir die Aussage  $F \in I$  nach  $a$  aufgelöst hatten, gelangten wir zu  $a \in J = [\bar{x} - r\sigma/\sqrt{n}, \bar{x} + r\sigma/\sqrt{n}]$ : Das ist ein Konfidenzintervall zum Irrtumsniveau  $\alpha$ .

---

<sup>10)</sup>Beachte:  $l = 0.2$  bedeutet, dass wir  $a$  bis auf eine Stelle nach dem Komma genau lokalisieren wollen.

Es wird erforderlich sein, einige weitere Verteilungen im Zusammenhang mit der Normalverteilung zu untersuchen. Sie spielen in der Statistik an verschiedenen Stellen eine wichtige Rolle, in diesem Buch werden wir sie hier für die Bestimmung von Konfidenzintervallen und im nächsten Kapitel in der Testtheorie benötigen. Wir werden die  *$\chi$ -Quadrat-Verteilungen* und die *t-Verteilungen* behandeln. Welche Rolle sie in diesem Zusammenhang spielen, wird danach ausführlich beschrieben werden.

### Die $\chi$ -Quadrat-Verteilungen

Sei  $X$  eine  $N(0, 1)$ -verteilte Zufallsvariable. Die Werte von  $X^2$  sind dann größer oder gleich Null und liegen tendenziell „nahe bei Null“, da die  $X$ -Werte meist nicht besonders groß sind.

Das wollen wir nun genauer untersuchen. Zunächst betrachten wir  $|X|$ . Da  $X$  die Dichtefunktion  $f(x) = (1/\sqrt{2\pi})e^{-x^2/2}$  hat, muss man zur Bestimmung der Dichtefunktion von  $|X|$  die Werte von  $f$ , die für negative  $x$  zuständig sind, nur „umklappen“. (Vgl. Übungsaufgabe 10.5.1.) So folgt, dass  $|X|$  die auf  $[0, +\infty[$  definierte Dichtefunktion  $x \mapsto (2/\sqrt{2\pi})e^{-x^2/2}$  hat. Die Dichtefunktion für  $|X|^2$  ergibt sich daraus mit Hilfe von Satz 3.2.3. Für die Zufallsvariable  $X$  in diesem Satz ist  $x \mapsto x^2$  einzusetzen, als Dichtefunktion für  $|X|^2$  erhalten wir

$$x \mapsto \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{x}} e^{-x/2},$$

definiert für  $x > 0$ . Diese Funktion wird bei Null singulär, solche Phänomene haben wir aber auch schon bei anderen Beispielen – etwa der Verteilung auf Seite 78 – gesehen. Da  $X^2 = |X|^2$  gilt, ist damit auch die Verteilung von  $X^2$  bekannt<sup>11)</sup>. Wir fassen zusammen:

**Lemma 10.5.1.** *Ist  $X$   $N(0, 1)$ -verteilt, so ist  $X^2$  gemäß der auf  $]0, +\infty[$  definierten Dichtefunktion  $(1/\sqrt{2\pi x})e^{-x/2}$  verteilt.*

Nun kommt Satz 4.6.5 ins Spiel. Danach entstehen die Dichten von unabhängigen Summen von Zufallsvariablen durch Faltung der Dichten der Summanden. Aufgrund des vorstehenden Lemmas heißt das, dass die Summen von Quadraten von  $N(0, 1)$ -verteilten Zufallsvariablen eine Dichte haben, die durch Faltung von Funktionen des Typs  $(1/\sqrt{2\pi x})e^{-x/2}$  bestimmt werden kann. Es ist überraschend, dass es dafür eine geschlossene Formel gibt. Als Hilfsmittel zur Auswertung der Faltungsintegrale wird der Transformationsatz für Gebietsintegrale (vgl. Anhang, Seite 361) verwendet. Die technischen Einzelheiten sind verwickelt, im nachstehenden Satz wird nur das Endergebnis angegeben<sup>12)</sup>:

<sup>11)</sup>Der Umweg über  $|X|$  war deswegen notwendig, weil Satz 3.2.3 nur dann angewendet werden kann, wenn man induzierte Verteilungen für *monotone* Zufallsvariable ausrechnen möchte. Da das für  $x \mapsto x^2$  auf  $\mathbb{R}$  nicht gilt, wohl aber auf  $[0, +\infty[$ , haben wir uns in einem Zwischenschritt mit  $|X|$  beschäftigt.

<sup>12)</sup>Beweise zu den Ergebnissen dieses Abschnitts findet man in der angegebenen weiterführenden Literatur, z.B. im Buch von Georgii.

**Satz 10.5.2.** Es sei  $n \in \mathbb{N}$ , und  $X_1, \dots, X_n$  seien unabhängige  $N(0, 1)$ -verteilte Zufallsvariable. Wir definieren  $X$  als  $X_1^2 + \dots + X_n^2$ . Dann hat  $X$  die auf  $]0, +\infty[$  definierte Dichtefunktion  $c_n x^{n/2-1} e^{-x/2}$ , wobei  $c_n$  eine Konstante ist<sup>13)</sup>.

Die so definierte Verteilung wird als die  $\chi$ -Quadrat-Verteilung mit  $n$  Freiheitsgraden (kurz  $\chi_n^2$ -Verteilung) bezeichnet.

Nachstehend sind die Dichten der  $\chi$ -Quadrat-Verteilungen für einige  $n$  skizziert:

→  
Programm!

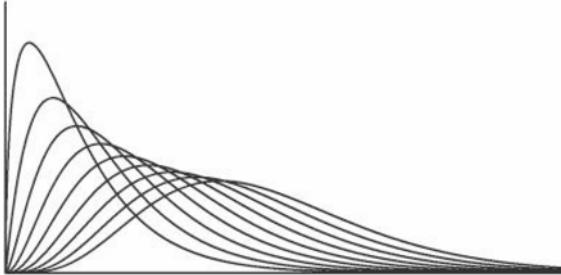


Bild 10.5.1: Die Dichten der  $\chi_n^2$ -Verteilungen für  $n = 3, \dots, 11$ .

Da wir uns für diese Verteilungen nur im Zusammenhang mit Konfidenzbereichen interessieren, geht es nicht um Fragen wie „Wie wahrscheinlich ist es, dass eine  $\chi_{10}^2$ -verteilte Zufallsvariable einen Wert in  $[3, 5]$  annimmt?“ Eher möchte man wissen: „Wie findet man ein möglichst kleines Intervall  $I$ , so dass eine  $\chi_{10}^2$ -verteilte Zufallsvariable nur mit Wahrscheinlichkeit 0.05 einen Wert außerhalb von  $I$  annimmt?“. Zur Beantwortung dieser Fragen gibt es Tabellen, eine finden Sie am Ende des Buches auf Seite 366. Aus ihr können die interessierenden Informationen leicht abgelesen werden:

- $X$  sei  $\chi_{20}^2$ -verteilt. Wie findet man ein geeignetes Intervall  $I$ , so dass  $X$  nur mit Wahrscheinlichkeit 0.01 nicht zu  $I$  gehört? Dazu muss man in der Tabelle das Intervall zu  $n = 10$  und  $\alpha = 0.01$  ablesen, man erhält  $I = [7.43, 40.00]$ .
- Angenommen, es geht um 25 Freiheitsgrade. Der Tabelle ist zu entnehmen, dass eine  $\chi_{25}^2$ -verteilte Zufallsvariable  $X$  mit Wahrscheinlichkeit 0.99 einen Wert in  $[10.52, 46.93]$  annehmen wird.

### Die $t$ -Verteilungen

Diesmal betrachten wir unabhängige  $N(0, 1)$ -verteilte Zufallsvariable  $X$  und  $Y_1, \dots, Y_n$ . Wegen  $\sigma^2(Y_i) = 1$  sind dann die  $Y_i^2$  „im Mittel“ gleich Eins, die

<sup>13)</sup>Der konkrete Wert ist unerheblich. Es soll nur sichergestellt sein, dass das Integral der Funktion über  $]0, +\infty[$  gleich Eins ist. Es lässt sich übrigens zeigen, dass  $c_n = 1/(\Gamma(n/2)2^{n/2})$ , wobei  $\Gamma$  die Gammafunktion bezeichnet. Aber das spielt – wie schon erwähnt – eigentlich keine Rolle.

Zufallsvariable  $(\sum_{i=1}^n Y_i^2)/n$  sollte sich daher (für genügend große  $n$ ) wenig von der Konstanten Eins unterscheiden. Deswegen ist es wenig überraschend, dass  $X/\sqrt{\sum Y_i^2/n}$  „fast“ gleich einer Standard-Normalverteilung ist. Wirklich lässt sich zeigen:

**Satz 10.5.3.** Die Zufallsvariable  $X/\sqrt{\sum_{i=1}^n Y_i^2/n}$  hat eine Dichtefunktion  $\tau_n : \mathbb{R} \rightarrow \mathbb{R}$ . Es ist

$$\tau_n(x) = d_n \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2},$$

wobei  $d_n$  eine Konstante ist<sup>14)</sup>.

Die durch  $\tau_n$  definierte Verteilung heißt die Studentsche  $t$ -Verteilung mit  $n$  Freiheitsgraden (kurz:  $t_n$ -Verteilung).

Dass  $\tau_n$  „beinahe“ die Dichte der Standard-Normalverteilung ist, sieht man so ein: Für beliebige  $y$  ist doch  $\lim_n (1+y/n)^n = e^y$ . Und deswegen gilt

$$\begin{aligned} \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2} &= \frac{1}{\sqrt{(1+x^2/n)^n}} \frac{1}{\sqrt{1+x^2/n}} \\ &\rightarrow \frac{1}{\sqrt{e^{x^2}}} \\ &= e^{-x^2/2}. \end{aligned}$$

Da die  $\tau_n$  und  $e^{-x^2/2}/\sqrt{2\pi}$  Dichtefunktionen sind, ergibt sich auch noch  $d_n \rightarrow 1/\sqrt{2\pi}$ .

Hier eine Skizze einiger Dichten der  $t$ -Verteilungen. Sie konvergieren sehr schnell gegen die Standard-Normalverteilung.

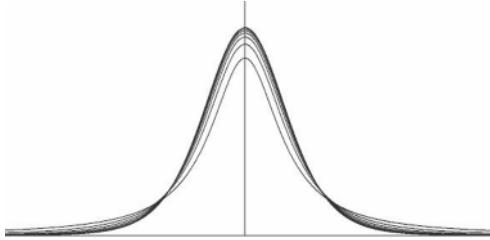


Bild 10.5.2: Die Dichten der  $t$ -Verteilungen für  $n = 2, \dots, 7$ .

Ähnlich wie bei den  $\chi$ -Quadrat-Verteilungen ist man überwiegend daran interessiert, für Irrtumswahrscheinlichkeiten  $\alpha$  Intervalle so bestimmen zu können, dass die Wahrscheinlichkeit unter  $t_n$  gleich  $1 - \alpha$  ist. Für einige Werte von  $n$  und  $\alpha$  können solche Intervalle aus der Tabelle im Anhang (vgl. Seite 367) abgelesen werden.

---

<sup>14)</sup>Durch die Multiplikation mit  $d_n$  wird garantiert, dass  $\int_{\mathbb{R}} \tau_n(x) dx = 1$  gilt.

Beispiel 1: Es sei  $n = 5$  und  $\alpha = 0.05$ . In der Tabelle findet man das zugehörige Intervall  $[-2.75, 2.75]$ , dafür gilt  $\mathbb{P}_{t_5}([-2.75, 2.75]) = 0.95 = 1 - \alpha$ .

Beispiel 2: Bei der Vorgabe von  $\alpha = 0.02$  und  $n = 10$  erhält man das Intervall  $[-2.764, 2.764]$ : Es ist  $\mathbb{P}([-2.764, 2.764]) = 1 - \alpha = 0.98$ .

Für die Standardnormalverteilung wird die Wahrscheinlichkeit 0.98 schon bei einem kleineren Intervall erreicht: Es ist  $\mathbb{P}([-2.326, 2.326]) = 0.98$ . Das zeigt, dass die Approximation von  $N(0, 1)$  durch  $t_n$  für  $n = 10$  noch nicht besonders gut ist.

Mit Hilfe der  $\chi$ -Quadrat-und der  $t$ -Verteilungen können die Verteilungen beschrieben werden, die sich bei der Berechnung von Stichprobenmittel und Stichprobenvarianz ergeben, wenn die Stichprobe  $x_1, \dots, x_n$  durch Abfrage einer Normalverteilung entstanden ist.

Um das zu präzisieren, betrachten wir unabhängige Zufallsvariable  $X_1, \dots, X_n$ , die  $N(a, \sigma^2)$ -verteilt sein sollen. Dabei ist  $a \in \mathbb{R}$  und  $\sigma^2 > 0$ . Mit den  $X_i$  bilden wir die Zufallsvariable  $M := (X_1 + \dots + X_n)/n$ : Sie entspricht dem Übergang von den  $X_1, \dots, X_n$  zum Stichprobenmittel. Und wenn man die Stichprobenvarianz ebenfalls als Zufallsvariable auffassen möchte, muss man  $\sum_{i=1}^n (X_i - M)^2/(n-1)$  betrachten. Diese Zufallsvariable wollen wir  $V^*$  nennen. Es gilt dann:

**Satz 10.5.4.** (i)  $M$  ist  $N(a, \sigma^2/n)$ -verteilt.

(ii)  $\frac{n-1}{\sigma^2}V^*$  ist  $\chi^2_{n-1}$ -verteilt.

(iii)  $\frac{\sqrt{n}(M-a)}{\sqrt{V^*}}$  ist  $t_{n-1}$ -verteilt.

Dabei ist die Aussage (i) im Wesentlichen das Wurzel- $n$ -Gesetz. In (ii) (bzw. (iii)) ist überraschend, dass die fragliche Verteilung nicht von  $a$  (bzw. von  $\sigma^2$ ) abhängt. Die Beweise dieser Aussagen sind recht verwickelt und sollen hier nicht geführt werden.

### Bestimmung von Konfidenzintervallen

Nach diesen Vorbereitungen können wir die zu Beginn des Abschnitts beschriebene Strategie zum Auffinden von Konfidenzintervallen erfolgreich realisieren. Es folgen die für die Anwendungen wichtigsten Beispiele<sup>15)</sup>.

Es ist wichtig daran zu erinnern, dass es in allen Fällen um das Schätzen von Parametern im Zusammenhang mit der Normalverteilung geht. Für andere Situationen sind unsere Ergebnisse nicht anwendbar<sup>16)</sup>.

#### A. Erwartungswert einer Normalverteilung bei bekannter Streuung

<sup>15)</sup>Der Vollständigkeit halber wird die am Ende des letzten Abschnitts diskutierte Situation unter „A.“ auch noch einmal aufgeführt.

<sup>16)</sup>Jedenfalls nicht unmittelbar. Manchmal kommt die Normalverteilung aber über den zentralen Grenzwertsatz ins Spiel. Dann muss im Einzelfall begründet werden, warum es gerechtfertigt ist, die auftretenden Zufallsvariablen durch normalverteilte Variablen zu approximieren.

*Problem:* Das statistische Modell bestehe aus allen  $N(a, \sigma^2)$ . Es ist ein Konfidenzintervall für  $a$  zum Irrtumsniveau  $\alpha$  aufgrund von  $n$  Messungen zu finden.  $\sigma$  ist bekannt.

*Lösung:* Bestimme  $r$  so, dass  $\mathbb{P}_{N(0,1)}([-r, +r]) = 1 - \alpha$ . Berechne dann nach der Messung der  $x_1, \dots, x_n$  das Stichprobenmittel  $\bar{x}$  und wähle das Intervall  $[\bar{x} - r\sigma/\sqrt{n}, \bar{x} + r\sigma/\sqrt{n}]$  als Konfidenzintervall für  $a$ . Mit Wahrscheinlichkeit  $1 - \alpha$  wird das eine richtige Prognose sein.

*Begründung:* Eine Begründung wurde am Ende des letzten Abschnitts gegeben.

*Ein Beispiel:* s.o., Seite 305

### B. Erwartungswert einer Normalverteilung bei unbekannter Streuung

*Problem:* Diesmal besteht das Modell aus allen  $N(a, \sigma^2)$ , wobei  $\sigma$  unbekannt ist. Gesucht ist ein Konfidenzintervall zum Irrtumsniveau  $\alpha$  für  $a$  aus einer Stichprobe des Umfangs  $n$ . (Beispiel: Eine wirkliche Länge soll aus einer mehrfachen Längenmessung mit einem unbekannten Messgerät ermittelt werden.)

*Lösung:* Berechne aus der Stichprobe  $x_1, \dots, x_n$  das Stichprobenmittel  $\bar{x}$  und die Stichprobenstreuung  $s_x = \sqrt{V_x}$  wie in Abschnitt 9.3.

Bestimme mit Hilfe einer Tabelle der  $t$ -Verteilungen ein  $r$  so, dass das Intervall  $[-r, r]$  unter  $t_{n-1}$  die Wahrscheinlichkeit  $1 - \alpha$  hat. (Achtung:  $t_{n-1}$ , nicht  $t_n$ !) Löse dann noch  $\sqrt{n}(\bar{x} - a)/s_x \in [-r, r]$  nach  $a$  auf: Diese Bedingung ist gleichwertig zu  $a \in [\bar{x} - r \cdot s_x/\sqrt{n}, \bar{x} + r \cdot s_x/\sqrt{n}]$ , folglich ist dieses Intervall das gesuchte Konfidenzintervall.

*Begründung:* Die „wirkliche“ Zufallsvariable ist  $N(a, \sigma^2)$ -verteilt, nach Satz 10.5.4 ist dann  $\sqrt{n}(\bar{x} - a)/s_x t_{n-1}$ -verteilt. Folglich liegt – nach Wahl von  $r$  – die Zahl  $\sqrt{n}(\bar{x} - a)/s_x$  mit Wahrscheinlichkeit  $1 - \alpha$  in  $[-r, r]$ . Nun muss man nur noch nach  $a$  auflösen.

*Beispiel:* Es sei  $\alpha = 0.05$ ,  $n = 10$ ,  $\bar{x} = 3.1$  und  $V_x = 12.3$ . Wir müssen in der Tabelle mit 9 (!) Freiheitsgraden in der Spalte  $\alpha = 0.05$  nachsehen. Wir erhalten das Intervall  $[-2.262, 2.262]$ . Wir berechnen noch  $2.262\sqrt{12.3}/\sqrt{10} = 2.50$ . Als Konfidenzintervall ergibt sich

$$[3.1 - 2.50, 3.1 + 2.50] = [0.60, 5.60].$$

Das ist ziemlich groß, aber die Varianz ist ja auch erheblich.

*Kommentar dazu:* Die Ansätze in „A.“ und „B.“ sind sehr ähnlich. In „B.“ war das unbekannte  $\sigma$  durch die Stichprobenstreuung zu ersetzen, und statt mit der Standard-Normalverteilung musste man mit  $t_{n-1}$  arbeiten.

### C. Erwartungswertabstand zweier Normalverteilungen mit bekannter Streuung

*Problem:* Gegeben seien zwei Größen, die  $N(a_1, \sigma_1^2)$ - bzw.  $N(a_2, \sigma_2^2)$ -verteilt sind; dabei sind  $a_1$  und  $a_2$  unbekannt, aber  $\sigma_1$  und  $\sigma_2$  sind bekannt. Die erste wird  $m$ -mal, die zweite  $n$ -mal abgefragt; die Ergebnisse bezeichnen wir mit  $x_1, \dots, x_m$  bzw.  $y_1, \dots, y_n$ , die Stichprobenmittel mit  $\bar{x}$  und  $\bar{y}$ .

Wir wollen mit diesen Informationen ein Konfidenzintervall für  $a_1 - a_2$  zum Konfidenzniveau  $1 - \alpha$  finden. (Das tritt zum Beispiel dann auf, wenn man

wissen möchte, um wieviel erfolgreicher die eine Düngemethode als die andere ist.)

*Lösung:* Bestimme – mit Hilfe der Tafel der Standard-Normalverteilung – ein Intervall  $[-r, r]$ , das unter  $N(0, 1)$  die Wahrscheinlichkeit  $1 - \alpha$  hat. Setze

$$d := \frac{\sqrt{mn}}{\sqrt{n\sigma_1^2 + m\sigma_2^2}}.$$

Das gesuchte Konfidenzintervall für  $a_1 - a_2$  ist dann  $[\bar{x} - \bar{y} - r/d, \bar{x} - \bar{y} + r/d]$ .

*Begründung:* In Abschnitt 8.4 hatten wir gezeigt: Sind  $X, Y$  unabhängig und  $N(b_1, \sigma_1^2)$ - bzw.  $N(b_2, \sigma_2^2)$ -verteilt, so ist  $cX$   $N(cb_1, c^2\sigma_1^2)$ -verteilt und  $X + Y$   $N(b_1 + b_2, \sigma_1^2 + \sigma_2^2)$ -verteilt. Damit folgt: Die Verteilung von

$$[\bar{x} - \bar{y} - (a_1 - a_2)]d$$

ist  $N(0, 1)$ . Der Rest ist dann klar.

*Beispiel:*  $\alpha = 0.05$ ,  $m = 100$ ,  $n = 200$ ,  $\sigma_1 = 1$  und  $\sigma_2 = 2$ . Angenommen, wir erhalten  $\bar{x} - \bar{y} = 3$ . Da in diesem Fall  $d = \frac{\sqrt{20000}}{\sqrt{200 + 4 \cdot 100}} = 5.77$  ist und wir  $r = 1.96$  wählen können, folgt: Ein Konfidenzintervall für  $a_1 - a_2$  ist durch  $[3 - 1.96/5.77, 3 + 1.96/5.77] = [2.66, 3.34]$  gegeben.

#### D. Varianz einer Normalverteilung bei bekanntem Erwartungswert

*Problem:* Eine Normalverteilung  $N(a, \sigma^2)$  wird  $n$ -mal abgefragt (mit dem Ergebnis  $x_1, \dots, x_n$ ).  $a$  ist bekannt, ein Konfidenzintervall für  $\sigma^2$  zum Irrtumsniveau  $\alpha$  ist gesucht.

*Lösung:* Bestimme ein Intervall  $[a_1, a_2]$ , das unter  $\chi_n^2$  die Wahrscheinlichkeit  $1 - \alpha$  hat und berechne  $d := \sum(x_i - a)^2$ . Dann liegt  $d/\sigma^2$  mit Wahrscheinlichkeit  $1 - \alpha$  in  $[a_1, a_2]$ , d.h.  $[d/a_2, d/a_1]$  ist ein Konfidenzintervall für  $\sigma^2$ .

*Begründung* Die  $(x_i - a)/\sigma$  sind  $N(0, 1)$ -verteilt, folglich ist  $\sum(x_i - a)^2/\sigma^2$  nach Satz 10.5.2  $\chi_n^2$ -verteilt.

*Beispiel:* Es gebe 10 Abfragen, und  $\alpha = 0.05$  sei vorgegeben. Der Tabelle (abzulesen bei 10 Freiheitsgraden!) auf Seite 366 entnehmen wir, dass  $[a_1, a_2] = [3.25, 20.48]$  die geforderten Eigenschaften hat. Ist also etwa bei einer konkreten Messung das sich ergebende  $d$  gleich 123, so ergibt sich daraus das Konfidenzintervall  $[123/20.48, 123/3.25] = [6.00, 37.84]$  für  $\sigma^2$ .

#### E. Varianz einer Normalverteilung bei unbekanntem Erwartungswert

*Problem:* Wie eben, aber mit unbekanntem Erwartungswert.

*Lösung:* Bestimme ein Intervall  $[a_1, a_2]$ , das unter  $\chi_{n-1}^2$  die Wahrscheinlichkeit  $1 - \alpha$  hat und berechne  $\tilde{d} := (n - 1)V_x = \sum(x_i - \bar{x})^2$ . Dann liegt  $\tilde{d}/\sigma^2$  mit Wahrscheinlichkeit  $1 - \alpha$  in  $[a_1, a_2]$ , d.h.  $[\tilde{d}/a_2, \tilde{d}/a_1]$  ist ein Konfidenzintervall für  $\sigma^2$ .

*Begründung:* Hier wird Satz 10.5.4 (ii) angewendet, danach ist  $\tilde{d}/\sigma^2$   $\chi_{n-1}^2$ -verteilt.

*Beispiel:* Wir betrachten die Situation aus „D“, es sei aber  $a$  unbekannt. Diesmal müssen wir in der Tabelle bei 9 (!) Freiheitsgraden und  $\alpha = 0.05$  nachschauen, es ergibt sich das Intervall

$$[a_1, a_2] = [2.70, 19.02].$$

Das impliziert – falls wir wieder  $\tilde{d} = 123$  annehmen – das Konfidenzintervall

$$[123/19.02, 123/2.70] = [6.47, 45.55]$$

für  $\sigma^2$ . Dieses Intervall ist erwartungsgemäß größer als das in „E“, die Verschlechterung ist auf die fehlende Information über  $a$  zurückzuführen.

## 10.6 Verständnisfragen

### Zu Abschnitt 10.1

*Sachfragen*

**S1:** Was ist ein statistisches Modell?

**S2:** Was versteht man unter einem Schätzer?

*Methodenfragen*

**M1:** Eine statistische Fragestellung in eine statistisches Modell übersetzen können.

### Zu Abschnitt 10.2

*Sachfragen*

**S1:** Durch welche Eigenschaft ist ein erwartungstreuer Schätzer definiert?

**S2:** Was ist eine konsistente Schätzfolge?

**S3:** Wann heißt ein Schätzer ein Schätzer mit gleichmäßig bester Varianz?

**S4:** Wie erklärt sich der Nenner  $n - 1$  bei der Definition der Stichprobenvarianz?

*Methodenfragen*

**M1:** Nachprüfen können, ob ein Schätzer erwartungstreu ist.

**M2:** Nachprüfen können, ob ein Schätzer besser als ein anderer ist.

### Zu Abschnitt 10.3

*Sachfragen*

**S1:** Was versteht man unter einer Statistik? Wann heißt sie suffizient, wann vollständig?

**S2:** Was besagt der Satz von Lehmann-Scheffé?

**S3:** Was ist ein maximum-likelihood-Schätzer?

*Methodenfragen*

**M1:** Nachprüfen können, ob eine Statistik suffizient und vollständig ist, um auf diese Weise optimale Schätzer ermitteln zu können.

**M2:** Für vorgelegte Modelle maximum-likelihood-Schätzer bestimmen können.

### Zu Abschnitt 10.4

#### *Sachfragen*

**S1:** Was sind „Irrtumsniveau“ und „Konfidenzniveau“?

**S2:** Was ist ein Konfidenzbereich?

**S3:** Was ist die richtige Interpretation einer Aussage des Typs „Mit Wahrscheinlichkeit 0.95 liegt  $\gamma(\theta_0)$  in  $[1.2, 4.3]$ “?

**S4:** Wird ein Konfidenzbereich größer oder kleiner, wenn man die Irrtumswahrscheinlichkeit vergößert?

**S5:** Welche allgemeine Strategie gibt es, möglichst kleine Konfidenzbereiche zu berechnen?

#### *Methodenfragen*

**M1:** Konfidenzbereiche zu vorgegebenem Konfidenz- bzw. Irrtumsniveau bestimmen können.

### Zu Abschnitt 10.5

#### *Sachfragen*

**S1:** Wodurch sind die  $\chi^2$ -Verteilungen definiert?

**S2:** Bei welcher Kombination von Normalverteilungen treten die  $t_n$ -Verteilungen auf?

**S3:** Bei welcher Fragestellung werden die  $t$ -Verteilungen benötigt?

**S4:** Bei welcher Fragestellung werden die  $\chi^2$ -Verteilungen benötigt?

#### *Methodenfragen*

**M1:** Konfidenzintervalle im Zusammenhang mit Normalverteilungen  $N(a, \sigma^2)$  berechnen können, insbesondere in den Fällen:

- Konfidenzintervall für  $a$  bei bekanntem  $\sigma$ .
- Konfidenzintervall für  $a$  bei unbekanntem  $\sigma$ .
- Konfidenzintervall für  $\sigma$  bei bekanntem  $a$ .
- Konfidenzintervall für  $\sigma$  bei unbekanntem  $a$ .

## 10.7 Übungsaufgaben

### Zu Abschnitt 10.2

**Ü10.2.1** Wir betrachten das statistische Modell aller Gleichverteilungen auf allen Intervallen der Form  $[a, b]$  mit  $a < b$ . Es wird  $n$  Mal abgefragt. Beweisen oder widerlegen Sie:

- $x_1$  ist ein erwartungstreuer Schätzer für  $(a + b)/2$ .
- Das Stichprobenmittel ist ein erwartungstreuer Schätzer für  $(a + b)/2$ .

**Ü10.2.2** Prüfen Sie nach, in welchen der folgenden Beispiele der Schätzer  $d$  erwartungstreue ist:

a) Das statistische Modell bestehe aus allen Poisson-Verteilungen zu Parametern  $\lambda > 0$  auf  $\mathbb{R}$ , zu schätzen sei  $\lambda^2$ .  $d$  ist durch  $(x_1, \dots, x_n) \mapsto x_1^2$  gegeben.

b) Wie eben, aber  $d$  wird durch  $(x_1, \dots, x_n) \mapsto x_1 x_2$  ersetzt.

**Ü10.2.3** Gegeben seien ein statistisches Modell  $(\Omega, \mathcal{E}, \mathbb{P}_\theta)_{\theta \in \Theta}$ , eine Zufallsvariable  $X : \Omega \rightarrow \mathbb{R}$  (für die zu allen  $\mathbb{P}_\theta$  der Erwartungswert existiert) und die Abbildung  $\gamma : \Theta \rightarrow \mathbb{R}$ ,  $\gamma(\theta) = \mathbb{E}_{\mathbb{P}_\theta}(X)$ . Welche der folgenden Abbildungen  $d : \Omega^n \rightarrow \mathbb{R}$  sind erwartungstreue Schätzer für  $\gamma$ ?

a)  $d(x_1, \dots, x_n) = x_1$ .

b)  $d(x_1, \dots, x_n) = x_1 - x_2$ .

c)  $d(x_1, \dots, x_n) = \frac{1}{3}x_1 + \frac{2}{3}x_2$ .

**Ü10.2.4** In einem statistischen Modell  $(\Omega, \mathcal{E}, (WP_\theta))_{\theta \in \Theta}$  seien  $d$  und  $d'$  zwei gleichmäßig beste erwartungstreue Schätzer für  $\gamma : \Theta \rightarrow \mathbb{R}$ . Zeigen Sie: Für alle  $\theta \in \Theta$  ist  $\mathbb{P}_\theta(d = d') = 1$ .

### Zu Abschnitt 10.3

**Ü10.3.1** Sei  $T : \Omega^n \rightarrow \mathbb{R}$  eine suffiziente und vollständige Statistik.

a) Für welche  $c \in \mathbb{R}$  ist  $cT$  suffizient?

b) Für welche  $c \in \mathbb{R}$  ist  $cT$  vollständig?

**Ü10.3.2** Die *geometrische Verteilung* ist bekanntlich durch die Wahrscheinlichkeiten  $q^{k-1}(1-q)$  (für  $k \in \mathbb{N}$ ) definiert, wobei  $q \in [0, 1]$ . Sie werde  $n$ -mal abgefragt, die Ergebnisse seien  $k_1, \dots, k_n$ .

Finden Sie eine maximum-likelihood-Schätzung für  $q$  aus den  $k_1, \dots, k_n$ .

**Ü10.3.3** Es sei das statistische Modell der Gleichverteilungen auf  $\{a, \dots, 100\}$  gegeben, wobei  $a \in \{1, \dots, 100\}$ . Zeigen Sie, dass das Minimum der Abfragen ein maximum-likelihood-Schätzer für  $a$  ist, wenn  $n$  Mal abgefragt wird.

### Zu Abschnitt 10.4

**Ü10.4.1** Für jedes  $c \geq 0$  sei ein Wahrscheinlichkeitsmaß  $\mathbb{P}_c$  auf  $[0, 1]$  durch die Dichtefunktion  $f_c(x) = (c+1)x^c$  gegeben. Bestimmen Sie nach einmaliger Abfrage ein Konfidenzintervall für  $c$  mit Irrtumswahrscheinlichkeit  $\alpha = 0.05$ . Verwenden Sie dabei das in Satz 10.4.1 beschriebene Verfahren.

**Ü10.4.2** Das statistische Modell bestehe aus den Poissonverteilungen zu den Parametern  $\lambda > 0$ . Es wird einmal abgefragt, das Ergebnis sei  $k$ . Geben Sie ein Verfahren an, mit dem man ein möglichst kleines Konfidenzintervall zum Irrtumsniveau  $\alpha$  berechnen kann.

**Ü10.4.3** Finden Sie ein Beispiel für ein statistisches Modell, in dem die Konfidenzbereiche als einpunktige Mengen gewählt werden können.

### Zu Abschnitt 10.5

**Ü10.5.1** Es sei  $X$  eine reellwertige Zufallsvariable, so dass  $\mathbb{P}_X$  eine Dichtefunktion  $f$  hat. Definiert man dann  $Y$  als  $|X|$ , so hat  $Y$  die auf  $[0, +\infty[$  definierte

Dichtefunktion  $x \mapsto f(x) + f(-x)$ . (Dieses Ergebnis wird in Abschnitt 10.5 benötigt.)

**Ü10.5.2** Es sei  $X : \Omega \rightarrow \mathbb{R}$  normalverteilt, wobei weder der Erwartungswert noch die Varianz bekannt sind. Es soll ein Konfidenzintervall  $I$  zum Niveau  $1 - \alpha = 0.95$  für den Erwartungswert von  $X$  bestimmt werden. Wie oft muss  $X$  abgefragt werden, damit die Länge von  $I$  höchstens 0.04 ist?

**Ü10.5.3** In einem Versuch soll eine Federkonstante  $a$  bestimmt werden. Dazu werden 1000 Messungen durchgeführt, die unabhängige  $N(a, 0.2)$ -verteilte Werte liefern. Wir erhalten  $\bar{x} = 10.2$ . Bestimmen Sie Konfidenzintervalle mit Irrtumswahrscheinlichkeit  $\alpha = 0.05$  und  $\alpha = 0.01$  für  $a$ .

# Kapitel 11

## Entscheiden

In den vorigen Kapiteln haben wir uns mit dem Problem beschäftigt, aufgrund von Abfragen eines Wahrscheinlichkeitsraums *Prognosen über Zahlen* zu treffen: Die Zahl  $\gamma(\theta)$  wurde geschätzt oder es wurde ein Intervall angegeben, in dem sie mit einer vorab wählbaren Wahrscheinlichkeit liegt.

In diesem Kapitel geht es um *Entscheidungen*. Der Zufall wird abgefragt, und dann soll mit Hilfe des Statistikers eine Entscheidung getroffen werden, ob dieses oder jenes getan werden sollte: Soll man dieses Medikament zulassen oder verbieten? Soll man diese Lieferung annehmen oder die Annahme verweigern? Kann man bestätigen, dass Düngemethode A besser ist als Düngemethode B? Und so weiter.

In *Abschnitt 11.1* sprechen wir zunächst über *Hypothesen*, das sind einfach Annahmen über die Wirklichkeit. In *Abschnitt 11.2* wird präzisiert, was es bedeutet, dass man aufgrund von Zufallsexperimenten zu Entscheidungen kommen kann. Je nach Ansatz gibt es dann verschiedene Möglichkeiten zu sagen, was eine „gute Entscheidungsfunktion“ ist. *Abschnitt 11.3* beschäftigt sich mit einer detaillierten Diskussion von Situationen, in denen man nur die Wahl zwischen zwei Möglichkeiten hat. In *Abschnitt 11.4* diskutieren wir etwas detaillierter den Spezialfall normalverteilter Messungen, und die *Abschnitte 11.5 und 11.6* sind für Verständnisfragen und Übungsaufgaben vorgesehen.

### 11.1 Hypothesen

Im wirklichen Leben sind permanent irgendwelche Entscheidungen zu treffen, die meisten ergeben sich ganz spontan. Nehmen wir zum Beispiel die Frage, ob Sie heute einen *Schirm mitnehmen* sollten, wenn Sie das Haus verlassen. Grundlage Ihrer Entscheidung wird doch eine Einschätzung sein: Wird es heute regnen oder nicht?

Mal angenommen, Sie tippen auf Regen und nehmen den Schirm mit. Wenn es dann wirklich regnet, dann haben Sie richtig vorgesorgt. Andernfalls tragen Sie Ihren Schirm den ganzen Tag lang völlig nutzlos spazieren.

Sind Sie dagegen optimistisch und rechnen mit gutem Wetter, so werden Sie bei dem ersten Regenschauer ein Problem haben.

Das Fazit: Wenn man die Wahl zwischen zwei Alternativen hat, so kann sich diese Wahl nachträglich als günstig oder ungünstig erweisen, je nachdem, was denn nun wirklich passiert.

In der Statistik hat sich die Terminologie eingebürgert, eine Annahme über den wirklichen Zustand der Welt eine *Hypothese* zu nennen. Genauer spricht man von einer *Nullhypothese*, das Gegenteil heißt die *Alternativhypothese*. (In unserem Beispiel könnte man „Es bleibt trocken“ oder aber auch „Es wird regnen“ als Nullhypothese deklarieren. Die Alternativhypothese ist dann die jeweils andere Aussage.)

Hier einige typische Beispiele, die in der Statistik eine Rolle spielen:

1. Die Nullhypothese könnte lauten: Dies ist ein fairer Würfel. Die Alternativhypothese wäre dann: Der Würfel ist unfair.
2. Gegeben sei das aus allen Normalverteilungen  $N(a, \sigma^2)$  bestehende statistische Modell. Eine mögliche Nullhypothese wäre: Der Erwartungswert  $a$  dieser Normalverteilung ist gleich 12.
3. Wie vorstehend, aber mit der Nullhypothese  $\sigma^2 \leq 2$ .

Es ist nicht schwer, sich weitere Beispiele mit statistischen Modellen und aus dem wirklichen Leben auszudenken. Es gibt aber nun ein Problem mit der Einschätzung des Wahrheitsgehalts von Hypothesen:

*Man kann Fehler machen!*

Mal angenommen, wir nennen die Nullhypothese  $H_0$ . Wenn wir glauben, dass sie falsch ist und uns danach richten, können zwei Dinge passieren. Sie kann wirklich falsch sein, dann ist alles in Ordnung. Sie kann aber auch wahr sein, dann haben wir einen Fehler gemacht.

Wir könnten aber auch glauben, dass  $H_0$  wahr ist. Falls ja, ist wieder alles bestens, andernfalls haben wir falsch getippt und müssen eventuell mit Konsequenzen rechnen.

**Definition 11.1.1.** *Es seien eine Nullhypothese  $H_0$  und die Alternativhypothese  $H_1$  gegeben.*

- (i) *Nimmt man an, dass  $H_0$  nicht gilt (d.h., dass  $H_1$  gilt), obwohl  $H_0$  in Wirklichkeit richtig ist, so nennt man das einen Fehler erster Art.*
- (ii) *Nimmt man an, dass  $H_0$  gilt (d.h., dass  $H_1$  nicht gilt), obwohl  $H_0$  in Wirklichkeit falsch ist, so nennt man das einen Fehler zweiter Art.*

### Bemerkungen und Beispiele:

1. Das ist wirklich etwas verwirrend. Daher soll die Definition noch einmal in einer kleinen Tabelle wiederholt werden, die man im Folgenden immer vor Augen haben sollte:

|                | $H_0$ angenommen | $H_1$ angenommen |
|----------------|------------------|------------------|
| $H_0$ wirklich | o.k.!            | Fehler 1. Art    |
| $H_1$ wirklich | Fehler 2. Art    | o.k.!            |

2. Es ist im Grunde völlig beliebig, welche der Hypothesen  $H_0$  und welche  $H_1$  ist. Es hat sich aber die folgende *Konvention* eingebürgert:

*Die Hypothesen sind so zu bezeichnen, dass der Fehler erster Art schwerwiegender Konsequenzen als der Fehler zweiter Art hat.*

Typischerweise wird das folgende Beispiel in der Literatur angeboten. Man stellt sich eine Feuerwehrzentrale vor, soeben ist ein Anruf eingegangen: „Das mathematische Institut brennt!“

- $H_0$ : Der Anrufer hat Recht, es brennt wirklich.
- $H_1$ : Es war ein Witzbold, in Wirklichkeit brennt es gar nicht.

Der Fehler erster Art bestünde in diesem Beispiel darin anzunehmen, dass es nicht brennt und generativ aufzulegen, obwohl alles bereits in Flammen steht. Und einen Fehler zweiter Art würde die Feuerwehr begehen, wenn sie ausrückt und am Ziel nur überraschte Mathematiker findet.

Es ist klar, dass der Fehler erster Art in diesem Beispiel als der gravierendere einzuschätzen wäre, deswegen wäre die Bezeichnungsweise im Einklang mit der Konvention.

Es sollte allerdings betont werden, dass das in vielen Fällen durchaus nicht so offensichtlich ist, dass also unterschiedliche Betrachter durchaus verschiedener Meinung sein können, welcher der Fehler gravierender ist.

3. Hier einige Beispiele für Nullhypothesen aus dem täglichen Leben. Versuchen Sie erstens, in den einzelnen Fällen die Fehler erster Art und zweiter Art verbal zu beschreiben und beobachten Sie sich zweitens im Alltag, wann Sie – mehr oder weniger unbewusst – mit dem Abwägen dieser beiden Fehlerarten beschäftigt sind.

- Demnächst steht die Klausur an. Die Nullhypothese: Die Aufgaben werden wohl ganz harmlos sein.
- Herr R. kommt in die Disco und sieht eine schöne Frau, die er gerne ansprechen würde. Neben ihr steht allerdings ein ziemlich brutal aussehender männlicher Mensch. Seine Hypothese: Es ist ihr Bruder.
- Herr S. möchte einen Badezimmerschrank anbringen und muss ein Loch bohren. Seine Hypothese: Das Stromkabel läuft ganz woanders.
- Frau T. möchte mit der Regionalbahn fahren, es sind nur drei Stationen. Ihre Hypothese: Es wird kein Kontrolleur kommen (und deswegen brauche ich auch keine Fahrkarte).

Es ist zu betonen, dass es bei der Anwendung statistischer Verfahren auch oft um Hypothesen geht, bei denen eine falsche Entscheidung weit schwerwiegender Konsequenzen hat. Fehler erster Art können sich manchmal wirklich fatal auswirken:

- Die Zulassung eines neuen Medikaments wird beantragt.

$H_0$ : Es ist noch nicht ausgeschlossen, dass es gefährliche Nebenwirkungen geben kann, und deswegen sollte es noch weiter gepftzt werden.

- $H_0$ : Es ist sinnvoll, eine Vorsorgeuntersuchung zu machen.

- Jugendliche haben einen Raubüberfall im Supermarkt begangen. Die Polizei kommt dazu.

$H_0$ : Die Waffe der Verbrecher ist keine Spielzeugpistole.<sup>1)</sup>

- Ein Tter (Mord, Kindesmissbrauch, Vergewaltigung, . . .) hat eine langjhrige Freiheitsstrafe abgesessen.

$H_0$ : Er ist immer noch gefrlich, und deswegen sollte eine Sicherungsverwahrung angeordnet werden.

## 11.2 Testfunktionen

Es ist nun Zeit, etwas formaler zu werden. Da manche der vorstehend zur Illustration verwendeten Beispiele fr eine mathematische Behandlung nicht przise genug formulierbar sind, werden wir uns auf Situationen beschrnen, die sich mit Hilfe statistischer Modelle beschreiben lassen. Wir gehen wieder von einer Familie  $(\mathbb{P}_\theta)_{\theta \in \Theta}$  von Wahrscheinlichkeitsmaen aus, die alle auf dem gleichen Messraum  $(\Omega, \mathcal{E})$  definiert sind. Wir nehmen an, dass  $\Theta$  disjunkt in zwei nicht leere Teilmengen  $\Theta_0$  und  $\Theta_1$  zerlegt ist und dass die Nullhypothese der Aussage  $\theta \in \Theta_0$  und die Alternativhypothese der Aussage  $\theta \in \Theta_1$  entspricht.

Mchte man zum Beispiel als Nullhypothese die Aussage „Diese Mnze ist fair“ untersuchen, so geht es um die Bernoulliverteilungen auf  $\Omega = \{0, 1\}$  zu allen Erfolgswahrscheinlichkeiten  $p \in [0, 1] =: \Theta$ , und es ist  $\Theta_0 = \{0.5\}$  und  $\Theta_1 = [0, 1] \setminus \{0.5\}$  zu setzen.

Ganz analog lassen sich Hypothesen des Typs „Die Erfolgswahrscheinlichkeit ist  $\leq 0.2$ “ oder „Der Erwartungswert der normalverteilten Zufallsvariablen ist 66“ durch die geeignete Wahl eines statistischen Modells und Definition von  $\Theta_0$  und  $\Theta_1$  formalisieren.

Wie in der Schtztheorie wird ein  $\theta_0 \in \Theta$  ausgewhlt, und aus  $(\Omega, \mathcal{E}, \mathbb{P}_{\theta_0})$  werden  $n$  unabhangige Abfragen gezogen. Aufgrund des Ergebnisses  $(x_1, \dots, x_n) \in \Omega^n$  sollen wir uns fr  $H_0$  oder  $H_1$  entscheiden, d.h. es ist eine Abbildung

---

<sup>1)</sup>In diesem Beispiel haben sowohl ein Fehler erster Art als auch ein Fehler zweiter Art dramatische Konsequenzen.

$d : \Omega^n \rightarrow \{0, 1\}$  anzugeben. Man spricht von einer *Testfunktion* oder auch kürzer von einem *Test* oder von einer *Entscheidungsfunktion*<sup>2)</sup>.

Wenn  $d(x_1, \dots, x_n) = 0$  ist, so wird das so interpretiert, dass wir uns aufgrund der Stichprobe  $x_1, \dots, x_n$  für  $\theta_0 \in \Theta_0$  entscheiden. Und  $d(x_1, \dots, x_n) = 1$  führt entsprechend zu  $\theta_0 \in \Theta_1$ .

Man ist in der Statistik mit den Formulierungen allerdings sehr vorsichtig. Statt „ $H_0$  wird angenommen“ sagt man „ $H_0$  wird nicht abgelehnt“, und statt „ $H_1$  wird angenommen“ sagt man „ $H_0$  wird abgelehnt“. Das trifft das, was wirklich ausgesagt werden kann, auch besser. Man stelle sich zum Beispiel vor, dass eine Münze auf Fairness getestet wird. Bei 20 Versuchen hat sie 12 Mal Kopf und 8 Mal Zahl gezeigt. Dann ist eine Aussage „Die Hypothese, dass die Münze fair ist, kann nicht abgelehnt werden“ sicher angemessener als eine der Form „Diese Münze ist aufgrund des Tests als fair anzusehen“.

Da für  $d(x_1, \dots, x_n) = 1$  die Nullhypothese abgelehnt wird, heißt die Menge  $\{d = 1\}$  auch der *Verwerfungsbereich* (auch: kritischer Bereich) von  $d$ . Da  $d$  nur zwei Werte annimmt, ist  $d$  eindeutig durch die Definition des Verwerfungsbereichs festgelegt.

Für jeden Test  $d$  kann man die zu erwartenden Fehler erster und zweiter Art berechnen. Mal angenommen, es ist  $\theta_0 \in \Theta_0$ . Dann wird für die Abfrage  $(\Omega, \mathcal{E})$  mit dem Wahrscheinlichkeitsmaß  $\mathbb{P}_{\theta_0}$  versehen und  $n$  Mal abgefragt. Wir erhalten die Stichprobe  $(x_1, \dots, x_n)$ . Wenn  $d(x_1, \dots, x_n) = 0$  gilt, ist alles in Ordnung, wenn allerdings  $d(x_1, \dots, x_n) = 1$  ist, lehnen wir die Nullhypothese ab und begehen einen Fehler erster Art. Das zeigt: Für  $\theta_0 \in \Theta_0$  ist die Wahrscheinlichkeit für einen Fehler erster Art gleich

$$\mathbb{P}_{\theta_0^n}(\{(x_1, \dots, x_n) \mid d(x_1, \dots, x_n) = 1\}).$$

Kürzer kann man das als  $\mathbb{P}_{\theta_0^n}(\{d = 1\})$  schreiben, das ist gerade der Erwartungswert  $\mathbb{E}_{\theta_0^n}(d)$  der Zufallsvariablen  $d$  auf dem Raum  $(\Omega^n, \mathcal{E}_n, \mathbb{P}_{\theta_0^n})$ . Ganz analog kann man einsehen, dass für  $\theta_0 \in \Theta_1$  die Zahl  $1 - \mathbb{E}_{\theta_0^n}(d)$  die Wahrscheinlichkeit für einen Fehler zweiter Art ist. Das führt zu der folgenden

**Definition 11.2.1.** Ist  $d : \Omega_n \rightarrow \{0, 1\}$  ein Test, so heißt die auf  $\Theta$  durch  $G_d(\theta) := \mathbb{E}_{\theta^n}(d)$  definierte Funktion die zu  $d$  gehörige Gütfunktion.

Aus der vorstehenden Diskussion ist klar, dass man nach Tests  $d$  suchen sollte, für die  $G_d$  auf  $\Theta_0$  so nahe wie möglich bei Null und auf  $\Theta_1$  so nahe wie möglich bei Eins ist. Auch ist damit offensichtlich, dass man einen Test  $\tilde{d}$  besser als einen Test  $d$  ansehen wird, wenn  $G_{\tilde{d}} \leq G_d$  auf  $\Theta_0$  und  $G_{\tilde{d}} \geq G_d$  auf  $\Theta_1$  gilt.

Zur Illustration der neu eingeführten Begriffe wird jetzt ein einfaches Beispiel ausführlich diskutiert. Hier die Ausgangssituation:

Im Baumarkt gibt es schon im Januar Tulpenzwiebeln, zwei Mischungen sind im Angebot. Bei Variante 1 kauft man 10 rote Tulpen,

---

<sup>2)</sup>Die Bezeichnung „ $d$ “ der Funktion soll an „decision“ (Entscheidung) erinnern.

20 weiße Tulpen und 70 gelbe Tulpen, bei Variante 2 sind es dagegen 55 rote, 35 weiße und 10 gelbe. Leider sieht man den Zwiebeln die Farbe der zukünftigen Blüte nicht an.

Sie haben eine Packung gekauft, sich aber nicht gemerkt, aus welchem Korb Sie die genommen haben. Und der Zettel mit der Beschreibung ist unterwegs verloren gegangen.

Ihre Nullhypothese lautet: „Es handelt sich um Variante 1“. Zum Test bringen Sie eine Zwiebel im Schnelldurchgang im Gewächshaus zur Blüte. Welche Testfunktion sollte man sinnvollerweise auswählen?

Qualitativ ist zum Beispiel offensichtlich, dass das Ergebnis „gelb“ eher für die Nullhypothese spricht als das Ergebnis „rot“. Um alles quantitativ behandeln zu können, stellen wir die möglichen Testfunktionen systematisch zusammen. Dazu reicht es, den kritischen Bereich (auf dem  $d = 1$  ist, in dem die Nullhypothese also abgelehnt wird) anzugeben. Nachstehend sind alle Möglichkeiten für kritische Bereiche zusammen mit den Wahrscheinlichkeiten für Fehler erster und zweiter Art<sup>3)</sup> zusammengestellt. Bei uns ist  $\Omega = \{\text{rot, weiß, gelb}\}$ , und  $n$  ist gleich 1, da wir nur einmal die Blütenfarbe testen. Folglich gibt es so viele Tests, wie man Teilmengen von  $\Omega^1$  als kritischen Bereich auswählen kann. Das sind  $2^3 = 8$  Möglichkeiten.

| Nummer | kritischer Bereich | W. für Fehler 1. Art | W. für Fehler 2. Art |
|--------|--------------------|----------------------|----------------------|
| 1      | $\emptyset$        | 0                    | 1                    |
| 2      | {rot}              | 0.1                  | 0.45                 |
| 3      | {weiß}             | 0.2                  | 0.65                 |
| 4      | {gelb}             | 0.7                  | 0.9                  |
| 5      | {rot, weiß}        | 0.3                  | 0.1                  |
| 6      | {rot, gelb}        | 0.8                  | 0.35                 |
| 7      | {weiß, gelb}       | 0.9                  | 0.55                 |
| 8      | {rot, weiß, gelb}  | 1                    | 0.                   |

Zur Erläuterung dieser Tabelle betrachten wir als Beispiel die Rechnung zu Test Nummer 3 mit dem kritischen Bereich {weiß}. Wenn  $H_0$  richtig ist, wir es also mit Variante 1 zu tun haben, so entscheiden wir uns genau dann für Variante 2 (machen also einen Fehler erster Art), wenn wir die weiße Tulpenzwiebel gewählt hatten. Das passiert mit Wahrscheinlichkeit 0.2, denn 20 der 100 Tulpen sind weiß.

Liegt dagegen Variante 2 vor, so gibt es dann eine falsche Entscheidung (also einen Fehler zweiter Art), wenn die Tulpe rot oder gelb war. Das ist mit  $0.55 + 0.10 = 0.65$  Wahrscheinlichkeit zu erwarten.

Hier wird noch einmal klar, dass die Summe der Wahrscheinlichkeiten für die Fehler erster und zweiter im Allgemeinen nicht Eins ist: Es werden zwar Wahrscheinlichkeiten für komplementäre Ereignisse addiert (kritischer Bereich und Komplement), die Wahrscheinlichkeitsmaße sind aber verschieden.

Nun können Tests mit verschiedenen Eigenschaften leicht bestimmt werden. Zum Beispiel:

<sup>3)</sup>Das sind die Zahlen  $G_d(\theta)$  für die Fälle  $\theta = \text{„Variante 1“}$  bzw. die  $1 - G_d(\theta)$  für die  $\theta = \text{„Variante 2“}$ .

- Bei Test 1 macht man garantiert keinen Fehler erster Art, dafür handelt man sich einen maximalen Fehler zweiter Art ein.
- Möchte man die Wahrscheinlichkeit für einen Fehler erster Art durch 0.2 beschränken, so ist Test 2 unter den dafür in Frage kommenden Tests der beste, denn bei ihm ist die Wahrscheinlichkeit für einen Fehler zweiter Art minimal. (Das ist auch plausibel: Bei einer roten Blüte ist kaum zu erwarten, dass Variante 2 vorlag.)
- Die Summe aus den Wahrscheinlichkeiten für die beiden Fehler ist bei Test 5 am kleinsten. Das ist auch derjenige Test, bei dem das Maximum der beiden Wahrscheinlichkeiten minimal ist.

Wie im vorstehenden Beispiel gibt es auch in der allgemeinen Situation stets extreme Tests: Ist  $d$  die konstante Funktionen Null (niemals  $H_0$  ablehnen, der kritische Bereich ist leer), so gibt es nie Fehler erster Art, und bei  $d = 1$  (immer  $H_0$  ablehnen, der kritische Bereich ist  $\Omega^n$ ) werden garantiert Fehler erster Art eintreten. Solche Tests sind sicher nicht sinnvoll<sup>4)</sup>. Da die Wahrscheinlichkeiten für Fehler erster und zweiter Art im Allgemeinen durch keinen noch so geschickt gewählten Test beide gleichzeitig zum Verschwinden gebracht werden können, wird man sich einen sinnvollen Kompromiss überlegen müssen.

Ein wichtiges Beispiel für eine Minimalforderung an Tests ist die Festsetzung einer Schranke für die Wahrscheinlichkeiten eines Fehlers erster Art. Das ist wirklich ein plausibler Ansatz, wenn man etwa an Nullhypotesen des Typs „Dieses Medikament ist noch nicht ausreichend getestet“ denkt. Präzisiert wird das so:

**Definition 11.2.2.** Sei  $\alpha \in ]0, 1[$ . Ein Test  $d$  heißt Test zum Irrtumsniveau  $\alpha$  (oder zum Konfidenzniveau  $1 - \alpha$ ), wenn  $G_d(\theta) \leq \alpha$  für alle  $\theta \in \Theta_0$  gilt, wenn also die Wahrscheinlichkeit für Fehler erster Art gleichmäßig durch  $\alpha$  beschränkt ist.

Trivialerweise ist  $d = 0$  ein derartiger Test. Sinnvoll ist es jedoch eher, unter allen Tests zum Irrtumsniveau  $\alpha$  solche zu finden, bei denen die Wahrscheinlichkeit für Fehler zweiter Art auch klein ist:  $G_d$  soll auf  $\Theta_0$  durch  $\alpha$  beschränkt und auf  $\Theta_1$  „möglichst groß“ sein.

Dieses Problem werden wir in den nächsten Abschnitten wieder aufgreifen. Vorher soll noch kurz ein *anderer Ansatz* besprochen werden.

Wir werden der Einfachheit halber annehmen, dass der Raum  $\Omega$  in unserem statistischen Modell endlich und dass  $\Theta$  zweielementig ist:  $\Theta = \{0, 1\}$ . Das bedeutet, dass auf  $\Omega$  durch ein Wahrscheinlichkeitsmaß  $\mathbb{P}$  zu einem Wahrscheinlichkeitsraum gemacht wird, wobei  $\mathbb{P} \in \{\mathbb{P}_0, \mathbb{P}_1\}$ .  $H_0$  bzw  $H_1$  sollen durch  $\mathbb{P} = \mathbb{P}_0$  bzw.  $\mathbb{P} = \mathbb{P}_1$  definiert sein.

---

<sup>4)</sup>Eine Ausnahme ist die Feuerwehr. Sie hat die Anweisung, mit Sicherheit keine Fehler erster Art zu begehen. Sie fährt immer, wenn sie gerufen wird, auch wenn es sich noch so sehr wie ein Fehlalarm anhört.

Nun soll berücksichtigt werden, dass die vier Möglichkeiten „In Wirklichkeit  $H_0$  oder  $H_1$ ; Entscheidung für  $H_0$  oder  $H_1$ “ unterschiedliche Konsequenzen haben können. Das wird wie folgt quantifiziert:

Bezeichne mit  $L_{ij}$  die Bewertung des Verlustes, wenn in Wirklichkeit  $H_i$  gilt und man sich für  $H_j$  entscheidet ( $i, j = 0, 1$ ). Die  $2 \times 2$ -Matrix  $(L_{ij})_{i,j=0,1}$  heißt dann die *Verlustmatrix*.

Die  $L_{ij}$  dürfen *auch negativ* sein, das kommt dann vor, wenn eine richtige Entscheidung einen Gewinn bringt.

Abstrakt sieht es dann so aus ( $L_{10}$  etwa bezeichnet den Verlust bei einem Fehler zweiter Art):

|                | $H_0$ angenommen | $H_1$ angenommen |
|----------------|------------------|------------------|
| $H_0$ wirklich | $L_{00}$         | $L_{01}$         |
| $H_1$ wirklich | $L_{10}$         | $L_{11}$         |

Je nach Situation kann die Verlustmatrix sehr unterschiedlich aussehen. Für das Feuerwehrbeispiel<sup>5)</sup> etwa könnten die folgenden Zahlen (in Euro) realistisch sein:

|                | $H_0$ angenommen | $H_1$ angenommen |
|----------------|------------------|------------------|
| $H_0$ wirklich | 5000             | 10 000 000       |
| $H_1$ wirklich | 1000             | 0                |

(5000 Euro kostet ein normaler Einsatz, 1000 Euro sind bei einem Fehlalarm anzusetzen usw.)

Wir kehren zu der vor wenigen Zeilen beschriebenen Situation mit einem zweielementigen  $\Theta$  zurück, und wir nehmen an, dass man sich auf eine Verlustmatrix  $(L_{ij})$  geeinigt hat. Welchen Test  $d : \Omega^n \rightarrow \{0, 1\}$  sollte man wählen?

Sei  $d$  so ein Test. Je nachdem, welche der vier möglichen Situationen ( $H_i$  angenommen,  $H_j$  wahr;  $i, j = 0, 1$ ) eingetreten ist, wird das unterschiedliche Kosten verursachen. Definiert man  $\alpha_{ij}^d$  als die Wahrscheinlichkeit, dass man sich unter  $d$  für  $j$  entscheidet, wenn in Wirklichkeit  $H_i$  zutrifft<sup>6)</sup>, so ist

$$R_d(i) := \sum_{j=0,1} L_{ij} \alpha_{ij}^d$$

für  $i = 0, 1$  der zu erwartende Verlust, wenn wirklich  $H_i$  zutrifft.

Auf diese Weise definiert jeder Test  $d$  einen Punkt  $(R_d(0), R_d(1))$  im  $\mathbb{R}^2$ . Wenn  $\Omega$   $r$  Elemente hat, gibt es  $r^n$  Elemente in  $\Omega^n$  und folglich  $2^{(r^n)}$  mögliche

---

<sup>5)</sup>Vgl. Seite 319.

<sup>6)</sup>Es ist also  $\alpha_{ij}^d = \mathbb{P}_i(\{d = j\})$ .

Tests. Die Tupel  $(R_d(0), R_d(1))$  erzeugen also so etwas wie eine Punktwolke  $W$ . Doch welcher Test ist der beste?

Da man von zwei Punkten im  $\mathbb{R}^2$  im Allgemeinen nicht sagen kann, welches der größere ist, ist diese Frage nicht eindeutig zu beantworten. Es folgt die Beschreibung zweier Ansätze, um „bestmöglich“ zu definieren.

### *Bayes-Ansatz*

Beim Bayes-Ansatz nimmt man eine a-priori-Verteilung für das Auftreten von  $H_0$  und  $H_1$  an. Gegeben sind also zwei nichtnegative Zahlen  $p_0$  und  $p_1$  mit  $p_0 + p_1 = 1$ , wobei  $p_i$  die aufgrund vieler Erfahrungen geschätzte Wahrscheinlichkeit für das Eintreten von  $H_i$  ist. Geht man von diesen Wahrscheinlichkeiten aus, so ist

$$V_d := p_0 R_d(0) + p_1 R_d(1)$$

der Erwartungswert des Verlusts, wenn man mit dem Test  $d$  arbeitet.

Um zu entscheiden, welches  $d$  optimal ist, muss nur an die elementare geometrische Tatsache erinnert werden, dass die Menge aller  $(x, y)$ , für die  $p_0x + p_1y = c$  gilt, eine Gerade ist und dass diese Geraden für verschiedenes  $c$  parallel sind. Um eine optimale Lösung zu finden, muss man also das  $c$  in dieser Geradenschar so klein wie möglich wählen, also so, dass es gerade noch die Punktwolke  $W$  trifft. Die  $d$ , für die  $(R_d(0), R_d(1))$  auf der Geraden zum minimalen  $c$  liegt, sind dann bestmöglich und es ist dann  $V_d = c$ .

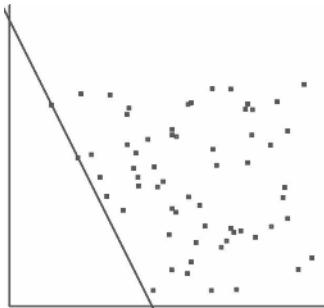


Bild 11.2.1: Optimale Entscheidungen: Bayes-Ansatz.

Durch das Bild wird klar, dass es manchmal mehrere  $d$  geben kann, die zum gleichen minimalen  $V_d$  führen.

Der Spezialfall  $p_0 = p_1 = 0.5$  entspricht übrigens dem Versuch, die Summe  $R_d(0) + R_d(1)$  zu minimieren.

### *Minimax-Lösungen*

Traut man den a-priori-Wahrscheinlichkeiten nicht, kann man auf eine sehr konservative Strategie ausweichen: Man versucht, das  $d$  so zu wählen, dass das

Maximum der Verluste unter Kontrolle bleibt. Genauer: Man betrachtet die Zahlen

$$\max\{R_d(0), R_d(1)\}$$

und bestimmt  $d$  so, dass diese Zahl minimal ist. (Man spricht von einer *Minimax-Lösung*.)

Auch das kann man sich veranschaulichen: Man fixiert zunächst ein „sehr kleines“  $r$  und betrachtet alle  $(x, y)$  mit  $\max\{x, y\} = r$ ; diese Menge soll  $\Delta_r$  heißen.  $\Delta_r$  besteht aus allen Punkten des  $\mathbb{R}^2$ , die auf den Halbgeraden liegen, die von  $(r, r)$  nach links bzw. nach unten gehen.

Lässt man nun  $r$  wachsen, so wird  $\Delta_r$  irgendwann einmal erstmalig die Punktfolge  $W$  berühren. Alle  $d$ , die zu getroffenen  $(R_d(0), R_d(1))$  gehören, sind Minimax-Lösungen.

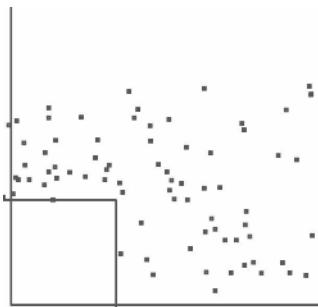


Bild 11.2.2: Optimale Entscheidungen: Minimax-Ansatz.

Das Fazit lautet also: Es ist nicht zu erwarten, dass es so etwas wie „den besten“ Test gibt. Je nach Problemstellung können unterschiedliche Optimalitätsforderungen sinnvoll sein und zu verschiedenen Ergebnissen führen.

### 11.3 Neyman-Pearson-Theorie

Als Vorbereitung zu den Untersuchungen dieses Abschnitts verallgemeinern wir die Definition „Testfunktion“. Bisher hatten wir doch unter einem Test eine Abbildung  $d : \Omega^n \rightarrow \{0, 1\}$  verstanden: Ist  $d(x_1, \dots, x_n) = 0$  bzw. = 1, so lautet die Entscheidung  $H_0$  bzw.  $H_1$ <sup>7)</sup>. Wir benötigen hier die folgende Verallgemeinerung:

**Definition 11.3.1.** *Wir legen ein statistisches Modell zugrunde, und wie zu Beginn des vorigen Abschnitts sei  $\Theta$  disjunkt in  $\Theta_0$  und  $\Theta_1$  zerlegt.*

*Ein stochastischer Test ist eine Zufallsvariable  $d : \Omega^n \rightarrow [0, 1]$ . Die Interpretation: Ist  $d(x_1, \dots, x_n) = p \in [0, 1]$ , so wird ein Bernoulliexperiment mit*

---

<sup>7)</sup>Genauer natürlich: „ $H_0$  wird nicht abgelehnt“ bzw. „ $H_0$  wird abgelehnt“.

Erfolgswahrscheinlichkeit  $p$  durchgeführt. Ist das Ergebnis 1 („Erfolg“) bzw. 0 („Misserfolg“), so lautet die Entscheidung  $H_1$  bzw.  $H_0$ .

Anders ausgedrückt: Man entscheidet sich nicht gleich nach Erhebung der Stichprobe, sondern erst, nachdem der Zufall noch einmal abgefragt wurde.

Bemerkungen: 1. Nimmt  $d$  nur die Werte 0 und 1 an, so geht die neue Definition offensichtlich in die schon bekannte über.

2. Ist  $\lambda \in [0, 1]$  und sind  $d_1, d_2$  stochastische Testfunktionen, so auch die Funktion  $\lambda d_1 + (1-\lambda)d_2$ . Die Menge der stochastischen Testfunktionen ist also *konvex*; für die 0-1-wertigen Tests gilt diese Aussage nicht.

3. Auch für stochastische Testfunktionen  $d$  lässt sich die Gütefunktion  $G_d$  wie in Definition 11.2.1 erklären. Wie in Abschnitt 11.2 sind Funktionen  $d$  interessant, für die  $G_d$  nahe bei Null auf  $\Theta_0$  und nahe bei Eins auf  $\Theta_1$  liegt.

4. Durchläuft  $d$  alle stochastischen Tests, so ist die Menge der  $(R_d(0), R_d(1))$ , die wir am Ende von Abschnitt 11.2 eingeführt haben, nicht mehr eine Punktfolge, sondern eine konvexe Teilmenge des  $\mathbb{R}^2$ . Genauer: Diese Menge ist die konvexe Hülle der vorher betrachteten Punktfolge. Daraus folgt dann, dass es für den Bayes-Ansatz ausreicht, die Punktfolge zu betrachten (denn eine lineare Abbildung nimmt das Minimum auf einer Ecke an), dass aber beim Minimax-Ansatz optimale Lösungen eventuell nur durch stochastische Tests realisiert werden können.

In diesem Abschnitt soll gezeigt werden, wie man im Fall einer Alternativenentscheidung (d.h. im Fall  $\mathbb{P} \in \{\mathbb{P}_0, \mathbb{P}_1\}$ ) einen optimalen stochastischen Test zu einem vorgegebenen Irrtumsniveau  $\alpha$  finden kann. Ohne Einschränkung bestehe die Stichprobe nur aus einem Element, das lässt sich durch Übergang von  $\Omega$  zu  $\Omega^n$  stets erreichen. Wir behandeln *zunächst den diskreten Fall* und diskutieren dann die Modifikationen, die im Fall von Maßen mit Dichten notwendig sind.

### Der Fall diskreter Räume

$\alpha$  sei vorgelegt. Wir wollen einen Test zur Irrtumswahrscheinlichkeit  $\alpha$  so konstruieren, dass die Wahrscheinlichkeit für einen Fehler zweiter Art so klein wie möglich ist.

Ein Test  $d$  – stochastische Tests werden gleich ins Spiel kommen – ist doch dadurch definiert, dass man sagt, wie  $K = \{d = 1\}$  (der kritische Bereich) aussehen soll. Die Bedingung „Wahrscheinlichkeit für Fehler erster Art kleiner gleich  $\alpha$ “ ist erfüllt, wenn  $\mathbb{P}_0(K) = \sum_{x \in K} \mathbb{P}_0(\{x\}) \leq \alpha$  gilt. Gleichzeitig soll  $\mathbb{P}_1(\Omega \setminus K) = \sum_{x \notin K} \mathbb{P}_1(\{x\})$ , die Wahrscheinlichkeit für Fehler zweiter Art, so klein wie möglich sein, was gleichwertig dazu ist, dass  $\mathbb{P}_1(K) = \sum_{x \in K} \mathbb{P}_1(\{x\})$  so groß wie möglich ist.

Es ist daher plausibel, dass man in  $K$  diejenigen  $x \in \Omega$  sammelt, für die der Quotient  $\mathbb{P}_1(\{x\})/\mathbb{P}_0(\{x\})$  möglichst groß ist. Und zwar so viele  $x$ , dass gerade noch die Bedingung  $\sum_{x \in K} \mathbb{P}_0(\{x\}) \leq \alpha$  erfüllt ist.

Es kann sein, dass man bei diesem „Einsammeln“ mit  $\sum_{x \in K} \mathbb{P}_0(\{x\})$  den Wert  $\alpha$  nicht genau trifft. Das wird auf elegante Weise dadurch gelöst, dass man auch stochastische Tests zulässt.

### Möglichst viele Äpfel!

Formal erinnert dieses Problem an die folgende Situation. Jemand möchte auf dem Markt möglichst viele Äpfel kaufen, er/sie hat  $E$  Euro zur Verfügung. Dann ist es doch sinnvoll, zunächst alle Äpfel an dem Stand zu kaufen, wo sie am billigsten sind, dann alle Äpfel dort, wo der nächsthöhere Preis gefordert wird, und so weiter: So lange, bis  $E$  Euro ausgegeben sind.

Dabei kann es natürlich sein, dass man beim letzten Einkauf nicht alle Äpfel kauft, die angeboten werden. Das entspricht in unserem Fall der Situation, dass nur ein stochastischer Test das optimale Resultat liefert.

Diese Idee soll nun präzisiert werden:

**Definition 11.3.2.** *Mit den vorstehenden Bezeichnungen definieren wir den Likelihood-Quotienten  $R : \Omega \rightarrow \mathbb{R}$  durch  $R(x) := \mathbb{P}_1(\{x\})/\mathbb{P}_0(\{x\})$  für  $x \in \Omega$ <sup>8)</sup>.*

*Ein stochastischer Test  $d$  heißt ein Neyman-Pearson-Test zum Niveau  $\alpha$ , falls gilt:*

- (i) *Der Erwartungswert von  $d$  unter  $\mathbb{P}_0$  ist  $\alpha$ . Insbesondere ist  $d$  also ein Test zum Irrtumsniveau  $\alpha$ .*
- (ii) *Es gibt eine Zahl  $c$ , so dass  $d(x) = 1$  (bzw. gleich 0) ist für  $R(x) > c$  (bzw.  $< c$ ). Auf der Menge  $\{R = c\}$  darf  $d$  beliebige Werte haben.*

Anders ausgedrückt heißt das: Der kritische Bereich enthält alle  $x$ , für die  $R > c$  ist, aber kein  $x$  mit  $R(x) < c$ . Die Bedingung, dass  $d$  ein Test zum Irrtumsniveau  $\alpha$  ist, wird dadurch erreicht, dass  $d$  auf  $\{R = c\}$  geeignete Werte in  $]0, 1[$  annimmt: Nur dort ist  $d$  eventuell wirklich „stochastisch“.

Wir werden nun beweisen, dass solche Tests in einem präzisierbaren Sinn optimal sind:

**Satz 11.3.3.** *Es gibt einen Neyman-Pearson-Test zum Niveau  $\alpha$ , und jeder derartige Test ist bestmöglich im folgenden Sinn: Andere stochastische Tests zum Niveau  $\alpha$  führen zu einem nicht kleineren Fehler zweiter Art. Außerdem gilt: Jeder beste Test ist ein Neyman-Pearson-Test.*

**Beweis:** Wir stellen uns die Elemente von  $\Omega = \{x_1, \dots\}$  so sortiert vor, dass  $R$  monoton fällt. Wir wählen dann ein  $c$ , so dass  $\mathbb{P}_0(\{x \mid R(x) > c\}) < \alpha$  und  $\mathbb{P}_0(\{x \mid R(x) \geq c\}) \geq \alpha$ . So ein  $c$  existiert, denn  $\sum_{i=1}^k \mathbb{P}(\{x_i\})$  wächst monoton in  $k$ <sup>9)</sup>.

---

<sup>8)</sup>Wir werden voraussetzen, dass alle  $\mathbb{P}_0(\{x\})$  und alle  $\mathbb{P}_1(\{x\})$  von Null verschieden sind. Es ist nämlich klar, dass die  $x$  mit  $\mathbb{P}_0(\{x\}) = 0$  zu  $K$  und die  $x$  mit  $\mathbb{P}_1(\{x\}) = 0$  zu  $\Omega \setminus K$  gehören müssen.

<sup>9)</sup>Der Fall  $k = 0$  ist zugelassen: Dann ist schon  $\mathbb{P}(\{x_1\}) \geq \alpha$ .

Definiere  $\Delta := \{R = c\}$ .  $d$  ist durch die Werte 1 (auf  $\{R > c\}$ ), 0 (auf  $\{R < c\}$ ) und  $\gamma$  (auf  $\Delta$ ) definiert. Dabei ist

$$\gamma := \frac{\alpha - \mathbb{P}_0(\{R > c\})}{\mathbb{P}_0(\Delta)}.$$

$d$  ist dann ein Neyman-Pearson-Test zum Niveau  $\alpha$ , denn

$$\mathbb{E}_{\mathbb{P}_0}(d) = \gamma \mathbb{P}_0(\Delta) + \mathbb{P}_0(R > c) = \alpha.$$

Sei nun  $\psi$  ein beliebiger Test zum Niveau  $\alpha$  und  $d$  ein entsprechender Neyman-Pearson-Test. Wir wollen zeigen, dass  $d$  nicht schlechter als  $\psi$  ist.

Setze  $a_i := \psi(x_i)$ . Mal angenommen, es wäre  $a_1 < 1$ . Wähle ein beliebiges  $j > 1$  und betrachte für „kleines“  $\varepsilon > 0$  einen Test  $\psi'$ : Der ist auf allen  $i \notin \{1, j\}$  wie  $\psi$  definiert, bei 1 hat er den Wert  $a_1 + \varepsilon$  und bei  $j$  den Wert  $a_j - \epsilon(\mathbb{P}_0(x_1)/\mathbb{P}_0(x_j))$ . (Die Zahl  $\varepsilon$  sollte so klein sein, dass  $a_1 + \varepsilon \leq 1$  gilt.) Auch  $\psi'$  ist ein Test zum Niveau  $\alpha$ , denn für die Wahrscheinlichkeit des Fehlers erster Art gilt

$$\begin{aligned} \mathbb{E}_{\mathbb{P}_0}(\psi') &= \sum_{i=1}^n \psi'(x_i) \mathbb{P}_0(\{x_i\}) \\ &= (a_1 + \varepsilon) \mathbb{P}_0(\{x_1\}) + \left( a_j - \varepsilon \frac{\mathbb{P}_0(\{x_1\})}{\mathbb{P}_0(\{x_j\})} \right) \mathbb{P}_0(\{x_j\}) + \\ &\quad + \sum_{i \notin \{1, j\}} \psi'(x_i) \mathbb{P}_0(\{x_i\}) \\ &= \sum_{i=1}^n a_i \mathbb{P}_0(\{x_i\}) \\ &= \alpha. \end{aligned}$$

Der Fehler zweiter Art ist aber eher kleiner geworden, das folgt daraus, dass  $R$  monoton fällt.

Kurz: Man kann von  $\psi$  zu einem nicht schlechteren Test übergehen, für den  $a_1 = 1$  ist. Genau so zeigt man, dass man  $a_2 = 1$ ,  $a_3 = 1$  usw. erreichen kann, ohne den Test zu verschlechtern, und zwar so lange, bis das  $\mathbb{P}_0$ -Maß von  $\{x_1, \dots, x_n\}$  den Wert  $\alpha$  nicht übersteigt. Ganz analog kann man im Fall  $a_n > 0$  einen besseren Test finden, indem man zu  $a_n = 0$  übergeht und  $d$  bei einem  $x_i$  mit  $i < n$  entsprechend erhöht.

Ein Test wird also besser, wenn man die  $x_i$  mit kleinem  $i$  nach Möglichkeit auf Eins und die  $x_i$  mit großem  $i$  nach Möglichkeit auf Null abbildet. Das geht so lange, bis sich die beiden Verbesserungsstrategien in der Mitte auf einer Menge  $\{R = c\}$  treffen. Damit liegt dann aber ein Neyman-Pearson-Test vor.

Insbesondere müssen Tests, die nicht mehr zu verbessern sind, Neymann-Pearson-Tests sein, und damit ist der Satz vollständig bewiesen  $\square$

Für den Beweis war es bequem,  $R$  als monoton fallend vorauszusetzen. Für die allgemeine Situation muss man wie folgt verfahren, um einen optimalen Test zu konstruieren:

- Bestimme die Mengen  $\Delta_b := \{R > b\}$  für  $b \in \mathbb{R}$ . Die Abbildung  $b \mapsto \Delta_b$  ist monoton fallend, für sehr große bzw. sehr kleine  $b$  ist  $\Delta_b = \emptyset$  bzw.  $\Delta_b = \Omega$ .
- Suche ein  $c$  so, dass  $\mathbb{P}_0(\Delta_c) < \alpha$  und  $\mathbb{P}_0(\{R \geq c\}) \geq \alpha$ . So ein  $c$  gibt es immer, es ist  $c = \sup\{b \mid \mathbb{P}_0(\Delta_b) \geq \alpha\} = \inf\{b \mid \mathbb{P}_0(\Delta_b) < \alpha\}$ .
- Definiere  $d$  wie im Beweis des Satzes: Setze also  $d(x) := 1$  für  $x \in \Delta_c$ ,  $d(x) := (\alpha - \mathbb{P}_0(\Delta_c)) / \mathbb{P}_0(\Delta)$  für  $x \in \Delta := \{R = c\}$  und  $d(x) := 0$  für die restlichen  $x$ .

*Beispiel 1:*  $p_0$  und  $p_1$  seien zwei Zahlen in  $]0, 1[$ , es gelte  $p_0 < p_1$ ; man denke an eine Münze, die aus einer von zwei Produktionslinien für gefälschte Münzen stammt, die Wahrscheinlichkeit für „Kopf“ kann  $p_0$  oder  $p_1$  sein. Die Münze wird nun  $n$ -mal geworfen, aus der Anzahl  $k$  der Erfolge soll man sich bei vorgegebenem Konfidenzniveau  $1 - \alpha$  für  $H_0 : p = p_0$  oder  $H_1 : p = p_1$  entscheiden.

Es geht also um zwei Binomialverteilungen auf  $\{0, \dots, n\}$ , die erste hat ihren „Buckel“ links von dem der zweiten. Praktischerweise sind die Quotienten schon sortiert, wir behaupten nämlich:

Die Abbildung  $k \mapsto R(k) := \frac{b(k, n; p_1)}{b(k, n; p_0)}$  ist monoton steigend.

(Zum Beweis betrachte man den Ausdruck  $R(k+1)/R(k)$ . Er ist – wie man leicht zeigt – genau dann  $> 1$ , wenn  $p_0 < p_1$  gilt.) Folglich sind die Mengen  $\Delta_b$  von der Form  $\{k, k+1, \dots, n\}$ . Wir müssen also ein  $k_0$  suchen, so dass gilt: Es ist  $\eta := \sum_{k=k_0}^n b(k, n; p_0) < \alpha$ , aber  $\sum_{k=k_0-1}^n b(k, n; p_0) \geq \alpha$ . Die Menge  $\Delta$  aus dem vorigen Beweis ist hier also die Menge  $\{k_0 - 1\}$ .

Wir definieren dann  $d$  wie folgt:

- Für  $k < k_0 - 1$  ist  $d(k) := 0$ .
- Für  $k \geq k_0$  ist  $d(k) := 1$ .
- $d(k_0 - 1)$  hat den Wert  $(\alpha - \eta) / b(k_0 - 1, n; p_0)$ .

Dann ist  $d$  ein Neyman-Pearson-Test zum Niveau  $\alpha$ : Der Erwartungswert unter  $\mathbb{P}_0$  von  $d$  ist wirklich

$$\begin{aligned} \sum_k d(k) b(k, n; p_0) &= d(k_0 - 1) b(k_0 - 1, n; p_0) + \sum_{k \geq k_0} d(k) b(k, n; p_0) \\ &= (\alpha - \eta) + \eta \\ &= \alpha. \end{aligned}$$

Aus dem Satz folgt, dass dieser Test optimal ist.

Man beachte also: Werden  $k$  Erfolge erzielt und ist  $k$  „klein“ bzw. „groß“, so entscheidet man sich für  $H_0$  (bzw.  $H_1$ ). Es gibt aber einen Grenzfall, nämlich  $k = k_0 - 1$ , da muss dann eine Zufallsentscheidung herangezogen werden.

Ist etwa  $p_0 = 0.5$  und  $p_1$  beliebig mit  $p_1 > 0.5$ , so wird ein Neyman-Pearson-Test  $d$  zum Niveau  $\alpha = 0.125$  für den Fall von  $n = 4$  Versuchen wie folgt gefunden.

Zunächst bestimmen wir das passende  $k_0$ . Wegen  $\sum_{k=4}^4 b(k, 4; 0.5) = 0.0625$  und  $\sum_{k=3}^4 b(k, 4; 0.5) = 0.3125$  ist  $k_0 = 4$  zu setzen. Es ist  $\eta = 0.0625$ , für  $d(3)$  ergibt sich daher der Wert  $(0.125 - 0.0625)/b(3, 4; 0.5) = 0.25$ .

Werden folglich bei 4 Versuchen  $k$  Erfolge beobachtet, so lautet die Empfehlung:

- Für  $k \leq 2$  ist  $d(k) = 0$ : Die Hypothese  $H_0$  wird nicht abgelehnt.
- Ist  $k = 3$ , so wird noch einmal ein Zufallsexperiment gestartet, das mit Wahrscheinlichkeit 0.25 eine 1 liefert. Nur in diesem Fall wird  $H_0$  abgelehnt.
- Bei 4 Erfolgen wird  $H_0$  abgelehnt.

*Das* ist der optimale Test zum Irrtumsniveau  $\alpha = 0.125$ .

*Beispiel 2:* Diesmal betrachten wir zwei Poissonverteilungen zu den Parametern  $\lambda_0$  (das entspricht der Nullhypothese  $H_0$ ) und  $\lambda_1$  (die Alternativhypothese  $H_1$ ). Wir wollen  $\lambda_1 < \lambda_0$  voraussetzen. Wie sieht ein optimaler Test aus, wenn wir  $n$  Abfragen mit den Ergebnissen  $x_1, \dots, x_n$  durchgeführt haben?

In Abschnitt 4.6 auf Seite 152 haben wir nachgewiesen, dass  $s := x_1 + \dots + x_n$  wieder poissonverteilt ist, und zwar unter  $\mathbb{P}_0$  zum Parameter  $n\lambda_0$  und unter  $\mathbb{P}_1$  zum Parameter  $n\lambda_1$ . Der Likelihoodkoeffizient  $R : \mathbb{N}_0 \rightarrow \mathbb{R}$  ist also

$$R(k) = \frac{p(k; n\lambda_1)}{p(k; n\lambda_0)} = \left( \frac{n\lambda_1}{n\lambda_0} \right)^k e^{n(\lambda_0 - \lambda_1)}.$$

Damit ist  $R$  diesmal eine fallende Funktion, und es folgt, dass die  $\Delta_b$  Mengen des Typs  $\{0, 1, \dots, k\}$  sind.

Wähle  $k_0$  so, dass  $\sum_{k=0}^{k_0} p(k; np_0) < \alpha$  und  $\sum_{k=0}^{k_0+1} p(k; np_0) \geq \alpha$  gilt. Wenn dann  $s \leq k_0$  ist, wird  $H_0$  abgelehnt, für  $s > k_0 + 1$  wird  $H_0$  nicht abgelehnt, und im Fall  $s = k_0 + 1$  wird eine Zufallsentscheidung fällig, die mit Wahrscheinlichkeit  $(\alpha - \sum_{k=0}^{k_0} p(k; np_0))/p(k_0 + 1; n\lambda_0)$  zur Ablehnung führt.

### Maße mit Dichten

Wir betrachten nun den Fall, dass  $\Omega$  ein Intervall in  $\mathbb{R}$  ist und die Maße  $\mathbb{P}_0$  und  $\mathbb{P}_1$  durch zwei Dichten  $f_0$  und  $f_1$  gegeben sind. Diesmal ist der Maximum-Likelihood-Quotient die Funktion

$$x \mapsto R(x) := \frac{f_1(x)}{f_0(x)}.$$

Die Definition eines Neyman-Pearson-Tests ist analog zum diskreten Fall:  $d$  heißt *Neyman-Pearson-Test* zum Irrtumsniveau  $\alpha$ , falls  $d$  unter  $\mathbb{P}_0$  den Erwartungswert  $\alpha$  hat und es ein  $c$  so gibt, dass  $d(x) = 1$  (bzw. gleich 0) ist für  $R(x) > c$  (bzw.  $< c$ ); in der Regel wird  $\{R = c\}$  eine Nullmenge sein, so dass es sich um einen deterministischen Test handelt.

Die Aussage des vorigen Satzes 11.3.3 gilt dann analog: Beste Tests sind Neyman-Pearson-Tests und umgekehrt. (Der Beweis im kontinuierlichen Fall kann ähnlich wie im diskreten Fall geführt werden; vgl. das Buch von Georgii, Abschnitt 10.2.)

Zur Konstruktion von  $d$  ist so zu verfahren:

- Bestimme für beliebige  $b \in \mathbb{R}$  die Menge  $\Delta_b := \{R \geq b\}$ .
- Suche ein  $c$  so, dass  $\int_{\Delta_c} f_0(x) dx = \alpha$ .
- Definiere  $d(x) := 1$  für  $x \in \Delta_c$  und  $d(x) := 0$  sonst.

*Beispiel 1:* Gegeben seien zwei Normalverteilungen  $N(a_0, 1)$  und  $N(a_1, 1)$  mit  $a_0 < a_1$ . Der Quotient der Dichtefunktionen, also

$$x \mapsto R(x) := \frac{\exp(-(x - a_1)^2/2)}{\exp(-(x - a_0)^2/2)},$$

stimmt bis auf einen positiven Faktor mit der monoton steigenden Funktion  $x \mapsto \exp((a_1 - a_0)x)$  überein. Die Mengen  $\Delta_b$  sind also von der Form  $[r, \infty[$ . Man muss das  $r$  so wählen, dass  $\mathbb{P}_{N(a_0, 1)}([r, \infty[) = \alpha$ , der Neyman-Pearson-Test ist dann die charakteristische Funktion dieses Intervalls<sup>10)</sup>.

Der so konstruierte Test hängt übrigens gar nicht von  $a_1$  ab, da  $R$  für alle  $a_1 > a_0$  monoton steigt. Die Wahrscheinlichkeit für einen Fehler zweiter Art ist aber umso größer, je näher  $a_1$  an  $a_0$  liegt.

Soll etwa „ $H_0$ : Die Verteilung ist  $N(0, 1)$ “ gegen „ $H_1$ : Die Verteilung ist  $N(3, 1)$ “ zum Irrtumsniveau  $\alpha = 0.02$  getestet werden, so muss man ein  $r$  wählen, so dass  $\mathbb{P}_{N(0, 1)}([r, \infty[) = 0.02$ . Mit Hilfe der Tafel von  $N(0, 1)$  erhält man  $r = 2.06$ . Die optimale Testvorschrift lautet also:  $H_0$  wird genau dann abgelehnt, wenn das Ergebnis der Zufallsabfrage größer oder gleich 2.06 ist.

Der Fehler zweiter Art ist hier

$$\mathbb{P}_{N(3, 1)} ]-\infty, 2.06] = \mathbb{P}_{N(0, 1)} ]-\infty, 2.06 - 3] = 0.1446.$$

---

<sup>10)</sup>Das bedeutet:  $H_0$  wird genau dann abgelehnt, wenn die Abfrage ein Ergebnis in  $[r, \infty[$  liefert.

*Beispiel 2:* Diesmal wollen wir  $N(0, 1)$  (die Nullhypothese) gegen  $N(0, 2)$  testen: Die Varianzen sind also unterschiedlich. Es ist ein Test zu erwarten, der  $H_0$  dann ablehnt, wenn das Ergebnis „zu groß“ ist. Der Likelihoodquotient ist gleich

$$\frac{f_1(x)}{f_0(x)} = \frac{\sqrt{2\pi}}{\sqrt{4\pi}} e^{x^2/2 - x^2/4},$$

das ist bis auf einen positiven Faktor die Funktion  $e^{x^2/4}$ . Folglich sind die Mengen  $\Delta_b$  von der Form  $I_r := ]-\infty, -r] \cup [r, +\infty[$ . Das richtige  $r$  findet man mit Hilfe der Bedingung  $\mathbb{P}_{N(0,1)}(I_r) = \alpha$ , das läuft wegen der Symmetrie der Verteilung auf  $\mathbb{P}_{N(0,1)}([-\infty, r]) = 1 - \alpha/2$  hinaus und kann deswegen leicht abgelesen werden.

Fazit: Lehne  $H_0$  ab, wenn das Ergebnis der Abfrage in  $I_r$  liegt.

*Beispiel 3:* Wie testet man zwei Exponentialverteilungen zu den Parametern  $\lambda_0, \lambda_1$  mit  $\lambda_0 < \lambda_1$  gegeneinander? Diesmal ist  $R(x) = (\lambda_1/\lambda_0)e^{(\lambda_0-\lambda_1)x}$  für  $x \geq 0$ . Das ist eine monoton fallende Funktion, die  $\Delta_b$  sind folglich Intervalle der Form  $[0, r]$ . Damit  $[0, r]$  der kritische Bereich wird, muss  $\lambda_0 \int_0^r e^{-\lambda_0 x} dx = 1 - e^{-r\lambda_0} = \alpha$  gelten, d.h.  $r = -\log(1-\alpha)/\lambda_0$ . Die Nullhypothese wird abgelehnt, wenn die Abfrage kleiner oder gleich  $r$  ausfällt. Das ist plausibel, denn die Verteilung ist für den Parameter  $\lambda_1$  stärker bei Null konzentriert als für  $\lambda_0$ .

## 11.4 Verständnisfragen

### Zu Abschnitt 11.1

#### *Sachfragen*

**S1:** Was bedeutet in der Statistik das Wort „Hypothese“? Wie verhalten sich Nullhypothese und Alternativhypothese zueinander?

**S2:** Was sagt man – etwas vorsichtiger – statt „Die Nullhypothese wird angenommen“?

**S3:** Was ist ein Fehler erster bzw. zweiter Art?

#### *Methodenfragen*

**M1:** Entscheiden können, was zu vorgegebener Nullhypothese die Fehler erster und zweiter Art sind. (Durch Selbstbeobachtung sollte man auch feststellen, dass man täglich viele Male zwischen Fehlern erster und zweiter Art abwägt.)

### Zu Abschnitt 11.2

#### *Sachfragen*

**S1:** Was ist eine Test- bzw. Entscheidungsfunktion?

**S2:** Wie ist die Gütfunktion eines Tests definiert?

**S3:** Was versteht man unter einem Test zum Irrtumsniveau  $\alpha$ ?

**S4:** Was beschreibt eine Verlustmatrix?

**S5:** Wie findet man eine optimale Testfunktion mit dem Bayes-Ansatz?

**S6:** Wie findet man eine optimale Testfunktion mit dem Minimax-Ansatz?

*Methodenfragen*

**M1:** Die Gütfunktion eines Tests interpretieren können.

**M2:** Optimale Tests mit dem Bayes-Ansatz ermitteln können.

**M3:** Optimale Tests mit dem Minimax-Ansatz ermitteln können.

### Zu Abschnitt 11.3

*Sachfragen*

**S1:** Was ist ein stochastischer Test? Inwiefern sind die vorher eingeführten Tests Spezialfälle?

**S2:** Durch welche Bedingungen ist ein Neyman-Pearson-Test definiert?

**S3:** Wie kann man im Fall von nur zwei Alternativen optimale Tests konstruieren?

*Methodenfragen*

**M1:** Optimale Tests mit Hilfe der Neyman-Pearson-Theorie entwickeln können.

## 11.5 Übungsaufgaben

### Zu Abschnitt 11.1

**Ü11.1.1** Was sind die Fehler erster und zweiter Art bei der Nullhypothese „Im nächsten Winter kommt keine Grippewelle. Ich brauche mich also nicht impfen zu lassen.“?

**Ü11.1.2** Es geht um die Hypothese „Am Freitag um Mitternacht kann man in Paris gefahrlos mit der Metro fahren.“ Sollte man sie als Nullhypothese oder als Alternativhypothese betrachten, um in Übereinstimmung mit der üblichen Konvention (der Fehler erster Art ist der schwerwiegendere) zu sein?

**Ü11.1.3** Finden Sie ein Beispiel aus den letzten 24 Stunden Ihres Lebens, bei dem Sie im Zusammenhang mit einer speziellen Nullhypothese abwägen mussten, ob der Fehler erster Art oder zweiter Art die unangenehmeren Konsequenzen hat.

### Zu Abschnitt 11.2

**Ü11.2.1** Es sei  $\Omega = \{1, 2, 3\}$ , und Wahrscheinlichkeitsmaße  $\mathbb{P}_0$  und  $\mathbb{P}_1$  auf  $\Omega$  seien durch die folgende Tabelle gegeben.

| $k$               | 1             | 2              | 3             |
|-------------------|---------------|----------------|---------------|
| $\mathbb{P}_0(k)$ | $\frac{2}{3}$ | $\frac{1}{12}$ | $\frac{1}{4}$ |
| $\mathbb{P}_1(k)$ | $\frac{1}{2}$ | $\frac{1}{3}$  | $\frac{1}{6}$ |

Weiter sei die Verlustmatrix

$$L = \begin{pmatrix} 5 & 7 \\ 10 & 1 \end{pmatrix}$$

vorgelegt.

- a) Skizzieren Sie die Punkte  $(R_d(0), R_d(1))$  für alle Tests  $d : \Omega \rightarrow \{0, 1\}$ .  
 b) Bestimmen Sie die Minimax-Lösung für das Testproblem.

- c) Bestimmen Sie die Bayes-Lösung für das Testproblem zu  $p_0 = p_1 = \frac{1}{2}$ .  
 d) Finden Sie  $p_0$  und  $p_1$ , so dass die Bayes-Lösung nicht eindeutig ist.

**Ü11.2.2** Finden Sie ein Beispiel, in dem eine Testfunktion so gefunden werden kann, dass die Gütfunktion auf  $\Theta_0$  exakt gleich 0 und auf  $\Theta_1$  exakt gleich 1 ist.

**Ü11.2.3** Es sei  $X : \Omega \rightarrow \mathbb{R}$  eine Zufallsvariable mit unbekanntem Erwartungswert  $\mu$  und bekannter Varianz  $\sigma^2$ . Sie wird  $n$ -mal abgefragt, wobei  $n$  „groß“ ist. Entwerfen Sie einen Test für die Nullhypothese „ $\mu \leq \mu_0$ “ zum Niveau  $\alpha$ .

**Ü11.2.4** Es sei die Familie aller Gleichverteilungen auf  $[a - 1, a + 1]$  für alle  $a \in \mathbb{R}$  gegeben. Entwerfen Sie einen Test zum Niveau  $\alpha$  für die Nullhypothese „ $a \leq a_0$ “.

### Zu Abschnitt 11.3

**Ü11.3.1** Hier ist die geometrische Version des Neyman-Pearson-Ergebnisses.

Zu  $\lambda_1, \dots, \lambda_n \in [0, 1]$ ,  $\mu_1, \dots, \mu_n \in [0, 1]$  und  $\alpha \in [0, 1]$  mit

$$\sum_{i=1}^n \lambda_i = \sum_{i=1}^n \mu_i = 1$$

sei das folgende Maximierungsproblem gegeben.

$$\begin{aligned} \text{Maximiere} \quad & \sum_{i=1}^n \mu_i x_i, \\ \text{so dass} \quad & 0 \leq x_i \leq 1 \text{ für alle } 1 \leq i \leq n, \\ & \sum_{i=1}^n \lambda_i x_i = \alpha. \end{aligned}$$

- a) Beweisen Sie die Existenz einer Lösung dieses Problems.  
 b) Zeigen Sie, dass eine Lösung  $x_1, \dots, x_n$  existiert, bei der es höchstens ein  $i$  mit  $x_i \notin \{0, 1\}$  gibt.

**Ü11.3.2** Es sei  $\Omega = \{1, \dots, 10\}$ , und Wahrscheinlichkeitsmaße  $\mathbb{P}_0$  und  $\mathbb{P}_1$  auf  $\Omega$  seien durch die folgende Tabelle gegeben.

| $k$               | 1               | 2               | 3               | 4              | 5              | 6              | 7               | 8              | 9               | 10              |
|-------------------|-----------------|-----------------|-----------------|----------------|----------------|----------------|-----------------|----------------|-----------------|-----------------|
| $\mathbb{P}_0(k)$ | $\frac{1}{2}$   | $\frac{1}{10}$  | $\frac{1}{5}$   | $\frac{1}{20}$ | $\frac{1}{25}$ | $\frac{1}{25}$ | $\frac{3}{100}$ | $\frac{1}{50}$ | $\frac{1}{100}$ | $\frac{1}{100}$ |
| $\mathbb{P}_1(k)$ | $\frac{1}{100}$ | $\frac{3}{100}$ | $\frac{3}{100}$ | $\frac{1}{25}$ | $\frac{1}{20}$ | $\frac{2}{25}$ | $\frac{3}{50}$  | $\frac{1}{10}$ | $\frac{1}{5}$   | $\frac{2}{5}$   |

Finden Sie alle Neyman-Pearson-Tests zum Niveau  $\alpha = 0.1$ .

**Ü11.3.3** Wir betrachten das statistische Modell aller Poisson-Verteilungen zu Parametern  $\lambda \in \mathbb{R}$ . Geben Sie einen Neyman-Pearson-Test

$$d : \{0, 1, 2, \dots\}^3 \rightarrow [0, 1]$$

zum Niveau  $\alpha = 0.1$  für die Nullhypothese „ $\lambda \leq 0.3$ “ an.

**Ü11.3.4** Es sei  $\Omega = \{0, \dots, 25\}$ , und  $\mathbb{P}_0$  sowie  $\mathbb{P}_1$  seien die folgenden Wahrscheinlichkeitsmaße:

$$\mathbb{P}_0(k) = b(25, k; 3/5), \quad \mathbb{P}_1(k) = b(25, k; 2/5).$$

Bestimmen Sie einen Neyman-Pearson-Test zum Niveau  $\alpha = 0.05$ .

**Ü11.3.5** Eine Münze sei vorgelegt, die Nullhypothese besagt, dass die Wahrscheinlichkeit  $p$  für „Kopf“ gleich 0.4 ist. Sie wird fünf Mal geworfen. Konzipieren Sie einen Neyman-Pearson-Test von  $H_0$  gegen  $H_1 : p = 0.7$  zum Niveau  $\alpha = 0.1$ .

## Kapitel 12

# Nichtparametrische Statistik

In den vorigen Kapiteln wurden Verfahren besprochen, die dazu dienten, Zahlen oder Vektoren zu schätzen oder damit zusammenhängende Vermutungen zu testen. Man spricht dann von *parametrischen Verfahren*. Viele wichtige Fragen sind aber dadurch nicht abgedeckt: Sind gewisse Zufallsvariable unabhängig? Verhält sich ein vorgelegter Würfel wirklich so wie behauptet? ...

In diesem Kapitel wollen wir kurz auf einige dieser Probleme eingehen. Dabei steht die Beschreibung der Verfahren im Vordergrund. Die zugehörigen Beweise sind meist recht technisch, in einigen Fällen wird es nur Skizzen geben.

*Die grundlegende Idee* ist dabei immer die gleiche. *Erstens* fasst man das, was zu untersuchen ist, als Nullhypothese auf. Und dann sucht man *zweitens* eine Statistik  $T$  – also eine Funktion, die der konkret beobachteten Stichprobe  $x_1, \dots, x_n$  eine Zahl  $T(x_1, \dots, x_n)$  zuordnet – die eine bekannte Verteilung hat, wenn die Nullhypothese wirklich zutrifft.

Dann kann man *drittens* leicht Tests zu einem vorgegebenen Irrtumsniveau  $\alpha$  entwerfen. Man muss ja nur eine möglichst kleine Teilmenge  $\Delta$  von  $\mathbb{R}$  so finden, dass unter der Annahme von  $H_0$  die Statistik  $T$  mit Wahrscheinlichkeit  $1 - \alpha$  in  $\Delta$  liegt. Lehnt man  $H_0$  genau dann ab, wenn  $T(x_1, \dots, x_n)$  nicht in  $\Delta$  liegt, so ist die Wahrscheinlichkeit für einen Fehler erster Art durch  $\alpha$  beschränkt.

Das Problem wird im Wesentlichen darin bestehen, eine geeignete Abbildung  $T$  zu finden.

In *Abschnitt 12.1* wird die Hypothese getestet, ob eine vorgelegte Wahrscheinlichkeit auf einem endlichen  $\Omega$  mit einer speziellen, genau beschriebenen übereinstimmt. Überraschenderweise führt das wieder auf die  $\chi^2$ -Verteilung, der

zugehörige Test heißt der  $\chi^2$ -*Anpassungstest*. Danach, in *Abschnitt 12.2*, wird es um  $\chi^2$ -*Tests auf Unabhängigkeit* gehen. Damit kann – einheitlich für alle Verteilungen – getestet werden, ob zwei Zufallsvariable, die jeweils nur endlich viele Werte annehmen können, unabhängig sind.

*Abschnitt 12.3* ist dann den *Rangtests* gewidmet. Die Hypothese, ob zwei Wahrscheinlichkeitsmaße auf  $\mathbb{R}$  gleich sind, kann dabei dadurch getestet werden, dass die Ränge von Stichproben miteinander verglichen werden.

Es folgt dann noch die Besprechung des *Kolmogoroff-Smirnoff-Tests*: Liefern Stichproben einen Anhaltspunkt dafür, ob eine bestimmte kontinuierliche Verteilung vorliegt? Das wird in *Abschnitt 12.4* diskutiert werden.

Am Ende des Kapitels findet man in den Abschnitten 12.5 und 12.6 Verständnisfragen und Übungsaufgaben.

## 12.1 Der $\chi^2$ -Anpassungstest

### Das Problem

Angenommen, jemand behauptet, dass ein Würfel fair ist oder auf ganz bestimmte Weise gefälscht. Wie könnte man das testen? Man wird ihn „genügend oft“ werfen und die Ergebnisse zählen. Sind die empirischen Häufigkeiten „nahe genug“ bei den behaupteten, ist alles in Ordnung, andernfalls sind Zweifel angebracht. Doch wie sollte man „nahe genug“ quantifizieren, um das testen zu können?

### Die Lösung

Sei ein Wahrscheinlichkeitsraum auf  $\{1, \dots, s\}$  durch Vorgabe der Zahlen  $p_1, \dots, p_s > 0$  mit  $p_1 + \dots + p_s = 1$  definiert. Wir fragen ihn  $n$ -mal ab, wobei  $n$  groß gegen  $s$  sein soll:  $h_n(1)$ -mal erscheint die 1,  $h_n(2)$ -mal die 2 usw. Es ist also  $h_n(1) + \dots + h_n(s) = n$ , und es ist zu erwarten, dass die  $h_n(i)/n$  in der Nähe von  $p_i$  sind: Abweichungen würden ein Indiz dafür sein, dass die Stichprobe von einem anderen Wahrscheinlichkeitsraum gezogen wurde. Die richtige Größe, um das zu testen, findet man im nachstehenden

**Satz 12.1.1.** ( $\chi^2$ -*Anpassungstest von Pearson, 1900<sup>1)</sup>)  
Definiere eine Zufallsvariable  $D$  durch*

$$D := n \sum_{i=1}^s p_i \left( \frac{h_n(i)/n}{p_i} - 1 \right)^2.$$

Dann ist  $D$  für große  $n$  näherungsweise  $\chi_{s-1}^2$ -verteilt.

Wie nutzt man das aus? Um zum Irrtumsniveau  $\alpha$  zu testen, ob die Stichprobe  $h_1, \dots, h_n$  mit den Wahrscheinlichkeiten  $p_1, \dots, p_s$  erzeugt wurde, muss man nur aus einer Tafel<sup>2)</sup> ein Intervall  $[r_1, r_2]$

<sup>1)</sup>Von K. Pearson, Vater von E. Pearson aus dem Neyman-Pearson-Lemma.

<sup>2)</sup>Zum Beispiel der im Anhang auf Seite 366.

mit  $\mathbb{P}_{\chi_{s-1}^2}([r_1, r_2]) = 1 - \alpha$  ablesen. Aus den  $h_1, \dots, h_n$  wird  $D$  berechnet, und die Nullhypothese wird genau dann abgelehnt, wenn  $D$  nicht in  $[r_1, r_2]$  liegt.

*Beweisskizze:* Schritt 1: In den Ergänzungen zu Kapitel 5, in Abschnitt 5.5, hatten wir einige Tatsachen über die *Multinomialverteilung* zusammengestellt. Für die hier zu behandelnde Situation folgt: Macht man  $n$  Experimente, so ist die Wahrscheinlichkeit dafür, dass es genau  $h_n(i)$  Mal das Ergebnis  $i$  gab, gleich  $M(h_n(1), \dots, h_n(s); p_1, \dots, p_s; n)$ , wobei

$$M(j_1, \dots, j_s; p_1, \dots, p_s; n) := \frac{n!}{j_1! \cdots j_s!} p_1^{j_1} \cdots p_s^{j_s}.$$

Schritt 2: Nun betrachten wir unabhängige Zufallsvariable  $S_1, \dots, S_s$ , dabei soll  $S_i$  poissonverteilt zum Parameter  $np_i$  sein. Mit  $N$  bezeichnen wir die Summe:  $N := S_1 + \cdots + S_s$ . Da Summen von unabhängigen Poissonverteilungen wieder poissonverteilt sind<sup>3)</sup>, ist  $N$  zum Parameter  $n$  poissonverteilt.

Wenn man das weiß, kann man leicht die bedingte Wahrscheinlichkeit

$$\mathbb{P}(S_i = j_i \text{ für } i = 1, \dots, s \mid N = n)$$

ausrechnen, sie stimmt mit  $M(j_1, \dots, j_s; p_1, \dots, p_s; n)$  überein<sup>4)</sup>. Falls man also den Ausgang einer  $n$ -maligen Abfrage aus unserem Wahrscheinlichkeitsraum auf eine originelle Weise simulieren möchte, so könnte man unabhängige poissonverteilte Zufallsvariable (zu den Parametern  $np_1, \dots, np_s$ ) abfragen, und zwar so oft, bis diese  $s$ -fache Abfrage zur Gesamtsumme  $n$  führt. Die Ausgaben können dann mit gutem Gewissen als  $h_n(1), \dots, h_n(s)$  verwendet werden.

Ob das wirklich empfehlenswert ist, ist allerdings fraglich. Einerseits muss man zwar nur  $s$  Mal den Zufall bemühen, doch andererseits kann man die Simulation nur dann verwenden, wenn die Summe der Ergebnisse gleich  $n$  ist.

Unser Interesse an diesem Zugang ist eher theoretisch begründet. Eine zu  $np$  poissonverteilte Zufallsvariable  $S$  kann doch als Summe von  $n$  unabhängigen Poisson- $p$ -verteilten Variablen aufgefasst werden (vgl. wieder Seite 152). Folglich sollte aufgrund des zentralen Grenzwertsatzes  $(S - np)/\sqrt{np}$  approximativ standard-normalverteilt sein. (Es ist nur zu beachten, dass eine Poisson- $\lambda$ -verteilte Variable Erwartungswert und Varianz  $\lambda$  hat.)

Schritt 3: Definiere  $S_i^* := (S_i - np_i)/\sqrt{np_i}$  für  $i = 1, \dots, s$ . Dann wird die Summe der Quadrate approximativ eine Quadratsumme von  $n$  standard-Normalverteilungen sein, dafür ist also (in etwa) eine  $\chi_s^2$ -Verteilung zu erwarten. Damit wir für die  $S_i$  die  $h_n(i)$  einsetzen dürfen, muss noch die Nebenbedingung  $N = n$  berücksichtigt werden. Das kann so umformuliert werden: Ist  $\phi : \mathbb{R}^n \rightarrow \mathbb{R}$  die lineare Abbildung  $(x_1, \dots, x_n) \mapsto \sum_{i=1}^s x_i \sqrt{np_i}$  und definiert man  $h_n^*(i) := (h_n(i) - np_i)/\sqrt{np_i}$  für  $i = 1, \dots, s$ , so liegt der Vektor  $(h_n^*(1), \dots, h_n^*(s))$  im Kern von  $\phi$ , also in einem 1-codimensionalen Unterraum des  $\mathbb{R}^n$ .

<sup>3)</sup>Vgl. Seite 152.

<sup>4)</sup>Vgl. Übungsaufgabe 12.1.1.

Zusammen: Der Vektor  $(S_1^*, \dots, S_s^*)$  ist so verteilt wie  $s$  unabhängige  $N(0, 1)$ -Verteilungen, und die  $(h_n^*(1), \dots, h_n^*(s))$  bestehen aus denjenigen  $(S_1^*, \dots, S_s^*)$ , die in einem 1-codimensionalen Unterraum liegen. Deswegen ist es nicht überraschend, dass sich die Quadratsumme der  $h_i^*$  so verhält wie die Summe aus  $s - 1$  unabhängigen  $N(0, 1)$ -Verteilungen. So ergibt sich eine  $\chi_{s-1}^2$ -Verteilung. (Das ist natürlich noch kein vollständiger Beweis, aber eine Präzisierung des hier erforderlichen Begriffs „bedingte Verteilung“ würde den Rahmen des Buches sprengen.)

*Schritt 4: Finale.* Nun müssen wir nur noch rechnen. Es wurde schon begründet, dass die Summe der quadrierten  $h_i^*$  näherungsweise  $\chi_{s-1}^2$ -verteilt ist. Und diese Summe ist gleich der Statistik  $D$  aus Satz 12.1.1:

$$\begin{aligned}\sum (h_n^*(i))^2 &= \sum \frac{(h_n(i) - np_i)^2}{np_i} \\ &= n \sum p_i \left( \frac{h_n(i)/n}{p_i} - 1 \right)^2 \\ &= D.\end{aligned}$$

Damit ist alles gezeigt. □

### Beispiele

1. In den Lehrbüchern gibt es ein beliebtes Beispiel: das Mendelsche Kreuzungs-experiment. Es stellt sich nämlich heraus, dass der  $\chi^2$ -Test, angewandt auf die Mendelschen Zahlen, eigentlich viel zu perfekt ausfällt. Zwischen den Zeilen steht dann immer der Verdacht, dass Mendel die Zahlen vielleicht ein wenig manipuliert hat: corriger la fortune ...

2. Ein Zufallsgenerator, der gleichverteilte Zahlen in  $\{1, 2, 3, 4, 5\}$  ausgeben soll, wird getestet. Er wird 50 Mal abgefragt, und die Anzahlen, mit denen 1, 2, 3, 4, 5 erscheint, seien 12, 13, 10, 8, 7. Ist er zum Niveau  $\alpha = 0.05$  als fair anzusehen?

Für  $D$  ergibt sich der Wert

$$D = 50 \sum_{i=1}^5 \frac{1}{5} \left( \frac{h_n(i)/5}{1/5} - 1 \right)^2 = 2.60.$$

Diese Zahl liegt in dem zu  $\alpha = 0.05$  gehörigen, aus der Tabelle bei  $n = 4$  abgelesenen Konfidenzintervall  $[0.48, 11.14]$ , und deswegen sollte die Hypothese „Der Generator erzeugt gleichverteilte Zahlen“ nicht abgelehnt werden.

3. Nun ist ein anderer Zufallsgenerator zu testen. Es wird behauptet, dass er das gleiche leistet wie der vorstehende. Bei 50 Abfragen treten die Zahlen 1, 2, 3, 4, 5 mit den Häufigkeiten 7, 4, 20, 10, 9 auf. Diesmal ist  $D = 14.60$ , und da diese Zahl nicht in  $[0.48, 11.14]$  liegt, wird die Nullhypothese zum Niveau  $\alpha = 0.05$  abgelehnt. (Sie wäre auch zum Niveau  $\alpha = 0.02$ , nicht aber zum Niveau  $\alpha = 0.01$  abgelehnt worden.)

## 12.2 Der $\chi^2$ -Test auf Unabhängigkeit

### Das Problem

Sind zwei vorgegebene Zufallsvariable  $X, Y$  unabhängig? Wir wollen annehmen, dass  $X$  und  $Y$  jeweils nur endlich viele Werte annehmen: Die Bilder von  $X$  sollen in  $\{1, \dots, k\}$  und die von  $Y$  in  $\{1, \dots, l\}$  liegen. Definiert man dann  $p_i := \mathbb{P}(\{X = i\})$  für  $i = 1, \dots, k$ ,  $q_j := \mathbb{P}(\{Y = j\})$  für  $j = 1, \dots, l$  und  $p_{ij} := \mathbb{P}(\{X = i, Y = j\})$ , so gilt:

- $p_i = \sum_j p_{ij}$  und  $q_j = \sum_i p_{ij}$ .
- $X, Y$  sind genau dann unabhängig, wenn  $p_{ij} = p_i q_j$  für alle  $i, j$  ist.

Es gibt viele wichtige Beispiele, bei denen die Frage nach der Unabhängigkeit interessant ist. Sind die Eigenschaften „Raucher“ und „Geschlecht“ unabhängig (hier ist  $k = l = 2$ )? Liegt Unabhängigkeit vor, wenn man die Gehaltsgruppe und die Uni vergleicht, an der der Abschluss gemacht wurde? Ist die politische Einstellung unabhängig von dem Bezirk, in dem man lebt? ...

Das einzige, was man messen kann, sind die relativen Häufigkeiten in einer Stichprobe. Man nimmt also eine Stichprobe vom Umfang  $n$  und zählt, wie sich die Merkmale verteilen (wie viele weibliche Raucher usw.). Wenn es  $h_{ij}$  Ergebnisse des Typs  $(i, j)$  gibt, so ist

- $\sum_{ij} h_{ij} = n$ ;
- $h_{ij}/n$  eine Schätzung für  $p_{ij}$ ;
- Unabhängigkeit ist dann eine plausible Vermutung, wenn  $h_{ij}/n$  in der Nähe des Produkts aus  $(\sum_j h_{ij})/n$  und  $(\sum_i h_{ij})/n$  liegt.

Wie kann man das richtig quantifizieren?

### Die Lösung

Die Idee des  $\chi^2$ -Tests auf Unabhängigkeit besteht darin, aus den  $h_{ij}$  eine Testgröße zu berechnen, die im Fall der Unabhängigkeit nach einer bekannten Verteilung verteilt ist. Das ist der Inhalt des folgenden Satzes:

**Satz 12.2.1.** *Mit den vorstehenden Bezeichnungen gilt: Setzt man  $h_i^1 := \sum_j h_{ij}$ ,  $h_j^2 := \sum_i h_{ij}$  und*

$$T := \sum_{ij} \frac{(h_{ij} - h_i^1 h_j^2 / n)^2}{h_i^1 h_j^2 / n},$$

*so ist  $T$  näherungsweise  $\chi^2_{(k-1)(l-1)}$ -verteilt.*

*Wie nutzt man das aus?* Wenn man die Unabhängigkeit der Testgrößen zum Irrtumsniveau  $\alpha$  testen möchte, so muss man nur aus

einer Tafel ein Intervall  $[r_1, r_2]$  mit  $\mathbb{P}_{\chi^2_{(k-1)(l-1)}}([r_1, r_2]) = 1 - \alpha$  ablesen. Aus den  $h_{ij}$  wird  $T$  berechnet, und die Nullhypothese „ $X, Y$  unabhängig“ wird genau dann abgelehnt, wenn  $D$  nicht in  $[r_1, r_2]$  liegt.

Der *Beweis* ist ähnlich wie der zum  $\chi^2$ -Anpassungstest: Man überlegt sich, dass die Statistik  $T$  (wenigstens näherungsweise) wie die Summe der Quadrate von  $(k-1)(l-1)$  unabhängigen Standardnormalverteilungen verteilt ist. Der zentrale Grenzwertsatz spielt eine wesentliche Rolle, die technischen Einzelheiten sind recht verwickelt. Man findet sie zum Beispiel im Buch von Georgii.

*Ein Beispiel:*

Wir nehmen an, dass  $X$  und  $Y$  jeweils die Werte 1, 2, 3 annehmen kann. Wir nehmen 90 Stichproben, die Ergebnisse sind in der folgenden Matrix zusammengefasst:

|         | $Y = 1$ | $Y = 2$ | $Y = 3$ |
|---------|---------|---------|---------|
| $X = 1$ | 8       | 12      | 10      |
| $X = 2$ | 14      | 8       | 14      |
| $X = 3$ | 6       | 10      | 8       |

Kann die Unabhängigkeit zum Niveau  $\alpha = 0.02$  abgelehnt werden? Das zugehörige Intervall zu  $\chi^2_{(k-1)(l-1)} = \chi^2_4$  ist gleich  $[0.30, 13.28]$ . Liegt die Statistik  $T$  in diesem Intervall? Die Rechnung ergibt  $T = 7.57$ , die Hypothese kann also nicht abgelehnt werden.

Hätte die Abfrage allerdings

|         | $Y = 1$ | $Y = 2$ | $Y = 3$ |
|---------|---------|---------|---------|
| $X = 1$ | 2       | 18      | 10      |
| $X = 2$ | 2       | 24      | 4       |
| $X = 3$ | 3       | 6       | 10      |

ergeben, wäre das zugehörige  $T$  gleich 14.90, und die Nullhypothese wäre folglich abgelehnt worden.

## 12.3 Rangtests

*Das Problem*

Wie kann man testen, ob zwei Stichproben aus der gleichen Verteilung gezogen wurden? Um entsprechende Tests zu motivieren, betrachten wir vorbereitend die folgende Situation: Es treten bei einem Mathematikwettbewerb zwei Teams an, Team A und Team B. Wenn dann Team A „deutlich besser“ ist, findet es niemand überraschend, dass die Platzierung (von schlecht nach gut) etwa so aussieht:

$B, A, B, B, B, A, B, B, B, B, A, A, A, A, A, A, A$ .

(Ausführlich: Der schlechteste Teilnehmer kommt aus Team  $B$ , der zweitschlechteste aus Team  $A$  usw.) Bei „ausgeglichenen“ Teams würde man eher Ergebnisse des Typs

$$A, B, B, A, B, A, A, B, B, B, A, B, A, A, B, A, A, B$$

erwarten. Kurz: Eine ausgeglichene Rangfolge ist ein Indiz für ein vergleichbares Potenzial.

Um das zu präzisieren, gehen wir von zwei reellwertigen Zufallsvariablen  $X, Y$  aus. Wir wollen gleich Stichproben vergleichen, und wir setzen voraus, dass die induzierten Wahrscheinlichkeiten eine stetige Dichte haben. Dadurch wird sichergestellt, dass nur mit Wahrscheinlichkeit Null eine Abfrage von  $X$  und von  $Y$  zum gleichen Ergebnis führt.

Nun werden Stichproben genommen:  $k$  unabhängige Abfragen  $x_1, \dots, x_k$  von  $X$  und  $l$  unabhängige Abfragen  $y_1, \dots, y_l$  von  $Y$ . Diese  $k + l$  Zahlen sind mit Wahrscheinlichkeit Eins paarweise verschieden und können folglich auf eindeutige Weise der Größe nach sortiert werden. Wir betrachten irgendein  $x_i$ . Die Anzahl der  $y_j$ , die kleiner als  $x_i$  sind, bezeichnen wir mit  $U_i$ , es ist damit  $U_i \in \{0, \dots, l\}$ . Die Summe  $U := U_1 + \dots + U_k$  heißt die  $U$ -Statistik.

Ein Beispiel: Es sei  $k = l = 3$ , die  $x_i$  seien durch 1.1, 4, 7.2 und die  $y_j$  durch 2.4, 3.1, 10 gegeben. Dann ist  $U = 0 + 2 + 2 = 4$ .

Nach Definition liegt  $U$  in  $\{0, 1, \dots, kl\}$ , und es ist plausibel, dass  $U$  „nahe bei 0“ bzw. „nahe bei  $kl$ “ liegen wird, wenn  $X$  „tendenziell viel kleiner“ bzw. „tendenziell viel größer“ als  $Y$  ist.

Wie groß sollte  $U$  sein, wenn  $X$  und  $Y$  die gleiche Verteilung haben? Wenn  $U$  „zu groß“ oder „zu klein“ ist, werden die Verteilungen wohl nicht übereingestimmt haben. Doch was soll das genau bedeuten?

Diese Frage stellt sich bei vielen Problemen aus der Praxis:

- Bleiben Patienten länger gesund, wenn sie mit der Heilmethode  $H_1$  statt mit  $H_2$  behandelt wurden?  $X$  bzw.  $Y$  ist hier die Zeitdauer, bis die Krankheit nach Behandlung gemäß  $H_1$  bzw. gemäß  $H_2$  wieder auftritt.
- Verdienen Männer mehr Geld als Frauen in der Metallindustrie?
- Führt das Schlafmittel  $S_1$  zu einer längeren Nachtruhe als das Schlafmittel  $S_2$ : Man schläft  $X$  Stunden mit  $S_1$  und  $Y$  Stunden mit  $S_2$ .

### *Die Lösung*

Unter der Annahme, dass  $X$  und  $Y$  die gleiche Verteilung haben, kann man die Verteilung von  $U$  berechnen:

**Satz 12.3.1.** *Es seien  $x_1, \dots, x_k$  und  $y_1, \dots, y_l$  Stichproben, die aus einer Verteilung mit einer stetigen Dichte gezogen wurden.  $U$  sei wie vorstehend definiert. Für  $0 \leq m \leq kl$  ist dann die Wahrscheinlichkeit, dass  $U = m$  gilt, durch*

$$\frac{N(m; k, l)}{\binom{n}{k}}$$

gegeben. Dabei ist  $n = k + l$ , und  $N(m; k, l)$  ist die Anzahl der Möglichkeiten, die Zahl  $m$  als

$$m = m_1 + \cdots + m_k \text{ mit } 0 \leq m_1 \leq \cdots \leq m_k \leq l$$

zu schreiben.

Wie nutzt man das aus? Sind  $X, Y$  zwei Zufallsvariable mit stetigen Verteilungen und besagt die Nullhypothese  $H_0$ , dass diese Verteilungen übereinstimmen, so kann man einen Test zum Irrtumsniveau  $\alpha$  dadurch konstruieren, dass  $H_0$  genau dann abgelehnt wird, wenn die  $U$ -Statistik nicht in  $\{m_*, \dots, m^*\}$  liegt. Dabei sind  $m_*, m^*$  natürliche Zahlen mit

$$\sum_{m=m_*}^{m^*} \frac{N(m; k, l)}{\binom{n}{k}} \geq 1 - \alpha.$$

*Beweisskizze:* Da die Verteilung eine stetige Dichte hat, sind die Zahlen  $x_1, \dots, x_k, y_1, \dots, y_l$  paarweise verschieden. Dadurch ist die Rangfolge eindeutig, und wir können annehmen, dass es sich um  $k + l$  unabhängige Abfragen der gleichen Zufallsvariablen  $X$  handelt. Deswegen schreiben wir statt  $x_1, \dots, x_k, y_1, \dots, y_l$  nun  $x_1, \dots, x_{k+l}$ .

Für  $i \neq j$  (mit  $i, j \in \{1, \dots, k+l\}$ ) hat  $x_i < x_j$  die gleiche Wahrscheinlichkeit wie  $x_j < x_i$ : Das folgt aus der Unabhängigkeit der Abfragen. Da jede Permutation der  $x_1, \dots, x_{k+l}$  durch Hintereinanderausführung von solchen Transpositionen erzeugt wird, heißt das, dass alle Permutationen gleichwahrscheinlich sind.

Damit läuft die Frage „Ist für  $x_1, \dots, x_{k+l}$  die  $U$ -Statistik in Bezug auf  $x_1, \dots, x_k$  gleich  $m$ ?“ gleichwertig zu der Frage „Ist für eine zufällig ausgewählte Permutation der Zahlen  $1, \dots, n+k$  die  $U$ -Statistik in Bezug auf  $1, 2, \dots, k$  gleich  $m$ ?“ Dabei haben alle Permutationen die gleiche Wahrscheinlichkeit.

Es liegt also ein Laplace Raum vor, und uns interessiert die Wahrscheinlichkeit des Ereignisses „Die Permutationen mit  $U = m$ “.

Es gibt  $n!$  Permutationen: Diese Zahl steht im Nenner der gesuchten Wahrscheinlichkeit. Im Zähler steht die Anzahl der Permutationen, die zu  $U = m$  führen, wenn wir die  $U$ -Statistik in Bezug auf  $1, 2, \dots, k$  ausrechnen. Wie viele sind das?

Zur Berechnung stellen wir uns  $k$  rote Kugeln vor, die wir vor uns in eine Reihe gelegt haben. Auf wie viele Weisen können wir  $l$  weiße Kugeln so zwischen die roten platzieren, dass links von der ersten  $m_1$  weiße, links von der zweiten (insgesamt)  $m_2$  weiße, ..., links von der  $k$ -ten (insgesamt)  $m_k$  weiße liegen und zusätzlich  $m_1 + \cdots + m_k = m$  gilt? Es gibt offensichtlich so viele Möglichkeiten, wie man  $0 \leq m_1 \leq \cdots \leq m_k \leq l$  mit  $m_1 + \cdots + m_k = m$  wählen kann. Das ist nach Definition die Zahl  $N(m; k, l)$ . Es bleibt noch, die Zahlen  $1, \dots, k$  den roten und die Zahlen  $k+1, \dots, k+l$  den weißen Kugeln zuzuordnen. Das geht auf

$k!$  bzw.  $l!$  verschiedene Weisen. Und so folgt: Genau  $k! l! N(m; k, l)$  Permutationen führen zu  $U = m$  in Bezug auf  $1, \dots, k$ . Für die gesuchte Wahrscheinlichkeit erhält man also

$$\frac{k! l! N(m; k, l)}{n!} = \frac{N(m; k, l)}{\binom{n}{k}}.$$

□

Man beachte, dass die  $U$ -Statistik nicht von der speziellen Verteilung abhängt. Man spricht von einem *Mann-Whitney-Test* oder auch einem *Wilcoxon-Test*, wenn aufgrund dieser Statistik die Hypothese, dass die Verteilungen für beide Stichproben gleich sind, zu einem vorgelegten Irrtumsniveau  $\alpha$  behandelt werden soll.

Als konkretes Beispiel betrachten wir den Fall  $k = l = 4$ . Dann ist  $n = 8$ , und folglich ist  $\binom{n}{k}$  gleich 70. Hier eine Tabelle der relevanten Zahlen:

→  
Programm!

| $m$ | $N(m; 4, 4)$ | $\mathbb{P}(\{U = m\})$ | $\mathbb{P}(\{U \leq m\})$ |
|-----|--------------|-------------------------|----------------------------|
| 0   | 1            | 0.0143                  | 0.0143                     |
| 1   | 1            | 0.0143                  | 0.0286                     |
| 2   | 2            | 0.0286                  | 0.0571                     |
| 3   | 3            | 0.0429                  | 0.1000                     |
| 4   | 5            | 0.0714                  | 0.1714                     |
| 5   | 5            | 0.0714                  | 0.2429                     |
| 6   | 7            | 0.1000                  | 0.3429                     |
| 7   | 7            | 0.1000                  | 0.4429                     |
| 8   | 8            | 0.1143                  | 0.5571                     |
| 9   | 7            | 0.1000                  | 0.6571                     |
| 10  | 7            | 0.1000                  | 0.7571                     |
| 11  | 5            | 0.0714                  | 0.8286                     |
| 12  | 5            | 0.0714                  | 0.9000                     |
| 13  | 3            | 0.0429                  | 0.9429                     |
| 14  | 2            | 0.0286                  | 0.9714                     |
| 15  | 1            | 0.0143                  | 0.9857                     |
| 16  | 1            | 0.0143                  | 1.0000                     |

Damit ist es leicht, Tests zu entwerfen. Wenn man die Nullhypothese zum Niveau  $\alpha$  testen möchte, dass  $X$  und  $Y$  die gleiche Verteilung haben, muss man nur ein Intervall  $I = \{m_*, m_* + 1, \dots, m^*\}$  ablesen, so dass  $\mathbb{P}(\{m \notin I\}) \leq \alpha$  gilt. Für  $\alpha = 0.2$  etwa kann man  $I = \{4, \dots, 12\}$  wählen. Fällt der  $U$ -Wert nicht in dieses Intervall, ist die Nullhypothese abzulehnen.

Um Tabellen wie die vorstehende zu berechnen, ist es nützlich, die Beziehungen zwischen den  $N(m; k, l)$  näher zu untersuchen. Nachdem man die Formel

$$N(m; k, l) = \sum_{j=0}^k N(m - j; j, l - 1)$$

bewiesen hat<sup>5)</sup>, kann man die  $N(m; k, l)$  rekursiv schnell ermitteln: zuerst alle  $N(m; k, l)$  mit  $l = 0$ , dann alle mit  $l = 1$  usw.

Für nicht zu große  $k$  und  $l$  können die hier gebrauchten Tabellen mit dem Computerprogramm auf der zum Buch gehörigen Internetseite berechnet werden. Das ist auch der Grund, warum darauf verzichtet wurde, sie im Anhang aufzunehmen: Sie hätten einen unverhältnismäßig großen Platz eingenommen.

## 12.4 Der Kolmogoroff-Smirnoff-Test

### *Das Problem*

Mit dem  $\chi^2$ -Anpassungstest konnte man die Hypothese behandeln, ob eine Stichprobe aus einer vorgegebenen Verteilung auf einem *endlichen* Wahrscheinlichkeitsraum gezogen wurde. In diesem Abschnitt geht es um die analoge Frage für den Fall von Wahrscheinlichkeitsverteilungen mit Dichten.

Gegeben sei also ein Wahrscheinlichkeitsmaß auf  $\mathbb{R}$ , das durch eine stetige Dichte  $f$  definiert ist. Dann ist die Verteilungsfunktion  $F$  eine differenzierbare Funktion von  $\mathbb{R}$  nach  $\mathbb{R}$ . Die Nullhypothese besteht darin, dass eine Stichprobe  $x_1, \dots, x_n$  von genau diesem Wahrscheinlichkeitsraum gezogen wurde.

Wie kann man das testen?

### *Die Lösung*

Die Idee beim Kolmogoroff-Smirnoff-Test besteht darin, die Stichprobe zur Definition der *empirischen Häufigkeitsverteilung* zu verwenden und die Ablehnung der Hypothese davon abhängig zu machen, wie weit diese von  $F$  entfernt ist.

Wir ziehen eine Stichprobe  $x_1, \dots, x_n$  und sortieren die  $x_i$  der Größe nach. Die so umsortierten Zahlen bezeichnen wir mit  $y_1, \dots, y_n$ . Dann wird eine Funktion  $F_n$  wie folgt erklärt:  $F_n(x)$  ist gleich 0 für  $x \in ]-\infty, y_1]$ , gleich  $i/n$  auf  $[y_i, y_{i+1}]$  für  $i = 1, \dots, n-1$  und gleich 1 auf  $[y_n, +\infty]$ . ( $F_n$  lässt sich auch ohne den Umweg über die  $y_n$  direkt durch  $F_n := \frac{1}{n} \sum_{i=1}^n \chi_{[x_i, +\infty]}(x)$  definieren.)

Wenn die Stichprobe wirklich aus dem gerade betrachteten Wahrscheinlichkeitsraum gezogen wurde, sollte  $F_n$  nicht allzu weit von  $F$  entfernt sein. Denn der Anteil der  $x_i$ , die unter einer Zahl  $a$  liegen, ist einerseits gleich  $F_n(a)$ , andererseits näherungsweise gleich der Wahrscheinlichkeit, dass ein  $x$  in  $]-\infty, a]$  erzeugt wurde. Und diese Wahrscheinlichkeit ist gleich  $\int_{-\infty}^a f(x) dx = F(x)$ .

Deswegen ist es naheliegend, sich um die durch

$$D_n := \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$$

definierte Zufallsvariable zu kümmern. Sie sollte klein sein, und wenn man die Verteilung berechnen könnte, hätte man eine Möglichkeit, die Hypothese „Die

---

<sup>5)</sup>Vgl. Übungsaufgabe 12.3.1.

Stichprobe wurde mit genau diesem Wahrscheinlichkeitsraum erzeugt“ zu vor-  
gegebenen Irrtumsniveaus zu behandeln.

Es folgen zwei Simulationsbeispiele. Es ist zu testen, ob die Gleichverteilung auf  $[0, 1]$  vorliegt. Wir fragen den entsprechenden Wahrscheinlichkeitsraum 50 Mal ab. Im ersten Beispiel liegt wirklich die Gleichverteilung vor. Im folgenden Bild sind  $F$  (durchgezogen) und  $F_n$  (die „Treppe“) eingezeichnet. Der maximale Abstand  $D_{50}$  ist gleich 0.108:

→  
**Programm!**

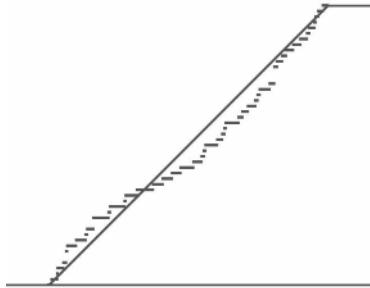


Bild 12.4.1: Kolmogoroff-Smirnoff-Test: Liegt die Gleichverteilung vor? (Ja!).

Im zweiten Beispiel wurden die 50 Stichproben aus dem Raum  $[0, 1]$  gezogen, der mit der Dichtefunktion  $2x$  versehen wurde. Wirklich weicht  $F_n$  deutlich von  $F$  ab:

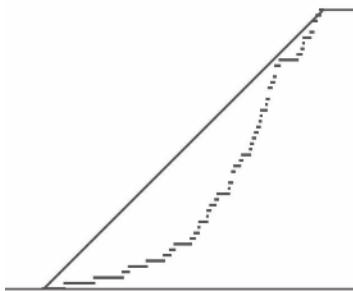


Bild 12.4.2: Kolmogoroff-Smirnoff-Test: Liegt die Gleichverteilung vor? (Nein!).

Das zugehörige  $D_n$  ist deutlich größer:  $D_n = 0.368$ .

Überraschenderweise gilt:

**Satz 12.4.1.** *Die Verteilung von  $D_n$  ist unabhängig von  $F$  und muss folglich nur ein einziges Mal bestimmt werden. Sie heißt die Kolmogoroff-Smirnoff-Verteilung.*

**Beweis:** Es ist für die Rechnungen unbequem, dass in der Definition von  $D_n$  ein Betrag auftritt. Deswegen machen wir einen kleinen Umweg. Wir definieren:

$$D_n^+ := \sup_{x \in \mathbb{R}} (F_n(x) - F(x)),$$

$$D_n^- := \sup_{x \in \mathbb{R}} (F(x) - F_n(x)) = - \inf_{x \in \mathbb{R}} (F_n(x) - F(x)).$$

Es ist klar, dass dann  $D_n = \max\{D_n^+, D_n^-\}$  gilt. Deswegen reicht es,  $D_n^+$  detaillierter zu untersuchen (die Rechnungen für  $D_n^-$  ergeben sich dann durch Übergang von „sup“ zu „inf“.)

Entscheidend sind die folgenden *Beobachtungen*:

- Angenommen, eine Zufallsvariable  $X$  ist so verteilt, dass  $P_X$  die Verteilungsfunktion  $F$  hat. Dann ist  $F(X)$  gleichverteilt in  $[0, 1]$ .

*Begründung:* Sei  $[a, b]$  ein Teilintervall von  $[0, 1]$ . Dann sind offensichtlich äquivalent:

- $F(X)$  liegt in  $[a, b]$ .
- $X$  liegt in  $[\alpha, \beta]$ ; dabei ist  $\alpha := F^{-1}(a)$  und  $\beta := F^{-1}(b)$ .

Deswegen ist die Wahrscheinlichkeit für das eben beschriebene Ereignis gleich

$$F(F^{-1}(b)) - F(F^{-1}(a)),$$

also gleich  $b - a$ . Das beweist die Behauptung.

- Zur Berechnung von  $D_n^+$  spielt es keine Rolle, wie die Stichprobe angeordnet ist: O.B.d.A. gilt  $x_1 < \dots < x_n$ .
- Für  $x$  zwischen  $x_i$  und  $x_{i+1}$  ist  $F_n(x) = i/n$ , und deswegen ist

$$\begin{aligned} D_n^+ &= \max_{0 \leq i \leq n} \sup_{x_i \leq x \leq x_{i+1}} \left( \frac{i}{n} - F(x) \right) \\ &= \max_{0 \leq i \leq n} \left( \frac{i}{n} - \inf_{x_i \leq x \leq x_{i+1}} F(x) \right) \\ &= \max_{0 \leq i \leq n} \left( \frac{i}{n} - F(x_i) \right). \end{aligned}$$

(Für das letzte Gleichheitszeichen wurde ausgenutzt, dass  $F$  monoton steigt.) Zusammen: Um  $D_n^+$  zu simulieren, erzeuge man  $n$  gleichverteilte Zahlen in  $[0, 1]$  und sortiere sie der Größe nach; wir nennen sie  $y_1, \dots, y_n$ . Bestimme dann

$$\max_{0 \leq i \leq n} \left( \frac{i}{n} - y_i \right).$$

Die Größe, die sich auf diese Weise ergibt, ist genau so verteilt wie  $D_n^+$ . Man beachte, dass diese Konstruktion *nicht mehr von  $F$  abhängt*. (Für  $D_n^-$  muss man in der vorstehenden Rechnung „max“ durch „min“ ersetzen.)

Damit ist der Satz bewiesen. □

**Ein Beispiel:** Wir wenden das Ergebnis auf die vor dem Satz beschriebenen Simulationen an (vgl. Bild 12.4.1 und Bild 12.4.2). Die Stichprobe hatte jeweile einen Umfang von 50 Werten, die kritischen  $D$ -Werte sind also wie folgt<sup>6)</sup>:

$$\alpha = 0.01 : D = 1.22/\sqrt{50} = 0.173; \quad \alpha = 0.025 : D = 1.22/\sqrt{50} = 0.192;$$

$$\alpha = 0.05 : D = 1.52/\sqrt{50} = 0.215.$$

Damit wäre die Nullhypothese „Die vorliegende Verteilung ist die Gleichverteilung“ im ersten Beispiel (Bild 12.4.1) sogar zum Niveau  $\alpha = 0.01$  nicht abgelehnt worden, denn es ist  $D_n = 0.108$ . Das zweite Beispiel hatte dagegen den  $D_n$ -Wert 0.368: Hier wäre die Nullhypothese schon für  $\alpha = 0.05$  verworfen worden.

## 12.5 Verständnisfragen

### Zu Abschnitt 12.1

*Sachfragen*

**S1:** In welchen Fällen verwendet man den  $\chi^2$ -Anpassungstest?

*Methodenfragen*

**M1:** Tests zur Hypothese, dass eine vorgelegte Verteilung gleich einer bestimmten Verteilung ist, zu verschiedenen Irrtumsniveaus entwerfen können.

### Zu Abschnitt 12.2

*Sachfragen*

**S1:** In welchen Fällen verwendet man den  $\chi^2$ -Unabhängigkeitstest?

*Methodenfragen*

**M1:** Tests zur Hypothese, dass zwei vorgelegte Zufallsvariable unabhängig sind, zu verschiedenen Irrtumsniveaus entwerfen können.

### Zu Abschnitt 12.3

*Sachfragen*

**S1:** Was ist ein Rangtest?

*Methodenfragen*

**M1:** Rangtests zu verschiedenen Irrtumsniveaus entwerfen können.

### Zu Abschnitt 12.4

*Sachfragen*

**S1:** In welchen Fällen verwendet man den Kolmogoroff-Smirnoff-Test.

---

<sup>6)</sup>Vgl. die Tabelle auf Seite 368.

*Methodenfragen*

**M1:** Entscheiden können, ob die Hypothese „Die vorgelegte Verteilung ist gleich der Verteilung der Zufallsvariablen  $X_0$ “ zum Irrtumsniveau  $\alpha$  abgelehnt werden sollte.

## 12.6 Übungsaufgaben

### Zu Abschnitt 12.1

**Ü12.1.1** Zeigen Sie, dass die bedingte Wahrscheinlichkeit

$$\mathbb{P}(S_i = j_i \text{ für } i = 1, \dots, s \mid N = n)$$

gleich  $M(j_1, \dots, j_s; p_1, \dots, p_s; n)$  ist (vgl. den Beweis von Satz 12.1.1).

**Ü12.1.2** Es soll getestet werden, ob ein vorgelegter Würfel fair ist. Er wird 600 Mal geworfen, die Ergebnisse 1, 2, 3, 4, 5 bzw. 6 treten dabei 50, 120, 120, 75, 102 bzw. 133 Mal auf. Sollte daraufhin die Hypothese „Der Würfel ist fair“ zum Irrtumsniveau  $\alpha = 0.05$  abgelehnt werden?

### Zu Abschnitt 12.2

**Ü12.2.1** Es werden 1000 Leute befragt, dabei geht es um den Zusammenhang zwischen „Sind Sie verheiratet?“ und „Kleiden Sie sich gern teuer und aufwändig ein?“ Das Ergebnis:

- 402 für „verheiratet, Kleidung aufwändig“.
- 306 für „verheiratet, Kleidung eher einfach“.
- 120 für „nicht verheiratet, Kleidung aufwändig“.
- 172 für „nicht verheiratet, Kleidung eher einfach“.

Sollte daraufhin die Hypothese, dass „verheiratet“ und „aufwändige Kleidung“ unabhängig sind, zum Fehlerniveau  $\alpha = 0.05$  abgelehnt werden?

### Zu Abschnitt 12.3

**Ü12.3.1** Beweisen Sie die Rekursionsformel

$$N(m; k, l) = \sum_{j=0}^k N(m - j; j, l - 1).$$

**Ü12.3.2** Zeigen Sie, dass  $N(m; k, l) = N(m; l, k)$  für alle  $m, k, l$  gilt.

**Ü12.3.3** Verwenden Sie die Tabelle in Abschnitt 12.3, um im Fall  $k = l = 4$  einen Konfidenzbereich zum Niveau  $\alpha = 0.1$  für die Nullhypothese „Die Verteilungen von  $X$  und  $Y$  sind gleich“ zu finden.

**Zu Abschnitt 12.4**

**Ü12.4.1** Es soll getestet werden, ob eine Verteilungsfunktion  $F$  vorliegt. Dazu wird eine Stichprobe aus 40 Werten erzeugt, und es ergibt sich  $D_n = 0.102$ . Wird die Hypothese daraufhin zum Niveau  $\alpha = 0.05$  abgelehnt?

**Ü12.4.2** Das gleiche Problem mit 600 Werten und  $\alpha = 0.25$ .

# Anhänge

## Mengenlehre

Aus der elementaren Mengenlehre muss man nur die üblichen Bezeichnungen kennen:

- „ $\subset$ “ bezeichnet die Inklusion. Die Inklusion muss nicht echt sein, so ist etwa  $M \subset M$  eine richtige Aussage.
- $A \setminus B$  steht für die *Mengendifferenz*:  $A \setminus B$  ist die Menge aller  $x$ , die zu  $A$ , aber nicht zu  $B$  gehören.
- Ist  $M$  eine Menge, so bezeichnen wir mit  $\mathcal{P}(M)$  die *Potenzmenge* von  $M$ , also diejenige Menge, die aus allen Teilmengen von  $M$  besteht.

## Äquivalenzrelationen

Zum Verständnis der Konstruktion in Abschnitt 1.7 einer Menge, die keine Borelmenge ist, sollte man wissen, was eine Äquivalenzrelation ist. Ist  $M$  eine Menge, so verstehen wir unter einer *Relation auf  $M$*  irgendeine Teilmenge  $R$  der Produktmenge  $M \times M$ . Dabei schreibt man für  $(x, y) \in R$  kürzer  $xRy$ . Man spricht von einer *Äquivalenzrelation*, wenn  $R$  die folgenden Bedingungen erfüllt:

- $R$  ist *reflexiv*, d.h. für alle  $x$  gilt  $xRx$ .
- $R$  ist *symmetrisch*: Aus  $xRy$  folgt stets  $yRx$ .
- $R$  ist *transitiv*, d.h. aus  $xRy$  und  $yRz$  darf man stets  $xRz$  folgern.

Jede Äquivalenzrelation führt zu einer disjunktten Zerlegung von  $M$  in Äquivalenzklassen. Definiert man nämlich für  $x \in M$  die zu  $x$  gehörige *Äquivalenzklasse* durch  $K_x := \{y \mid xRy\}$ , so sind zwei Äquivalenzklassen  $K_x$  und  $K_{x'}$  entweder disjunkt oder identisch, und  $M$  ist die Vereinigung der  $K_x$ .

## Auswahlaxiom

In Abschnitt 1.7 spielt auch das *Auswahlaxiom* eine wichtige Rolle. Es besagt: Ist  $M$  eine Menge und sind für  $i$  in einer Indexmenge  $I$  nichtleere Teilmengen  $M_i$

von  $M$  definiert, die paarweise disjunkt sind<sup>7)</sup>, so gibt es eine „Auswahlmenge“  $\Delta$ : Das ist eine Teilmenge von  $M$ , die aus jedem  $M_i$  genau ein Element enthält.

Stellt man sich die  $M_i$  als Schubladen vor, die irgendwelche Gegenstände enthalten, so besagt das Auswahlaxiom gerade, dass man so etwas wie einen Musterkoffer konstruieren kann, in dem jedes  $M_i$  durch genau ein Element repräsentiert ist.

Wie der Name schon sagt, ist das Auswahlaxiom ein *Axiom*. Man kann es zu den üblichen weniger weitgehenden Axiomen der Mengenlehre hinzunehmen oder auch nicht. Gleichwertig zum Auswahlaxiom sind *das Zornsche Lemma* (das ist auch ein Axiom!) und das Axiom, das sich jede Menge wohlordnen lässt.

Um das *Zornsche Lemma* formulieren zu können, definiert man vorher, was eine induktiv geordnete Menge ist. Das ist ein geordnete Menge  $(M, \leq)$ , in der jede Teilmenge, in der je zwei Elemente bzgl. „ $\leq$ “ vergleichbar sind, eine obere Schranke hat. Das Zornsche Lemma besagt dann: Jede nicht-leere induktiv geordnete Menge hat ein maximales Element. Es gibt also ein  $x \in M$ , so dass aus  $y \geq x$  stets  $y = x$  folgt.

Das Zornsche Lemma spielt bei vielen Existenzbeweisen eine fundamentale Rolle. Zum Beispiel verwendet man dieses Lemma in der linearen Algebra, um zu zeigen, dass jeder Vektorraum eine Basis hat.

Auch beim *Wohlordnungsaxiom* geht es um geordnete Mengen. Eine Ordnung „ $\leq$ “ heißt eine Wohlordnung, wenn jede nichtleere Teilmenge ein kleinstes Element enthält. Als typisches Beispiel denke man an die übliche Ordnung auf den natürlichen Zahlen. Das Wohlordnungsaxiom besagt dann, dass es auf jeder Menge eine Wohlordnung gibt.

## Vereinigungen von $\sigma$ -Algebren

Als eine der grundlegenden Definitionen wurden schon im ersten Kapitel  $\sigma$ -Algebren eingeführt. Die allermeisten Beweise im Zusammenhang mit solchen Algebren sind Selbstläufer: Für alle, die sich ein bisschen in der Theorie auskennen, machen sie keine Schwierigkeiten.

Eine bemerkenswerte Ausnahme bildet das nachstehende Ergebnis. Dabei geht es um (echt) aufsteigende Folgen von  $\sigma$ -Algebren. Beispiele dafür zu finden, dass die Vereinigung einer aufsteigende Folge von  $\sigma$ -Algebren nicht wieder eine  $\sigma$ -Algebra zu sein braucht, ist nicht besonders schwierig. (In diesem Buch ist das die Aufgabe 1.2.10.) Es ist aber alles andere als offensichtlich, dass das *nie* der Fall ist:

**Satz:** Es sei  $M$  eine Menge, und  $\mathcal{E}_1, \mathcal{E}_2, \dots$  seien  $\sigma$ -Algebren auf  $M$ . Ist dann  $\mathcal{E}_1 \subsetneq \mathcal{E}_2 \subsetneq \mathcal{E}_3 \subsetneq \dots$ , so ist  $\mathcal{E} := \bigcup_{n=1}^{\infty} \mathcal{E}_n$  keine  $\sigma$ -Algebra.

**Beweis:** Dieses Ergebnis wurde im American Mathematical Monthly 84 von den Autoren Allen Broughton und Barthel W. Huff veröffentlicht (1977, Seite 553 bis 554). Der dort gegebene Beweis ist allerdings lückenhaft. Es folgt eine alternative Herleitung. Das vorbereitende Ergebnis (bei uns ist das „Schritt 1“)

---

<sup>7)</sup>Für  $i \neq j$  ist also  $M_i \cap M_j = \emptyset$ .

ist schon im Mathematical Monthly zu finden, der Hauptbeweis („Schritt 2“) geht im Wesentlichen auf Maikel Nadolski zurück, einen Studenten an der FU Berlin, der die „Elementare Stochastik“ vor wenigen Semestern beim Autor dieses Buches hörte.

*Schritt 1:* Laut Voraussetzung gibt es für jedes  $n > 1$  ein  $E_n \in \mathcal{E}_n \setminus \mathcal{E}_{n-1}$ . Wir behaupten, dass man dabei annehmen darf, dass die  $E_n$  paarweise disjunkt sind. *Beweis dazu:* Wir beginnen mit einer Notation. Wenn  $\mathcal{F}$  eine  $\sigma$ -Algebra auf einer Menge  $N$  und  $G$  eine Teilmenge von  $N$  ist, so bezeichnen wir mit  $\mathcal{F}_G$  das Mengensystem

$$\mathcal{F}_G := \{F \cap G \mid F \in \mathcal{F}\}.$$

In Übungsaufgabe 1.2.2 sollte gezeigt werden, dass das eine  $\sigma$ -Algebra auf  $G$  ist.

Nun seien  $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \dots$   $\sigma$ -Algebren auf einer Menge  $N$ , so dass für unendlich viele  $n$  die  $\sigma$ -Algebra  $\mathcal{F}_n$  eine echte Teilmenge von  $\mathcal{F}_{n+1}$  ist. Wir wählen ein beliebiges  $G \in \mathcal{F}_1$ . Ist dann  $n \in \mathbb{N}$  und gilt  $(\mathcal{F}_n)_G = (\mathcal{F}_{n+1})_G$  sowie  $(\mathcal{F}_n)_{N \setminus G} = (\mathcal{F}_{n+1})_{N \setminus G}$ , so muss  $\mathcal{F}_n = \mathcal{F}_{n+1}$  gelten. (Denn jedes  $F \in \mathcal{F}_{n+1}$  kann als  $F = (F \cap G) \cup (F \cap (N \setminus G))$  geschrieben werden.) Aber unendlich oft ist das nicht der Fall, und deswegen gilt für  $G_0 = G$  oder  $G_0 = N \setminus G$ , dass in der aufsteigenden Folge der  $\sigma$ -Algebren  $(\mathcal{F}_1)_{G_0} \subset (\mathcal{F}_2)_{G_0} \subset \dots$  unendlich oft eine echte Inklusion steht.

Diese Vorbereitung nutzen wir, um die gesuchten  $E_n$  zu finden.

- Setze  $n_1 := 2$  und wähle ein  $E_1 \in \mathcal{E}_{n_1}$  mit  $E_1 \notin \mathcal{E}_{n_1-1}$ . Das ist nach Voraussetzung möglich. Aufgrund der Vorbereitung kann man dann  $F_1 = E_1$  oder  $F_1 = N \setminus E_1$  wählen, so dass unendlich viele echte Inklusionen in der aufsteigenden Folge  $(\mathcal{E}_n)_{F_1}$  stehen. Man beachte, dass in beiden Fällen  $F_1 \notin \mathcal{E}_{n_1-1}$  gilt.
- Wir konzentrieren uns nun auf die  $\sigma$ -Algebren  $(\mathcal{E}_n)_{F_1}$  auf der Menge  $F_1$ . Wir wählen ein  $n_2$ , so dass  $n_2 > n_1$  gilt und für das wir ein  $E \in (\mathcal{E}_{n_2})_{F_1}$  wählen können, das nicht in  $(\mathcal{E}_{n_2-1})_{F_1}$  liegt. Wieder nach Vorbereitung gilt für  $F_2 := E$  oder  $F_2 := F_1 \setminus E$ , dass die Einschränkungen der  $\sigma$ -Algebren auf  $F_2$  unendlich oft echt aufsteigen. Unabhängig von der Definition von  $F_2$  liegt diese Menge nicht in  $\mathcal{E}_{n_2-1}$ .
- Und so weiter.

Wir erhalten Indizes  $n_1 < n_2 < \dots$  und Mengen  $F_i \in \mathcal{E}_{n_i}$ , so dass  $F_1 \supset F_2 \supset \dots$  und  $F_i \notin \mathcal{E}_{n_i-1}$  (und damit auch  $F_i \notin \mathcal{E}_{n_{i-1}}$ ). Wir setzen noch  $E_i := F_{i-1} \setminus F_i$  für  $i > 1$ : Diese Mengen sind paarweise disjunkt, und es gilt  $E_i \in \mathcal{E}_{n_i} \setminus \mathcal{E}_{n_{i-1}}$ . Wenn wir also von  $\mathcal{E}_1, \mathcal{E}_2, \dots$  zu  $\mathcal{E}_{n_1}, \mathcal{E}_{n_2}, \dots$  übergehen, so haben wir Mengen so gefunden, wie in Schritt 1 behauptet. Und da die Vereinigung der  $\mathcal{E}_{n_i}$  gleich  $\mathcal{E}$  ist, können wir für den Nachweis, dass  $\mathcal{E}$  keine  $\sigma$ -Algebra ist, von den  $\mathcal{E}_n$  zu der Teilfolge  $\mathcal{E}_{n_i}$  übergehen.

Wir fassen zusammen:

Angenommen,  $\mathcal{E}$  ist die Vereinigung einer echt aufsteigenden Folge von  $\sigma$ -Algebren. Dann kann man  $\mathcal{E}$  auch als  $\bigcup_n \mathcal{E}_n$  schreiben, wobei gilt:

- Es ist  $\mathcal{E}_1 \subsetneq \mathcal{E}_2 \subsetneq \mathcal{E}_3 \subsetneq \dots$ .
- Man findet  $E_n \in \mathcal{E}_n \setminus \mathcal{E}_{n-1}$  für  $n \in \mathbb{N}$ , so dass die  $E_1, E_2, \dots$  paarweise disjunkt sind.

*Schritt 2: Hauptbeweis.* Wir werden annehmen, dass  $\mathcal{E}$  eine  $\sigma$ -Algebra ist, und daraus wird ein Widerspruch hergeleitet.

Wir fixieren eine bijektive Abbildung  $\phi$  von  $\mathbb{N} \times \mathbb{N}$  nach  $\mathbb{N}$ , damit können wir die Folge  $E_n$  als Doppelfolge schreiben: Für  $i, j \in \mathbb{N}$  wird  $B_{i,j}$  als  $E_{\phi(i,j)}$  definiert. Insbesondere sind diese Mengen paarweise disjunkt.

Wir denken uns die  $B_{i,j}$  als (unendliche) Matrix angeordnet:

$$\begin{pmatrix} B_{1,1} & B_{1,2} & B_{1,3} & B_{1,4} & \cdots \\ B_{2,1} & B_{2,2} & B_{2,3} & B_{2,4} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}.$$

In der  $i$ -ten Zeile stehen dann die Mengen  $B_{i,1}, B_{i,2}, \dots$ . Nach Annahme ( $\mathcal{E}$  ist  $\sigma$ -Algebra) gehört  $B_i := \bigcup_j B_{i,j}$  zu  $\mathcal{E}$ . Es gibt also ein  $n_i \in \mathbb{N}$  mit  $B_i \in \mathcal{E}_{n_i}$ . Wir wollen dabei annehmen, dass der Index  $n_i$  mindestens gleich  $i$  ist. Und da es bei der Vereinigung der  $B_{i,1}, B_{i,2}, \dots$  auf die Reihenfolge nicht ankommt, können wir es so einrichten, dass  $B_{i,1}$  nicht zu  $\mathcal{E}_{n_i}$  gehört. (Die  $\phi(i,1), \phi(i,2), \dots$  sind ja unendlich viele natürliche Zahlen. Eine wird bestimmt größer als  $n_i$  sein, und das zugehörige  $B_{i,j}$  schreiben wir an den Beginn der Zeile.)

Und nun das Finale. Wir vereinigen die  $B_{i,1}$ , also die Mengen in der ersten Spalte unserer Matrix. Diese Menge, wir wollen sie  $B$  nennen, liegt – wieder nach unserer Annahme – in  $\mathcal{E}$  und folglich in einem  $\mathcal{E}_{n_0}$ . Wir suchen irgendeine Zeile  $i_0$  mit  $i_0 \geq n_0$ . Dann gilt erstens, dass  $n_{i_0} \geq n_0$ , und deswegen gehört  $B$  zu  $\mathcal{E}_{n_{i_0}}$ . Zweitens liegt auch  $B_{i_0}$  in  $\mathcal{E}_{i_0}$ , und damit müsste auch der Durchschnitt  $B \cap B_{i_0}$  in dieser  $\sigma$ -Algebra liegen. Dieser Durchschnitt – Vereinigung über die Mengen der ersten Spalte geschnitten mit der Vereinigung über die Elemente der  $i_0$ -ten Spalte – ist aber gleich  $B_{i_0,1}$ . Hier ist es wesentlich, dass die  $B_{i,j}$  paarweise disjunkt sind. Auf diese Weise haben wir einen Widerspruch erhalten, denn laut Konstruktion sollte ja  $B_{i_0,1}$  nicht in  $\mathcal{E}_{i_0}$  liegen.  $\square$

## Maßtheorie

Die große Leistung von Kolmogoroff bestand darin, ein mathematisch präzise behandelbares Fundament für die Wahrscheinlichkeitstheorie zu schaffen: Statt mit ziemlich vagen Begriffen umgehen zu müssen kann alles ganz exakt im Rahmen der Maßtheorie untersucht werden.

Deswegen sollte man eigentlich vor einer Stochastikvorlesung eine Vorlesung zur Maßtheorie gehört haben. Das ist aber unrealistisch, und deswegen werden in

Anfängerveranstaltungen zur Stochastik – und auch in diesem Buch – maßtheoretische Begriffe eher sparsam verwendet. Wie man bei einem systematischen Aufbau vorgehen würde, soll nun in Stichworten beschrieben werden.

### Maßräume

1. Man führt zunächst den Begriff des *Messraums* ein: Ein Messraum ist eine Menge  $M$  zusammen mit einer  $\sigma$ -Algebra  $\mathcal{E}$  auf  $M$ . Der Messraum  $(M, \mathcal{E})$  wartet sozusagen darauf, dass auf  $\mathcal{E}$  ein Maß definiert wird.
2. Ein *Maß*  $\mu$  auf einem Messraum  $(M, \mathcal{E})$  ist eine Abbildung  $\mu : \mathcal{E} \rightarrow [0, +\infty]$  (ja, der Wert  $+\infty$  ist zugelassen<sup>8)</sup>!), so dass erstens  $\mu(\emptyset) = 0$  gilt und zweitens das Bild einer disjunktten Vereinigung gleich der Summe der Maße derjenigen Mengen ist, die zu dieser Vereinigung beigetragen haben:

$$\mu\left(\bigcup_{n=1}^{\infty} E_n\right) = \sum_{n=1}^{\infty} \mu(E_n), \text{ falls } E_n \cap E_m = \emptyset \text{ für } n \neq m.$$

(Man beachte, dass die rechts stehende Reihensumme in  $[0, +\infty]$  immer existiert, auch dann, wenn einige oder sogar alle Summanden gleich  $+\infty$  sind.)

$\mu$  heißt *endlich*, wenn  $\mu(M) < +\infty$  gilt, und im Fall  $\mu(M) = 1$  spricht man von einem *Wahrscheinlichkeitsmaß* auf  $M$ .

Das Tripel  $(M, \mathcal{E}, \mu)$  wird ein *Maßraum* genannt.

3. Das wichtigste Ergebnis, um die Existenz von konkreten Maßen mit vorgegebenen Eigenschaften zu garantieren, ist der *Satz von Carathéodory*. Vereinfacht ausgedrückt, besagt er: Ist  $\mathcal{E}_0$  eine Teilmenge der Potenzmenge von  $M$ , die unter endlichen Mengenoperationen abgeschlossen ist (Durchschnitte gehören wieder dazu usw.) und  $\mu_0 : \mathcal{E}_0 \rightarrow [0, +\infty]$  eine Abbildung, so gilt unter gewissen Voraussetzungen, dass man  $\mu_0$  zu einem Maß  $\mu$  auf die von  $\mathcal{E}_0$  erzeugte  $\sigma$ -Algebra fortsetzen kann.

Auf diese Weise wird zum Beispiel sichergestellt, dass es ein Maß  $\lambda$  auf den Borelmengen von  $\mathbb{R}$  gibt, so dass  $\lambda([a, b]) = b - a$  für alle Intervalle  $[a, b]$  gilt. Es ist eindeutig bestimmt, man nennt es das *Borel-Lebesgue-Maß*.

### Das Integral

Es sei  $(M, \mathcal{E}, \mu)$  ein Maßraum und  $f : M \rightarrow [-\infty, +\infty]$  eine Abbildung. Das Integral  $\int_M f(x) d\mu$  soll dann so etwas sein wie eine durch  $\mu$  gewichtete  $f$ -Messung. Um es einzuführen, verfährt man in mehreren Schritten:

- Damit alles, was definiert werden soll, auch wirklich erklärt ist, beschränkt man sich auf solche  $f$ , bei denen  $f^{-1}(B)$  für jede Borelmenge  $B$  in  $\mathbb{R}$  ein Element von  $\mathcal{E}$  ist. Solche Abbildungen heißen *messbar*. Die in Kapitel 3 eingeführten Zufallsvariablen sind also gerade die messbaren Abbildungen auf dem gerade behandelten Wahrscheinlichkeitsraum. Ab hier geht es nur noch um messbare  $f$ . Die Faustregel: Alle Funktionen, die bei konkreten Modellierungen eine Rolle spielen, sind messbar.

<sup>8)</sup>Wie sollte man sonst sagen, was die Länge von  $\mathbb{R}$  oder der Flächeninhalt des  $\mathbb{R}^2$  ist?

- Zunächst wird das Integral für „ganz einfache“  $f$  erklärt, so genannte *Treppenfunktionen*: Wenn sich  $f$  als  $\sum_{n=1}^{\infty} a_n \chi_{E_n}$  mit  $a_1, a_2, \dots \in [0, +\infty]$  und paarweise disjunkten  $E_1, E_2, \dots$  schreiben lässt, setzt man

$$\int_M f(x) d\mu(x) := \sum_{n=1}^{\infty} a_n \mu(E_n).$$

Dabei werden Produkte der Form  $0 \cdot \infty$  oder  $\infty \cdot 0$  als Null definiert.

Man muss dann zeigen, dass dieser Ausdruck nur von  $f$ , nicht aber von der konkreten Darstellung durch die  $a_n$  und die  $E_n$  abhängt.

- Danach überlegt man sich vorbereitend, dass es für jedes messbare  $f$  eine Folge  $\tau_1 \leq \tau_2 \dots$  von Treppenfunktionen gibt, so dass  $f(x) = \sup_n \tau_n(x)$  für alle  $x$  gilt. Damit ist es naheliegend, das Integral über  $f$  durch

$$\int_M f(x) d\mu(x) := \sup_n \int_M \tau_n(x) d\mu(x) \in [0, +\infty]$$

zu definieren. So wird es wirklich gemacht, es ist allerdings recht aufwändig nachzuweisen, dass das so definierte Integral nur von  $f$  (und nicht von der Wahl der Folge  $(\tau_n)$ ) abhängt.

- Wenn  $f : M \rightarrow [-\infty, +\infty]$  eine beliebige messbare Funktion ist, schreibt man  $f = f_1 - f_2$  mit Funktionen  $f_1, f_2$ , die nur nichtnegative Werte annehmen. Auf jeden Fall existieren dann  $\int_M f_1(x) d\mu(x)$  und  $\int_M f_2(x) d\mu(x)$  in  $[0, +\infty]$ . Das Integral  $\int_M f(x) d\mu(x)$  soll als Differenz dieser Werte erklärt werden, doch das wird problematisch, wenn man es dabei mit  $+\infty$  zu tun hat. Damit man damit keine Probleme bekommt, definiert man:  $f$  heißt *integrabel*, wenn sowohl  $\int_M f_1(x) d\mu(x)$  als auch  $\int_M f_2(x) d\mu(x)$  endlich ist, und in diesem Fall setzt man

$$\int_M f(x) d\mu(x) := \int_M f_1(x) d\mu(x) - \int_M f_2(x) d\mu(x).$$

All das spielt eine wichtige Rolle für die Behandlung von Zufallsvariablen in Kapitel 3, denn der Erwartungswert einer Zufallsvariablen ist nichts weiter als ein Spezialfall des allgemeinen Integralbegriffs aus der Maßtheorie.

### Der Satz von der monotonen Konvergenz

Es gibt eine Reihe von zum Teil recht schwierig zu beweisenden Tatsachen aus der Maßtheorie, die von Bedeutung werden, wenn man die Wahrscheinlichkeitstheorie systematisch fortführen möchte. Eines dieser Ergebnisse wurde hier schon (am Ende von Abschnitt 8.2 und in Abschnitt 8.4) verwendet: Es gilt der

*Satz von der monotonen Konvergenz:* Es seien  $f_1, f_2, \dots : M \rightarrow [0, +\infty]$  messbare Funktionen mit  $f_1 \leq f_2 \leq \dots$ . Definiert man dann die Funktion  $\sup_n f_n$  punktweise, also durch  $(\sup_n f_n)(x) := \sup_n f_n(x)$ , so gilt

$$\int_M (\sup_n f_n)(x) d\mu(x) = \sup_n \int_M f_n(x) d\mu(x).$$

## Das Skalarprodukt auf dem $\mathbb{R}^n$

In den Kapiteln zur Statistik ist es wichtig, sich an einige Tatsachen im Zusammenhang mit dem Skalarprodukt auf dem  $\mathbb{R}^n$  zu erinnern:

- Für  $x, y \in \mathbb{R}^n$  wird das Skalarprodukt  $\langle x, y \rangle$  wie folgt definiert: Schreibt man  $x = (x_1, \dots, x_n)$  und  $y = (y_1, \dots, y_n)$ , so ist

$$\langle x, y \rangle := x_1 y_1 + \dots + x_n y_n.$$

- Durch  $\|x\| := \sqrt{\langle x, x \rangle}$  wird dann eine Norm (die *euklidische Norm*) auf dem  $\mathbb{R}^n$  erklärt.
- Zwei Vektoren  $x, y$  heißen *orthogonal*, wenn  $\langle x, y \rangle = 0$  gilt. In diesem Fall ist

$$\|x + y\|^2 = \|x\|^2 + \|y\|^2$$

(Satz des Pythagoras in euklidischen Räumen).

- Ist  $U$  ein Unterraum des  $\mathbb{R}^n$  und  $x_0 \in \mathbb{R}^n$ , so hat ein  $y \in U$  genau dann den kleinstmöglichen Abstand zu  $x_0$  unter allen Elementen aus  $U$ , wenn  $x_0 - y$  orthogonal zu allen Elementen aus  $U$  ist.

## Analysis

In diesem Anhang sind einige Ergebnisse aus der Analysis zusammengestellt, die in diesem Buch eine Rolle spielen.

### Unbedingte und absolute Konvergenz

Es seien  $a_1, a_2, \dots$  reelle Zahlen. Bekanntlich sagt man dann, dass die Reihensumme  $\sum_{n=1}^{\infty} a_n$  existiert (oder dass die Reihe konvergiert) und den Wert  $S$  hat, wenn die Folge der Partialsummen, also die Folge  $(a_1 + \dots + a_n)_n$ , gegen  $S$  konvergent ist. Man sollte sich an die folgenden Tatsachen erinnern:

- Absolute Konvergenz impliziert Konvergenz: Wenn  $\sum_n |a_n|$  in  $\mathbb{R}$  existiert, so existiert auch  $\sum_{n=1}^{\infty} a_n$ .
- Ist die Reihe der  $(a_n)$  absolut konvergent, so ist sie auch *unbedingt konvergent*: Jede Umordnung der  $a_1, a_2, \dots$  hat die gleiche Reihensumme.

Als Folgerung ergibt sich noch, dass man in einer absolut konvergenten Reihe auf ganz beliebige Weise Klammern setzen kann. Etwas präziser kann das so formuliert werden:

- $\sum_n a_n$  sei eine absolut konvergente Reihe und es sei  $1 = n_1 < n_2 < \dots$ .
- Definiere  $b_k := a_{n_k} + \dots + a_{n_{k+1}-1}$ .
- Dann ist  $\sum_n a_n = \sum_k b_k$ .  
(Zum Beispiel ist  $a_1 + a_2 + \dots = (a_1 + a_2) + (a_3 + a_4) + \dots$ )

### Ungeordnete Summation

Jeder weiß, wie  $\sum_{i=1}^n a_i$  und  $\sum_{i=1}^\infty a_i$  erklärt sind, wenn die  $a_1, a_2, \dots$  reelle Zahlen sind. Doch was soll  $\sum_{i \in E} a_i$  bedeuten.

In diesem Buch tritt dieses Problem bei der Definition von Wahrscheinlichkeiten auf: Was ist  $\sum_{\omega \in E} p_\omega$ , wenn  $E$  eine abzählbare Menge ist und die  $p_\omega$  die Wahrscheinlichkeiten der Ereignisse  $\{\omega\}$  sind?

Diese Summe ist so definiert: Zähle die Elemente aus  $E$  auf irgendeine Weise auf. Schreibe also  $E = \{\omega_1, \dots, \omega_n\}$  im endlichen bzw.  $E = \{\omega_1, \omega_2, \dots\}$  im abzählbaren Fall. Der Ausdruck  $\sum_{\omega \in E} p_\omega$  ist dann als endliche Summe  $\sum_{i=1}^n p_{\omega_i}$  bzw. als Reihe  $\sum_{i=1}^\infty p_{\omega_i}$  definiert. Zwei Tatsachen sind zu beachten. Erstens existiert die Reihensumme, denn die Partialsummen sind monoton wachsend und nach oben durch Eins beschränkt. Zweitens ist der Wert der so entstehenden Summe bzw. Reihe unabhängig davon, wie man  $E$  durchnummeriert hat. Im endlichen Fall folgt das sofort aus der Kommutativität der Addition, im abzählbaren Fall liegt es daran, dass Umordnungen absolut konvergenter Reihen zum gleichen Wert konvergieren. (Vgl. den Unterpunkt „Absolute und unbedingte Konvergenz“.) Folglich ist der Ausdruck  $\sum_{\omega \in E} p_\omega$  stets wohldefiniert. Es ist noch zu ergänzen, dass  $\sum_{\omega \in \emptyset} p_\omega$  als 0 definiert ist.

### Supremum und Infimum

Techniken, die mit Supremum und Infimum zusammenhängen, spielen eine wichtige Rolle in der Analysis. Wir behandeln hier nur den Fall des Supremums: Aussagen für das Infimum erhält man, wenn man stets  $\geq, \leq, <, >$  durch  $\leq, \geq, >, <$  ersetzt. Man sollte folgendes wissen:

- Ist  $A$  eine nicht leere Teilmenge von  $\mathbb{R}$ , so heißt ein  $x \in \mathbb{R}$  *Supremum* von  $A$ , wenn erstens  $y \leq x$  für alle  $y \in A$  gilt und man zweitens (für  $x' \in \mathbb{R}$ ) aus „ $y \leq x'$  für alle  $y \in A$ “ stets  $x \leq x'$  folgern kann.  $x$  ist also die bestmögliche obere Schranke.
- Das Supremum von  $A$  ist, falls es existiert, eindeutig bestimmt. Man schreibt dafür  $\sup A$ .
- $\sup A$  existiert genau dann, wenn  $A$  nicht leer und nach oben beschränkt ist.

- Wenn  $\sup A$  existiert, so weiß man: Für jedes  $\varepsilon > 0$  existiert ein  $y \in A$  mit  $y > \sup A - \varepsilon$ .

Der Transformationssatz für Gebietsintegrale

Es sei  $B$  eine „einfache“ Teilmenge des  $\mathbb{R}^n$  (eine Kugel, ein Hyperquader o.ä.) und  $g : B \rightarrow \mathbb{R}$  eine nichtnegative Funktion. Das (Volumen-)Integral  $\int_B g(\mathbf{x}) d\mathbf{x}$  kann man dann als  $n+1$ -dimensionales Volumen zwischen  $B$  und dem Graphen  $\{(\mathbf{x}, g(\mathbf{x})) \mid \mathbf{x} \in B\} \subset \mathbb{R}^{n+1}$  interpretieren. (Ist zum Beispiel  $B$  die Kreisscheibe mit dem Radius  $r_0$  und  $g(x, y) := \sqrt{r_0^2 - x^2 - y^2}$ , so ist  $\int_B g(\mathbf{x}) d\mathbf{x}$  das Volumen einer Halbkugel mit Radius  $r_0$ .)

Ist  $B$  ein Produkt von Intervallen, so kann das Volumenintegral durch iterierte eindimensionale Integration ausgerechnet werden. Hat etwa  $B$  die Form  $[a, b] \times [c, d] \subset \mathbb{R}^2$ , so ist

$$\int_B g(\mathbf{x}) d\mathbf{x} = \int_a^b \left( \int_c^d g(x, y) dy \right) dx.$$

(Achtung: Im linken Integral bedeutet das  $\mathbf{x}$  einen zweidimensionalen Vektor, im rechten steht  $x$  für eine eindimensionale Variable.)

Durch den Transformationssatz für Gebietsintegrale kann die Berechnung von  $\int_B g(\mathbf{x}) d\mathbf{x}$  oft auf eine „einfache“ iterierte Integration zurückgeführt werden. Hier die genaue Formulierung:

*Transformationssatz:* Es seien  $B, C \subset \mathbb{R}^n$  und  $\phi : C \rightarrow B$  bijektiv und differenzierbar. Dann ist  $\int_B g(\mathbf{x}) d\mathbf{x} = \int_C g \circ \phi(\mathbf{y}) |\det J_\phi(\mathbf{y})| d\mathbf{y}$ . (Hier bezeichnet  $|\det J_\phi(\mathbf{y})|$  den Betrag der Jacobideterminante von  $\phi$  an der Stelle  $\mathbf{y}$ .) Das nutzt in der Regel nur dann etwas, wenn  $C$  ein Produkt von – evtl. unbeschränkten – Intervallen ist und das rechts stehende Integral berechnet werden kann. Die Kunst besteht darin,  $C$  und  $\phi$  richtig zu wählen, d.h.  $B$  richtig zu parametrisieren.

*Ein Beispiel.* Wir parametrisieren die Kreisscheibe mit Radius  $r_0$  im  $\mathbb{R}^2$  mit Polarkoordinaten. Wir setzen also  $C := [0, r_0] \times [0, 2\pi]$ , und  $\phi$  wird durch  $\phi(r, \varphi) := (r \cos \varphi, r \sin \varphi)$  erklärt. Für die vor wenigen Zeilen definierte Funktion  $g(x, y) := \sqrt{r_0^2 - x^2 - y^2}$  wird man auf das Integral  $\int_0^{r_0} \left( \int_0^\pi r \sqrt{r_0^2 - r^2} d\varphi \right) dr$  geführt, denn die Jacobideterminante bei  $(r, \varphi)$  ist gleich  $r$  und es gilt  $g \circ \phi(r, \varphi) = \sqrt{r_0^2 - r^2}$ . Die Auswertung ist einfach:

$$\begin{aligned} \int_0^{r_0} \left( \int_0^\pi r \sqrt{r_0^2 - r^2} d\varphi \right) dr &= \int_0^{r_0} 2\pi \sqrt{r_0^2 - r^2} dr \\ &= -\frac{2\pi}{3} (r_0^2 - r^2)^{3/2} \Big|_0^{r_0} \\ &= \frac{2\pi}{3} r_0^3. \end{aligned}$$

Und das ist – wie aufgrund der Problemstellung zu erwarten war – das Volumen einer halben Kugel mit Radius  $r_0$ .

Das Integral über die Funktion  $e^{-x^2/2}$

Dieses Integral kann mit einem kleinen Trick berechnet werden. Dazu betrachten wir die durch  $(x, y) \mapsto e^{-(x^2+y^2)/2}$  definierte Funktion  $g : \mathbb{R}^2 \rightarrow \mathbb{R}$ . Ihr Graph, also die Menge  $\{(x, y, g(x, y)) \mid (x, y) \in \mathbb{R}^2\}$  sieht wie ein Sombrero aus<sup>9)</sup>. Wie groß ist das Volumen  $V_S$  zwischen Sombrero und  $(x, y)$ -Ebene?

$V_S$  ist doch das Gebietsintegral  $\int_{\mathbb{R}^2} g(\mathbf{x}) d\mathbf{x}$ . Einerseits ist es durch iterierte Integration leicht zu berechnen, denn der Integrand ist das Produkt aus zwei Funktionen, wobei die eine nur von  $x$  und die andere nur von  $y$  abhängt:

$$\begin{aligned}\int_{\mathbb{R}^2} g(\mathbf{x}) d\mathbf{x} &= \int_{\mathbb{R}} \left( \int_{\mathbb{R}} e^{-(x^2+y^2)/2} dy \right) dx \\ &= \int_{\mathbb{R}} \left( \int_{\mathbb{R}} e^{-x^2/2} e^{-y^2/2} dy \right) dx \\ &= \left( \int_{\mathbb{R}} e^{-y^2/2} dy \right) \left( \int_{\mathbb{R}} e^{-x^2/2} dx \right) \\ &= \left( \int_{\mathbb{R}} e^{-x^2/2} dx \right)^2.\end{aligned}$$

Das liegt daran, dass die Zahl  $\int_{\mathbb{R}} e^{-y^2/2} dy$  für die  $x$ -Integration eine Konstante ist und dass die Variablennamen für den Wert des Integrals keine Rolle spielen.

Andererseits können wir  $B = \mathbb{R}^2$  auch durch Polarkoordinaten parametrisieren. Dazu setzen wir  $C = [0, +\infty[ \times [0, 2\pi]$ , und  $\phi$  wird wieder durch  $\phi(r, \varphi) := (r \cos \varphi, r \sin \varphi)$  definiert. Aufgrund des Transformationssatzes ist

$$\begin{aligned}\int_{\mathbb{R}^2} g(\mathbf{x}) d\mathbf{x} &= \int_C g \circ \phi(\mathbf{y}) |\det J_\phi(\mathbf{y})| d\mathbf{y} \\ &= \int_0^\infty \left( \int_0^{2\pi} r e^{-r^2/2} d\varphi \right) dr \\ &= \int_0^\infty 2\pi r e^{-r^2/2} dr \\ &= -2\pi e^{-r^2/2} \Big|_0^\infty \\ &= 2\pi.\end{aligned}$$

So folgt, dass das Quadrat der gesuchten Zahl  $\int_{\mathbb{R}} e^{-x^2/2} dx$  gleich  $2\pi$  ist, und damit ist gezeigt, dass wirklich  $\int_{\mathbb{R}} e^{-x^2/2} dx = \sqrt{2\pi}$  gilt. So hängen überraschender Weise die  $e$ -Funktion und die Kreiszahl  $\pi$  zusammen.

## Tabellen

In der Wahrscheinlichkeitsrechnung werden oft Werte von Funktionen benötigt, die man nicht durch einen geschlossenen Ausdruck darstellen kann. Ein prominentes Beispiel ist die Dichte der Normalverteilung:

---

<sup>9)</sup>Dieser Sombrero ist uns auf Seite 60 schon einmal begegnet.

Um zum Beispiel  $\mathbb{P}_{N(0,1)}(\{-\infty, a]\})$  auszurechnen, muss man das Integral  $(1/\sqrt{2\pi}) \int_{-\infty}^a e^{-x^2/2} dx$  bestimmen, doch  $e^{-x^2/2}$  hat keine geschlossenen darstellbare Stammfunktion<sup>10)</sup>.

Deswegen werden die interessierenden Ausdrücke numerisch ausgewertet. Sie stehen als *Tabellen* zur Verfügung oder werden, wenn man die geeigneten Programme hat, im Bruchteil einer Sekunde ausgerechnet. Tabellen werden seit Jahrhunderten verwendet. Heute steht alles, was gebraucht wird, im Internet zur Verfügung, und deswegen wäre diese Abteilung des Anhangs eigentlich entbehrlich. Da aber einige Tabellen-Informationen in diesem Buch verwendet werden, wurden trotzdem einige aufgenommen, damit man das Lesen nicht für Zusatzrecherchen unterbrechen muss.

Viel ausführlichere Tabellen findet man – natürlich – im Internet. Zum Beispiel gibt es eine ausführliche Tabelle der  $\chi^2$ -Verteilungen in

[http://de.wikibooks.org/wiki/Mathematik:\\_Statistik:\\_Tabelle\\_der\\_Chis-Quadrat-Verteilung](http://de.wikibooks.org/wiki/Mathematik:_Statistik:_Tabelle_der_Chis-Quadrat-Verteilung),

und Tabellen zu den  $t$ -Verteilungen werden in

[http://de.wikipedia.org/wiki/Studentsche\\_t-Verteilung#Tabelle\\_einiger\\_t-Quantile](http://de.wikipedia.org/wiki/Studentsche_t-Verteilung#Tabelle_einiger_t-Quantile)

zur Verfügung gestellt.

### Tabelle der Standardnormalverteilung

*Wann braucht man sie?* Immer dann, wenn man wissen möchte, mit welcher Wahrscheinlichkeit eine standard-normalverteilte Zufallsvariable Werte in einem Intervall  $[a, b]$  annimmt. Mit dieser Information lässt sich die entsprechende Frage dann auch für  $N(a, \sigma^2)$ -verteilte Zufallsvariable beantworten.

---

<sup>10)</sup>Das ist ein berühmter Satz von Liouville aus dem 19. Jahrhundert. Die genaue Formulierung und den Beweis findet man in meinem Buch „Analysis 2“.

Die Tabelle

|      | 0      | 1      | 2      | 3      | 4      | 5      | 6      | 7      | 8      | 9      |
|------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| -3.0 | 0.0013 | 0.0010 | 0.0007 | 0.0005 | 0.0012 | 0.0003 | 0.0002 | 0.0002 | 0.0001 | 0.0000 |
| -2.9 | 0.0019 | 0.0018 | 0.0018 | 0.0017 | 0.0016 | 0.0016 | 0.0015 | 0.0015 | 0.0014 | 0.0014 |
| -2.8 | 0.0026 | 0.0025 | 0.0024 | 0.0023 | 0.0023 | 0.0022 | 0.0021 | 0.0021 | 0.0020 | 0.0019 |
| -2.7 | 0.0035 | 0.0034 | 0.0033 | 0.0032 | 0.0031 | 0.0030 | 0.0029 | 0.0028 | 0.0027 | 0.0026 |
| -2.6 | 0.0047 | 0.0045 | 0.0044 | 0.0043 | 0.0041 | 0.0040 | 0.0039 | 0.0038 | 0.0037 | 0.0036 |
| -2.5 | 0.0062 | 0.0060 | 0.0059 | 0.0057 | 0.0055 | 0.0054 | 0.0052 | 0.0051 | 0.0049 | 0.0048 |
| -2.4 | 0.0082 | 0.0080 | 0.0078 | 0.0075 | 0.0073 | 0.0071 | 0.0069 | 0.0068 | 0.0066 | 0.0064 |
| -2.3 | 0.0107 | 0.0104 | 0.0102 | 0.0099 | 0.0096 | 0.0094 | 0.0091 | 0.0089 | 0.0087 | 0.0084 |
| -2.2 | 0.0139 | 0.0136 | 0.0132 | 0.0129 | 0.0125 | 0.0122 | 0.0119 | 0.0116 | 0.0113 | 0.0110 |
| -2.1 | 0.0179 | 0.0174 | 0.0170 | 0.0166 | 0.0162 | 0.0158 | 0.0154 | 0.0150 | 0.0146 | 0.0143 |
| -2.0 | 0.0228 | 0.0222 | 0.0217 | 0.0212 | 0.0207 | 0.0202 | 0.0197 | 0.0192 | 0.0188 | 0.0183 |
| -1.9 | 0.0287 | 0.0281 | 0.0274 | 0.0268 | 0.0262 | 0.0256 | 0.0250 | 0.0244 | 0.0239 | 0.0233 |
| -1.8 | 0.0359 | 0.0351 | 0.0344 | 0.0336 | 0.0329 | 0.0322 | 0.0314 | 0.0307 | 0.0301 | 0.0294 |
| -1.7 | 0.0446 | 0.0436 | 0.0427 | 0.0418 | 0.0409 | 0.0401 | 0.0392 | 0.0384 | 0.0375 | 0.0367 |
| -1.6 | 0.0548 | 0.0537 | 0.0526 | 0.0516 | 0.0505 | 0.0495 | 0.0485 | 0.0475 | 0.0465 | 0.0455 |
| -1.5 | 0.0668 | 0.0655 | 0.0643 | 0.0630 | 0.0618 | 0.0606 | 0.0594 | 0.0582 | 0.0571 | 0.0559 |
| -1.4 | 0.0808 | 0.0793 | 0.0778 | 0.0764 | 0.0749 | 0.0735 | 0.0721 | 0.0708 | 0.0694 | 0.0681 |
| -1.3 | 0.0968 | 0.0951 | 0.0934 | 0.0918 | 0.0901 | 0.0885 | 0.0869 | 0.0853 | 0.0838 | 0.0823 |
| -1.2 | 0.1151 | 0.1131 | 0.1112 | 0.1093 | 0.1075 | 0.1056 | 0.1038 | 0.1020 | 0.1003 | 0.0985 |
| -1.1 | 0.1357 | 0.1335 | 0.1314 | 0.1292 | 0.1271 | 0.1251 | 0.1230 | 0.1210 | 0.1190 | 0.1170 |
| -1.0 | 0.1587 | 0.1562 | 0.1539 | 0.1515 | 0.1492 | 0.1469 | 0.1446 | 0.1423 | 0.1401 | 0.1379 |
| -0.9 | 0.1841 | 0.1814 | 0.1788 | 0.1762 | 0.1736 | 0.1711 | 0.1685 | 0.1660 | 0.1635 | 0.1611 |
| -0.8 | 0.2119 | 0.2090 | 0.2061 | 0.2033 | 0.2005 | 0.1977 | 0.1949 | 0.1922 | 0.1894 | 0.1867 |
| -0.7 | 0.2420 | 0.2389 | 0.2358 | 0.2327 | 0.2296 | 0.2266 | 0.2236 | 0.2207 | 0.2177 | 0.2148 |
| -0.6 | 0.2743 | 0.2709 | 0.2676 | 0.2643 | 0.2611 | 0.2578 | 0.2546 | 0.2514 | 0.2483 | 0.2451 |
| -0.5 | 0.3085 | 0.3050 | 0.3015 | 0.2981 | 0.2946 | 0.2912 | 0.2877 | 0.2843 | 0.2810 | 0.2776 |
| -0.4 | 0.3446 | 0.3409 | 0.3372 | 0.3336 | 0.3300 | 0.3264 | 0.3228 | 0.3192 | 0.3156 | 0.3121 |
| -0.3 | 0.3821 | 0.3783 | 0.3745 | 0.3707 | 0.3669 | 0.3632 | 0.3594 | 0.3557 | 0.3520 | 0.3483 |
| -0.2 | 0.4207 | 0.4168 | 0.4129 | 0.4090 | 0.4052 | 0.4013 | 0.3974 | 0.3936 | 0.3897 | 0.3859 |
| -0.1 | 0.4602 | 0.4562 | 0.4522 | 0.4483 | 0.4443 | 0.4404 | 0.4364 | 0.4325 | 0.4286 | 0.4247 |
| -0.0 | 0.5000 | 0.4960 | 0.4920 | 0.4880 | 0.4840 | 0.4801 | 0.4761 | 0.4721 | 0.4681 | 0.4641 |

|     | 0      | 1      | 2      | 3      | 4      | 5      | 6      | 7      | 8      | 9      |
|-----|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 0.0 | 0.5000 | 0.5040 | 0.5080 | 0.5120 | 0.5160 | 0.5199 | 0.5239 | 0.5279 | 0.5319 | 0.5359 |
| 0.1 | 0.5398 | 0.5438 | 0.5478 | 0.5517 | 0.5557 | 0.5596 | 0.5636 | 0.5675 | 0.5714 | 0.5753 |
| 0.2 | 0.5793 | 0.5832 | 0.5871 | 0.5910 | 0.5948 | 0.5987 | 0.6026 | 0.6064 | 0.6103 | 0.6141 |
| 0.3 | 0.6179 | 0.6217 | 0.6255 | 0.6293 | 0.6331 | 0.6368 | 0.6406 | 0.6443 | 0.6480 | 0.6517 |
| 0.4 | 0.6554 | 0.6591 | 0.6628 | 0.6664 | 0.6700 | 0.6736 | 0.6772 | 0.6808 | 0.6844 | 0.6879 |
| 0.5 | 0.6915 | 0.6950 | 0.6985 | 0.7019 | 0.7054 | 0.7088 | 0.7123 | 0.7157 | 0.7190 | 0.7224 |
| 0.6 | 0.7257 | 0.7291 | 0.7324 | 0.7357 | 0.7389 | 0.7422 | 0.7454 | 0.7486 | 0.7518 | 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 | 0.8943 | 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 | 0.9773 | 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 | 0.9990 | 0.9993 | 0.9995 | 0.9997 | 0.9998 | 0.9998 | 0.9999 | 0.9999 | 1.0000 |

Wie berechnet man sie? Durch numerische Integration.

Wie liest man sie? Das wurde auf Seite 52 ausführlich erläutert.

Ein Beispiel: Wie groß ist  $\mathbb{P}([-1.21, 0.88])$ ? Der Tabelle entnehmen wir, dass  $\mathbb{P}(]-\infty, 0.88]) = 0.8106$  und  $\mathbb{P}(]-\infty, -1.21]) = 0.1131$ . Folglich ist

$$\mathbb{P}([-1.21, 0.88]) = 0.8106 - 0.1131 = 0.6975 :$$

Mit fast 70 Prozent Wahrscheinlichkeit liegt der Wert einer  $N(0, 1)$ -verteilten Zufallsvariablen in  $[-1.21, 0.88]$ .

**Konfidenzintervalle bei Binomialverteilungen**

*Wann braucht man solche Tabellen?* Dann, wenn man aufgrund der Anzahl der Erfolge bei  $n$  Versuchen ein Konfidenzintervall für eine unbekannte Wahrscheinlichkeit finden möchte (vgl. Abschnitt 10.4).

→  
Programm!

*Die Tabellen* kann man sich mit dem auf der Internetseite zu diesem Buch zur Verfügung gestellten Programm ausgeben lassen

Wie werden sie berechnet? Angenommen,  $n$  und  $\alpha$  sind vorgegeben. Dann möchte man doch bei konkreter Erfolgsanzahl  $k$  ein Konfidenzintervall  $[p_*, p^*] \subset [0, 1]$  finden, so dass das wirkliche  $p$  mit Wahrscheinlichkeit  $1 - \alpha$  in diesem Intervall liegt.

Man kann, wie in Abschnitt 10.4 beschrieben, zur Bestimmung von  $[p_*, p^*]$  so vorgehen:

- Berechne für  $p$  in einer genügend feinen Unterteilung von  $[0, 1]$  (z. B.  $p = i/1000$  für  $i = 0, \dots, 1000$ ) ein möglichst kleines Intervall  $[n(p), m(p)]$  in  $\mathbb{N}$  mit der Eigenschaft  $\sum_{k=n(p)}^{m(p)} b(k, n; p) \geq 1 - \alpha$ .
- Definiere dann, bei vorgegebenem  $k$ , das Intervall  $[p_*, p^*]$  durch

$$[p_*, p^*] := \{p \mid k \in [n(p), m(p)]\}.$$

(Genauer:  $p_*$  bzw.  $p^*$  ist das Minimum bzw. das Maximum der berechneten  $n(p)$  bzw.  $m(p)$ , für die  $k \in [n(p), m(p)]$  gilt.)

*Ein Beispiel* wurde auf Seite 306 behandelt.

**Einige Konfidenzintervalle für die  $\chi$ -Quadrat-Verteilungen**

*Wann braucht man diese Tabelle?* Diese Frage wird in Abschnitt 10.5 beantwortet: Wenn man für die Varianz einer Normalverteilung aus  $n$  Beobachtungen ein Konfidenzintervall konstruieren möchte, spielt  $\chi_n^2$  oder  $\chi_{n-1}^2$  eine Rolle, je nachdem, ob man den Erwartungswert kennt oder nicht. Das ist auch der Grund dafür, dass nachstehend nur die Werte für einige Zahlen  $n$  und  $n - 1$  angegeben sind, wobei  $n$  durch 5 teilbar ist.

*Die Tabelle*

| $n$ | $\alpha = 0.01$ | $\alpha = 0.02$ | $\alpha = 0.5$ |
|-----|-----------------|-----------------|----------------|
| 4   | [0.21, 14.86]   | [0.30, 13.28]   | [0.48, 11.14]  |
| 5   | [0.41, 16.75]   | [0.55, 15.09]   | [0.83, 12.83]  |
| 9   | [1.73, 23.59]   | [2.09, 21.67]   | [2.70, 19.02]  |
| 10  | [2.16, 23.58]   | [2.56, 21.66]   | [3.25, 20.48]  |
| 14  | [4.07, 31.32]   | [4.66, 29.14]   | [5.62, 26.12]  |
| 15  | [4.60, 32.80]   | [5.22, 30.58]   | [6.26, 27.49]  |
| 19  | [6.84, 38.58]   | [7.63, 36.19]   | [8.90, 32.85]  |
| 20  | [7.43, 40.00]   | [8.26, 37.57]   | [9.59, 34.17]  |
| 24  | [9.89, 45.59]   | [10.86, 42.98]  | [12.40, 39.36] |
| 25  | [10.52, 46.93]  | [11.52, 44.31]  | [13.12, 40.65] |
| 29  | [13.12, 52.34]  | [14.26, 49.59]  | [16.05, 45.72] |
| 30  | [20.70, 53.67]  | [14.95, 50.89]  | [16.79, 46.98] |

Wie berechnet man sie? Durch numerische Integration.

*Wie liest man sie?* Man geht in die Spalte, in der der gewünschte Freiheitsgrad steht, und dann in die Zeile, die zu dem zum Problem gehörige Freiheitsgrad gehört.

*Ein Beispiel:* Ein Konfidenzintervall zu  $\alpha = 0.02$  und 24 Freiheitsgraden ist durch [10.86, 42.98] gegeben. Das bedeutet: Dieses Intervall hat die Eigenschaft, dass eine  $\chi^2_{24}$ -verteilte Zufallsvariable mit Wahrscheinlichkeit  $1 - 0.02 = 0.98$  ihre Werte dort annimmt.

#### Einige Konfidenzintervalle der $t$ -Verteilungen

*Wann braucht man diese Tabelle?* Auch hierfür sollte man Abschnitt 10.5 konsultieren: Die  $t$ -Verteilungen werden dann wichtig, wenn man Konfidenzintervalle für den Erwartungswert einer Normalverteilung bei unbekannter Streuung finden möchte.

*Die Tabelle*

| $n$ | $\alpha = 0.01$ | $\alpha = 0.02$ | $\alpha = 0.05$ |
|-----|-----------------|-----------------|-----------------|
| 4   | [-4.60, 4.60]   | [-3.75, 3.75]   | [-2.78, 2.78]   |
| 9   | [-3.25, 3.25]   | [-2.82, 2.82]   | [-2.26, 2.26]   |
| 14  | [-2.98, 2.98]   | [-2.62, 2.62]   | [-2.15, 2.15]   |
| 19  | [-2.80, 2.80]   | [-2.54, 2.54]   | [-2.09, 2.09]   |
| 24  | [-2.78, 2.78]   | [-2.49, 2.49]   | [-2.06, 2.06]   |
| 29  | [-2.76, 2.76]   | [-2.46, 2.46]   | [-2.05, 2.05]   |

Wie berechnet man sie? Durch numerische Integration.

*Wie liest man sie?* Man geht in die Spalte, in der der gewünschte Freiheitsgrad steht, und dann in die Zeile, die zu dem zum Problem gehörige Freiheitsgrad gehört.

*Ein Beispiel* ist auf Seite 312 zu finden.

### Die Wilcoxon-Verteilungen

*Wann braucht man die nachstehende Tabelle?* Wenn man mit einem Rangtest entscheiden möchte, ob zwei Stichproben gemäß der gleichen Verteilung erzeugt wurden.

#### Die Tabelle

Man kann die Tabellen mit dem Programm berechnen, das auf der zum Buch gehörigen Internetseite zu finden ist.

Wie wurden diese Tabellen berechnet? Mit Hilfe der am Ende von Abschnitt 12.3 beschriebenen Rekursionsformel.

#### Wie liest man sie?

Man gibt  $k$  und  $l$  ein und lässt die zugehörige Tabelle berechnen. Dann sucht man bei vorgegebenem  $\alpha$  ein Intervall  $I = \{m_*, m_* + 1, \dots, m^*\}$ , so dass  $\mathbb{P}(\{m \notin I\}) \leq \alpha$  gilt. Im konkreten Einzelfall wird die Nullhypothese genau dann abgelehnt, wenn das zugehörige  $m$  nicht in  $I$  liegt.

*Ein Beispiel* wurde in Abschnitt 12.3 behandelt.

### Einige Tabellenwerte zum Kolmogoroff-Smirnoff-Test

*Wann braucht man die nachstehende Tabelle?* Wenn man den Kolmogoroff-Smirnoff-Test anwenden möchte: Wurde die vorgelegte Stichprobe gemäß einer bestimmten Verteilung erzeugt?

#### Die Tabelle

| $n$    | $\alpha = 0.01$ | $\alpha = 0.025$ | $\alpha = 0.05$ |
|--------|-----------------|------------------|-----------------|
| 5      | 0.51            | 0.56             | 0.63            |
| 10     | 0.37            | 0.41             | 0.46            |
| 15     | 0.30            | 0.34             | 0.38            |
| 20     | 0.26            | 0.29             | 0.33            |
| 25     | 0.24            | 0.26             | 0.30            |
| 30     | 0.22            | 0.24             | 0.27            |
| 35     | 0.20            | 0.22             | 0.25            |
| 40     | 0.19            | 0.21             | 0.24            |
| $> 40$ | $1.22/\sqrt{n}$ | $1.36/\sqrt{n}$  | $1.52/\sqrt{n}$ |

Wie berechnet man sie? Durch numerische Integration.

#### Wie liest man sie?

*Ein Beispiel* wurde am Ende von Abschnitt 12.4 diskutiert.

## Die Computerprogramme zum Buch

Auf der Internetseite

<http://www.springer-spektrum.de/Buch/978-3-8348-1939-0/Elementare-Stochastik.html>  
sind Ergänzungen zu diesem Buch zu finden. Unter anderem gibt es dort ein Computerprogramm, das jeder für nichtkommerzielle Zwecke frei herunterladen kann. Es handelt sich um eine exe-Datei, die auf allen Windows-Rechnern problemlos laufen sollte.

Am einfachsten ist es, die Datei `behrends_programm.zip` in einen neuen Ordner zu kopieren und zu entpacken. Dann sollte man `liesmich.txt` lesen, und danach kann es mit Klicken auf die exe-Datei losgehen.

Das Programm bietet an:

- *Verteilungen (Werte).* Die Wahrscheinlichkeiten für die hier behandelten diskreten Räume (Binomial, Poisson usw.).
- *Verteilungen (Skizzen).* Einige Wahrscheinlichkeitsdichten können gezeichnet werden.
- *Simulationen.* Es werden Zufallszahlen zu verschiedenen Verteilungen erzeugt: Gleichverteilung, Normalverteilung, Poissonverteilung, usw.
- *Projekte.* Hier können Monte-Carlo-Verfahren getestet werden, insbesondere die  $\pi$ -Berechnung à la Buffon. Auch kann man den Affen an die Schreibmaschine setzen, sich die Funktionen im Kolmogoroff-Smirnoff-Test skizzieren lassen usw.
- *Paradoxien.* Durch Simulation kann man sich davon überzeugen, dass sich die theoretischen Ergebnisse zum Geburtstagsparadoxon und zum Über-einstimmungsparadoxon reproduzieren lassen.
- *Zentraler Grenzwertsatz.* Das ist der Favorit des Autors unter den Programmen: Bei Überlagerung entsteht immer die Glockenkurve!
- *Tabellen.* Hier werden Tabellen in Realzeit berechnet, die in den Kapiteln zur Statistik benötigt werden.

## Literatur

### A. Literatur zur elementaren Stochastik

Hier weisen wir nur auf eine kleine Auswahl aus dem aktuellen Angebot hin:

**Bewersdorff, Jörg.** Statistik - wie und warum sie funktioniert. Vieweg+Teubner, 2011

Statistik wird hier sehr elementar erläutert. Zu empfehlen für alle, denen der Statistik-Teil dieses Buches nicht ausführlich genug oder zu mathematisch ist.

**Fischer, Gerd.** Stochastik einmal anders. Vieweg 2005

Eine sehr ausführliche Darstellung, die allerdings bei der exakten Darstellung des mathematischen Hintergrunds nicht sehr weit geht.

**Georgii, Hans-Otto.** Stochastik (4. Auflage). de Gruyter, 2009.

Eine empfehlenswerte Einführung, die wesentlich kompakter geschrieben ist als das vorliegende Buch.

### B. Zum Weiterlesen

**Behrends, Ehrhard.** An Introduction to Markov Chains with Special Emphasis on Rapid Mixing. Vieweg, 1998.

Ein spezielles Gebiet der (etwas fortgeschrittenen) Wahrscheinlichkeitstheorie wird hier ausführlich dargestellt. So genannte Markovketten sind Modelle für das Verhalten des Zufalls, wenn auch eine zeitliche Entwicklung eine Rolle spielt.

**Klenke, Achim.** Wahrscheinlichkeitstheorie. Springer, 2005.

Hier findet man die Weiterführung der (nicht mehr so elementaren) Wahrscheinlichkeitstheorie, es gibt auch erste Ergebnisse zu Markovprozessen und stochastischen Differentialgleichungen.

**Meintrup, David - Schäffler, Stefan.** Stochastik. Springer 2005.

Eine empfehlenswerte Gesamtdarstellung der Stochastik, die bis zum Ito-Integral geht. Es gibt auch viele Anwendungen.

### C. Zum besseren Kennenlernen: Maß- und Integrationstheorie

**Behrends, Ehrhard.** Maß- und Integrationstheorie. Springer, 1983.

Das ist ein vor langer Zeit geschriebenes Buch, in dem man alles findet, was für die Stochastik aus der Maß- und Integrationstheorie relevant ist. Es ist allerdings nur noch antiquarisch verfügbar.

**Elstrodt, Jürgen.** Maß- und Integrationstheorie. Springer, 1996.

Mein Favorit für eine gute Darstellung der Maß- und Integrationstheorie. Alles, was man für die Stochastik benötigt, ist enthalten, auch gibt es viele historische Anmerkungen.

#### D. Zum besseren Kennenlernen: Analysis

**Behrends, Ehrhard.** Analysis 1 und 2 (5. Auflage). Vieweg+Teubner, 2011.  
Hier findet man alles ausführlich dargestellt, was in dem Anhang zur Analysis kurz zusammengestellt wurde.

#### D. Zum besseren Kennenlernen: Lineare Algebra

**Fischer, Gerd.** Lineare Algebra. Vieweg+Teubner 2010

Ein Klassiker. Hier kann man alles nachlesen, was im vorliegenden Buch zu Räumen mit Skalarprodukt relevant ist.

**Liesen, Jörg und Mehrmann, Volker.** Lineare Algebra. Vieweg+Teubner, 2011

Eine Neuerscheinung, die auch sehr empfohlen werden kann.

#### E. Weitere, eher populäre Bücher des Autors

**Aigner, Martin und Behrends, Ehrhard (Herausgeber).** Alles Mathematik. Vieweg+Teubner, 2009.

Ein Buch für alle, die an den verschiedenen Aspekten der Mathematik interessiert sind. (Auch auf Englisch verfügbar.)

**Behrends, Ehrhard.** Fünf Minuten Mathematik. Vieweg+Teubner, 2009.

Das ist eine ausführlichere Version der Kolumne, die zwei Jahre lang jede Woche in der Zeitung DIE WELT erschienen. (Auch auf Englisch, Französisch und Japanisch verfügbar.)

**Behrends, Ehrhard, Gritzmann, Peter, und Ziegler, Günter (Herausgeber).**  $\pi$  und Co. - Kaleidoskop der Mathematik. Springer, 2008.

Dieses Buch wurde zum Jahr der Mathematik 2008 herausgegeben. Es enthält viele Artikel zu verschiedenen Aspekten der Mathematik. Seine „offizielle“ Funktion: Es ist der Abiturpreis der Deutschen Mathematikervereinigung. An jedem Gymnasium Deutschlands erhält der beste Abiturient / die beste Abiturientin ein Exemplar dieses Buches.

# Register

- $N(a, \sigma^2)$ , 51  
 $\#E$ , 10  
 $\Omega$ , 5  
 $\mathcal{E}$ , 8  
 $\mathbb{P}(E)$ , 6  
 $\mathbb{P}_X$ , 76  
 $\mathcal{D}$ , 23  
 $\mathcal{P}$ , 7  
 $\phi_{\mathbb{P}}$ , 108  
 $\sigma$ -Algebra, 8, 11  
 $\sigma$ -Algebra der Borelmengen, 17  
 $\sigma$ -Algebra, erzeugte, 14  
 $\sigma$ -additiv, 9  
 $\sigma(\mathcal{M})$ , 14  
 $\{X \in F\}$ , 73  
 $\mathcal{D}(\mathcal{M}) \subset \mathcal{D}$ , 23  
 $\binom{n}{k}$ , 93  
abgeschlossen, 18  
abzählbar, 8  
Äquivalenzrelation, 353  
Affe an der Schreibmaschine, 225  
Alternativhypothese, 318  
Auswahlaxiom, 27, 353  
Bayes, 124  
bedingte Wahrscheinlichkeit, 116  
Bernoulliraum, 40  
Binomialkoeffizient, 94  
Binomialverteilung, 164  
Borel-Cantelli, Lemmata von, 222  
Borel-Lebesgue-Maß, 26, 357  
Borelmenge, 17  
Borelmengen im  $\mathbb{R}^n$ , 19  
Borelmengen, Erzeuger der, 19  
Boxplots, 272  
Buffonsches Nadelexperiment, 47  
Carathéodory, 357  
Cauchyverteilung, 106  
charakteristische Funktion, 259  
Chiquadrattest auf Unabhängigkeit, 341  
Computer, viii  
Dichtefunktion, 43  
diskrete gedächtnislose Wartezeiten, 198  
diskreter Raum, 38  
Dynkinsystem, 23  
Elementarereignis, 5  
Entemadi, Satz von, 235  
Entscheidungsfunktion, 321  
Ereignis, 6  
erwartungstreuer Schätzer, 291  
Erwartungswert, 80, 81, 83, 86  
Erwartungswert der Exponentialverteilung, 83  
Erwartungswert der hypergeometrischen Verteilung, 102  
Erwartungswert der Normalverteilung, 84  
erzeugende Funktion, 108  
euklidische Norm, 359  
Eulersche Zahl, 171  
Exponantialverteilung, Maxima, 195  
Exponentialverteilung, 50, 190  
Exponentialverteilung, Minima, 198  
Exponentialverteilung, Summen, 193  
Faltung, 149, 152  
fast sichere Konvergenz, 209  
Fehler erster Art, 318  
Fehler zweiter Art, 318  
Freiheitsgrad, 307

- Gütefunktion, 321  
Geburtstagsparadoxon, 97  
gedächtnislos, 188, 198  
gedächtnislose Wartezeit, 188  
geometrische Verteilung, 42, 167, 199  
Gleichverteilung, 45
- Histogramm, 271  
Homepage, viii  
hypergeometrische Verteilung, 98  
Hypothese, 318
- Indikatorfunktion, 73  
induzierter Wahrscheinlichkeitsraum, 76  
induziertes Wahrscheinlichkeitsmaß, 76
- Infimum, 360  
Inklusion-Exklusion-Satz, 95  
integrabel, 358  
Irrtumsniveau, 323  
Irrtumswahrscheinlichkeit, 302  
iterierter Logarithmus, 253
- Klonsatz, 141  
Kolmogoroff-Smirnoff-Test, 346  
Kombinatorik, 91  
Konfidenzbereichsschätzung, 302  
Konfidenzintervall, 303  
Konfidenzniveau, 302, 323  
Konsistenz, 296  
Kontingenztafel, 272  
Konvergenz in Verteilung, 211  
Konvergenz in Wahrscheinlichkeit, 208  
Konvergenz, fast sichere, 209  
Korrelationskoeffizient, 277  
kritischer Bereich, 321
- Laplaceraum, 40  
Lemmata von Borel-Cantelli, 222  
Likelihood-Quotient, 328  
Limes superior, 222  
Lotto, 99
- Mann-Whitney-Test, 345  
Markov-Ungleichung, 230  
maximum-likelihood-Schätzung, 300  
Maßraum, 357  
Median, 275
- Mengendifferenz, 353  
Mengensystem, 7  
Merkmale, qualitative, 270  
Merkmale, quantitative, 270  
messbar, 357  
Messraum, 357  
Monte-Carlo, 47
- negative Binomialverteilung, 166  
Neyman-Pearson-Test, 328  
Normalverteilung, 51  
Nullhypothese, 318
- offen, 17  
orthogonal, 359
- Paradoxie, 98  
Partition, 9  
Poissonverteilung, 41, 194  
Popstarbildchen-Problem, 104  
Potenzmenge, 7, 353  
Prävalenz, 125  
Pseudozufallszahl, 64  
Punktmaß, 33  
Pythagoras, 359
- Quartil, 272
- Rangmerkmale, 270  
Rangtests, 342  
reflexiv, 353  
Regressionsgerade, 279  
Relation, 353
- Satz vom iterierten Logarithmus, 253  
Satz von Bayes, 124  
Satz von de Moivre-Laplace, 179  
Satz von der totalen Wahrscheinlichkeit, 123  
Satz von Lehmann-Scheffé, 299  
Schätzer, 290  
Schätzer, erwartungstreuer, 291  
Schätzfunktion, 290  
schnitt-stabil, 23  
schwaches Gesetz der großen Zahlen, 232  
Sensitivität, 125

- sigma-Algebra, 8
- Simpsonparadoxon, 281
- Simulation, 53
- Simulation: abzählbare Wahrscheinlichkeitsräume, 55
- Simulation: Bernoulliraum, 56
- Simulation: endliche Wahrscheinlichkeitsräume, 55
- Simulation: geometrische Verteilung, 57
- Simulation: Laplaceräume, 55
- Simulation: Poissonverteilung, 56
- Simulation: Räume mit Dichtefunktionen, 57
- Skalarprodukt, 359
- Spezifität, 125
- starkes Gesetz der großen Zahlen, 235
- Statistik, 298
  - statistisches Modell, 290
  - Stichprobe, 273, 289
  - Stichprobenmittel, 273
  - Stichprobenstreuung, 275
  - Stichprobenvarianz, 274
  - Stirlingformel, 175
  - stochastischer Test, 326
  - Streuung, 88
  - suffiziente Statistik, 298
  - Supremum, 360
  - symmetrisch, 353
- Test, 321
  - Test zum Irrtumsniveau  $\alpha$ , 323
  - Test zum Konfidenzniveau  $1 - \alpha$ , 323
  - Testfunktion, 321
  - Tortendiagramm, 271
  - totale Wahrscheinlichkeit, 123
  - transitiv, 353
  - Treppenfunktion, 358
  - Tschebyscheff-Ungleichung, 229
- Übereinstimmungsparadoxon, 103
- U-Statistik, 343
- unabhängig, 120, 130, 131, 135, 139
- unbedingte Konvergenz, 359
- ungeordnete Summation, 360
- Varianz, 88
- Verlustmatrix, 324
- Verteilungsfunktion einer Zufallsvariable, 76
- Verwerfungsbereich, 321
- vollständige Statistik, 299
- Würfel, 10
- Wahrscheinlichkeit, 6, 9
- Wahrscheinlichkeitsdichte, 43
- Wahrscheinlichkeitsmaß, 9
- Wahrscheinlichkeitsraum, 10
- Wallisprodukt, 174
- Wartezeit, 188
- Wilcoxon-Test, 345
- Wohlordnungsaxiom, 354
- Wurzel- $n$ -Gesetz, 147
- zentraler Grenzwertsatz, 242, 246
- Ziegenprobem, 126
- Zornsches Lemma, 354
- Zufallsautomat, 5
- Zufallsvariable, 72