

Statistik und ihre Anwendungen

Daniel Bättig

# Angewandte Datenanalyse

Der Bayes'sche Weg

*2. Auflage*



Springer Spektrum

---

*Reihenherausgeber*

Prof. Dr. Holger Dette · Prof. Dr. Wolfgang Härdle

# Statistik und ihre Anwendungen

Weitere Bände dieser Reihe finden Sie unter  
<http://www.springer.com/series/5100>

---

Daniel Bättig

# Angewandte Datenanalyse

Der Bayes'sche Weg

2., überarbeitete und erweiterte Auflage



Springer Spektrum

Daniel Bättig  
Institut für Risiko- und Extremwertanalyse  
Berner Fachhochschule  
Burgdorf, Schweiz

Statistik und ihre Anwendungen  
ISBN 978-3-662-54219-4  
DOI 10.1007/978-3-662-54220-0

ISBN 978-3-662-54220-0 (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  
© Springer-Verlag GmbH Deutschland 2015, 2017  
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. Der Verlag, die Autoren und die Herausgeber gehen davon aus, dass die Angaben und Informationen in diesem Werk zum Zeitpunkt der Veröffentlichung vollständig und korrekt sind. Weder der Verlag noch die Autoren oder die Herausgeber übernehmen, ausdrücklich oder implizit, Gewähr für den Inhalt des Werkes, etwaige Fehler oder Äußerungen. Der Verlag bleibt im Hinblick auf geografische Zuordnungen und Gebietsbezeichnungen in veröffentlichten Karten und Institutionsadressen neutral.

Planung: Iris Ruhmann

Gedruckt auf säurefreiem und chlorfrei gebleichtem Papier.

Springer Spektrum ist Teil von Springer Nature  
Die eingetragene Gesellschaft ist Springer-Verlag GmbH Germany  
Die Anschrift der Gesellschaft ist: Heidelberger Platz 3, 14197 Berlin, Germany

*Für Franziska*

---

# Vorwort

---

## Information, Unsicherheit und Statistik

Menschen sind interessiert daran, zukünftige Ereignisse einschätzen zu können. Dass dieses Anliegen komplex ist, ist einleuchtend und soll nachfolgend angedacht werden. Die Fahrzeit des Personenzugs, der am nächsten Tag um acht Uhr von Bern nach Zürich fährt, ist nicht exakt prognostizierbar. Dies ist so, weil Informationen zu den herrschenden Wetterbedingungen, zu den Verhaltensweisen der Passagiere und zum Verkehrsaufkommen auf dem Schienennetz nicht vollständig erfassbar sind. Der Weltmarktpreis für ein Kilo Weizen am 1. Dezember des nächsten Jahrs kann nur unsicher prognostiziert werden, weil Informationen zu den Anbauflächen, zum Wetter oder zur Inflation fehlen. Die Lebensdauern von Menschen zu bestimmen ist schwierig: Informationen zu Lebensdauern aufgrund der Körperkonstitution, des Lebensortes, der Lebensgewohnheiten u. a. m. fehlen, um eine präzise Rechnung zu machen.

In einem Produktionsprozess können wegen sich ändernden Bedingungen – Arbeitsteams, die wechseln, Rohstoffe, die in der Qualität streuen – keine Fernsehgeräte produziert werden, die eine identische Lebensdauer haben. Um verlässliche Aussagen zu nicht direkt messbaren Größen, wie die durchschnittliche Lebensdauer eines Fernsehgeräts einer Produktionsserie, zu machen, ist Information oder Wissen notwendig. Viele nicht direkt messbare Größen müssen Ingenieurinnen und Ingenieure bestimmen. So kann der Druck in einer Kammer nur indirekt mit Apparaturen gemessen werden. Wegen variierender Bedingungen und wegen Messungenauigkeiten der Apparaturen erhält man Messwerte, die um den gesuchten Druck mehr oder weniger streuen.

Spezifische Information zu (a) zukünftigen Werten von unsicheren Größen oder (b) zu nicht direkt messbaren Größen kann man mit Messungen, Zahlen und Daten, sowie mit Sachwissen erlangen. Ein Blick in die ersten geschriebenen Dokumente der Menschheit zeigt, dass das geordnete Zusammenstellen von Zahlen und Daten eine langjährige Tradition hat. So stellen die ältesten bekannten Schrifttafeln, die „Tafeln von Uruk“ aus dem 4. Jahrtausend vor Christus, Auszüge über die soziale Organisation einer Bevölkerungsgruppe dar. Man erfährt, dass die religiöse Gemeinschaft des Tempels Lagash unter anderem aus 18 Bäckern, 31 Brauern, 7 Sklaven bestand.

Die Zeit der modernen Statistik begann in der zweiten Hälfte des 18. Jahrhunderts, als grosse und recht komplexe Datensätze von Nationalstaaten untersucht wurden. Volkszählungen für die Erhebung von Steuern oder für die Rekrutierung von Heeren waren für die Staaten wichtig. Zahlungsbilanzen zwischen Staaten wurden betrachtet und analysiert. Derartige Daten in Tabellen darzustellen, um besondere Merkmale hervorzuheben, war keine geeignete Methode mehr. Grafische Methoden, um grosse Datenmengen darzustellen, wurden erfunden: Stabdiagramme, Histogramme und Grafiken, um Zeitreihen zu visualisieren. Im 20. Jahrhundert wurden im Rahmen der Massenproduktion in der Industrie, die durch die automatisierte Produktion von Autos, Fernsehgeräten, Medikamenten, Chips und integrierten Schaltkreisen gekennzeichnet ist, vielfältige Arten von grafischen Darstellungen erfunden, die eine schnelle und effiziente Analyse der Produktion ermöglichen. Diese Darstellungen spielen bei der Qualitätskontrolle eine wichtige Rolle. Als Beispiele seien Kontrollkarten und Box & Whisker Plots erwähnt.

Die moderne Statistik ist die Wissenschaft, die einerseits Methoden aufzeigt, wie Daten oder Messwerte effizient gesammelt werden sollten. Man spricht von der *Versuchsplanung* (engl. *Design of Experiments*). Insbesondere versucht man bei minimalem Aufwand einen maximalen Ertrag an Informationen zu erhalten. Andererseits erklärt die Statistik, wie mit der Information aus Daten und Messwerten nicht direkt messbare Größen berechnet oder zukünftige Werte unsicherer Größen prognostiziert werden können. Wie plausibel solche Rechnungen oder Prognosen sind, wird in der statistischen Arbeit mit einer Wahrscheinlichkeit ausgedrückt. So will eine Ärztin wissen, mit welcher Wahrscheinlichkeit ein Medikament bei einer Person wirken wird. Oder ein Produzent möchte berechnen, wie lang die durchschnittliche Lebensdauer seiner hergestellten Fernsehgeräte ist und wie zuverlässig – formuliert mit einer Wahrscheinlichkeit – eine solche Aussage ist. Das Resultat einer derartigen Rechnung könnte so aussehen: „Mit einer Wahrscheinlichkeit von 90 % beträgt die durchschnittliche Lebensdauer der Fernsehgeräte zwischen 10 und 12 Jahren.“ Oder eine Geologin möchte prognostizieren, wie lange man auf das nächste schwere Erdbeben warten muss und wie sicher diese Angabe ist. Es war Pierre-Simon Laplace (1749–1827), der Anfang des 19. Jahrhunderts auf die Idee kam, Wahrscheinlichkeiten zu benutzen, um Plausibilitäten zu nicht direkt messbaren Größen aus Astronomie, Natur- und Sozialwissenschaften auszudrücken. Unter Statistikern ist aber umstritten, was eine Wahrscheinlichkeit von beispielsweise 90 % bedeutet und wie sie aus Daten berechnet werden soll:

„It is unanimously agreed that statistics depends somehow on probability. But, as to what probability is and how it is connected with statistics, there has seldom been such complete disagreement and breakdown of communication since the Tower of Babel.“

L. J. Savage: *The Foundations of Statistics*, Dover Publications, Inc. New York, 1972, S. 2.

In diesem Buch sind Wahrscheinlichkeiten dazu da, um zu messen, wie *plausibel* Aussagen sind. So bedeutet die Aussage „Mit einer Wahrscheinlichkeit von 90 % beträgt die durchschnittliche Lebensdauer der Fernsehgeräte zwischen 10 und 12 Jahren“, dass man bereit ist, 90 zu 10 Franken zu wetten, dass die durchschnittliche Lebensdauer der

Fernsehgeräte zwischen 10 und 12 Jahren ist. Man spricht auch von der Bayes'schen Wahrscheinlichkeitsrechnung. Sehr verbreitet sind auch andere *frequentistische* Interpretationen. Diese werten Wahrscheinlichkeiten als Langzeit-Häufigkeiten. Dies ist vor allem dann interessant, wenn aus Probandengruppen auf Parameter einer Gesamtpopulation gerechnet wird. Die Fragen, die sich dabei stellen, sind: Wie wäre die statistische Rechnung ausgefallen, wenn man eine andere Probandengruppe ausgewählt hätte? Was wäre passiert, wenn man das Experiment wiederholt hätte? Man spricht hier von der *Stichprobenunsicherheit* (engl. *random error*) der Rechnung. Solche Rechnungen stehen nicht im Zentrum des Buchs. In Ingenieur- und Wirtschaftswissenschaften sind sie meist wenig interessant. Man hat Daten wie Messungen zu Defekten, zu Temperaturen, zu Unfallzahlen oder zu Schadensummen. Man kümmert sich nicht um „imaginäre Wiederholungen“, sondern es sollen Plausibilitäten zu Parametern gerechnet, Prognosen gemacht, sowie Entscheide getroffen werden.

---

## Inhalt und Leserschaft

Der Inhalt des Buches zeigt vor allem, wie Daten oder allgemeiner, wie Information benutzt werden kann, um *nicht direkt messbare Größen* zu bestimmen, Prognosen zu *zukünftigen Werten von unsicheren Größen* zu rechnen, *Regressionsmodelle* zu konstruieren und solche Modelle zu *vergleichen* und *auszuwählen*. Hier Beispiele dazu:

### Nicht direkt messbare Grösse rechnen

Eine Person möchte ihr Gewicht bestimmen. Sie steht dazu viermal auf eine Personenwaage und liest folgende Zahlen ab:

$$75,5 \text{ kg}, \quad 74,8 \text{ kg}, \quad 75,2 \text{ kg}, \quad 75,7 \text{ kg}$$

Die angezeigten Gewichte variieren, weil das Luftvolumen im Körper ändert, die Person nicht immer gleich ruhig auf der Waage steht und die Waage selber Messunsicherheiten hat. Das Gewicht der Person ist daher nicht direkt messbar. In der Hoffnung, die zufälligen Fehler zu minimieren, wird die Person vielleicht das arithmetische Mittel der Messungen betrachten. Dieses beträgt

$$\frac{75,5 \text{ kg} + 74,8 \text{ kg} + 75,2 \text{ kg} + 75,7 \text{ kg}}{4} = 75,3 \text{ kg}$$

Die Person könnte sagen: „Mein Gewicht beträgt 75,3 kg.“ Die Frage stellt sich: Wie *genau* und wie *plausibel* ist diese Angabe? Dank statistischen Werkzeugen kann man auf diese Frage mit Aussagen, wie „Die Wahrscheinlichkeit, dass mein Gewicht zwischen 75,2 und 75,4 kg liegt, beträgt 95 %“, antworten. Oder was vielleicht eine äquivalente

Aussage ist: „Ich wette 95 zu 5 Franken, dass mein Gewicht zwischen 75,2 und 75,4 kg beträgt.“ Diese Arbeit bezeichnet man als *schliessende Statistik* (engl. *statistical inference*).

## Zukünftige Werte einer unsicheren Grösse prognostizieren

Eine Person, die am nächsten Morgen um acht Uhr den Schnellzug von Bern nach Zürich nimmt, möchte wissen, wie lang die Fahrzeit des Zugs sein wird. Die Fahrzeit eines Schnellzugs von Bern nach Zürich ist eine komplexe Grösse, die von vielen Faktoren abhängt. Die Person wird daher wegen fehlender Information zu den Faktoren den morgigen Wert dieser Grösse nicht berechnen können. Sie benutzt daher Informationen aus dem Fahrplan und Daten, wie beispielsweise drei gemessene Fahrzeiten der letzten Woche von 56'57", 59'53" und 57'38". Mit denen wird sie versuchen, die Fahrzeit zu prognostizieren. Mit statistischen Werkzeugen und Wahrscheinlichkeiten kann ausgedrückt werden, wie plausibel dies ist: „Die Wahrscheinlichkeit, dass die Fahrzeit des Zugs, der morgen um acht Uhr nach von Bern nach Zürich fährt, mehr als 65 Minuten sein wird, beträgt 5 %.“ Man sagt, dass man eine *Prognose* (engl. *prediction*) gerechnet hat.

## Regressions- und Klassifikationsmodelle

Viele Untersuchungen versuchen aus einer ersten Grösse einen *unsicheren Wert* einer zweiten Grösse zu prognostizieren. So möchte jemand den Preis eines Gebrauchtwagens aus dem Kilometerstand des Wagens berechnen. Oder ein Arzt will das Lungenvolumen aus dem Alter eines Patienten bestimmen. Dazu braucht man Daten aus Versuchsgruppen. Auch hier können Prognosen gerechnet werden: „Der Preis des Gebrauchtwagens liegt mit einer Wahrscheinlichkeit von 95 % zwischen 5000 CHF und 7000 CHF, wenn der Kilometerstand 100 000 Kilometer beträgt.“ Man spricht in solchen Fällen auch vom *statistischem Lernen* (engl. *statistical learning*). Dazu benutzt man *Regressions- und Klassifikationsmodelle*. Bei grossen, komplexen Datenmengen („Big Data“) können verschiedene Modelle Zusammenhänge zwischen Grössen beschreiben. Statistische Werkzeuge helfen optimale Modelle zu finden. Man spricht von *Modellselektion*.

## Leserschaft und Bayes'sche Statistik

Das Buch richtet sich an Studierende, die in angewandten Wissenschaften, wie Ingenieur-, Natur- und Wirtschaftswissenschaften einen Bachelor- oder Mastergrad abschliessen wollen. Es wird die Bayes'sche Statistik vorgestellt, um Problemstellungen, wie sie oben erwähnt wurden, wissenschaftlich zu diskutieren. Sie arbeitet im wesentlichen mit einem *einzigem* Werkzeug, der Regel von Bayes. Die Regel erklärt, wie man Informationen aus Daten und Zusatzinformationen verarbeiten kann. Die Zusatzinformation kann frü-

heres Wissen sein. Man kann also auch auf Datensammlungen von anderen Forschern zurückgreifen. Um die Regel von Bayes anzuwenden, muss man weiter formulieren, wie Messwerte oder Daten streuen. Man sagt, dass man ein Wahrscheinlichkeitsmodell wählen muss. Dies legt offen, auf welchen Annahmen die statistischen Rechnungen basieren. Ist dies getan, können Wahrscheinlichkeiten zu nicht direkt messbaren Grössen, zu zukünftigen Werten von unsicheren Grössen oder zu Hypothesen meist schnell gerechnet werden. Dies ist – auch bei komplizierteren Wahrscheinlichkeitsmodellen – dank den heutigen Computern mit Simulationsprogrammen möglich. Die benutzten Modelle und Resultate können meist auch gut erklärt werden.

---

## Zusatzmaterial: Programmcode und Lösungen

Es ist nicht möglich Beobachtungen oder Messungen ohne Statistikprogramme auszuwerten. Programmcodes für die Statistiksoftware R und für die im Buch benutzten Monte-Carlo-Simulationen mit `rstan` sind als Zusatzmaterial online unter

[www.irex.bfh.ch](http://www.irex.bfh.ch)

im Verzeichnis *Bücher* oder unter [www.baettig.one](http://www.baettig.one) verfügbar. R ist eine unter [www.r-project.org](http://www.r-project.org) frei erhältliche Statistiksoftware. `rstan` ist ein Paket, das auf die Software STAN zugreift, um Wahrscheinlichkeitsmodelle „abzutasten“. Informationen dazu findet man unter [www.mc-stan.org](http://www.mc-stan.org).

Ausgewählte Resultate und Lösungen zu den Reflexionsaufgaben und die im Buch benutzten Datensätze sind ebenfalls als Zusatzmaterial online bei den oben angegebenen Adressen erhältlich.

## Dank

Der Autor dankt den Studierenden in Chemie, Maschinentechnik und Elektrotechnik der Berner Fachhochschule in Burgdorf, die mit ihrer kritischen Lektüre geholfen haben, diesen Text zu gestalten. Ein Dank geht auch an das Departement Civil Engineering der Oregon State University (OSU), wo der Autor im Frühling 2006 die ersten Bausteine des Texts legen konnte. Sehr hilfreich waren die Diskussionen mit den Statistikerinnen und Statistikern des IDA an der Universität Linköping, wo der Autor während des Frühlingssemesters 2013 weilte. Darum spricht der Autor ein grosses Dankeschön an Mattias Villani, Patrick Waldmann und Oleg Sysoev von der Universität Linköping aus. Schliesslich geht auch ein Dank an Franziska Bitter Bättig. Ihre kritische Durchsicht des Textes und ihre sprachlichen und didaktischen Anregungen haben den Entstehungsprozess des vorliegenden Buchs begleitet.

Burgdorf, Januar 2017

Daniel Bättig

---

# Inhaltsverzeichnis

<b>1</b>	<b>Eine Einführung und ein Überblick</b>	1
1.1	Bayes: Lernen aus Information	1
1.2	Experimente und Beobachtungen	8
1.3	Statistik und Qualitätskontrolle	12
Reflexion	.....	17
Literatur	.....	23
<b>2</b>	<b>Wie man Versuche planen kann</b>	25
2.1	Worauf bezieht sich eine Grösse?	26
2.2	Grössen aus Physik und Technik	28
2.3	Grössen und ihre Typ B Unsicherheit	29
2.4	Grössen müssen messbar sein	29
2.5	Faktoren und Niveaus bestimmen	31
2.6	Paretdiagramme	36
2.7	Systematische Fehler	40
Reflexion	.....	43
Literatur	.....	46
<b>3</b>	<b>Messen und Kontrollieren</b>	47
3.1	Randomisierung: die Stichprobenwahl	48
3.2	Wiederholung: Wie viele Messungen?	51
3.3	Kontrolle: Experiment oder Beobachtung?	53
3.4	Statistische Kontrolle und Vertauschbarkeit	58
3.5	Qualitätsmanagement: Personen rangieren?	65
3.6	Memorandum zur Datensammlung	67
3.7	Weiterführende Literatur zu Kap. 2 und diesem Kapitel	68
Reflexion	.....	69
Literatur	.....	73

<b>4 Das Fundament: Wahrscheinlichkeiten</b>	75
4.1 Die drei Rechengesetze	75
4.2 Eine andere Interpretation der Wahrscheinlichkeit	86
4.3 Wahrscheinlichkeitsmodelle	88
4.4 Wie kann man Informationen aus Wahrscheinlichkeitsmodellen zusammenfassen?	96
4.5 Monte-Carlo-Simulationen	100
4.6 Weiterführende Literatur zu diesem Kapitel	108
Reflexion	109
Literatur	116
<b>5 Nicht direkt messbare Größen bestimmen</b>	117
5.1 Die Regel von Bayes	117
5.2 Berechnen eines Anteils	121
5.3 Wie hängt das Resultat von der Vorinformation ab?	127
5.4 Versuchsplanung und Unabhängigkeit	131
Reflexion	134
Literatur	136
<b>6 Mehrere Größen und Korrelation</b>	137
6.1 Das gemeinsame Modell	138
6.2 Randverteilungen mit Monte-Carlo-Simulationen rechnen	140
6.3 Verbundene Größen und Korrelation	141
6.4 Autokorrelation und Unabhängigkeit	145
Reflexion	150
Literatur	155
<b>7 Messwerte prognostizieren</b>	157
7.1 Objekte, die in zwei Kategorien auftreten	157
7.2 Ein Verfahren, um Messwerte zu prognostizieren	162
Reflexion	165
Literatur	166
<b>8 Modellwahl: Information und Entropie</b>	167
8.1 Das Problem: Modell und Vorwissen	168
8.2 Transformation und minimale Vorinformation	169
8.3 Unordnung und relative Entropie	175
8.4 Der Erwartungswert als Information	181
Reflexion	183
Literatur	186

<b>9</b>	<b>Zwei Modelle zu positiven Größen</b>	187
9.1	Die Exponentialverteilung	187
9.2	Die Poissonverteilung	196
9.3	Zusammenfassung	204
Reflexion		204
Literatur		208
<b>10</b>	<b>Streuung und Normalverteilung</b>	209
10.1	Die Streuung als Information	209
10.2	Die Normalverteilung	212
10.3	Normalverteilung als Datenmodell	217
Reflexion		230
Literatur		234
<b>11</b>	<b>Explorative Datenanalyse</b>	235
11.1	Erste Beispiele zu grafischen Darstellungen	235
11.2	Darstellen, wie Daten verteilt sind	242
11.3	Ausreisser und Extremwerte	254
11.4	Robustere Datenmodelle	261
11.5	Weiterführende Literatur	265
Reflexion		265
Literatur		268
<b>12</b>	<b>Regressionsmodelle</b>	271
12.1	Streudiagramme	272
12.2	Beispiele von Regressionsmodellen	280
Reflexion		287
Literatur		290
<b>13</b>	<b>Regressionsmodelle: Parameter und Prognosen</b>	291
13.1	Beispiele mit der Normalverteilung	291
13.2	Die Methode der kleinsten Quadrate	301
13.3	Kritische Überlegungen	306
13.4	Prior, Likelihood, Posterior	315
13.5	Verallgemeinerte lineare Modelle	317
Reflexion		320
Literatur		326

<b>14</b>	<b>Standardfehler, Ranglisten und Modelle</b>	329
14.1	Die Methode von Laplace	330
14.2	Die $\delta$ -Methode	334
14.3	Eine gefahrenrächtige Gleichung	339
14.4	Struktur und hierarchische Modelle	345
Reflexion		351
Literatur		355
<b>15</b>	<b>Plausibilität von Modellen und von Hypothesen</b>	357
15.1	Plausibilität von Modellen	357
15.2	Plausibilität von Hypothesen	372
Reflexion		377
Literatur		380
<b>A</b>	<b>Formeln bei minimaler Vorinformation</b>	381
A.1	Datenmodell Exponentialverteilung	381
A.2	Datenmodell Poissonverteilung	382
A.3	Datenmodell Normalverteilung	383
	<b>Sachverzeichnis</b>	389

*„Auf die Plätze!“ schrie die Königin mit Donnerstimme, und sogleich rannte alles blind drauflos und stolperte übereinander; nach einer Weile aber hatten sich alle ordentlich aufgestellt, und das Spiel begann.*  
Lewis Carroll, Alice im Wunderland (Insel Taschenbuch, 1973, S. 85)

## Zusammenfassung

Das Kapitel zeigt Werkzeuge, die in diesem Buch benutzt werden, um Daten zu analysieren. So erhält die Leserin oder der Leser einen ersten Eindruck, wie solche Werkzeuge funktionieren, ohne zu erwarten, dass die Details zu diesem Zeitpunkt verstanden werden. Insbesondere messen die Werkzeuge, wie *plausibel* Aussagen sind. Vorgestellt werden dabei die zwei wichtigsten Rechenregeln, um Plausibilitäten zu bestimmen. Einerseits ist dies die Regel von Bayes, die es erlaubt, Aussagen zu nicht direkt messbaren Größen zu quantifizieren. Andererseits ist dies das Gesetz der Marginalisierung, mit dem man versuchen kann, zukünftige Beobachtungen einer unsicheren Größe zu prognostizieren. Auch erfährt der Leser oder die Leserin, wie man die Statistik im Bereich der Qualitätssicherung einsetzen kann.

## 1.1 Bayes: Lernen aus Information

Sicherlich haben Sie schon Aussagen formuliert, wie „Morgen wird es in der Stadt Bern sicher stark regnen.“, „Die Aktien der Firma XY werden nächstes Jahr um mindestens CHF 100.– steigen.“ oder „Der Schnellzug nach Zürich von heute Abend 20 Uhr wird wahrscheinlich pünktlich am Zielbahnhof ankommen.“ Diese Aussagen sind, da genaue Informationen zu den Ereignissen fehlen, „unsicher“. In zwei der drei Aussagen vermitteln die Adjektive „sicher“ und „wahrscheinlich“ dies. Mit der Wahrscheinlichkeitsrechnung quantifiziert man, wie plausibel solche Aussagen sind. Ein Beispiel dazu ist: „Ich glaube mit einer Wahrscheinlichkeit von 95 %, dass es morgen in der Stadt Bern stark regnen

wird.“ Das folgende Beispiel zeigt, wie man Wahrscheinlichkeiten wissenschaftlich berechnen kann, wenn man mit unsicheren Aussagen zu tun hat:

**Beispiel 1.1 (OptiMAL)** Malaria ist eine lebensbedrohende Krankheit. Sie zeichnet sich unter anderem bei betroffenen Personen durch hohe, periodisch auftretende Fieberschübe aus. Krankheiten lassen sich meist nur indirekt und nicht mit Sicherheit erkennen. Ärztinnen messen dazu Größen, wie Körpertemperaturen, Blutbilder, benutzen Kardiogramme oder medizinische Tests. OptiMAL ist ein solcher Test, um Malaria schnell zu „detektieren“. Er kann nicht zu 100 % garantieren, dass er positiv ausschlägt, wenn eine Person Malaria hat. Klinische Versuche in [7] haben gezeigt, dass bei 96 Personen, die Malaria hatten, der OptiMAL-Test 89 mal positiv ausschlug:  $89/96 = 0,927 = 92,7\%$ . Diese Zahl nennt man die *Sensitivität* oder *Richtigpositiv-Rate* des Tests. Sie ist also die Wahrscheinlichkeit, dass der Test positiv ausfällt, *gegeben* die Person hat Malaria. Dies kann man so schreiben:

$$\mathbb{P}(\text{Test positiv} \mid \text{Patient hat Malaria}) = 0,927 \quad (1.1)$$

Der vertikale Strich in dieser Schreibweise heisst „gegeben“. „Test positiv | Patient hat Malaria“ liest man also: Test ist positiv, gegeben die Person hat Malaria. Analog sollte ein medizinischer Test möglichst negativ reagieren, wenn die untersuchte Person nicht von der Krankheit betroffen ist. Wie gut ein Test dies tut, sagt die *Spezifität* (auch *Richtignegativ-Rate*). Diese ist die Wahrscheinlichkeit, dass ein Test negativ ausfällt, *gegeben* die Person ist nicht krank. Bei der erwähnten Studie wurde der OptiMAL-Test bei 106 Personen ohne Malaria angewendet. Bei 104 Personen war das Testresultat auch negativ. Somit ist Spezifität  $= 104/106 = 0,981$ :

$$\mathbb{P}(\text{Test negativ} \mid \text{Patient hat keine Malaria}) = 0,981$$

Eine Ärztin kann nun den OptiMAL-Test an einer Person anwenden. Der Test schlage positiv aus. Wie gross ist die Wahrscheinlichkeit, dass der Patient Malaria hat? Gesucht ist also

$$\mathbb{P}(\text{Patient hat Malaria} \mid \text{Test positiv}) = ?$$

Beachten Sie, dass die Rollen von *Patient hat Malaria* und *Test positiv* im Vergleich zu Gleichung (1.1) vertauscht sind! Man sagt daher auch: Die von der Ärztin gesuchte Wahrscheinlichkeit ist eine *inverse* Wahrscheinlichkeit. Beide Wahrscheinlichkeiten können sehr unterschiedlich sein.

So ist  $\mathbb{P}(\text{Frau} \mid \text{schwanger}) = 100\%$ , da schwangere Personen Frauen sind. Die dazu inverse Wahrscheinlichkeit  $\mathbb{P}(\text{schwanger} \mid \text{Frau})$  ist etwa 3 %: Nur eine Minderheit von Frauen ist schwanger.

Es scheint vernünftig, dass die zweite Wahrscheinlichkeit nicht direkt aus der ersten berechnet werden kann. Die Ärztin kann also ihre gesuchte Wahrscheinlichkeit nicht direkt aus der Sensitivität oder Spezifität bestimmen. Sie braucht dazu zusätzliche In-

formation. Sie muss dazu *vor* dem Test beurteilen, wie wahrscheinlich ihr Patient von Malaria betroffen ist: Wie verlaufen die Fieberschübe der Person? Hat die Person eine Malaria prophylaxe durchgeführt? Mit dieser Vorinformation – nennen wir sie  $\mathcal{K}$  – kann sie eine Einschätzung aussprechen: „Ich glaube mit einer Wahrscheinlichkeit von 30 %, dass mein Patient Malaria hat.“ Der Sachverhalt drückt ihre Vorinformation aus:<sup>1</sup>  $\mathbb{P}(\text{Malaria} | \mathcal{K}) = 0,3$ . Wahrscheinlichkeiten werden im medizinischen Bereich gerne wie bei Wettbüros durch *Chancen* (engl. *odds*) ausgedrückt. Eine Prozentzahl wie 30 % heißt „dreissig auf hundert“, oder teilt 100 Objekte im Verhältnis von 30 : 70 auf. Dies ist die Chance, dass der Patient Malaria hat:

$$\mathbb{O}(\text{Malaria} | \mathcal{K}) = 30 : 70$$

Das Resultat des Malaria-Tests sei positiv. Diese Information bewirkt, dass die Chance für den Patienten, Malaria zu haben, erhöht wird. In der Tat besagt die Regel von Bayes, dass<sup>2</sup>

$$\mathbb{O}(\text{Malaria} | \text{Test positiv}) = K \cdot \mathbb{O}(\text{Malaria} | \mathcal{K})$$

Dabei ist der Faktor  $K$  bei positivem Malaria-Test gleich der Sensitivität dividiert durch 1 – Spezifität. Der Ärztin erhält

$$\mathbb{O}(\text{Malaria} | \text{Test positiv}) = \frac{\text{Sensitivität}}{1 - \text{Spezifität}} \cdot \frac{30}{70} = \frac{0,927}{1 - 0,981} \cdot \frac{30}{70} = 20,91$$

Die Chance, dass eine Malariaerkrankung besteht, ist beim Patienten also 20,91 oder mit einem Bruch ausgedrückt 2091:100. Daraus lässt die Wahrscheinlichkeit, dass der Patient Malaria hat, ausrechnen

$$\mathbb{P}(\text{Malaria} | \text{Test positiv}) = \frac{2091}{2091 + 100} = 0,95 = 95\%$$

Die Ärztin wird daher davon ausgehen, dass der Patient von Malaria betroffen ist. Hätte der Malaria-Test ein negatives Resultat geliefert, so erhält man mit der Regel von Bayes

$$\mathbb{O}(\text{Malaria} | \text{Test negativ}) = \frac{1 - \text{Sensitivität}}{\text{Spezifität}} \cdot \underbrace{\mathbb{O}(\text{Malaria} | \mathcal{K})}_{\text{Vorwissen}}$$

Der Ärztin hätte in diesem Fall eine Wahrscheinlichkeit von nur noch 3,1 % erhalten, dass der Patient Malaria hat.  $\square$

---

<sup>1</sup> Wichtig ist: die Wahrscheinlichkeit von 30 % ist Ausdruck bei gegebenem Wissen der Ärztin. Eine zweite Ärztin wird mit ihrer Vorinformation oder ihrer Erfahrung eine andere Wahrscheinlichkeit setzen. Wahrscheinlichkeiten hängen also von der gegebenen Information ab.

<sup>2</sup> Die Regel von Bayes wird später im Buch erklärt.

Das obige Beispiel zeigt, wie die Wahrscheinlichkeitsrechnung benutzt werden kann, um Aussagen, die unsicher sind, zu aktualisieren. Insbesondere ist interessant, dass dazu neue Information aus Daten – das Resultat des medizinischen Tests – zusammen mit Vorinformation verwendet wird:

$$\mathbb{O}(\text{krank} \mid \text{neue Information}) = \underbrace{\text{Faktor}}_{\text{Aus Daten}} \cdot \mathbb{O}(\text{krank} \mid \text{alte Information})$$

Den Faktor nennt man den *Likelihood-Quotienten*. Die Gleichung ist ein Spezialfall der Regel von Bayes. Der Faktor besteht aus zur gesuchten Wahrscheinlichkeit inversen, einfacheren Wahrscheinlichkeiten: der Sensitivität und der Spezifität. Bei positivem Test ist er Sensitivität/(1 – Spezifität) und bei negativem Test lautet er (1 – Sensitivität)/Spezifität.

Das nächste Beispiel illustriert, dass auch im Ingenieurwesen mit dem gleichen Verfahren gearbeitet werden kann.

**Beispiel 1.2 (Messgeräte)** Mit technischen Geräten lassen sich Frequenzen von Wechselströmen bestimmen. Durch Schwankungen in der Produktion, der in den Geräten eingebauten Bestandteile, werden die Geräte bei einem Wechselstromkreis mit Frequenz von 4 Hz nicht exakt eine Frequenz von 4 Hz anzeigen. Eine Frequenz eines Stromkreises ist daher nicht direkt messbar. In der Gebrauchsanleitung des Geräts finden sich dazu Angaben, wie sie Tab. 1.1 zeigt. Die Tabelle sagt, dass gemessene Frequenzen im Bereich von 4 Hz (bei einer Spannung zwischen 100 mV und 300 mV) in einer Ordnung von  $\pm 0,10\%$ , also um  $\pm 0,004$  Hz schwanken. Aus ihr lässt sich deshalb bestimmen, wo angezeigte Frequenzen liegen werden, wenn die Frequenz  $f$  4 Hz beträgt. Mathematisch wird dies mit einer Wahrscheinlichkeit ausgedrückt:

$$\mathbb{P}(\text{angezeigte Frequenzen im Bereich } XY \mid f = 4 \text{ Hz}) \quad (1.2)$$

Man nennt dies ein Modell für Messwerte. Es bezeichnet mit einer Wahrscheinlichkeit, wie gemessene Frequenzen um eine Frequenz *streu*en. Man nennt es auch das *Datenmodell* oder das *Streumodell*. Personen, die dieses Gerät benutzen, möchten die (nicht direkt

**Tab. 1.1** Genauigkeitsangaben aus einer Gebrauchsanleitung

Spannung	Frequenz	+-% des Messwerts
100 mV	3–5 Hz	0,10
bis	5–10 Hz	0,05
300 mV	10–40 Hz	0,03

messbare) Frequenz eines Stromkreises aus gemessenen Frequenzen bestimmen. Dies ist die inverse Wahrscheinlichkeit zur obigen Wahrscheinlichkeit:

$$\mathbb{P}(f = 4 \text{ Hz} \mid \text{angezeigte Frequenzen im Bereich } XY)$$

Wie beim Beispiel zu Malaria, lässt sich diese Wahrscheinlichkeit mit der Regel von Bayes berechnen. Man braucht dazu Vorinformation und die dazu inverse Wahrscheinlichkeit aus der Gebrauchsanleitung:

$$\underbrace{\mathbb{P}(f = 4 \text{ Hz} \mid \text{angezeigte Frequenzen})}_{\text{Aussage zur Frequenz, gegeben Daten}} = \underbrace{\text{Faktor}}_{\text{Aus Daten}} \cdot \underbrace{\mathbb{P}(f = 4 \text{ Hz})}_{\text{Aus Vorinformation}}$$

Der Faktor ganz rechts besagt, wie plausibel die Frequenz bei 4 Hz liegt, wenn keine Messdaten vorhanden sind. Aus der Konstruktion des Stromkreises wird eine Ingenieurin vielleicht dank ihrem technischen Wissen sagen, dass  $\mathbb{P}(f = 4 \text{ Hz}) = 0,8$  ist. Der erste Faktor auf der rechten Seite nennt man den *Likelihood*-Faktor. Berechnet wird er aus dem Streumodell von Gleichung (1.2). Ist der Likelihood-Faktor klein, so sind die Daten kaum mit der Frequenz von 4 Hz verträglich. Dies wird der Fall sein, wenn man 4,105 oder 3,800 Hz gemessen hat. Wenn man 4,001 oder 3,997 Hz misst, wird der Faktor gross. Der Wert der linken Seite der Gleichung wird deshalb aktualisiert und grösser. □

Die Regel von Bayes verbindet also Vorinformation mit den Daten, um neu zu werten, wie plausibel Aussagen sind. Hier ein weiteres Beispiel dazu aus dem Bereich der Geowissenschaft:

**Beispiel 1.3 (Zeit zwischen starken Erdbeben)** Die Wissenschaft interessiert sich für die Zeitabstände zwischen starken Erdbeben. Tab. 1.2 zeigt Zeitpunkte, Orte und Stärke aller 29 Erdbeben ohne Nachbeben mit einer Stärke von mindestens 8 zwischen dem 1. Januar 1969 und dem 31. Dezember 2007. Die Zeitabstände zwischen aufeinanderfolgenden Erdbeben nennt man *Wartezeiten*. Sie sind in Tab. 1.2 in der Spalte rechtsaußen angezeigt. Zwischen dem ersten und zweiten Erdbeben sind 261,64 Tage vergangen. Die längste Zeit ist 2300,14 Tage und die kürzeste beträgt 3,89 Tage. Zwei Arten von Fragen interessieren hier meistens:

- (a) Kann man *nicht direkt messbare Größen*, wie die durchschnittliche Zeit zwischen zwei zukünftigen, starken Erdbeben, *berechnen*?
- (b) Ist es möglich, die (einzelne) Zeit bis zum nächsten starken Erdbeben zu *prognostizieren*?

Bei der Frage (a) scheint es „vernünftig“, die durchschnittliche Zeit zwischen zukünftigen, starken Erdbeben mit dem arithmetischen Mittel der beobachteten Zeiten zu bestimmen.

**Tab. 1.2** Zeitpunkte, Orte und Stärke aller 29 Erdbeben ohne Nachbeben mit einer Stärke von mindestens 8 zwischen dem 1. Januar 1969 und dem 31. Dezember 2007 (Iris-Consortium, Purdue University)

Stärke	Datum	Breitengrad	Längengrad	Wartezeit (in Tagen)
8,1	1969/08/11 21:27:33	43,20	147,60	
8,5	1970/04/30 12:51:29	13,95	-93,27	261,64
8,0	1976/08/16 16:11:10	7,30	123,60	2300,14
8,2	1979/12/12 07:59:07	1,00	-78,00	1212,66
8,1	1985/09/19 13:17:44	18,02	-102,75	2108,22
8,5	1986/11/14 21:19:48	21,60	123,00	421,33
8,4	1990/07/16 07:26:32	15,20	120,90	1339,42
8,1	1991/04/22 21:56:51	9,96	-83,05	280,60
8,1	1994/10/04 13:22:55	43,77	147,32	1260,64
8,8	1996/02/17 05:59:35	-1,01	136,99	500,69
9,3	1996/08/18 10:47:01	-7,72	129,98	183,20
8,8	1997/01/13 16:09:35	33,39	-115,92	148,22
8,4	1999/04/12 22:10:53	16,55	-94,80	819,25
8,0	1999/08/18 01:02:37	40,77	30,59	127,12
8,7	1999/09/07 11:57:29	37,40	24,04	20,45
8,2	1999/09/30 16:31:31	15,96	-95,92	23,19
8,9	1999/10/04 13:59:06	-3,24	-78,19	3,89
9,9	1999/10/14 10:59:35	-16,56	-62,76	9,88
8,7	1999/11/26 05:34:08	18,44	-95,49	42,77
8,3	1999/12/23 00:03:09	-0,95	-76,72	26,77
8,5	2000/05/18 14:28:26	15,54	-94,80	147,60
8,0	2001/01/26 03:16:40	23,42	70,23	252,54
8,3	2001/06/23 20:33:14	-16,27	-73,64	148,72
8,2	2003/05/30 04:46:34	52,43	142,79	705,34
8,7	2003/08/25 06:28:26	13,39	-91,31	87,07
8,3	2003/09/25 19:50:06	42,21	143,84	31,56
9,0	2004/12/26 00:58:50	3,09	94,26	457,21
8,2	2005/03/28 16:09:31	2,03	97,04	92,64
8,5	2007/09/12 11:10:26	-4,44	101,37	897,79

Diese beträgt

$$\text{durch. Zeit} = \frac{261,64 + 2300,14 + 1212,66 + \dots + 897,79}{28} \text{ Tage} = 496,81 \text{ Tage}$$

Teilt man diese Rechnung anderen Personen mit, wird oft die Frage gestellt, wie *genau* die Rechnung ist. Man kann antworten: „Die Genauigkeit beträgt  $\pm 2\%$ .“, was zu einer Frage führen kann, wie sicher diese Genauigkeit ist (siehe dazu [5]). Die durchschnittliche Zeit  $\mu$  zwischen zukünftigen, starken Erdbeben mit einer Zahl zu beschreiben, ist also

unbefriedigend. Man sollte zumindest angeben, wie genau und wie plausibel das Resultat ist. Eine solche Angabe wäre: „Die durchschnittliche zukünftige Wartezeit liegt, mit einer Wahrscheinlichkeit von 90 %, zwischen 450 und 550 Tagen.“ Gesucht ist deshalb eine Wahrscheinlichkeit der Form:

$$\mathbb{P}(\mu \text{ liegt im Bereich } XY \mid \text{beobachtete Zeiten})$$

Wie im vorigen Beispiel ist die zur gesuchten Wahrscheinlichkeit inverse Wahrscheinlichkeit einfacher berechenbar:

$$\mathbb{P}(\text{Zeiten liegen im Bereich } AB \mid \mu) \quad (1.3)$$

Diese Wahrscheinlichkeit beschreibt, wie Zeiten zwischen starken Erdbeben um die durchschnittliche Wartezeit  $\mu$  streuen. Man sagt, dass sie das *Datenmodell* oder das *Streumodell* für die Wartezeiten darstellt. Dieses kann mit Gesetzen zur Informationsverarbeitung oder zur Physik fixiert werden. Wie beim Beispiel zum Messgerät, lässt sich dann zu vorgegebenem  $\mu$  sagen, wie plausibel die beobachteten Zeiten sind. Ist etwa  $\mu = 1$  Tag, so ist  $\mathbb{P}(\text{Zeit} = 2300,14 \text{ Tage} \mid \mu)$  eher klein. Man berechnet, wie beim Beispiel zu Malaria, die zur Gleichung (1.3) inverse Wahrscheinlichkeit mit der Regel von Bayes. Diese verbindet Vorwissen zu  $\mu$  und die Information aus den Daten:

$$\underbrace{\mathbb{P}(\mu \mid \text{Daten})}_{\substack{\text{Aussage zu } \mu \text{ mit Daten}}} = \underbrace{\text{Likelihood}}_{\substack{\text{Daten}}} \cdot \underbrace{\mathbb{P}(\mu \mid \mathcal{K})}_{\substack{\text{Aussage zu } \mu \text{ ohne Daten}}}$$

Wir wollen hier von einer unerfahrenen Person ausgehen, die wenig Vorwissen  $\mathcal{K}$  hat. Sie nimmt an, dass  $\mu$  irgendwo mit gleicher Wahrscheinlichkeit liegen kann. Den Likelihood-Faktor berechnet man aus den Messwerten mit dem Streumodell aus Gleichung (1.3). Wie dies durchgeführt wird, wird später beschrieben. Wertet man die obige Gleichung komplett aus, erhält man: Gegeben die Daten, ist die durchschnittliche Wartezeit  $\mu$  mit einer Wahrscheinlichkeit von 50 % zwischen 444 und 573 Tagen. Es besteht eine Wahrscheinlichkeit von 95 %, dass sie zwischen 354 und 747 Tagen liegt.

Man kann auch versuchen, die Frage (b) nach einer Prognose zum nächsten Zeitpunkt eines Erdbebens zu beantworten. Die Antwort gibt die Gleichung (1.3). Leider kennt man  $\mu$  nicht präzis. Man weiss ja aus (a) nur, wo  $\mu$  mit hoher Wahrscheinlichkeit liegt. Daher lässt sich  $\mu$  nicht einfach in die Gleichung (1.3) einsetzen. Man braucht dazu das Gesetz der Marginalisierung. Es besagt, dass man alle möglichen Werte  $\mu$  in diese Gleichung einsetzt und diese nach ihrer Wahrscheinlichkeit gewichtet mittelt. In einer Formel ausgedrückt ist also

$$\mathbb{P}(\text{Wartezeit} \mid \text{Daten}) = \sum_{\mu} \underbrace{\mathbb{P}(\text{Wartezeit} \mid \mu)}_{\substack{\text{Streumodell}}} \cdot \underbrace{\mathbb{P}(\mu \mid \text{Daten})}_{\substack{\text{aus Daten}}}$$

Die Summe kann man bestimmen.<sup>3</sup> Man erhält beispielsweise: Die Wahrscheinlichkeit ist 85,7 %, dass die Zeit bis zum nächsten starken Erdbeben höchstens 1000 Tage sein wird.

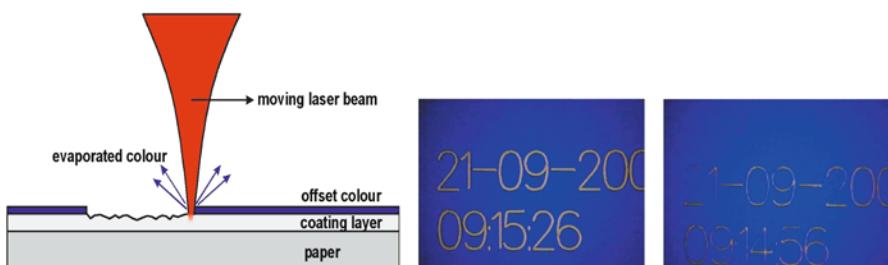
In einem letzten Schritt der Rechnung wird man noch zu klären versuchen, ob das berechnete Resultat gut ist. Dazu wird man messen, wie gut das Streumodell (1.3) zu den Daten passt. Die Rechnungen zeigen schliesslich, wie drei Faktoren geschickt miteinander verbunden werden, um aus Daten zu lernen: Das Datenmodell zu Wartezeiten, die Daten, sowie die vorhandene Vorinformation  $\mathcal{K}$  der Person.  $\square$

## 1.2 Experimente und Beobachtungen

Daten sind wertvoll, um Aussagen zu nicht direkt messbaren Grössen oder zu zukünftigen Werten von unsicheren Grössen zu machen. Um verlässliche und nachvollziehbare Resultate zu erhalten, scheint es sinnvoll, die Daten in einem kontrollierbaren Rahmen zu sammeln.

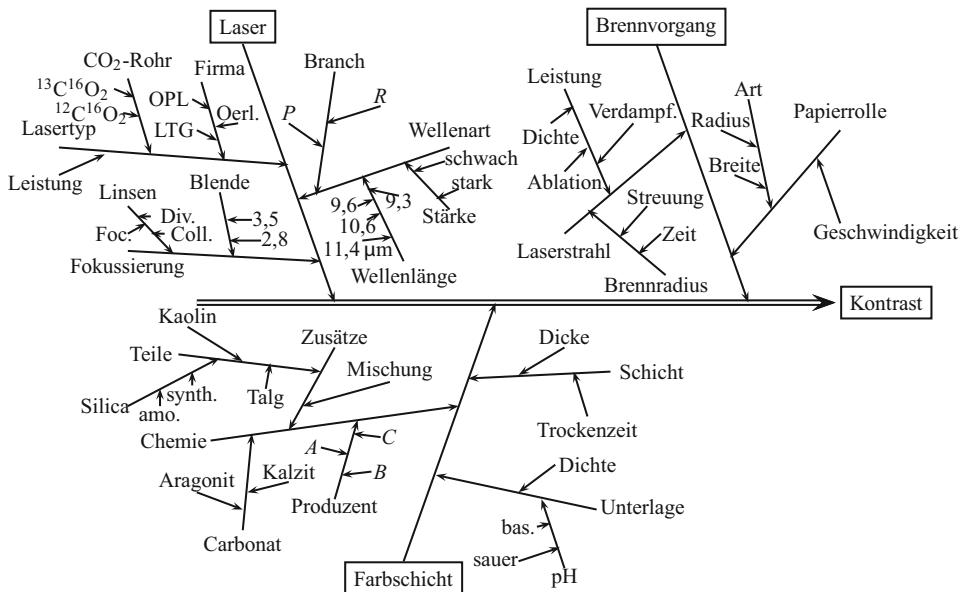
**Beispiel 1.4 (Lasermarkierung)** Um Packungen zu identifizieren, können sie mit Lasern markiert werden (siehe Abb. 1.1). Dabei werden aufgetragene Farbschichten durch einen Laserstrahl entfernt. Die beiden weiteren Bilder in Abb. 1.1 zeigen Resultate des Verfahrens aus [3]. Markiert wurden zwei mit Calcium-Karbonatfarbe beschichtete Papiere mit einem Laser mit einer Wellenlänge von 10,6  $\mu\text{m}$ . Ein Papier enthielt 80 %-Karbonat (links) und eines 100 %-Karbonat (rechts).

Das Ziel eines Teams von Physikern und Ingenieuren war es, den Kontrast der Markierung in Funktion der vorgegebenen Laserleistung zu berechnen. Der Kontrast hängt von vielen Faktoren, wie Laserart, Art des Brennvorgangs und Art der verwendeten Farbschicht ab. Ein *Ursache-Wirkungs-Diagramm* erlaubt es zu visualisieren, wie die erwähnten Faktoren auf den Kontrast wirken. Dies zeigt Abb. 1.2. Mit dem Ursache-Wirkungs-Diagramm lässt sich der Kontrast in einer kontrollierten Umgebung messen.  $\square$



**Abb. 1.1** Lasermarkierung und zwei mit Calcium-Karbonatfarbe beschichtete Papiere, markiert mit Laserstrahl

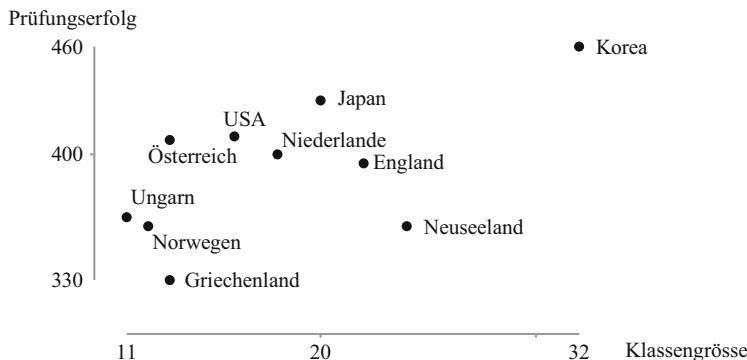
<sup>3</sup> Die Rechnung wird später erklärt.



**Abb. 1.2** Ein Ursache-Wirkungs-Diagramm

Ein Ursache-Wirkungs-Diagramm erlaubt dem Experimentierteam, kontrolliert Daten zu sammeln, indem es bestimmt, welche Faktoren berücksichtigt und variiert werden und welche Faktoren festgehalten werden. *Experimente* sind Untersuchungen, bei denen Werte mit kontrollierten Faktoren gemessen werden. Es ist also bei der Lasermarkierung möglich, den Kontrast in Funktion der Wellenlänge zu messen, wenn die anderen Faktoren im Ursache-Wirkungs-Diagramm konstant gehalten werden. Man bezeichnet den Kontrast als die *abhängige Variable* (engl. *dependent variable*) und die Wellenlänge als die *unabhängige Variable* (engl. *independent variable*). Mit Experimenten kann man also kausale Zusammenhänge zwischen zwei Variablen herstellen.

**Beispiel 1.5 (Prüfungserfolg)** In vielen Ländern interessiert man sich für die „optimale“ Klassengröße. Kleinere Klassen werden deshalb geschätzt, weil sich die Lehrperson den einzelnen Schülerinnen und Schülern mehr widmen kann. Damit steigen jedoch der Bedarf an Lehrpersonen und die Bildungskosten. Die Organisation für wirtschaftliche Zusammenarbeit (OECD) hat die Frage zu beantworten versucht, ob der Prüfungserfolg von der Klassengröße abhängt. Um dies zu tun, dienen Daten wie sie Abb. 1.3 in einem Streudiagramm zeigt. Beobachtet wurden die Klassengrößen und der Lernerfolg bei zehn Ländern im Jahr 2005. Die Abbildung illustriert, dass mit zunehmender Klassengröße der Prüfungserfolg tendenziell zuzunehmen scheint. Eine entscheidende Frage ist: gilt dies auch für alle Klassen in allen Ländern? Eine Rechnung auf alle Länder ist schwierig: Die Daten stammen nicht aus einem Experiment. Faktoren, wie Motivation der Schülerin-



**Abb. 1.3** Klassengrösse und Lernerfolg bei zehn Ländern, OECD Untersuchung *Education at Glance* aus dem Jahr 2005

nen und Schüler, soziales Umfeld, Lehr- und Lernmethoden, Schwierigkeit der Prüfungen wurden bei der Studie nicht kontrolliert. Der Prüfungserfolg von Korea könnte daher wegen dieser Faktoren und nicht wegen der Klassengrösse erklärbar sein. □

Es ist also schwierig aus *Beobachtungen*, das heisst aus Daten, die nicht kontrolliert gesammelt wurden, kausale Schlüsse zu ziehen. Die untersuchte Grösse könnte hier von Faktoren abhängen, die unbekannt sind. So ist es möglich, dass eine abhängige Variable  $Z$  (Prüfungserfolg), scheinbar von  $X$  (Klassengrösse) abhängt, in Wahrheit das Resultat einer dritten unkontrollierten Variable  $Y$  (Lernmethoden) ist. In einem solchen Fall spricht man von einem *Konfundierungs- oder Vermengungseffekt* (engl. *confounded variables*).<sup>4</sup>

Grafische Darstellungen, wie die Figur zu den Schülerdaten, erlauben es, Daten klar, präzise und effizient darzustellen. Um Daten zu verarbeiten oder grafisch zu illustrieren, sollten die gemessenen Werte miteinander vergleichbar sein. Hier ein Beispiel dazu:

**Beispiel 1.6 (Unwetterschäden)** Das Bundesamt für Statistik der Schweiz sammelt Schadensummen von Unwettern, um aussergewöhnliche Unwetterkatastrophen zu beschreiben und Versicherungsprämien zu überprüfen. Schadensummen von Unwetterkatastrophen sind Grössen, die von vielen Faktoren abhängen und damit stark variieren. Bei Schadensummen zu zukünftigen Unwettern können wiederum zwei Arten von Fragen von Interesse sein:

<sup>4</sup> Konfundieren heisst vermengen oder durcheinander geraten.

**Tab. 1.3** Schadensummen  
(in Mio. Franken) wegen  
Unwettern in der Schweiz  
(Statistisches Jahrbuch der  
Schweiz, 1997)

Jahr	Summe	Jahr	Summe	Jahr	Summe
1977	255	1984	105	1991	45
1978	525	1985	50	1992	60
1979	50	1986	120	1993	900
1980	20	1987	1230	1994	200
1981	50	1988	125	1995	70
1982	40	1989	10	1996	35
1983	60	1990	295	1997	195

**Tab. 1.4** Preisentwicklung,  
Daten der Gebäudeversiche-  
rung des Kantons Luzern  
(Statistisches Jahrbuch der  
Schweiz, 1990 und 1999)

Jahr	Index	Jahr	Index	Jahr	Index
1977	88,5	1984	118,7	1991	152,9
1978	90,7	1985	121,2	1992	153,8
1979	93,3	1986	124,4	1993	147,8
1980	100,0	1987	127,0	1994	148,0
1981	107,5	1988	130,6	1995	151,8
1982	114,6	1989	135,8	1996	149,7
1983	117,3	1990	145,2	1997	141,9

- (a) Kann man *nicht direkt messbare Grössen* zu Schadensummen, wie die durchschnittliche Schadensumme der nächsten fünf Jahre, *rechnen*? Wie genau und wie sicher ist eine solche Angabe?
- (b) Kann man einzelne zukünftige Schadensummen *prognostizieren*? Etwa: Wie gross ist die Wahrscheinlichkeit, dass nächstes Jahr eine Schadensumme von mindestens einer Milliarde Franken auftritt?

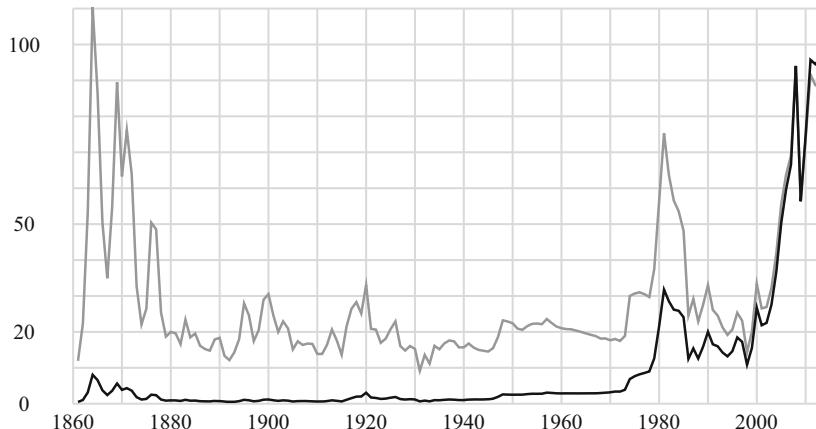
Um solche Fragen zu beantworten, müssen Daten zu Schadensummen vorliegen. Tab. 1.3 zeigt die während der Jahre 1977 bis 1997 festgestellten Schadensummen der Unwetter in der Schweiz. Will man die Schadensummen aus verschiedenen Jahren vergleichen, so müssen die Beobachtungen dieselbe Messeinheit besitzen. Dies ist hier nicht der Fall. Der Wert des Frankens hat nämlich zwischen 1977 und 1996 inflationsbedingt abgenommen.

Schadensummen von Unwettern bestehen grösstenteils aus Schäden an Gebäuden. Um auf reale Franken umzurechnen, betrachtet man daher hier die Preisentwicklung zur Erstellung von Wohnbauten in Tab. 1.4.

Die Schadensumme des Jahres 1987 beträgt damit in Mio. Franken des Jahres 1980

$$1230 \text{ Mio. Franken} \cdot 100,0 / 127,0 = 969 \text{ Mio. Franken}$$

Die Schadensumme des Jahres 1978 hat 1980 einen Wert von  $525 \times 100,0 / 90,7 = 579$  Mio. Franken. Die Daten mit Schadensummen in realen Millionen Franken des Jahres 1980, dargestellt in Tab. 1.5, dienen als Grundlage der Untersuchung. Die Schadensummen der Jahre 1978 und 1993 sind etwa gleich gross.  $\square$



**Abb. 1.4** Entwicklung des Preises in US Dollar für ein Barrel Erdöl (U.S. Energy Information Administration, 2014): graue Linie in realen Dollar des Jahres 2008, schwarze Linie nicht inflationsbereinigt

**Tab. 1.5** Schadensummen in realen Millionen Franken des Jahres 1980

Jahr	Summe	Jahr	Summe	Jahr	Summe
1977	288	1984	88	1991	29
1978	579	1985	41	1992	39
1979	54	1986	96	1993	609
1980	20	1987	969	1994	135
1981	47	1988	96	1995	46
1982	35	1989	7	1996	23
1983	51	1990	203	1997	137

Abb. 1.4 zeigt ein weiteres Beispiel. Es ist die Preisentwicklung von 1861 bis 2013 des Preises für ein Barrel Rohöl einmal inflationsbereinigt in Dollar des Jahres 2008 und einmal – weniger sinnvoll – in absoluten Zahlen.

### 1.3 Statistik und Qualitätskontrolle

Kein Produktionsprozess kann so eingestellt werden, dass alle produzierten Waren identisch sind: Verschleiss von Antriebsteilen oder Schwankungen von Temperatur und Feuchtigkeit führen zu geringen Abweichungen bei den produzierten Waren. Es ist aber wichtig, dass ein Produktions- oder Fertigungsprozess optimiert wird und innerhalb gewisser Grenzen kontrolliert abläuft. Dies geschieht in der Regel wie folgt:

- (1) Ein Produktionsprozess wird gestartet und man versucht, ihn stabil laufen zu lassen. Stabil bedeutet wissenschaftlich, dass der Prozess unter *statistischer Kontrolle* (engl. *statistical control*) ist.
- (2) Ist der Produktionsprozess unter statistischer Kontrolle, lassen sich mit den in Abschn. 1.1 vorgestellten Werkzeugen – Wahrscheinlichkeitsmodell, Daten, Vorwissen und der Regel von Bayes – Kennzahlen der Produktion berechnen. Es lässt sich auch prognostizieren, in welchem Bereich zukünftige Produktionsgrößen liegen werden.
- (3) Anschliessend kann man versuchen, die Qualität der Produktion zu erhöhen. Eine höhere Qualität führt zu weniger Ausschuss und damit zu tieferen Gesamtkosten der Produktion. Auch ist es wichtig, den stabilisierten Prozess weiter zu überwachen.

Die *statistische Prozessregelung* oder die *statistische Prozesssteuerung* (engl. *statistical process control* (SPC)) definiert, (a) was ein stabiler Prozess ist, sagt, (b) wie ein solcher unter statistischer Kontrolle gebracht werden kann und (c) wie man die Qualität eines Produktionsprozesses erhöhen kann.<sup>5</sup> Hier eine Illustration dazu:

**Beispiel 1.7 (Nicht keimende Blumenzwiebeln)** Eine Gärtnerei produziert jeden Tag 40 Kisten mit je 100 Blumenzwiebeln. Erfahrungen der Gärtnerei zeigen, dass der Produktionsprozess nur wirtschaftlich erfolgreich ist, wenn praktisch alle ausgelieferten Blumenkisten höchstens 15 % nicht keimende Blumenzwiebeln beinhalten. Blumenkisten, welche diese Spezifikation nicht erfüllen, führen zu unnötigen Kosten durch eventuelle Garantieansprüche. Um die Produktion zu überwachen, wird jeden Tag eine Kiste der Produktion kontrolliert. Die Blumenzwiebeln in der Kiste werden gepflanzt und der Anteil (in %) nicht keimender Blumenzwiebeln wird gezählt. Tab. 1.6 zeigt einen Ausschnitt aus der Kontrolle während 50 Tagen.<sup>6</sup> Der Produktionsprozess bei der Gärtnerei läuft si-

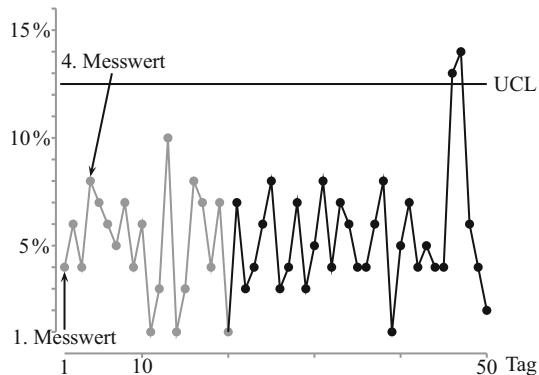
**Tab. 1.6** Kontrollblatt aus der Produktion (Die Daten werden horizontal gelesen.)

4	6	4	8	7	6	5	7	4	6
1	3	10	1	3	8	7	4	7	1
7	3	4	6	8	3	4	7	3	5
8	4	7	6	4	4	6	8	1	5
7	4	5	4	4	13	14	6	4	2

<sup>5</sup> Was statistische Prozessregelung genau umfasst, findet man in [1]. Beispiele, wie die statistische Prozessregulierung erfolgreich bei Unternehmen eingesetzt wurde, finden sich in [6].

<sup>6</sup> Das Beispiel wurde von E. Wyler (Berner FH) zur Verfügung gestellt. Es entstand durch eine Computersimulation.

**Abb. 1.5** Kontrollkarte der Produktion



cher nicht stabil, wenn der Anteil der nicht keimenden Blumenzwiebeln zunimmt. Auch „aussergewöhnlich“ hohe Messwerte können auf eine instabile Produktion hinweisen. Mit einer Kontrollkarte (Abb. 1.5) lassen sich die Messwerte aus der Produktion gegen die Zeitachse aufzeichnen. Die Messkarte zeigt, wie erwartet, dass die Messwerte wegen variierender Produktionsbedingungen, aber auch wegen Glück oder Pech, streuen. Auffallend sind zwei Arten von Streuungen: erstens Messwerte, die oberhalb einer oberen Kontrollgrenze UCL (engl. *upper control limit*) liegen und zweitens Messwerte, die unterhalb der Kontrollgrenze liegen. In der abgebildeten Kontrollkarte wurde die Grenze aus den ersten zwanzig kontrollierten Blumenzwiebelkisten geschätzt. Man erhält hier:<sup>7</sup>

$$\text{UCL} \approx 12,5\%$$

Messwerte, die unterhalb der UCL-Grenze streuen, führt man auf „Zufallsphänomene“ zurück. Sie sind ein Ausdruck der normalen Streuung der Messgrösse auf Grund von nie konstant haltbaren Produktionsbedingungen oder des Produktionssystems selbst. Der Wert 10 % am 13. Tag ist nicht aussergewöhnlich, da er unterhalb der Kontrollgrenze liegt. Statistisch kann man sagen: dieser Wert ist nicht deutlich höher als der Wert 1 % am elften Tag. Der Wert 13 % vom 46. Tag ist deutlich höher als der Wert 10 %! □

Das Beispiel illustriert, dass ein professionelles Qualitätsmanagement nicht auf einzelne Messwerte reagiert. Vielmehr werden Regeln zur Kontrolle aufgestellt, die auf Trends oder Zyklen von Messwerten oder auf dem Unterschied zwischen „normaler“ Streuung und aussergewöhnlicher, spezieller Streuung der Messwerten basieren.

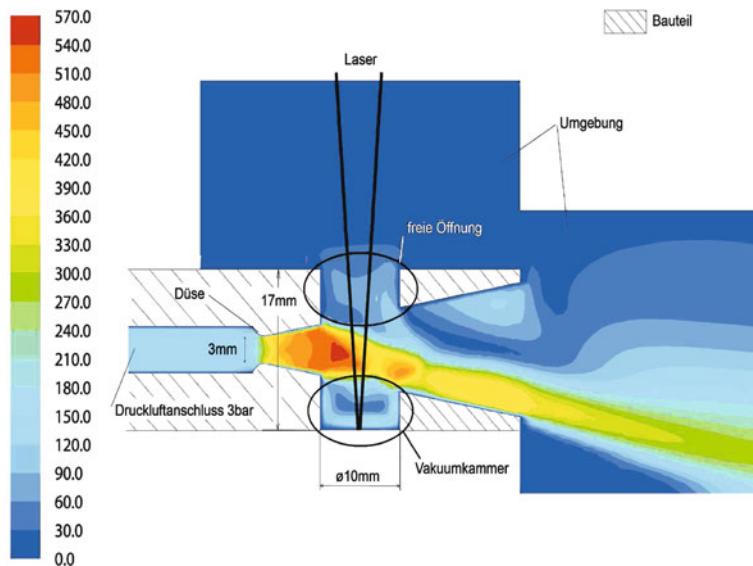
<sup>7</sup> Wie dies gerechnet wurde, wird später erklärt.

Wie kann die Gärtnerei ihre Produktion stabil halten oder gar verbessern? Die entscheidende Beobachtung ist, dass dazu nicht nur der durchschnittliche Anteil nicht keimender Blumenzwiebeln kontrolliert, sondern zusätzlich die Streuung der Produktion möglichst klein gehalten werden muss! Bei einer kleinen Streuung können Aussagen zur Anzahl nicht keimender Blumenzwiebeln präziser formuliert werden als bei grosser Streuung. Auch ist die Wahrscheinlichkeit, dass bei einer kontrollierten Kiste mehr als 15 % nicht keimende Blumenzwiebeln festgestellt werden, bei kleiner Streuung gering. Daher wird auch gesagt:

*Qualität ist umgekehrt proportional zur Streuung oder zur Unsicherheit.*

Hier ein weiteres Beispiel einer Untersuchung aus einem technischen Bereich, bei dem ein Prozess beurteilt werden musste:

**Beispiel 1.8 (Druck in einer Vakuumkammer)** Abb. 1.6 zeigt die Auslegung einer Luftkammer, in die Luft mit einem Druck von 3 bar geblasen wird. Auf der linken Seite befindet sich der Druckluftanschluss. Die Luft fliesst nach rechts und entweicht der Kammer aus der grossen Öffnung. Im oberen Teil der Kammer befindet sich eine kleine Öffnung. Auch dort entweicht Luft; im unteren Teil, in der Vakuumkammer, bildet sich ein Unterdruck. Die Grafik zeigt den Geschwindigkeitsverlauf der Luft in m/s in und ausserhalb der Kammer.



**Abb. 1.6** Die Druckauslegung in einer Vakuumkammer (M. Grossenbacher, Berner Fachhochschule, Bachelorarbeit, 2006)

**Tab. 1.7** 20 Messungen des Unterdrucks (in bar) in der Vakuumkammer (Die Reihenfolge der Messwerte liest man vertikal entlang der Spalten.)

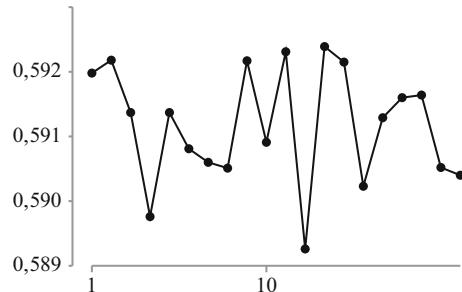
0,59198	0,59137	0,59217	0,59239	0,59160
0,59218	0,59081	0,59091	0,59215	0,59164
0,59137	0,59060	0,59231	0,59023	0,59052
0,58976	0,59051	0,58926	0,59129	0,59040

Entwicklerinnen der Kammer mussten den Unterdruck in der Kammer bestimmen. Der Unterdruck kann aber nicht direkt gemessen werden. Einerseits variieren Messungen des Unterdrucks wegen kleinen Luftturbulenzen in der Kammer und andererseits haben Messinstrumente zufällige Messfehler. Messungen des Unterdrucks in der Vakuumkammer werden also um diesen streuen. Diese Überlegungen führen dazu – wie bei den Wartezeiten zwischen starken Erdbeben – zwei Arten von Fragen zu stellen:

- (a) Kann eine *nicht direkt messbare Grösse*, wie der Unterdruck in der Vakuumkammer, *berechnet* werden? Wie genau und plausibel ist eine solche Angabe?
- (b) Können weitere, gemessene Druckwerte in der Kammer *prognostiziert* werden? Ein Beispiel: Wie gross ist die Wahrscheinlichkeit, dass ein weiterer Messwert beim Betrieb der Vakuumkammer zwischen 0,590 bar und 0,593 bar liegen wird?

Beide Fragen können nur beantwortet werden, wenn Messungen zur Vakuumkammer vorliegen. Tab. 1.7 zeigt 20 Messungen des Unterdrucks. Wie erwartet streuen die Messwerte um den gesuchten Druck. Wie im ersten Teil des Kapitels gesagt, können die beiden obigen Fragen mit der Regel von Bayes und dem Gesetz der Marginalisierung beantwortet werden. Dazu braucht man ein Datenmodell, das besagt, wie Messwerte um den gesuchten Druck streuen und Vorinformation. Bevor man dies tut, ist es sinnvoll zu beurteilen, ob das Experiment unter statistischer Kontrolle liegt. Abb. 1.7 zeigt die Messwerte in der Reihenfolge ihrer Messung. Das Streudiagramm zeigt weder Trends nach höheren oder tieferen Messwerten. Zudem sind keine aussergewöhnlichen Messwerte vorhanden. Der Versuch ist unter statistischer Kontrolle. □

**Abb. 1.7** Die Messwerte des Drucks in der Vakuumkammer (mit normalem Streuverhalten) gegen die Zeitachse geplottet



## Reflexion

**1.1** Ein Ärztin glaubt, dass der Patient  $X$  Malaria hat. Sie schätzt dabei die Chance mit  $\mathbb{O}(X \text{ hat Malaria} \mid \text{Vorwissen } \mathcal{K}) = 20 : 80$ .

- (a) Der OptiMAL-Test schlägt positiv aus. Mit welcher (aktualisierter) Chance muss die Ärztin rechnen, dass der Patient von Malaria betroffen ist? Welche Chance erhält die Ärztin, wenn der Test negativ ausfällt? Beschreiben Sie die berechneten Chancen auch mit Wahrscheinlichkeiten.
- (b) Was passiert, wenn die Ärztin einen positiven OptiMAL-Test erhält, wenn sie davon ausgeht, dass der Patient eine Chance von nur 1 : 99 hat, Malaria zu haben?
- (c) Mit dem Test OptiMAL, der eine hohe Sensitivität und Spezifität hat, könnte man alle Tropenreisenden, die in die Schweiz einreisen, auf Malaria untersuchen. Man nennt dieses Verfahren ein *Screening* oder eine „Fishing expedition“ (einen Fischzug). Ist dies sinnvoll? Gehen Sie davon aus, dass jeder tausendste Tropenreisende an Malaria erkrankt.<sup>8</sup>

**1.2** Eine Equipe sucht einen Flugschreiber auf dem Meeresgrund in einem Gebiet von  $10 \text{ km}^2$ . Sie geht davon aus, dass der Flugschreiber sich dort mit einer Chance von 1 : 1 (oder einer Wahrscheinlichkeit von 50 %) befindet. Um den Flugschreiber zu finden, wird eine Methode benutzt, die eine Richtigpositiv-Rate von 0,6 und eine Richtignegativ-Rate von 0,95 hat.<sup>9</sup>

- (a) Erklären Sie, was Richtigpositiv-Rate und Richtignegativ-Rate hier bedeuten.
- (b) Der Einsatz der Methode liefert keinen Hinweis auf den Flugschreiber. Wie gross ist die aktualisierte Chance, dass der Flugschreiber sich im Gebiet befindet? Wie lautet diese Chance mit einer Wahrscheinlichkeit?
- (c) Eine zweite, unabhängige Methode mit gleicher Sensitivität und Spezifität wird nun eingesetzt. Auch diese Methode liefert keinen Hinweis auf den Flugschreiber. Wie gross ist die aktualisierte Chance, dass der Flugschreiber sich im Gebiet befindet? Wie lautet diese Chance mit einer Wahrscheinlichkeit?

---

<sup>8</sup> Das Phänomen, dass Screenings bei medizinischen Krankheiten, die nur kleine Teile der Bevölkerung betreffen, nicht sinnvoll sind, wird oft vergessen. Man hört etwa, dass alle Frauen regelmäßig auf Brustkrebs zu untersuchen sind, oder alle sexuell aktiven Leute sich einem Aids-Test unterziehen sollten. Abgesehen von den horrenden Kosten spricht auch die in (c) gemachte Rechnung gegen so umfassende Abklärungen.

<sup>9</sup> Einen Überblick zu Strategien, um „verlorene“ Objekte mit grösstmöglicher Wahrscheinlichkeit und möglichst schnell zu finden, findet man in [2], [8], [9] und [10].

**1.3** Eine Person geht davon aus, dass es morgen an ihrem Wohnsitz mit einer Wahrscheinlichkeit von 60 % regnen wird.

- (a) Mit welcher Chance rechnet die Person, dass es morgen an ihrem Wohnsitz regnen wird?
- (b) Die Person geht davon aus, dass der Wetterdienst bei einer Regenmeldung eine Richtigpositiv-Rate von 0,9 und eine Richtignegativ-Rate von 0,8 hat. Erklären Sie, was diese zwei Zahlen hier genau bedeuten.
- (c) Eine Stunde später meldet der Wetterdienst, dass es morgen tatsächlich regnen wird. Von welcher aktualisierten Chance und Wahrscheinlichkeit muss die Person ausgehen, dass es morgen an ihrem Wohnsitz regnen wird?

**1.4** Kommissar Huber verdächtigt 100 Personen einer Kleinstadt, einen Mord begangen zu haben. Alle Personen verneinen den Verdacht. Um den Täter unter den 100 Personen zu finden, will der Kommissar einen Lügendetektor einsetzen. Dieser hat nach [4] eine Sensitivität von 0,88 und eine Spezifität von 0,86.

- (a) Erklären Sie, was Sensitivität und Spezifität hier bedeuten.
- (b) Wie gross setzt der Kommissar die Wahrscheinlichkeit, dass eine Person den Mord begangen hat? Wie lautet diese Wahrscheinlichkeit mit einer Chance?
- (c) Ist der Einsatz des Lügendetektors sinnvoll? Berechnen Sie dazu die Chance und anschliessend die Wahrscheinlichkeit, dass die Person den Mord begangen hat, wenn der Lügendetektor ausschlägt.

**1.5** Der PSA-4.0-Test wird in der Medizin verwendet, um bei Vorsorgeuntersuchungen Prostatakrebs zu entdecken. Nach mediXSchweiz (Gesundheitsdossier Prostatavergrösserung, 2009) ist die Spezifität des Tests ist 0,93. Die Sensitivität ist jedoch lediglich 0,2.

- (a) Ein Arzt vermutet nach einem Gespräch mit einem Mann, dass dieser Mann Prostatakrebs hat. Er setzt die Chance auf 2 : 3. Wie gross ist die aktualisierte Chance, dass der Mann Prostatakrebs hat, wenn der Test positiv ausschlägt?
- (b) Man geht davon aus, dass von 1000 Männern im Alter zwischen 50 und 70 Jahren 25 Prostatakrebs haben. Ist es für einen Arzt sinnvoll den PSA-4.0-Test bei all seinen männlichen Patienten im Alter zwischen 50 und 70 Jahren einzusetzen?

**1.6** Sie benutzen das folgende einfache Instrument, um bei einer Person Malaria zu detektieren: Sie werfen eine Münze, fällt sie auf „Kopf“, so sagen Sie, dass die Person Malaria hat. Fällt die Münze auf „Zahl“, so ist das Testresultat negativ.

- (a) Wie lautet die Sensitivität und die Spezifität des Instruments?
- (b) Was liefert die Regel von Bayes bei diesem Instrument? Erstaunt Sie das Resultat?

**1.7** Eine Firma produziert elektrische Teile. Die Produktionschefin geht davon aus, dass rund 1 % der produzierten Teile defekt sind. Weiter kann sie die produzierten Teile mit einer Apparatur testen. Dabei besteht eine Wahrscheinlichkeit von 0,975, dass ein von der Apparatur als defekt bezeichnetes Teil auch defekt ist und eine Wahrscheinlichkeit von 0,999, dass ein als nicht defekt bezeichnetes Teil nicht defekt ist.

- (a) Wie lauten die Sensitivität und die Spezifität des Tests mit der Apparatur?
- (b) Wie gross ist für die Produktionschefin die Chance, das ein zufällig gewähltes Teil defekt ist?
- (c) Die Produktionschefin wählt ein elektrisches Teil der Produktion. Die Apparatur meldet es defekt. Wie gross ist (1) die aktualisierte Chance und (2) die aktualisierte Wahrscheinlichkeit, dass das produzierte Teil wirklich defekt ist?
- (d) Die Apparatur meldet bei einem produzierten elektrischen Teil, dass es nicht defekt ist. Wie gross ist die Chance, dass das produzierte Teil wirklich nicht defekt ist?

**1.8** Aus der Gratiszeitung 20-Minuten vom 18. Oktober 2004:

**Unfall beim Freizeitsport** Die Zahl der Freizeitunfälle hat zugenommen. Wir wollten darum wissen, ob unsere Leser sich schon einmal beim Freizeitsport verletzt haben. 1031 Teilnehmer (52 %) der 20-Minuten-Internetumfrage beantworteten die Frage mit Nein. Teilnehmertotal: 2003.

Welches grundsätzliche Problem birgt eine solche Internet-Umfrage?

**1.9** Tab. 1.8 zeigt den von der Regierung der USA festgelegten Minimallohn pro Stunde aus den Jahren 1960 bis 1995, die Arbeitgeber ihren Arbeitern zahlen müssen.

- (a) Zeigen Sie mit einer geeigneten Grafik, wie sich der Minimallohn entwickelt hat. Benutzen Sie zur Umrechnung auf reale Dollar von 1960 Tab. 1.9.
- (b) Wie würde die Grafik aussehen, wenn nur die nicht-inflationsbereinigten Minimallöhne pro Stunde dargestellt würden? Kommentieren Sie die erhaltene Grafik.
- (c) Das Medianeinkommen pro Haushalt betrug in den USA im Jahr 1980 \$ 17 710. Dies bedeutet, dass 50 % der Haushalte in den USA ein Einkommen von weniger als \$ 17 710 hatten und 50 % ein Einkommen über diesem Wert. Im Jahre 1999 betrug

**Tab. 1.8** Minimallohn pro Stunde in Dollar aus den Jahren 1960 bis 1995 in den USA

Jahr	1960	1965	1970	1975	1980	1985	1990	1995
Minimallohn	1,00	1,25	1,60	2,10	3,10	3,35	3,80	4,25

**Tab. 1.9** Die Preisentwicklung der Konsumentenpreise in den USA

Jahr	1960	1965	1970	1975	1980	1985	1990	1995
Preisindex (CPI)	29,6	31,5	38,8	53,8	82,4	107,6	130,7	152,4

**Tab. 1.10** Durchschnittlicher Verdienst von Verwaltungsratsmitgliedern und Qualität (Heidrick & Struggles, 2009, Daten aus Le Monde, 26.3.2009)

Land	durchschn. Verdienst	Qualität
Schweiz	194 000	64
Deutschland	110 000	39
Spanien	108 000	52
England	108 000	77
Italien	79 000	53
Portugal	68 000	41
Holland	67 000	71
Belgien	65 000	47
Dänemark	60 000	37
Schweden	54 000	66
Finnland	50 000	62
Frankreich	48 000	60
Österreich	25 000	36

das Medianeinkommen \$ 40 816. Um wie viel Prozent hat das Medianeinkommen zugenommen? (Hinweis: Für das Jahr 1999 beträgt der CPI 166,6.)

**1.10** Die Beratungsfirma Heidrick & Struggles hat untersucht, ob die Lohnhöhe von Verwaltungsratsmitgliedern mit der Qualität ihrer Arbeit, basierend auf 41 Kriterien, zusammenhängt. Resultate von 13 Ländern in Europa finden sich in Tab. 1.10.

Stellen Sie die Messpunkte in einem Streudiagramm dar. Sind aussergewöhnliche Beobachtungen vorhanden? Handelt es sich um Messwerte aus einem Experiment oder um Beobachtungen?

**1.11** Tab. 1.11 zeigt die Ausgaben der öffentlichen Hand für Forschung und Entwicklung von neun Regionen in Europa. Ist es möglich, eine Rangliste der Regionen nach ihren Ausgaben für Forschung und Entwicklung zu machen?

**Tab. 1.11** Ausgaben für Forschung und Entwicklung (BIP ist das Bruttoinlandsprodukt, aus: Eurostat)

Region	Jahr	in Mio. Euro	in % des BIP
Braunschweig (D)	1997	1675	4,84
Ile-de-France (F)	1998	12416	3,43
Midi-Pyrénées (F)	1998	1803	3,70
Oberbayern (D)	1997	5911	4,38
Stuttgart (D)	1997	5045	4,79
Pohjois-Suomi (Fin)	1998	410	3,82
Rheinessen-Pfalz (D)	1997	1527	3,50
Tübingen (D)	1997	1608	4,05
Uusimaa (Fin)	1998	1571	3,73

**Tab. 1.12** Jahresdurchschnittspreis (in CHF) von Benzin (Bleifrei 95) und Preisentwicklung in der Schweiz (Landesindex der Konsumentenpreise) in den Jahren 1970–2010

Jahr	1970	1980	1990	2000	2010
Benzin	0,59	1,16	1,10	1,40	1,64
Preisindex	66,9	108,6	151,6	183,8	200,3

**1.12** Die Staaten der Europäischen Union publizieren jeden Monat ihre Arbeitslosenquoten. Versuchen Sie herauszufinden, ob die Quoten bei allen Staaten gleich gemessen werden und damit die Daten verglichen werden können.

**1.13** Tab. 1.12 zeigt den Jahresdurchschnittspreis (in CHF) von Benzin (Bleifrei 95) und die Preisentwicklung in der Schweiz (Landesindex der Konsumentenpreise) in den Jahren 1970–2010.

- (a) Zeigen Sie mit einer geeigneten Grafik, wie sich der Benzinprix entwickelt hat.
- (b) Um wie viel Prozent hat der Benzinprix vom Jahr 1970 bis zum Jahr 2010 zu- oder abgenommen?

**1.14** In Zeitungen und Zeitschriften finden sich viele Analysen zu Daten. Finden Sie heraus: Handelt es sich jeweils um Messwerte aus Experimenten oder Daten aus Beobachtungen?

**1.15** Chloridgehalte von Lösungen können mit einer Methode, die auf elektrische Ladungen in Molekülen basiert, bestimmt werden. Hier zehn Messungen (in mol/m<sup>3</sup>) einer Kalium-Chloridlösung:

108,7 110,9 101,1 102,5 100,1 101,9 105,8 106,0 104,1 105,1

Die Messungen variieren, da die Umwelteinflüsse auf die Experimentierkammer nicht konstant gehalten werden können und die Messinstrumente Ungenauigkeiten haben. Hat es Trends oder zyklische Muster in den Daten? War das Experiment unter statistischer Kontrolle?

**1.16** Die Firma Tillamook-Cheese in Oregon (USA) produziert quaderförmige Frischkäsekörper mit einer mittleren Masse von etwa 19 kg. Die Körper werden gereift und anschliessend für den Handel in kleine Portionen von 500–1000 Gramm zerschnitten. Zur Qualitätskontrolle müssen neben der Zusammensetzung der Körper auch die durchschnittliche Masse und die Streuung der Massen einer Tagesproduktion berechnet werden. Tab. 1.13 zeigt 20 Messungen von Massen von Frischkäsekörpern.

Überprüfen Sie, ob die Messwerte keine Trends zeigen. Deuten die Messwerte darauf hin, dass die Firma ihre Produktion unter statistischer Kontrolle hat?

**Tab. 1.13** 20 Messungen von Frischkäsekörpern vom 18. Juli 2008 (Angaben in amerikanischen Pfund, Reihenfolge der Daten entlang der Spalten)

42,31	42,51	42,46	42,52	42,04	42,44	42,10	42,11	42,36	42,14
42,28	42,37	42,24	42,02	42,17	42,24	42,08	41,85	42,41	42,47

**1.17** Zeichnen Sie ein Streudiagramm der Schadensummen der Unwetter in der Schweiz während den Jahren 1977 bis 1997. Hat es Trends oder sind zyklische Muster aus den Daten ableitbar?

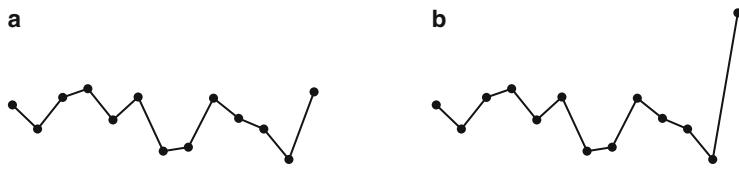
**1.18** Eine Messmethode, um eine nicht direkt messbare Grösse  $K$  zu ermitteln, funktioniere „korrekt“: Mit einer Wahrscheinlichkeit von 50 % liegen Messwerte unterhalb von  $K$  und mit einer Wahrscheinlichkeit von 50 % oberhalb von  $K$ .

- (a) Wie gross ist die Wahrscheinlichkeit, dass bei zwei Messungen beide Messwerte grösser als  $K$  sein werden?
- (b) Berechnen Sie die Wahrscheinlichkeit, dass bei drei Messungen der kleinste Messwert kleiner und der grösste Messwert grösser als  $K$  sein wird. (*Tipp:* Listen Sie alle möglichen Fälle auf.)

**1.19** Abb. 1.8 zeigt zwei Messreihen einer Prozessgrösse. Normale, variierende Produktionsbedingungen bewirken, dass in Abb. 1.8a der letzte Messwert höher als der zweitletzte Messwert ist. In Abb. 1.8b ist dies kaum der Fall.

Im Schweizer Fernsehen wird jeden Abend versucht, zu erklären, warum der Börsenindex höher oder tiefer als am Vortag ist. Wie ist dies zu bewerten?

**1.20** Die Gemeindepolizei Köniz führte im Jahr 2008 auf Strassen Geschwindigkeitskontrollen durch. Insbesondere ist es für sie interessant zu wissen, wie gross der Anteil der zu schnell fahrenden Fahrzeuge ist. So wurden gemäss dem Anzeiger Region Bern vom 8.2.2008 765 Fahrzeuge kontrolliert, 60 davon fuhren zu schnell. Sind die Daten Beobachtungen oder Messwerte eines Experiments?



**Abb. 1.8** Zwei verschiedene Messreihen: **a** mit normaler Streuung, **b** mit einem aussergewöhnlichen Messwert

## Literatur

1. D. S. Chambers, D. J. Wheeler, *Understanding Statistical Process Control* (SPC-Press, Inc., 1992)
2. S. Davey, N. Gordon, I. Holland, M. Rutten, J. Williams, *Bayesian Methods in the Search for MH 370* (Springer Verlag, 2016)
3. P. A. C. Gane, M. Buri, D. C. Spielmann, B. Neuenschwander, H. Scheidiger, D. Bättig, Mechanism of Post-Print Laser Marking on Coated Substrates: Factors Controlling Ink Ablation in the Application of High Brightness Calcium Carbonate. *Journal of Graphic Technology* **1** (2002)
4. J. Gastwirth, The statistical precision of medical screening procedures. *Statistical Science* **3**, 213–222 (1987)
5. E. T. Jaynes, *Papers on Probability, Statistics and Statistical Physics* (Kluwer Academic Publishers, 1989), herausgegeben von R. D. Rosenkrantz, S. 155
6. M. J. Kiemle, S. R. Schmidt, R. J. Berdine, *Basic Statistics, Tools for Continuous Improvement*, 4th ed. (Air Academy Press & Associates, LLC, 2000)
7. C. J. Palmer, J. F. Lindo, W. I. Klaskala, J. A. Quesada, R. Kaminski, M. K. Baum, A. L. Ager, Evaluation of the OptiMAL test for rapid diagnosis of Plasmodium vivax and Plasmodium falciparum malaria. In *J. Clin. Microbiol.* **36**(1) (1998)
8. L. D. Stone, The proces of search planning: current approaches and continuouing problems. *Oper. Res.* **31**, 207–233 (1983)
9. L. D. Stone, T. M. Kratzke, C. M. Keller, J. P. Strumpfer, Search for the Wreckage of Air France AF 447. *Statistical Science* **29**, 69–80 (2014)
10. L. D. Stone, J. O. Royset, A. R. Washburn, *Optimal Search for Moving Targets* (Springer Verlag, New York, 2016)

# Wie man Versuche planen kann

2

„Am vierzehnten März – ich glaube es wenigstens –“, sagte er.  
„Am fünfzehnten“, sagte der Schnapphase.  
„Am sechzehnten“, sagte die Haselmaus.  
„Schreibt euch das auf“, sagte der König zu den Schöffen, und sie schrieben eifrig alle drei Daten auf ihre Tafeln, zählten sie zusammen und rechneten sie in Pfund und Zentner um.  
*Lewis Carroll, Alice im Wunderland (Insel Taschenbuch, 1973, S. 113)*

## Zusammenfassung

In Gebieten der angewandten Wissenschaften, in der Medizin und in der Welt der Technik können Auswirkungen von Ereignissen, von Medikamenten oder von produzierten Gütern auf die Umwelt und den Menschen kaum allein durch theoretische Überlegungen modelliert und analysiert werden. So muss eine Versicherung Daten sammeln, um verlässliche Prognosen machen zu können, wie viele zukünftige, grosse Schadenssummen sie pro Jahr bezahlen muss. Auch in Produktionsprozessen können Verbesserungen und Effizienzsteigerungen nur erzielt werden, wenn Daten vorhanden sind. Dass Daten die Grundlage bilden, um Wissen zu generieren, zeigt ein bekannter Slogan aus dem Qualitätsmanagement:

*In God we trust, all others bring data.*

Daten erlauben, Wissen zu einer nicht direkt messbaren Grösse oder zu zukünftigen Werten einer unsicheren Grösse aufzubauen. Wie Versuche oder Experimente geplant werden, um hochstehendes Datenmaterial zu erhalten, wird in diesem Kapitel vorgestellt.

## 2.1 Worauf bezieht sich eine Grösse?

Effizient Daten sammeln kann man nur, wenn die *Zielsetzung* der Untersuchung festgelegt ist. Welche Messgrösse interessiert und worauf bezieht sie sich?<sup>1</sup> Die durchschnittliche Fahrzeit von Zügen, die von Bern nach Zürich fahren, wird verschieden aussehen, wenn man alle Züge des Jahres 2020, alle Züge während der morgendlichen Stosszeiten von sieben bis zehn Uhr oder alle Züge während der Sommermonate von Juni bis August betrachtet. Eine Prognose zu Schadenereignissen und Schadensummen auf Grund von Feuer hängt davon ab, welches Zeitfenster und welche Gegend betrachtet wird: Eine Prognose für Schadensummen in der Schweiz in den nächsten zehn Jahren wird anders sein als eine Prognose zu Schadensummen zum nächsten Halbjahr im Staat Kalifornien. Bei Qualitätskontrollen einer Fabrik, die Fernseher produziert, interessieren die Lebensdauer und die Leuchtkraft der Fernseher aus der Gesamtproduktion.

Grössen, wie in den obigen Beispielen, beziehen sich auf eine Gesamtheit, der *Grundgesamtheit* (engl. *sample space* oder *population*). Bei einer Untersuchung zu Lebensdauern von Fernsehern des Typs KDL-46Z4500 der Firma Sony ist die Grundgesamtheit die Gesamtproduktion aller Fernseher dieses Typs. Möchte man die durchschnittliche Fahrzeit von Zügen bestimmen, die im nächsten Jahr von Bern nach Zürich fahren, so setzt sich die Grundgesamtheit aus allen Zügen zusammen, die diese Fahrstrecke im nächsten Jahr zurücklegen. Ob die tägliche Einnahme von Aspirin das Risiko von nicht-tödlichen Herzinfarkten reduziert, hängt davon ab, ob es sich bei der Grundgesamtheit um Patienten mit kardiovaskulären Risiken, Personen im Alter von 80 bis 90 Jahren oder Bankangestellte im Alter von 20 bis 30 Jahren handelt. Für eine Gebäudeversicherung könnten die erwartbaren Schadensummen von allen Unwettern in Europa in den nächsten zehn Jahren, der Grundgesamtheit, von Interesse sein.

**Beispiel 2.1 (Pflanzenvielfalt)** Um zu dokumentieren, wie sich die Pflanzenvielfalt in der Schweiz entwickelt, wurden im Rahmen des Biodiversitätsmonitoring 2009 der Schweizerischen Eidgenossenschaft Artenzahlen pro Parzelle von Pflanzentypen untersucht. Bei der Artenzahl des Pflanzentyps Z9 möchte man berechnen, wie sich die durchschnittlichen Artenzahlen in den Zonen Ackerland, Alpweiden, Grünland, Siedlungen, Wald und in ungenutzten Zonen unterscheiden. Man interessiert sich daher für die Artenzahlen Z9 bezüglich aller Parzellen in der Schweiz, der Grundgesamtheit. □

Es ist üblich, die Anzahl Elemente in der Grundgesamtheit mit  $N$  zu bezeichnen. Die Anzahl  $N$  ist manchmal genau bekannt, oftmals ist jedoch nur ihre Grössenordnung schätzbar. Ist  $N$  sehr gross, wie beispielsweise bei Prognosen zum Wahlverhalten von

---

<sup>1</sup> Dabei unterscheidet K. Ishikawa, ein Begründer des japanischen Qualitätsmanagements, in [5] folgende, sich überschneidende Kategorien: (a) Daten, um aktuelle Situationen zu verstehen, (b) Daten, um Beziehungen zwischen Grössen zu analysieren, (c) Daten zur Prozesskontrolle, (d) Daten zur Prozessregulierung und (e) Daten, um (einfache) Entscheide zu treffen.

Personen in einem Land, setzt man auch gern  $N = \infty$ . Man sagt dann, dass die Grundgesamtheit *konzeptionell* gewählt wird.

Nur in wenigen Untersuchungen, wie Volkszählungen, wird bei endlichen Populationen die gesamte Grundgesamtheit erfasst. Man spricht in diesem Fall von einer *Vollerhebung* (engl. *Zensus*). Vollerhebungen sind in der Regel teuer und sehr zeitaufwändig. Im Bereich der Produktion von Massengütern können Qualitätsgruppen selten Vollerhebungen durchführen, da Messungen oft für die untersuchten Objekte schädlich oder gar zerstörerisch sind. In Beispiel 1.7 der Blumenzwiebeln in Kap. 1 müssen die Zwiebeln gepflanzt werden, um zu testen, ob sie keimfähig sind. Um eine unbekannte Grösse aus der Produktion zu analysieren, braucht man daher nur eine Gruppe von Messwerten, eine *Stichprobe* (engl. *batch* oder *sample*). Die Anzahl Elemente der Stichprobe nennt man den *Umfang der Stichprobe*. Dieser wird in der Regel mit  $n$  bezeichnet. Bei Untersuchungen von Personen bezeichnet man die Personen der Stichprobe als die *Probanden* eines Versuchs.

**Beispiel 2.2 (Bleigehalt in Weinen)** Das Beispiel zeigt eine Untersuchung aus einem Lebensmittellabor in der Schweiz. Zu hohe Bleigehalte in Lebensmitteln sind für Menschen schädlich. Der Gesetzgeber hat daher in der Schweiz Grenzwerte für den Bleigehalt in Lebensmitteln festgelegt. Die gesetzlich *erlaubte* Bleimenge für Weine, die in der Schweiz verkauft werden, beträgt 100 µg/kg und der gesetzlich *tolerierte* Grenzwert lautet 300 µg/kg. Behördliche Kontrollstellen können nicht alle angebotenen Weine untersuchen. Sie müssen gesuchte Grössen der Weine im Handel mit Stichproben berechnen. Solche Grössen können der Anteil der Weine im Handel mit einem Bleigehalt von mehr als 300 µg/kg oder deren durchschnittlicher Bleigehalt sein.

Das Lebensmittellabor Baselland mass dazu im Jahr 1994 den Bleigehalt von 128 Rotweinen aus der EU, aus Nord- und Südamerika sowie aus Australien und Neuseeland, die auf dem schweizerischen Markt erhältlich sind. Die Messwerte aus [1] finden sich in Tab. 2.1. Die Grundgesamtheit – der Untersuchungsgegenstand der Studie – besteht aus al-

**Tab. 2.1** Bleigehalt (in µg/kg) von 128 Rotweinen aus der EU, aus Nord- und Südamerika und aus Australien und Neuseeland, die auf dem schweizerischen Markt erhältlich sind

18,0	154,0	36,0	525,0	31,0	29,0	50,0	58,2	74,0	47,0	29,1
93,0	56,0	39,3	53,0	852,0	632,0	34,0	40,0	60,0	87,0	102,0
13,0	14,0	78,9	65,5	94,0	20,0	39,0	436,0	37,0	40,0	11,0
56,3	14,0	48,0	60,0	33,0	62,0	25,0	55,0	32,1	16,0	7,0
49,0	77,0	70,9	81,0	40,0	60,0	90,0	18,0	26,4	55,5	101,1
124,0	95,0	46,0	59,0	31,6	101,0	45,3	46,0	34,0	48,0	61,0
31,4	43,8	47,0	75,0	52,0	71,0	770,0	42,0	4,4	44,0	37,0
43,0	57,0	118,9	457,0	73,0	29,3	152,0	51,0	51,0	37,8	35,0
21,9	48,2	31,0	40,5	71,0	61,0	74,0	117,0	230,0	95,0	
57,0	54,3	49,7	74,0	69,0	139,0	59,0	34,0	66,0	48,0	
39,2	42,4	28,0	42,6	84,0	56,0	44,0	70,0	12,0	21,0	
64,0	49,0	46,6	21,0	47,0	71,0	87,0	90,0	57,0	13,9	



**Abb. 2.1** Stichprobe, um die Artenvielfalt in der Schweiz zu bestimmen (Biodiversitäts-Monitoring Schweiz BDM, 2009)

len ausländischen Weinen, die in der Schweiz im Jahr 1994 verkauft wurden. Sie so gross, dass sie konzeptionell gewählt wird ( $N = \infty$ ). Der Stichprobenumfang  $n$  beträgt 128. □

**Beispiel 2.3 (Pflanzenvielfalt)** Bei der Untersuchung zur Pflanzenvielfalt von Beispiel 2.1 besteht die Stichprobe aus 1222 Messpunkten, die je zehn Quadratmeter gross sind. Abb. 2.1 visualisiert diese. □

Messwerte oder Beobachtungen können Zahlen oder Zahlentupel sein, wie etwa der Bleigehalt in Weinen, die Anzahl Herzinfarkte oder Schadensummen. Die Datenwerte bei der Anzahl nicht keimender Blumenzwiebeln sind ganze Zahlen: Man spricht von *diskreten* Werten. Nicht diskrete Werte bezeichnet man als *stetige* Messwerte. Werte von Grössen, die nicht mittels Zahlen miteinander verglichen werden können, wie zum Beispiel Farben, Gruppen (Männer, Frauen, Kinder) nennt man *kategorial*. Kategoriale Grössen können manchmal *geordnet* werden. So können Kaffeesorten nach den Kriterien „sehr schlecht“, „schlecht“, „gut“ oder „ausgezeichnet“ beurteilt werden. Man unterscheidet neben stetigen, diskreten oder kategorialen Werten auch die quantitative Vielfalt der Werte: Messungen mit einem Merkmal heissen *univariat*, Messungen mit zwei Merkmalen *bivariat* und solche mit mehreren Merkmalen *multivariat*. Das Beispiel 1.5 zur Untersuchung des Prüfungserfolgs zeigt bivariate Messwerte: gemessen wurden die Merkmale „Prüfungserfolg“ und „Klassengrösse“.

## 2.2 Grössen aus Physik und Technik

Bei Grössen aus der Physik und der Technik ist es meist nicht sinnvoll, von Stichproben und Grundgesamtheiten zu sprechen. Der Ingenieur will Grössen, wie Längen oder Gewichte, von Objekten berechnen. Dazu hat er Messungen, die um die gesuchten Grössen streuen. Man könnte hier als Grundgesamtheit alle möglichen, unendlich vielen, hypothetischen Messungen betrachten. Die Messungen sind dann eine „Stichprobe“ aus der Grundgesamtheit. Wir wollen in diesem Buch davon Abstand nehmen.

**Beispiel 2.4 (Zeit zwischen starken Erdbeben)** Beim Beispiel 1.3 zu den Zeiten zwischen aufeinanderfolgenden, starken Erdbeben interessiert man sich für die zukünftige, durchschnittliche Wartezeit  $\mu$ . Dies will man mit den beobachteten 28 Wartezeiten aus Tab. 1.2 berechnen. Man kann die Daten als Stichprobe der hypothetischen oder imaginären „Grundgesamtheit“ aller möglichen vergangenen und zukünftigen Wartezeiten betrachten. Vernünftiger scheint es, ohne diese Sehensweise aus den Daten  $\mu$  zu berechnen (und anzugeben, was dazu benutzt wurde, und zu sagen, wie genau und wie plausibel das Resultat ist).  $\square$

---

## 2.3 Größen und ihre Typ B Unsicherheit

Um Werte von gesuchten Größen zu erhalten, werden Messgeräte eingesetzt. Messgeräte besitzen eine *Präzision* (engl. *precision*). Eine einfache, grobe Regel ist:

Wird ein nicht diskreter Messwert über eine digitale Anzeige aufgenommen, so entspricht die letzte angezeigte Stelle der Messunsicherheit. Man spricht auch von der Typ B Unsicherheit des Messwerts.

**Beispiel 2.5 (Bleigehalt in Weinen)** In der Untersuchung misst man den Bleigehalt. Der Bleigehalt ist eine Zahl grösser Null mit Einheit  $\mu\text{g}/\text{kg}$ . Der Bleigehalt der 128 Weine wurde mit einer Präzision von  $\pm 0,1 \mu\text{g}/\text{kg}$  gemessen. Dies kann so geschrieben werden:  $\Delta\text{Pb} = \pm 0,1 \mu\text{g}/\text{kg}$ .  $\square$

**Beispiel 2.6 (Druck in einer Vakuumkammer)** Beim Beispiel der Vakuumkammer, vorgestellt in Beispiel 1.8, misst man den Unterdruck  $p$ . Der Unterdruck wurde mit einem Instrument mit einer Präzision von  $\pm 10^{-4}$  bar gemessen:  $\Delta p = \pm 10^{-4}$  bar. Die letzte Ziffer in den Messresultaten, wie in 0,59081 bar, ist eine Typ B Unsicherheit.  $\square$

---

## 2.4 Größen müssen messbar sein

Wichtig ist es, Größen in operationeller Art messen zu können. Spezifikationen zu Massengütern, die mit Stichproben überprüft werden, sollten nachvollziehbar sein. Spezifikationen wie „rund“, „schief“, „gut“, „sicher“ oder „nachhaltig“ haben keine messbare Bedeutung, solange sie nicht klar definiert werden. So sind Fahrradfelgen nicht nach dem Kriterium „rund“ messbar: niemand kann alle Punkte auf einer Felge messen, um zu testen, ob alle Punkte den gleichen Abstand zur Achse haben. Sammelt man Daten zu nicht festgelegten Größen, erzeugt man nur Datenmüll.

Messbare Größen entstehen durch *operationelle Angaben* von physikalischen, chemischen oder technischen Messungen oder von Entscheidungen, die mit Ja oder Nein beantwortet werden können.

Physikalische, chemische oder technische Größen sind etwa Volumen, Längen, Gewichte, Konzentrationen, Lichtstärken, Festigkeiten, Drücke, Lebensdauern oder Geschwindigkeiten. Entscheidungen, die mit Ja oder Nein beantwortet werden können, sind etwa Auswirkungen von Medikamenten (Herzinfarkt / kein Herzinfarkt). Der amerikanische Statistiker und Qualitätsexperte W. E. Deming illustriert in [2] an einer Stoffetikette mit der Angabe „50 % Wolle“, wie eine Größe schlecht definiert sein kann:

The label on a blanket reads ‚50 per cent wool‘. What does it mean? You probably don't care much what it means. You are more interested in color, texture, and price than in the content. However, some people do care what the label means. The Federal Trade Commission does, but with what operation meaning?

Suppose that you tell me that you wish to purchase a blanket that is 50 per cent wool, and that I sell to you the blanket shown in the figure, one half being all wool and the other half all cotton:

All wool	All cotton
----------	------------

This blanket ist 50 per cent wool, by one definition. But you may, for your purpose, prefer another definition: you may say that 50 per cent wool means something different to you. If so, then what? You may say that you meant for the wool to be dispersed throughout the blanket. You could come through with an operational definition like this:

Cut 10 holes in the blanket, 1 or 1.5 cm in diameter, centered at random numbers. Number the holes 1 to 10. Hand these 10 pieces to your chemist for a test. He will follow prescribed rules. Ask him to record  $x_i$ , the proportion wool by weight for hole  $i$ . Compute  $\bar{x}$ , the average of 10 proportions. Criteria:

$$\bar{x} \geq 0.50 \text{ and } x_{\max} - x_{\min} \leq 0.02$$

If the sample fails on either criterion, the blanket fails to meet your specification.

Das Beispiel zeigt auch, dass Größen oft in Zusammenarbeit von verschiedenen involvierten Personen festgelegt werden. Bei der Stoffetikette sind dies der Produzent und der Chemiker. Auch müssen sich Personen und Amtsstellen, Produzenten (Lieferanten) und Konsumenten (Abnehmer), ein medizinisches Entwicklungslabor und seine Kontrollstelle, Lebensmittelproduzenten und Kontrollstellen, sowie eine Qualitätskontrollstelle und eine Produktionseinheit auf die genaue operationelle Definition der messbaren Größen einigen. Dies muss manchmal in mehreren, zeitaufwändigen Diskussionsrunden mit den beteiligten Personen festgelegt werden.

**Beispiel 2.7 (Lohnstrukturerhebung)** Das Bundesamt für Statistik der Schweiz erhebt die Lohnstruktur bei privaten und öffentlichen Unternehmen, sowie bei Verwaltungseinheiten von Bund, Kantonen und Gemeinden. Sie beruht jeweils auf den Löhnen des Monats Oktober. Gemessen wird dabei ein standardisierter Bruttomonatslohn, der (operational) definiert ist:

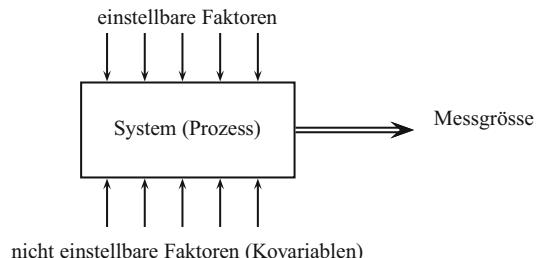
- (1) Um den Vergleich zwischen Vollzeit- und Teilzeitbeschäftigten zu ermöglichen, werden die erhobenen Beträge auf standardisierte Bruttomonatslöhne umgerechnet, d. h. auf eine einheitliche Arbeitszeit von 4,33 Wochen zu 40 Stunden.
- (2) Die berücksichtigten Lohnkomponenten sind: Bruttolohn im Monat Oktober (inkl. Arbeitnehmerbeiträge an die Sozialversicherung, Naturalleistungen, regelmässig ausbezahlte Prämien-, Umsatz- oder Provisionsanteile), Entschädigung für Schicht-, Nacht- und Sonntagsarbeit, 1/12 des 13. Monatslohns und 1/12 der jährlichen Sonderzahlungen.
- (3) Nicht berücksichtigt werden die Familien- und Kinderzulagen. □

## 2.5 Faktoren und Niveaus bestimmen

Grössen können in komplexer Weise von mehreren *Faktoren* (engl. *factors*), *Merkmalen* (engl. *features*) oder *Attributen* (engl. *attributes*) abhängen. Diejenigen Faktoren, die einen grossen Einfluss auf eine Grösse haben, sollten gemessen werden. Andere Faktoren haben kaum Wirkungen auf diese und sind vielleicht vernachlässigbar. Bei Experimenten werden wichtige Faktoren auf kontrollierte Werte eingestellt. Faktoren, die während Versuchen nicht kontrolliert eingestellt sind, bezeichnet man auch als *Kovariablen* (engl. *covariate* oder *covariable*). Die Nomenklatur illustriert Abb. 2.2. Um aussagekräftige und vertrauenswürdige Resultate zu erhalten, die für andere Personen nachvollziehbar sind, ist es wünschenswert, sich ein gutes Bild der Faktoren zu erarbeiten und die Werte der Faktoren zu kontrollieren.

**Beispiel 2.8 (Mondholz)** Verschiedene Personengruppen denken, dass Holz, das während Vollmondphasen geschlagen wird, sehr gute Holzqualitäten besitzt. Eine Zahl, die

**Abb. 2.2** Messgrösse, abhängend von Kovariablen und einstellbaren Faktoren



die Holzqualität quantifiziert, ist die Schrumpfung des Holzes durch Austrocknen. Man nennt sie das Schwindmass.

Eine Versuchsanordnung schlägt vor, während eines Jahres einmal pro Woche in einem Wald Holz zu schlagen und das Schwindmass zu messen. Daten zu nicht kontrollierbaren Faktoren (Kovariablen), wie Wachstumsphase, Temperatur, Feuchtigkeit und Krankheitsbefall, die einen grossen Einfluss auf die Holzqualität haben könnten, werden jedoch nicht gesammelt. Da viele bedeutende Kovariablen vorhanden sind, werden die Messwerte stark streuen. Schwerwiegend ist jedoch, dass bei den Vollmondphasen zufällig auch eine hohe Luftfeuchtigkeit vorhanden sein könnte. Da Daten dazu fehlen, ist der Einfluss der Kovariablen auf das Schwindmass nicht bestimmbar. Dies kann zu einer unvorsehbaren *Verzerrung* (engl. *Bias*) oder zu einem *systematischen Fehler* des Resultats führen. Die Untersuchung wird damit problematisch. □

Faktoren beschreiben *qualitative* Merkmale, die auf eine Grösse wirken. Einkommen von Personen können von Faktoren wie Alter und Ausbildungsniveau abhängen. Eine chemische Ausschüttung kann von Faktoren wie Temperatur und Konzentration eines Katalysators abhängen. Beim Faktor Ausbildungsniveau könnten sechs Werte auftreten: obligatorische Schule, Lehre, Matur, Bachelor, Master oder Doktor. Der Faktor Temperatur kann kontinuierliche Zahlenwerte belegen, beispielsweise Temperaturschritte von 25 °C zwischen 25°C und 100°C. Man bezeichnet die festgelegten Werte von kontrollierten Faktoren als *Niveaus* (engl. *levels*). Bei der Temperatur sind dies die Niveaus 25–50°C, 51–75°C und 76–100°C.

**Beispiel 2.9 (Chloridgehalt)** Es gibt mehrere Verfahren, den Chloridgehalt von Kalium-Chlorid-Lösungen zu bestimmen. Eine Methode bestimmt diesen mit einem chemisch-mechanischen Verfahren. Diese Methode ist abhängig von Faktoren wie Umgebung und ausführender Person. Die Umgebung hängt ab von Faktoren wie Temperatur der Lösung, Durchmischung, Umgebungsdruck und Sauberkeit der verwendeten Apparaturen. Niveaus der Temperatur könnten die vier Intervalle 6–10°C, 11–15°C, 16–20°C und 21–25°C sein. Hier die Resultate (in mol/L) der gemessenen Chloridgehalte eines Experiments:<sup>2</sup>

0,1028 0,1043 0,1023 0,1036 0,1033 0,1015 0,1021

Da die Durchmischung (eine Kovariable) nicht konstant gehalten werden kann und die verwendeten Messinstrumente Ungenauigkeiten haben, sind die Messresultate nicht immer gleich. Sie streuen um den gesuchten, nicht direkt messbaren Chloridgehalt. Drei Faktoren wurden hier zu einem festen Niveau „eingefroren“: Labor = Labor der Chemieabteilung der Berner Fachhochschule, Temperatur = 20 °C, Person = Person XX. Die Resultate aus dem Experiment gelten daher eigentlich nur für das Labor der Chemieabteilung der Berner Fachhochschule bei den oben beschriebenen Bedingungen. Um eine grössere Aussagekraft des Resultats zu erhalten, könnte es sinnvoll sein, andere Labors zu beteiligen. □

---

<sup>2</sup> Daten aus dem chemischen Labor der Berner Fachhochschule.

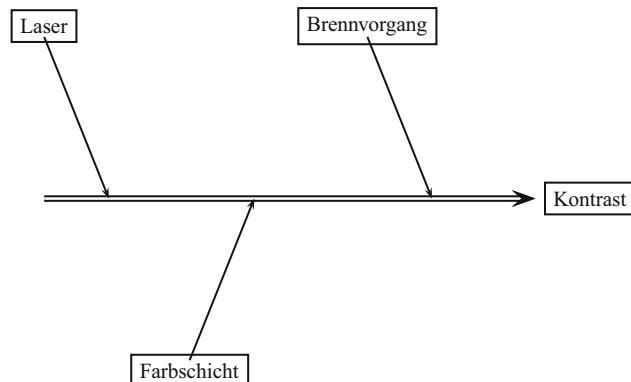
**Beispiel 2.10 (Mollusken)** Im Rahmen eines Biodiversitätsmonitorings wurde im Kanton Aargau zwischen 1996–2001 die Zahl der verschiedenen Molluskenarten erhoben. Artenzahlen können von verschiedenen Kovariablen wie Vegetationstyp, Klimazone, Höhe über Meer oder Beschaffenheit des Bodens abhängen. Frühere Untersuchungen ergaben, dass für die Topologie des Kantons Aargau nur der Faktor Vegetationstyp entscheidend auf die Artenzahlen der Mollusken wirkt. Die Molluskenarten wurden daher nur mit dem Faktor Vegetationstyp bestehend aus den Niveaus Wald, Grünland und Ackerland gemessen. □

Beliebt ist es, Faktoren und zugehörige Niveaus von Messgrößen mit *Ursache-Wirkungs-Diagrammen* (engl. *cause-and-effect-diagrams* [CE-diagram]) zu beschreiben. Diese sind bei Untersuchungen zur Qualität von Prozessen sehr verbreitet. Hier ein Beispiel dazu:

**Beispiel 2.11 (Lasermarkierung)** Wie im Kap. 1 in Beispiel 1.4 erklärt, können Packungen mit Lasern markiert werden. Dabei werden aufgetragene Farbschichten durch einen Laserstrahl entfernt. Wichtig ist es, eine möglichst kontrastreiche Markierung zu erhalten. Ein Ursache-Wirkungsdiagramm für die Zielgröße „Kontrast“ erstellt sich wie folgt: In einem ersten Schritt wird die Zielgröße in einem Kasten notiert, links davon wird ein langer breiter Pfeil gezeichnet:

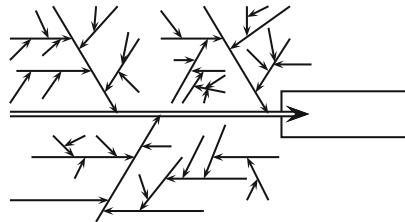


Im zweiten Schritt werden die *wenigen* Hauptfaktoren, welche auf die Messgrösse wirken, in Kästchen notiert. Sie bilden zusammen mit Pfeilen Äste des Ursache-Wirkungs-Diagramms. Dies zeigt Abb. 2.3. Zuletzt werden alle Faktoren, die auf die Hauptfaktoren wirken, entlang der Äste mit Pfeilen eingezeichnet. Die notierten Faktoren können

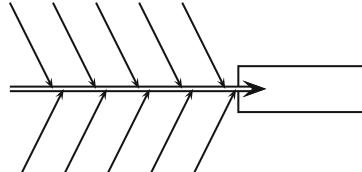


**Abb. 2.3** Die Hauptfaktoren: Laser, Farbschicht, sowie die Art des Brennvorgangs mit der der Laserstrahl die Farbschicht wegbrennt

**Abb. 2.4** Ein sorgfältig erstelltes CE-Diagramm



**Abb. 2.5** Ein CE-Diagramm aus oberflächlichem Wissen



weiter zerlegt werden, so dass komplizierte Verästelungen entstehen. Das Resultat ist das Ursache-Wirkungs-Diagramm, das in Abb. 1.2 des Kap. 1 dargestellt ist.

Das Ursache-Wirkungs-Diagramm erlaubt es, den Kontrast zu analysieren. Man kann die Faktoren bestimmen, die während der Datensammlung festgehalten werden und damit kontrolliert sind. Im Ursache-Wirkungs-Diagramm können auch die Kovariablen – die unkontrollierten Faktoren – markiert werden. Die variierenden Werte der Kovariablen, die den Kontrast wesentlich beeinflussen, müssen während des Versuchs aufgenommen werden. □

Größen hängen in der Regel in komplexer Art von mehreren Faktoren, Unterfaktoren und Kovariablen ab. Ursache-Wirkungs-Diagramme sind bei sorgfältiger und vertiefter Untersuchung daher eher kompliziert. Abb. 2.4 zeigt ein solches CE-Diagramm. Ein Ursache-Wirkungs-Diagramm wie in Abb. 2.5 weist meistens darauf hin, dass das Wissen über den zu untersuchenden Gegenstand oberflächlich ist.

Ursache-Wirkungs-Diagramme erlauben es, Experimente kontrolliert ablaufen zu lassen. Es ist aber nicht überraschend, dass aus Kosten- und Zeitgründen kaum alle Faktoren und zugehörigen Niveaus einer Grösse berücksichtigt werden. Man fixiert daher unbedeutende Faktoren meist auf einem Niveau. Lässt man zu viele Faktoren auf einem Niveau fixiert, ist die Experimentierumgebung zu klein und die Untersuchung beschränkt sich auf eine zu spezielle Situation. Man sagt, dass der *Wirkungsraum* (engl. *inference space*) der Untersuchung zu klein wird. Die Kunst, Experimente einzurichten, besteht darin, diejenigen wenigen Faktoren zu bestimmen und kontrolliert variieren zu lassen, die einen dominierenden Einfluss auf die untersuchte Grösse haben und die anderen Faktoren auf einem Niveau einzufrieren, sodass die Resultate trotzdem für einen grossen Wirkungsraum aussagekräftig sind.

**Tab. 2.2** Atmungsaktivität von Stoffen: Gramm Wasserdampf pro 24 Stunden und pro m<sup>2</sup> Stoff

Stoff	JIS L 1099	ASTM F 2298
eVent (Polyester)	27 825,6	6162,5
Gore-Tex XCR	21 193,6	3193,3
Sympatex	11 669,6	2960,1
Epic	6852,0	3238,5
Marmot Dry Touch	12 618,8	3769,5

**Beispiel 2.12 (Allwetterjacken)** Im Outdoor-Markt werden Allwetterjacken angeboten, die wasserdicht sind. Bei „hermetisch“ abschliessender Kleidung wird der anfallende Feuchtigkeits- und Hitzestau durch die körperliche Aktivität meist als sehr unangenehm empfunden. Daher werden Allwetterjacken aus atmungsaktiven Materialen wie PTFE-Laminaten (wie Gore-Tex oder eVent), Polyurethan-Laminaten, Propore oder Tyvek hergestellt. Es gibt verschiedene standardisierte Tests, um die Atmungsaktivität von solchen Materialien zu messen, wie die *JIS L 1099 Desiccant Inverted Cup*-Methode und die *ASTM F 2298 DMPC Diffusion Test*-Methode. Gemessen wird dabei, wie viel Gramm Wasserdampf pro 24 Stunden und pro m<sup>2</sup> Stoff durch das Material fliessen kann. Tab. 2.2 zeigt Resultate aus [6].

Es ist wichtig zu verstehen, dass Atmungsaktivitäts-Tests unter Laborverhältnissen an Stoffflächen stattfinden. Es ist daher schwierig, aus diesem kleinen Wirkungsraum auf die Atmungsfähigkeit von Allwetterjacken zu schliessen. Hier eine Bemerkung dazu von A. Dixon aus [3]:

According to two thought leaders in the field of fabric performance testing, Dr. Elizabeth McCullough (co-director for the Institute for Environmental Research, Kansas State University) and Dr. Phillip Gibson (US Army Soldier Systems Center), test results on fabrics are just that: tests on fabric. At best, these tests can only approximate the field performance garments. McCullough and Gibson both point out that more sophisticated tests that use moving and sweating mannequins clothed with realistic apparel systems and subjected to blowing wind are more accurate indicators of clothing performance, but still fall short of predicting actual field performance on human subjects. Mannequin tests are also expensive, difficult to execute, and provide data that can be interpreted with great latitude. Consequently, they are not likely to gain favor in breathability standardization anytime soon, and will remain primarily as a research tool.

Even if testing were performed on human subjects, many variables must be taken into account: individual metabolic variability, individual perspiration level, personal fitness, activity level, what garments are worn under the shell, shell venting characteristics (e.g. pit-zips), garment fit, whether or not the shell ‘pumps’ air (which is governed by fit, ventilation, and body motion), the type of activity performed, wind speed and direction, outside temperature, precipitation levels, etc. This list could go on! Clearly, there is no standard method or measure that can be used to predict the comfort level of rainwear garments on a human subject in realistic field conditions. There are simply too many variables. □

## 2.6 Paretodiagramme

Untersuchungen zu Größen, die von verschiedenen Faktoren mit mehreren Niveaus abhängen, können sehr anspruchsvoll und aufwändig sein. Schon bei einer Größe  $X$ , die nur von drei Faktoren  $A, B, C$  mit je zwei Niveaus beeinflusst ist

$$X = F(A, B, C), \quad \text{mit } A = \pm 1, B = \pm 1, C = \pm 1$$

sind schon, wie Tab. 2.3 zeigt, acht verschiedene Versuchsanordnungen möglich. Ein Versuchsplan, der alle möglichen Kombinationen von Niveaus berücksichtigt, nennt man einen *vollen faktoriellen Plan* (engl. *factorial design*). Allgemein müssen bei einer Größe mit  $m$  Faktoren zu je zwei Niveaus bei einem faktoriellen Plan mindestens

$$n = 2^m$$

Versuche durchgeführt werden. Der Aufwand wächst damit exponentiell mit der Anzahl Faktoren.<sup>3</sup>

Eine Strategie, die in der Praxis verbreitet ist, um die Auswirkung von Faktoren auf eine Größe zu bestimmen, besteht darin, sukzessiv die Wirkung einzelner Faktoren zu messen. Im obigen Beispiel der Größe, die von drei Faktoren  $A, B$  und  $C$  abhängt, würde man zuerst die Auswirkung des Faktors  $A$  erfassen, indem man Messungen mit variierenden Niveaus von  $A$ , aber festgefahrenen Niveaus von  $B$  und  $C$  (zum Beispiel  $B = C = 1$ ) durchführt:

$A$	$B$	$C$	Nummer Versuch	Größe $X$
-1	+1	+1	1	$x_1$
+1	+1	+1	2	$x_2$

Aus dem Vergleich der Werte  $x_1$  und  $x_2$  wird man das „beste“ Niveau  $A$  bestimmen. Dies könnte hier bei  $A = -1$  sein. In einem zweiten Schritt wird man die Auswirkung von  $B$  erfassen, indem man Messungen bei festgelegten Niveaus von  $A$  und  $C$  (Niveau von  $A$

**Tab. 2.3** Voller faktorieller Plan bei drei Faktoren mit je zwei Niveaus

Versuch	1	2	3	4	5	6	7	8
A	-1	+1	-1	+1	-1	+1	-1	+1
B	-1	-1	+1	+1	-1	-1	+1	+1
C	-1	-1	-1	-1	+1	+1	+1	+1

<sup>3</sup> Haben viele Faktoren entscheidenden Einfluss auf die Messgröße, erlauben Techniken aus der Versuchsplanung, wie *teilstatistische Pläne*, den Aufwand der Datensammlung effizient und möglichst klein zu halten. Informationen dazu finden sich in der Literaturliste am Ende des Kap. 3.

bei festgestelltem „bestem“ Niveau) durchführt:

<b>A</b>	<b>B</b>	<b>C</b>	Nummer Versuch	Grösse X
-1	<b>-1</b>	+1	3	$x_3$
-1	<b>+1</b>	+1	4	$x_4$

Aus dem Vergleich der Werte  $x_3$  und  $x_4$  wird man das „beste“ Niveau von  $B$  bestimmen (beispielsweise bei  $B = +1$ ). Schlussendlich wird man noch den Faktor  $C$  optimieren (Niveaus von  $A$  und  $B$  bei festgestellten „besten“ Niveaus):

<b>A</b>	<b>B</b>	<b>C</b>	Nummer Versuch	Grösse X
-1	+1	<b>-1</b>	5	$x_5$
-1	+1	<b>+1</b>	6	$x_6$

Diese *Ein-Faktor-pro-Versuch* (engl. *factor-at-a-time*)-Analyse ist jedoch kaum effizient, noch berücksichtigt sie mögliche *Interaktionen* (engl. *interaction*) zwischen den Faktoren. Interaktionen zwischen den Faktoren sind aber üblich. Viele Statistiker raten daher von der Ein-Faktor-pro-Versuch-Analyse ab.

Das folgende Beispiel zeigt einen faktoriellen Plan bei einer Grösse, die von drei Faktoren mit je zwei Niveaus abhängt. Dabei wird illustriert, wie mit Hilfe von *Pareto-diagrammen* grob beurteilt werden kann, welche Faktoren und welche Interaktionen einen grossen Einfluss auf die Grösse haben.

**Beispiel 2.13 (Kanalwärmetauscher)** Um aus Abwasser Wärme rückzugewinnen, werden Wärmetauschelemente in Abwasserkanälen eingebaut (siehe Abb. 2.6). Die Elemente müssen einerseits möglichst billig und andererseits so gebaut sein, dass ihre Verformungen nicht zu gross werden. Die Verformung hängt bei Wärmetauschern, die von einer Schweizer Firma produziert werden, von drei Faktoren ab: der Dicke  $S$  des Siggenblechs,

**Abb. 2.6** Abwasserrohr mit Wärmetauscher im Innern des Rohrs (Bachelorarbeit, Berner Fachhochschule, Burgdorf, 2009)



**Tab. 2.4** Resultat aus dem vollen faktoriellen Plan beim Kanalwärmetauscher

<i>S</i>	<i>D</i>	<i>R</i>	Verformung (in mm)			Arith. Mittel
-1	-1	-1	8,83	8,97	8,90	8,90
+1	-1	-1	3,17	2,88	2,86	2,97
-1	+1	-1	4,78	5,06	4,94	4,93
+1	+1	-1	1,74	1,75	1,64	1,71
-1	-1	+1	9,07	9,16	9,13	9,12
+1	-1	+1	2,63	2,74	2,48	2,62
-1	+1	+1	5,24	5,30	5,47	5,34
+1	+1	+1	1,46	1,53	1,44	1,48

der Dicke *D* des Deckblechs und der Breite *R* des Rinnendeckblechs. Die Faktoren wurden bezüglich je zwei Niveaus variiert. Tab. 2.4 zeigt das Resultat mit je drei Messungen zu einem vollen faktoriellen Plan – also zu allen  $2^3 = 8$  möglichen verschiedenen Experimentalbedingungen aus [11]. Dabei bedeutet:  $S = +1$ : 2 mm,  $S = -1$ : 1 mm;  $D = +1$ : 3 mm,  $D = -1$ : 2 mm;  $R = +1$ : 1200 mm,  $R = -1$ : 700 mm. Die Tabelle zeigt, dass die Messwerte trotz festgehaltenen Faktoren *S*, *D* und *R* streuen. Dies liegt daran, dass die Experimentierumgebung nicht konstant gehalten werden kann und keine identische Wärmetauschelemente produziert werden können. Daher wirken Kovariablen auf die Verformung. Diese Schwankungen der Messungen nennt man die *Typ A Unsicherheiten*.

Die Auswirkungen der verschiedenen Niveaus auf die Verformung wurden mit dem arithmetischen Mittel in der letzten Spalte der Tabelle zusammengefasst. Bei grosser Dicke des Siggenblechs ( $S = +1$ ) ergaben sich die arithmetischen Mittel

$$2,97 \quad 1,71 \quad 2,62 \quad 1,48$$

und bei kleiner Dicke ( $S = -1$ ) erhält man die arithmetischen Mittel

$$8,90 \quad 4,93 \quad 9,12 \quad 5,34$$

Die Auswirkung des Faktors Dicke des Siggenblechs auf die Verformung kann nun bestimmt werden: Ein Mass ist die Differenz zwischen der mittleren Auswirkung bei  $S = +1$  und der mittleren Auswirkung bei  $S = -1$ :

$$\begin{aligned} \text{Effekt}_S &= (\text{arith. Mittel bei } S = +1) - (\text{arith. Mittel bei } S = -1) \\ &= \text{arith. Mittel von } (2,97; 1,71; 2,62; 1,48) \\ &\quad - \text{arith. Mittel von } (8,90; 4,93; 9,12; 5,34) \\ &= \frac{2,97 + 1,71 + 2,62 + 1,48}{4} - \frac{8,90 + 4,93 + 9,12 + 5,34}{4} = -4,88 \end{aligned}$$

Analog ergeben sich für die Auswirkungen der Faktoren Deck- und Rinnenblech die Werte Effekt<sub>D</sub> = -2,54 und Effekt<sub>R</sub> = 0,01.

Neben den einzelnen Faktoren ist auch der Effekt der Interaktion von Faktoren zu berücksichtigen. Die Auswirkung des kombinierten Effekts von gleichgeschaltetem Siggen- und Deckblech ( $S = D = +1$  oder  $S = D = -1$ , d.h.  $S \cdot D = 1$ ) zu nicht-gleichgeschaltetem Siggen- und Deckblech ( $S = +1, D = -1$  oder  $S = -1, D = +1$ , d.h.  $S \cdot D = -1$ ) lässt sich messen: Man erhält bei  $S \cdot D = 1$  die arithmetischen Mittel (wiederum abgelesen aus der letzten Spalte der Tab. 2.4):

$$S \cdot D = 1 : \quad 8,90 \quad 1,71 \quad 9,12 \quad 1,48$$

und bei  $S \cdot D = -1$  die arithmetischen Mittel

$$S \cdot D = -1 : \quad 2,97 \quad 4,93 \quad 2,62 \quad 5,34$$

Die Auswirkung des Effekts des Zusammenwirkens der Faktoren Siggen- und Deckblech kann damit geschätzt werden:

$$\begin{aligned} \text{Effekt}_{S:D} &= (\text{arith. Mittel bei } S \cdot D = 1) - (\text{arith. Mittel bei } S \cdot D = -1) \\ &= \frac{8,90 + 1,71 + 9,12 + 1,48}{4} - \frac{2,97 + 4,93 + 2,62 + 5,34}{4} = 1,33 \end{aligned}$$

Auf analoge Art erhält man für die mittlere Auswirkung der Interaktionen Siggen- und Rinnenblech bzw. Deck- und Rinnenblech die folgenden Werte:

$$\text{Effekt}_{S:R} = -0,30, \quad \text{Effekt}_{D:R} = 0,08$$

Der Effekt der Interaktion aller drei Faktoren ist:

$$\begin{aligned} \text{Effekt}_{S:D:R} &= (\text{arith. Mittel bei } S \cdot D \cdot R = +1) \\ &\quad - (\text{arith. Mittel bei } S \cdot D \cdot R = -1) \\ &= \text{arith. Mittel von } (2,97; 4,93; 9,12; 1,48) \\ &\quad - \text{arith. Mittel von } (8,90; 1,71; 2,62; 5,34) = -0,02 \end{aligned}$$

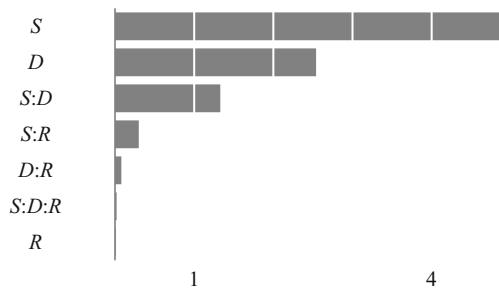
Mit einem Paretodiagramm, das in Abb. 2.7 visualisiert ist, lassen sich die berechneten Auswirkungen darstellen. Dazu werden die *Beträge* der erhaltenen Effekte, also

$$4,88 \quad 2,54 \quad 0,01 \quad 1,33 \quad 0,30 \quad 0,08 \quad 0,02$$

der Grösse nach geordnet und anschliessend mit einem Balkendiagramm dargestellt. Das *Paretodiagramm* zeigt den Effekt der Faktoren und ihrer Interaktionen auf die Messgrösse. Dabei werden die Vorzeichen der Effekte nicht dargestellt.  $\square$

Bei faktoriellen Plänen wächst der Aufwand eines Experiments exponentiell mit der Anzahl der Faktoren. Mit teilstatischen Plänen, wie sie in vielen Büchern zur Versuchsanalyse vorgestellt werden, lässt sich der Aufwand verkleinern. Es gibt auch Situationen,

**Abb. 2.7** Paretodiagramm:  
die Verformung wird vor allem  
von den Faktoren  $S$  und  $D$   
und ihrer Interaktion  $S : D$   
beeinflusst



wo man weiss, dass gewisse Interaktionen zwischen Faktoren kaum auf die untersuchte Grösse wirken. Im obigen Beispiel zum Kanalwärmetauscher hätte eine Experte vielleicht schon erahnt, dass der Faktor Rinnenblech ( $R$ ) alleine kaum auf die Verformung wirkt, sondern nur die Faktoren Dicke Siggenblech ( $S$ ) und Deckblech ( $D$ ), sowie die Interaktionen  $S : D$  und  $S : R$ . Dies erlaubt es, den faktoriellen Plan kleiner zu halten.<sup>4</sup>

## 2.7 Systematische Fehler

Wenn man Grössen messen will, sind methodische oder *systematische Messfehler* zu minimieren. Systematische Messfehler entstehen beispielsweise, wenn Daten falsch abgelesen werden. Gerade Ablesefehler bei Gleitkommadarstellungen, wie sie bei Computerausgaben Standard sind, sind üblich. So kann ein Computerausdruck der Form  $-1.33875e-3$  fälschlicherweise als  $-1,33875$  oder  $1,33875 \cdot 10^{-3}$  gelesen werden. Systematische Fehler treten zum Beispiel auf, wenn in Fragebögen Fragen falsch oder nicht beantwortet werden. So wurde in einer Untersuchung 500 Lesern der Zeitschrift „Folio“ Fragen gestellt: 27 % der Teilnehmer behaupteten, dass sie die Rubrik „Guter Rat“ regelmässig lesen. Diese Rubrik gab es aber nicht.<sup>5</sup> Vorsicht ist daher ein wichtiges Instrument, wenn man sich mit Umfragen beschäftigt.

Hat man mit Menschen zu tun, können systematische Fehler durch die Interaktion der Personen, die die Untersuchung durchführen, mit den Probanden entstehen. In medizinischen Studien entstehen systematische Fehler, wenn der Effekt von Scheinmedikamenten (der *Placebo-Effekt*) nicht berücksichtigt wird:

**Beispiel 2.14 (Magenschmerzen)** Im Journal of the American Medical Association (siehe [7]) zeigt eine Untersuchung, dass das Abkühlen des Magens während einer Stunde mit einer speziellen Flüssigkeit Magenschmerzen reduziert. Es handelt sich hier um eine Studie, die eine Beziehung zwischen dem Abkühlen des Magens und der Schmerzminderung

<sup>4</sup> Für weitergehende Betrachtungen zu faktoriellen Plänen und zu Auswirkungen von Faktoren und Interaktionen vgl. die Literaturangaben in Kap. 3 in Abschn. 3.7.

<sup>5</sup> NZZ Folio, 01/2006.

nachweisen will. Die Versuchsanordnung bestand einfach darin, bei einer Gruppe von Probanden den Magen mit der Flüssigkeit abzukühlen und den Effekt auf die Magenschmerzen zu messen. Die Studie ist jedoch wertlos, da viele Patienten auch auf eine beliebige Therapie oder Scheinmedikamentation reagieren. Die beobachtete Schmerzminderung könnte auch aus folgenden Gründen erfolgen: wegen des Vertrauens der Probanden in die neue Therapie oder in die behandelnden Ärzte, oder auch einfach, weil Magenschmerzen auch ohne medizinische Einflussnahme kleiner werden können. □

Damit der Einfluss der untersuchenden Personen bei Probanden nicht zu verfälschten und bedeutungslosen Resultaten führt, sollten Versuche mit Kontrollgruppen gemacht werden. So schildert der Statistiker E. Tufte (siehe [10] und [12]), dass ein bedeutender Chirurg in einem Vortrag beschrieb, wie er bei vielen Patienten Gefässoperationen mit einer neuen Methode erfolgreich durchgeführt habe. Ein Student stellte die Frage nach der Kontrolle der Versuche. Hätte er nur die Hälfte der Patienten operiert, dann wäre dies der sichere Tod für die Hälfte der Patienten gewesen, so die Antwort des Chirurgen. Die schüchterne Frage des Studenten darauf: „Welche Hälfte?“

**Beispiel 2.15 (Aspirin)** Bei Patienten mit kardiovaskulären Risiken, wie Herzinfarkten, wird vermutet, dass, wegen des biologischen Mechanismus von Aspirin, die tägliche Einnahme von tiefdosierten Aspirinmengen die Gefahr von nicht-tödlichen Herzinfarkten vermindert. Wie stark die Herzinfarkte durch die Aspirineinnahme reduziert werden, kann mit dem Verhältnis<sup>6</sup>

$$V_{\text{Herz}} = \frac{\text{Rate von Herzinfarkten bei Aspirineinnahme}}{\text{Rate von Herzinfarkten ohne Aspirineinnahme}}$$

gemessen werden. Ist  $V_{\text{Herz}} < 1$ , so reduziert die tägliche Einnahme von Aspirin die Möglichkeit eines Infarkts. Das Verhältnis  $V_{\text{Herz}}$  bezieht sich auf die Zielgruppe: alle Patienten mit kardiovaskulären Risiken, eine unendlich grosse Gruppe von Personen! Um  $V_{\text{Herz}}$  zu berechnen, wurden in mehreren Studien, der Second International Study of Infarct Survival (ISIS-2, siehe [4]), mehr als 17 000 Patienten untersucht. Dazu wurden zwei Gruppen gebildet. Einer Gruppe wurde Aspirin verabreicht, der anderen (der Kontrollgruppe) ein Placebo, d. h. eine Pille, die von den Probanden nicht von Aspirin zu unterscheiden war. Die Placeboeinnahme garantiert die „Blindheit“ der Probanden (einfache blinde Studie), reduziert den Placebo-Effekt und verhindert systematische Fehler. Die betreuenden Ärzte wussten ebenfalls nicht, in welcher Gruppe sich eine Versuchsperson befand (doppelt blinde Studie). Dies reduziert den Effekt der Messenden auf die Probanden. Die Einteilung der Patienten in zwei Gruppen erfolgte in zufälliger Weise, um die Gefahr zu minimieren, dass systematisch stark herzinfarktgefährdete Personen nur in einer Gruppe lagen. □

---

<sup>6</sup>Eine Rate ist meist ein durch eine Prozentzahl ausgedrücktes Verhältnis zwischen zwei statistischen Größen, das die Häufigkeit eines bestimmten Geschehens angibt. (DUDEN, Deutsches Universalwörterbuch 1989).

In Untersuchungen mit Lebewesen sind meist nur Experimente mit zufällig ausgewählten *Kontrollgruppen*, die *doppelt blind* verglichen werden, sinnvoll. Man spricht von einer *doppelt-blinden* und *zufalls-kontrollierten* Studie.

Auch bei technischen Untersuchungen können Messinstrumente auf Messgrößen wirken. So müssen Apparate, die Spannungen oder Ströme in elektrischen Kreisen messen, technisch aufwändig konstruiert werden, damit keine systematisch falsche Messwerte entstehen. Systematische Fehler treten auch auf, wenn bei Messungen mit nicht geeichten oder falsch kalibrierten Messgeräten gearbeitet wird. Das Kalibrieren von Messgeräten, um systematische Fehler zu vermeiden, ist sehr wichtig. Es verlangt hohes Können, Geschick oder einen grossen Zeitaufwand und darf keinesfalls unterschätzt werden.

**Beispiel 2.16 (Mollusken)** Im Rahmen eines Biodiversitätsmonitorings (siehe Beispiel 2.10) wurden auf verschiedenen Flächen die Zahl der Molluskenarten gezählt. Es gibt allerdings verschiedene Gründe, weshalb die Zählung der Arten mit zunehmender Dauer des Monitorings systematische Fehler hat: Landwirte, die die untersuchten Flächen kennen, lassen sich eventuell bezüglich der Landnutzung beeinflussen. Feldleute, die die Mollusken zählen, sind derart mit den Messflächen vertraut, dass sie die Molluskenzahl nicht mehr unbefangen erheben. Um diesen Effekten zu begegnen, kann man die bestehenden Messflächen allmählich aufgeben und durch neue, „unverbrauchte“ Messflächen ersetzen. Eine zweite Möglichkeit ist, parallel zu den untersuchten Flächen eine zweite, kleinere und geheime Stichprobe zu untersuchen, die zur Kontrolle der eigentlichen Stichprobe dient. □

Bei Umfragen können systematische Fehler, wie oben schon erwähnt, auch entstehen, wenn Fragebögen nicht vollständig beantwortet werden. Hier ist es wichtig, dass nach der Wahl der Stichprobe Massnahmen getroffen werden, dass alle Leute der Stichprobe ihre Fragebogen vollständig beantworten. In der Regel erhält man die beste Antwortrate durch persönliche Interviews, die zweitbeste Antwortrate durch Telefoninterviews und die dritt-besten Rate durch per Post versandte Fragebögen. Ein Plan muss ebenfalls ausgearbeitet werden, um Personen der Stichprobe, die nicht antworten, mehrmals zu kontaktieren. Nur eine Nichtantwortrate von vielleicht 5 oder 10 % garantiert, dass die gesammelten Daten nutzbar sind und systematische Fehler nicht zu gross werden.

Systematische Fehler und Streuungen von Messwerten hängen vom Aufbau der Experimente, Umfragen und Untersuchungen ab. Ohne kontrollierten Ablauf von Experimenten, ohne Bestimmung von Faktoren oder Kontrollgruppen, erhält man grosse Unsicherheiten und oft verzerrte Resultate. Man sagt: Eine Untersuchung

heisst *verzerrt* (engl. *biased*), wenn sie systematisch bestimmte Messwerte favorisiert.

Die oben erwähnte medizinische Studie zur Linderung von Magenschmerzen ist verzerrt, da sie den Placebo-Effekt nicht berücksichtigt.

---

## Reflexion

**2.1** Ein Betreiber einer Eisenbahnlinie vom Bahnhof *A* nach Bahnhof *B* erklärt, dass seine Züge die Strecke mit einer Geschwindigkeit von 110 km/h befahren, dass die Reisezeit 3 Stunden und 20 Minuten beträgt, und dass 95 % der Züge pünktlich fahren. Erklären Sie, warum diese Angaben so eher sinnlos sind.

**2.2** Auszug aus der Verordnung über technische Anforderungen an Strassenfahrzeuge in der Schweiz:

Fahrräder müssen mit zwei kräftigen Bremsen versehen sein, von denen die eine auf das Vorder- und die andere auf das Hinterrad wirkt.

Warum nützt diese Anforderung einem Hersteller von Fahrrädern wenig?

**2.3** Der Wetterdienst meldet, dass es morgen an Ihrem Wohnort regnen wird. Wie könnte man diese Aussage operationell messbar überprüfen?

**2.4** Sie wollen mit einem Experiment herausfinden, welche Faktoren die Qualität von Kartoffelstock beeinflussen. Legen Sie dazu eine Messgrösse fest, die die Qualität des Kartoffelstocks misst. Beschreiben Sie, wie Sie diese Grösse operationell messen. Welches sind die Hauptfaktoren, die auf die Messgrösse wirken? Stellen Sie ein detailliertes Ursache-Wirkungs-Diagramm für die gewählte Messgrösse auf.

**2.5** Sie wollen mit einem Experiment herausfinden, welche Faktoren den c-Ton bei einer Frequenz von 442 Hz bei Altblockflöten bestimmen. Welches sind die Hauptfaktoren, die auf die Messgrösse, den c-Ton, wirken? Stellen Sie ein Ursache-Wirkungs-Diagramm für die gewählte Messgrösse auf.

**2.6** Ingenieure möchten herausfinden, wie gross der Bremsweg des Fahrzeugtyps XY ist. Wie kann man operationell den Bremsweg messen? Welches sind die Hauptfaktoren, die auf den Bremsweg wirken? Stellen Sie dazu ein Ursache-Wirkungs-Diagramm auf.

**Tab. 2.5** Resultate aus einem vollen faktoriellen Plan

Testrun	$Z$	$T$	Dicke Veredlung				
2	-1	-1	116,1	116,9	112,6	118,7	114,9
1	+1	-1	116,5	115,5	119,2	114,7	118,3
3	-1	+1	106,7	107,5	105,9	107,1	106,5
4	+1	+1	123,2	125,1	124,5	124,0	124,7

**2.7** Eine Messgrösse hänge von vier Faktoren mit je zwei Niveaus ab. Stellen Sie einen vollen faktoriellen Plan auf. Legen Sie anschliessend durch Würfeln die Reihenfolge der einzelnen Versuche fest.

**2.8** Eine Firma veredelt Platten mit Nickel. Die Dicke der Veredlungsschicht hängt von den zwei Faktoren Veredlungszeit  $Z$  und der Temperatur  $T$  ab. Um den Effekt dieser Faktoren auf die Dicke der Veredlungsschicht zu studieren, wurden die Faktoren auf zwei Niveaus + und - gesetzt (Dabei bedeuten:  $Z = +1$ : 12 s,  $Z = -1$ : 4 s;  $T = +1$ : 32 °C,  $T = -1$ : 16 °C.) Tab. 2.5 zeigt die Resultate eines Versuchs mit je fünf Experimenten zu den vier möglichen verschiedenen Versuchsanordnungen.

- (a) Illustrieren Sie mit einem Paretodiagramm, wie die Faktoren und die Interaktion der Faktoren auf die Dicke der Veredlung wirken.
- (b) Die Effekte der Faktoren und ihrer Interaktionen lassen sich mit Statistikprogrammen schnell berechnen. Tun Sie dies mit Ihrem favorisierten Statistikprogramm!

**2.9** Die Lebensdauer eines Schneidewerkzeugs soll in Abhängigkeit der Krümmung der Klinge ( $K$ ), der Länge der Klinge ( $L$ ) und des Schneidewinkels ( $\alpha$ ) untersucht werden. Bei diesen drei Faktoren wurden zwei Niveaus festgelegt (hier mit + und - bezeichnet) und je zehn Messungen durchgeführt. Tab. 2.6 zeigt die Resultate. Stellen Sie ein Pareto-diagramm auf, um die Faktoren und Interaktionen zu bestimmen, die die grösste Wirkung auf die Lebensdauer des Schneidewerkzeugs haben.

**Tab. 2.6** Durchschnittliche Lebensdauern in Funktion der Krümmung  $K$ , der Länge der Klinge  $L$  und des Schneidewinkels  $\alpha$ 

Testrun	$K$	$L$	$\alpha$	arith. Mittel Lebensdauer (in h)
3	-1	-1	-1	46,30
5	+1	-1	-1	19,20
8	-1	+1	-1	13,81
1	+1	+1	-1	9,45
7	-1	-1	+1	11,25
2	+1	-1	+1	4,68
6	-1	+1	+1	57,29
4	+1	+1	+1	51,33

**2.10** Ein Tennisverein besteht aus zehn Mitgliedern. Die Mitglieder spielen vor allem bei Sonnenschein in T-Shirts Tennis. Drei Mitglieder des Vereins entwickelten im letzten Jahr an den Armen Hautkrebs. Erklären Sie, warum diese Beobachtung kein guter Nachweis ist, dass Sonnenstrahlen auf nackter Haut Krebs verursachen kann.

**2.11** Die aus vielen Beobachtungen zusammengesetzte Beurteilung [9] von Schweizer Lehrerinnen und Lehrern besagt, dass die schulsprachlichen Leistungen in Lesen und Schreiben bei vielen mehrsprachigen Kindern in den ersten vier Schuljahren stark ansteigen, dann bis zur achten Klasse auf dem Niveau von 4. bis 5. Klässlern stagnieren. Kann man aus dieser Beobachtung schliessen, dass die Lernmethoden ab der 5. Klasse in den Schulen verbessert werden sollten?

**2.12** Patienten für eine Operation zu anästhesieren, ist nicht problemlos. Es besteht immer ein Risiko, dass Patienten wegen der Anästhesie sterben. In der Studie [8] aus den USA wurde versucht, eine Beziehung zwischen dem gewähltem Anästhesieprodukt und der Todesrate herauszufinden. Hier die dabei beobachteten Todesraten aus 34 Spitäler mit total 850 000 Operationen von vier Anästhesieprodukten *A*, *B*, *C* und *D*:

Produkt	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>
Todesrate	1,7 %	1,7 %	3,4 %	1,9 %

Man sieht, dass das Produkt *C* gefährlich scheint. Eine präzisere Untersuchung stellte aber fest, dass das Produkt *C* keine höhere Todesrate besitzt als die anderen Produkte. Erklären Sie, warum die obige Tabelle irreführend ist.

**2.13** In der Zeitung „Sonntag“ vom 8.2.2009 wurde mit der Schlagzeile „*Die Spital-Liste: Wo es am meisten Tote gibt*“ die in Tab. 2.7 dargestellte Liste zu Sterberaten in Schweizer Spitäler veröffentlicht. Erklären Sie, warum das Bundesamt für Statistik darauf hinweist, dass es nicht ratsam ist, die dargestellten Zahlen zu vergleichen.

**Tab. 2.7** Sterberaten an den Zentrumsspitäler (Universitätsspitäler und grosse Kantonsspitäler mit 9000 bis 30 000 Behandlungsfällen pro Jahr), 2006

Genf	1,3%
Wallis	1,7%
Bern	1,8%
St. Gallen, Basel-Stadt	2,0%
DURCHSCHNITT	2,0%
Zürich, Graubünden, Aargau	2,1%
Waadt, Tessin, Basel-Landschaft	2,2%
Thurgau	2,3%
Neuenburg, Luzern	2,5%
Solothurn	2,9%
Freiburg	3,2%

## Literatur

1. D. Andrey, H. Beuggert, M. Ceschi, C. Corvi, M. De Rossa, A. Herrmann, B. Klein, N. Probst-Hensch, Monitoring-Programm „Schwermetalle in Lebensmitteln“. Mitt. Gebiete Lebensm. Hyg. **83**, 711–736 (1992)
2. W. E. Deming, *Out of the Crisis* (Massachusetts Institute of Technology, 1986)
3. A. Dixon, Waterproof Breathable Fabric Technologies, A Comprehensive Primer and State of the Market Technology Review. Backpackinglight.com, ISSN 1537-0364 (2004)
4. C. H. Hennekens, Aspirin in Chronic Cardiovascular Disease and Acute Myocardial Infarction. Clin. Cardiol. **13**, V-62–66 (1990)
5. K. Ishikawa, *Guide to Quality Control* (Asian Productivity Organization, 1982)
6. E. A. McCullough, M. Kwon and H. Shim, A comparison of standard methods for measuring water vapour permeability of fabrics. Institute Of Physics Publishing, Meas. Sci. Technol. **14**, 1402–1408 (2003)
7. D. S. Moore, G. P. McCabe, *Introduction to the Practice of Statistics* (W. H. Freeman and Company, New York, 2002)
8. L. E. Moses, F. Mosteller, Safety on Anesthetics. in *Statistics: A Guide to the Unknown*, 3rd edn., ed. by J. M. Tanur et al. (Wadsworth, 1989)
9. R. Müller, N. Dittmann-Domenichini, Die Entwicklung schulisch-standardsprachlicher Kompetenzen in der Volksschule. Eine Quasi-Längsschnittstudie. Linguistik online 32, 3/07, ISSN 1615-3014 (2007)
10. E. E. Peacock, Jr., University of Arizona College of Medicine, in Medical World News, September 1, S. 45 (1972)
11. M. Reinhard, Herstellungskostenoptimierung von Kanalwärmetauschern. Bachelorarbeit, Maschinentechnik, Berner Fachhochschule, Burgdorf (2009)
12. E. Tufte, *Data Analysis for Politics and Policy* (Englewood Cliff, New Jersey, Prentice-Hall 1974)

*Der Brachvogel legte zuerst die Rennbahn fest, eine Art Kreis („auf die genaue Form kommt es nicht an“, sagte er), und die Mitspieler mussten sich irgendwo auf der Bahn aufstellen, wie es sich gerade traf. Es gab kein „Eins – zwei – drei – los!“, sondern jeder begann zu laufen, wann er wollte, und hörte auf, wie es ihm einfiel, so dass gar nicht so leicht zu entscheiden war, wann der Wettkampf eigentlich zu Ende war.*  
*Lewis Carroll, Alice im Wunderland (Insel Taschenbuch, 1973, S. 30)*

## Zusammenfassung

Im vorigen Kapitel wird erklärt, wie man Versuche planen kann. Ursache-Wirkungs-Diagramme helfen zu überlegen, welche Faktoren auf eine Grösse wirken. Dies erlaubt es, Versuche kontrolliert ablaufen zu lassen und dabei Werte von wichtigen Kovariablen zu sammeln. Geeichte Instrumente und Kontrollgruppen bei Versuchen, bei denen Menschen beteiligt sind, helfen weiter systematische Fehler zu minimieren. Nach solchen Vorbereitungen ist es möglich gezielt, Daten zu sammeln. Wie soll man aber Probanden oder bei Qualitätskontrollen Objekte auswählen? Weiter ist es wichtig, möglichst effizient und kostengünstig Daten zu sammeln. Dazu gehört, dass nicht unnötig viele Messwerte aufgenommen werden. Wie geht man vor, wenn man Größen misst? Wie stellt man fest, ob die erhaltenen Messwerte wirklich unter statistischer Kontrolle waren? *Randomisierung* (engl. *randomization*), *Wiederholung* (engl. *replication*) und *Kontrolle* (engl. *controlling*) sind Methoden, die eine Antwort auf diese Fragen geben. (Der Genetiker und Statistiker R. A. Fisher ist der Urheber der wissenschaftlichen Versuchsplanung (engl. Design of Experiments). Zur Versuchsplanung gehören Prinzipien wie Wiederholung, Randomisierung und Kontrolle.)

### 3.1 Randomisierung: die Stichprobenwahl

Stichproben aus Populationen oder Produktionsserien mit Zufallsprozessen oder, wie man sagt, mit Randomisierung auszuwählen, stellt einen zentralen Bestandteil der statistischen Versuchsplanung dar. Randomisierung bedeutet, dass sowohl die Auswahl der Stichprobe als auch die Reihenfolge von Versuchen durch Zufall bestimmt wird. Hier Beispiele dazu:

**Beispiel 3.1 (Plattenveredlung)** Eine Firma produziert metallene Platten, die mit Nickel veredelt werden. Die Vernickelung kann mit zwei verschiedenen Verfahren  $X$  und  $Y$  erfolgen. Erwünscht ist dabei eine möglichst dicke durchschnittliche Nickelschicht. Mittels eines Experiments mit 20 Platten soll bestimmt werden, welches Verfahren besser ist, indem 10 Platten mit dem einen und 10 Platten mit dem anderen Verfahren veredelt werden.

Die Platteneigenschaften sind wegen schwankender Produktionsbedingungen Streuungen unterworfen. Systematische Fehler können beim Experiment entstehen, wenn von den 20 Platten die besseren zehn alle mit dem Verfahren  $X$  und die schlechteren zehn mit dem Verfahren  $Y$  getestet werden. Die Ingenieurin wird, um diese Gefahr zu minimieren, die 20 Platten des Experiments mit einem *Zufallsprozess* auswählen.  $\square$

**Beispiel 3.2 (Aspirin)** Um zu messen, ob die tägliche Einnahme von Aspirin das Herzinfarktrisiko reduziert, wurden Patienten mit kardiovaskulären Risiken in zufälliger Art in zwei Gruppen aufgeteilt. Dies geschah, um die Gefahr zu minimieren, dass sich stark herzinfarktgefährdete Personen systematisch nur in einer Gruppe befanden.  $\square$

Einfachere statistische Verfahren verlangen, dass Messungen sich gegenseitig nicht „beeinflussen“ oder austauschbar sind. Randomisierung kann hier helfen. Wenn man eine Stichprobe dadurch bestimmt, dass mit einem Zufallsmechanismus Elemente aus der Grundgesamtheit gezogen werden, spricht man von *zufälligem Ziehen* oder *Randomisieren* (engl. *random sampling*). Beim Randomisieren ist die Wahrscheinlichkeit bekannt, dass ein Element in der Stichprobe liegt. Ist dies nicht der Fall, können Untersuchungen zu Grundgesamtheiten verzerrt sein und bedeutungslos werden.

**Beispiel 3.3 (Internet-Umfrage)** In der folgenden Meldung aus der Gratiszeitung „20-Minuten“ vom 18. Oktober 2004 ist diese Wahrscheinlichkeit nicht bekannt:

*Unfall beim Freizeitsport:* Die Zahl der Freizeitunfälle hat zugenommen. Wir wollten darum wissen, ob unsere Leser sich schon einmal beim Freizeitsport verletzt haben. 1031 Teilnehmer (52 %) der 20-Minuten-Internetumfrage beantworteten die Frage mit Nein. Teilnehmer total: 2003.

Obwohl die Datenmenge gross ist, dürften die Daten von geringer Qualität sein. Die Wahrscheinlichkeit, dass eine Person aus der Gesamtbevölkerung an der Umfrage teilnimmt, ist für die Journalistinnen oder Journalisten nicht kontrollierbar. Personen, die sich

angesprochen fühlen, melden sich eher als solche, die diese Umfrage nicht interessiert. Das Resultat dürfte verzerrt sein.  $\square$

Die Randomisierung, die über das blosse Auswählen einer Stichprobe geht, spielt im industriellen Umfeld eine Rolle. Gerade in Situationen, in denen eine Grösse nur grob gemessen werden kann (z.B. wenn Effekte in der Größenordnung der Streuung gemessen werden), wenn die Experimente fast abgeschlossen sind, wenn Angestellte ohne statistische Kenntnisse mit Experimenten skeptische Vorgesetzte überzeugen wollen oder wenn ernsthafte, vielleicht lebensgefährdende Verfahren untersucht werden, wird nur Randomisierung eine überzeugende statistische Arbeit zulassen. Randomisierung ist hier besonders wichtig, um zu verhindern, dass die gemessenen Effekte nicht nur durch eine willkürliche Experimentieranordnung entstanden sind (siehe dazu [3]).

Natürlich garantiert das zufällige Ziehen nicht immer, dass eine „gute“ Stichprobe entsteht:

**Beispiel 3.4 (Pech)** Jemand will das Durchschnittsgewicht einer Personengruppe mit 1000 Personen aus einer Stichprobe von 10 Personen berechnen. Es kann sein, dass durch zufälliges Ziehen die zehn Personen mit dem grössten Gewicht ausgewählt werden und damit das Durchschnittsgewicht stark überschätzt wird.<sup>1</sup>  $\square$

**Beispiel 3.5 („Repräsentative“ Stichprobe)** In Zeitungen und Zeitschriften finden sich Rechnungen zu Grössen, die auf „repräsentativen“ (sic!) Stichproben basieren. Die Statistiker W. A. Wallis und H. V. Roberts betonen aber in [13], dass es unmöglich ist, eine Stichprobe so auszuwählen, dass sie die Grundgesamtheit bezüglich eines zu untersuchenden – und damit nicht bekannten – Merkmals identisch oder repräsentativ abbildet. Wenn eine Stichprobe repräsentativ ist, bedeutet dies, dass man die Grundgesamtheit bezüglich des Merkmals kennt und damit eine Stichprobe eigentlich nicht notwendig ist. „Repräsentative“ Untersuchungen sind daher zu hinterfragen: Wie ist die Stichprobe gewählt worden? Das Schweizer Fernsehen publizierte beispielsweise eine Umfrage im Januar 2013 wie folgt:

Durchgeführt wurde die Umfrage im Auftrag der SRG SSR vom Forschungsinstitut gfs.bern. Befragt wurden 1217 Personen. Davon geben 39 Prozent an, bestimmt zur Urne gehen zu wollen. Weitere 31 Prozent wollen sich eher an der Abstimmung beteiligen. Die Stichprobe ist sprachregional gewichtet und *repräsentativ* für die Schweizer Stimmberchtigten.

Das Forschungsinstitut, das die Umfrage durchführte, nennt präzis, wie die Stichprobe randomisiert gebildet wurde: sprachregional geschichtete, zweistufige Zufallsauswahl (Haushalte, dann Bewohner) mit Adressen aus zusammengesetzten Telefonverzeichnissen, randomisiert nach der Geburtsmethode, geschichtet nach Sprachregionen, gewichtet

---

<sup>1</sup> Die Wahrscheinlichkeit für das Eintreffen ist aber sehr klein!

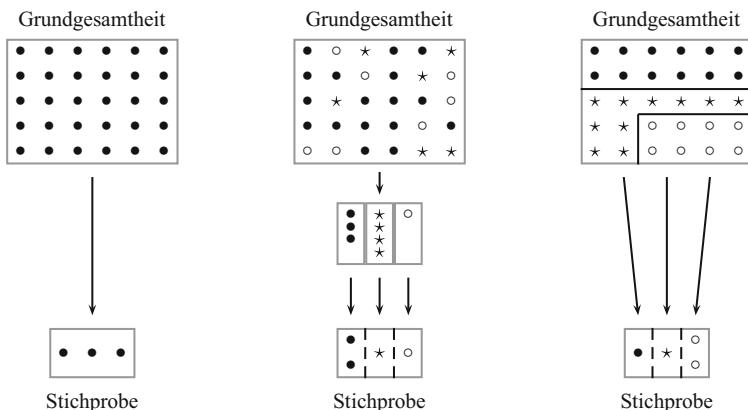
nach Sprache und Parteiaffinität, erhoben mit CATI (Computer Assisted Telephone Interview) während des Zeitraums vom 8. Februar 2013 bis 15. Februar 2013.  $\square$

Am besten verwendet man Computerprogramme, um Stichproben durch zufälliges Ziehen zu bestimmen. Dabei unterscheidet man zwei Fälle: *Ziehen ohne Zurücklegen* und *Ziehen mit Zurücklegen*. Besitzt die Grundgesamtheit beispielsweise 1000 Elemente, die von 1 bis 1000 durchnummieriert werden, so bedeutet zufälliges Ziehen ohne Zurücklegen, dass wie beim Lottospiel eine Zahl nach der anderen gezogen wird. Damit können keine Zahlen aus 1 bis 1000 mehr als einmal vorkommen. Das nächste Beispiel zeigt, was Ziehen mit Zurücklegen bedeutet:

**Beispiel 3.6 (Nicht keimende Blumenzwiebeln)** Beim Beispiel 1.7 der Gärtnerei wurde während 50 Tagen je eine Kiste aus der Tagesproduktion von 40 durchnummierierten Kisten durch zufälliges Ziehen mit Zurücklegen der Seriennummern 1 bis 40 bestimmt. Dies bedeutet, dass jede gezogene Zahl sofort nach ihrer Ziehung zurückgelegt wird. Es ist damit möglich, dass an verschiedenen Tagen die gleiche Seriennummer kontrolliert werden kann. Mit einem Statistikprogramm erhält man zum Beispiel die Seriennummern:

25	11	1	18	5	9	37	11	24	20	23	39	1	20	21	13	37
3	8	1	26	24	2	29	25	14	1	6	5	5	12	11	8	29
15	39	2	38	19	17	27	2	32	5	5	29	25	18	20	19	

Am ersten Tag wird also die Kiste 25, am zweiten Tag die Kiste 11 und so weiter kontrolliert.  $\square$



**Abb. 3.1** Drei Arten von Ziehen: Simple random sampling (SRS), Multistage random sampling und stratifiziertes zufälliges Ziehen

Die Abb. 3.1 zeigt drei Typen von zufälligem Ziehen bei endlichen Grundgesamtheiten:

- (1) *Einfaches zufälliges Ziehen* (engl. *simple random sampling*): Bei diesem Verfahren wählt man eine Stichprobe durch zufälliges Ziehen (mit oder ohne Zurücklegen) nach dem oben vorgestellten Verfahren. Beim Beispiel 3.1 der Plattenveredlung erfolgt die Auswahl der Platten durch zweimalige einfache Randomisierung. Zuerst werden aus der Gesamtproduktion durch einfaches zufälliges Ziehen 20 Platten zur Untersuchung bestimmt. Anschliessend werden die Platten durch einfaches zufälliges Ziehen in zwei Gruppen zerlegt.
- (2) *Zweistufiges zufälliges Ziehen* (engl. *multistage sampling design*): Besteht die Grundgesamtheit aus verschiedenen Objekten, wird in einer ersten Stufe durch zufälliges Ziehen eine Stichprobe gewählt. Die Stichprobe wird aussortiert und davon werden Stichproben von jedem auftretenden Objekt genommen. Man spricht hier auch von einer Studie, bei der *Blöcke* (engl. *blocks*) gebildet werden.
- (3) *Stratifiziertes zufälliges Ziehen*: Die Grundgesamtheit wird nach Kategorien in Strata getrennt. Von jedem Stratum wird mit zufälligem Ziehen eine Stichprobe ausgewählt. Dieses Verfahren fand Anwendung bei der Untersuchung zu der Zahl der Molluskenarten in Beispiel 2.10.

Bei konzeptionell gewählten Grundgesamtheiten bestehen mehrere Möglichkeiten, Stichprobenwerte durch zufälliges Ziehen zu sammeln. Eine besteht darin, die Untersuchung in ein endliches Zeitfenster zu setzen und daraus Stichproben durch zufälliges Ziehen von Zeiten zu bilden. Eine zweite Möglichkeit ergibt sich, indem man eine endliche Grundgesamtheit durch Spezifikationen bildet und daraus Stichproben sammelt.

In der Physik und der Technik hat man oft Messungen oder Beobachtungen, um damit nicht direkt messbare Grössen zu berechnen. So will man beim Beispiel 1.3 aus 28 beobachteten Wartezeiten die durchschnittliche Zeit zwischen zukünftigen, aufeinanderfolgenden, starken Erdbeben bestimmen. Die Daten sind keine Stichprobe aus einer hypothetischen Grundgesamtheit. Das Prinzip des Randomisierens ist hier also nicht anwendbar.

---

## 3.2 Wiederholung: Wie viele Messungen?

Je mehr Messungen man hat, desto mehr Information besitzt in der Regel die datensammelnde Person. Es ist deshalb nicht erstaunlich, dass dann nicht direkt messbare Grössen präziser berechnet werden können.<sup>2</sup> Andererseits führen viele Messungen zu wachsenden Kosten eines Experiments. Es ist daher sinnvoll zu entscheiden, wie viele Messungen man machen soll.

---

<sup>2</sup> Umfasst eine Stichprobe beispielsweise die gesamte Grundgesamtheit, so ist die Grösse bis auf die Typ B Unsicherheit genau bestimmt.

**Beispiel 3.7 (Pflanzenvielfalt)** Das Beispiel 2.1 ist eine Untersuchung, die zum Ziel hat, die durchschnittlichen Artenzahlen pro Parzelle der Zone Ackerland in der Schweiz zu bestimmen. Mit einer Stichprobe aus dem Jahr 2001, bestehend aus acht Parzellen, soll dies getan werden:<sup>3</sup>

326 111 332 129 151 230 243 157

Wie in Abschn. 1.1 in Kap. 1 erklärt ist, lässt sich mit einem Datenmodell, das besagt wie die Daten streuen, aus der vorhandenen Vorinformation lernen, in welchem Bereich und mit welcher Wahrscheinlichkeit die durchschnittliche Artenzahl  $\mu_{\text{Acker}}$  in den Ackerlandparzellen liegt. Beispielsweise mit einer Aussage der Form

$$\mathbb{P}(\mu_{\text{Acker}} \text{ ist zwischen } 170 \text{ und } 240 \mid \text{Modell, Daten, Vorinformation}) = 60\%$$

Anders geschrieben hat man: Mit einer Wahrscheinlichkeit von 60 % ist  $\mu_{\text{Acker}} = 205 \pm 35$ . Man sagt, dass das Resultat einen *Fehler 1. Art* von 40 % und einen *Fehler 2. Art* von  $\pm 35$  besitzt. Besitzen die Messwerte systematische Fehler, so hat man

$$\mu_{\text{Acker}} = 205 - \text{systematischer Fehler} \pm 35$$

Bei solchen Fehlerabschätzungen spricht man auch von der *Genauigkeit* (engl. *accuracy*). Sie besteht aus dem systematischen Fehler und dem „statistischen“ Fehler von hier  $\pm 35$ . Als *Präzision* (engl. *precision*) bezeichnet man den zweiten Fehler von  $\pm 35$ .  $\square$

Die Präzision bei statistischen Rechnungen verhält sich oft wie folgt in Funktion der Anzahl Messungen:

Wird aus  $n$  Messungen auf eine nicht direkt messbare Grösse  $\mu$  gerechnet, kann man die Wahrscheinlichkeit  $\gamma$  bestimmen, dass  $\mu = x \pm \Delta x$  ist. Sind die  $n$  Messungen kontrolliert erfasst worden und so, dass sie sich nicht gegenseitig beeinflussen, dann ist die Präzision  $\pm \Delta x$  oft proportional zu  $1/\sqrt{n}$ .<sup>4</sup> Die Gefahr, dass sich Messwerte aus Grundgesamtheiten gegenseitig beeinflussen, ist am kleinsten, wenn mit randomisierten Stichproben gearbeitet wird.

Die Eigenschaft, dass die Präzision oft proportional zu  $1/\sqrt{n}$  ist, bildet einen Eckpfeiler der Statistik: Je häufiger also ein Experiment derart wiederholt wird, dass sich die Messungen nicht gegenseitig beeinflussen, umso präziser kann eine nicht direkt messbare Grösse bestimmt werden. Der Faktor  $1/\sqrt{n}$  bedeutet: Hat man mit 10 Messwerten eine Grösse auf  $\pm 10^{-1}$  bestimmt, so sind  $10^2 = 100$  mehr Messungen – also 1000 Messungen! – nötig, um die Grösse mit einer Präzision von  $\pm 10^{-2}$  zu berechnen. Die Kosten einer Untersuchung wachsen damit proportional zum Quadrat der verlangten Präzision.

<sup>3</sup> Biodiversitätsmonitoring Schweiz BDM, Hintermann & Weber AG, Reinach (Basel), Herbst 2007.

<sup>4</sup> Präzisere Formulierungen finden sich in späteren Kapiteln.

**Beispiel 3.8 (Pflanzenvielfalt)** Bei der Untersuchung zur Pflanzenvielfalt ist die durchschnittliche Artenzahl  $\mu_{\text{Acker}}$  pro Parzelle für die Parzellen der Schweiz mit einer Wahrscheinlichkeit von 60 % durch  $\mu_{\text{Acker}} = 205 \pm 35$  gegeben. Die Präzision kann durch mehr Messdaten verbessert werden. Eine *einzige* Messung mit einem systematischen Fehler kann diese Genauigkeit zerstören. Es ist daher wichtiger, wenige Messungen ohne systematische Fehler, als viele Messungen mit systematischen Fehlern durchzuführen!  $\square$

*Eine Warnung:* Erfahrungen von bedeutenden Statistikern haben gezeigt, dass bei sehr vielen Messwerten Abhängigkeiten zwischen den Messwerten kaum vermeidbar sind. Warum dies so ist, ist schwierig zu beantworten und der Statistiker F. Hampel sagt in [5]:

It just appears to be a fact of life.

Abhängigkeiten zwischen Messwerten sind vor allem spürbar, wenn der Wirkungsraum des Experiments klein ist: beispielsweise, wenn die Messungen durch die gleiche Person oder das gleiche Labor ausgeführt werden. Bei der Pflanzenvielfalt kann es sein, dass benachbarte Parzellen ähnliche Artenzahlen haben. Die Präzision der Rechnung ist dann nicht proportional zu  $1/\sqrt{n}$ , sondern proportional zu  $1/n^\alpha$  mit  $\alpha$  zwischen Null und 0,5. Im schlimmsten Fall, wenn  $\alpha = 0$  ist, erhält man trotz mehr Messwerten keine höhere Präzision! Man sammelt so sinnlos Daten.

---

### 3.3 Kontrolle: Experiment oder Beobachtung?

Mit *kontrollierten* Versuchen lassen sich nicht nur systematische Fehler klein halten, sondern sie erlauben es erst, kausale Zusammenhänge zwischen Messgrößen zu zeigen. Dabei unterscheidet man zwei Arten, wie Daten gesammelt werden:

*Experimente* (engl. *experiments*) sind Untersuchungen, bei denen Daten mit kontrollierten Faktoren und bei vorhandenen Grundgesamtheiten durch Randomisierungsverfahren entstanden sind. Ist dies nicht der Fall, so spricht man von *Beobachtungen* (engl. *observations*).

Die Messungen zur Vakuumkammer in Beispiel 1.8, gemessene Kontraste bei Lasermarkierungen und Atmungsfähigkeiten von Stoffmaterialien (Beispiel 2.12) sind Daten von Experimenten. Die Werte der Unwetterschäden und die Wartezeiten zwischen grossen Erdbeben aus Kap. 1 sind Beobachtungen.

**Beispiel 3.9 (Stromleitungen und Alzheimer)** Während den Jahren 2000–2005 wurde in [6] bei allen Personen der Schweiz, die in der Nähe von Starkstromleitungen lebten,

untersucht, ob sie vermehrt an Alzheimer erkrankten als die restliche Bevölkerung der Schweiz. Die Stichprobe umfasste mehrere Millionen Leute. Aus den Daten folgte, dass Personengruppen, die mindestens 15 Jahre weniger als 50 m von Stromleitungen mit einer Stärke von 200–380 kV leben, mit einer Wahrscheinlichkeit von 95 % ein zwischen 21 % und 233 % höheres Risiko haben, an Alzheimer zu erkranken. Die Daten sind Beobachtungen: Daher ist es schwierig (oder sogar unmöglich) aus der Studie auf eine *kausale* Wirkung von Stromleitungen auf ein erhöhtes Alzheimer-Risiko zu schliessen. Wohnungen in nächster Nähe von Starkstromleitungen befinden sich außerhalb von Städten und Dörfern und werden oft billig vermietet. Es sind daher Leute mit kleineren Einkommen, die in solchen Wohnsituationen leben. Personen mit kleineren Einkommen ernähren sich eventuell weniger gesundheitsbewusst als Personen mit höherem Einkommen. Es könnte sein, dass das Alzheimer-Risiko wegen des Ernährungsverhaltens der Personen (einer Kovariablen) zu erklären ist. □

Im Gegensatz zu technischen Experimenten ist es bei Versuchen mit Menschen extrem schwierig, wenn nicht unmöglich, alle wichtigen Faktoren, die auf eine Grösse wirken, zu kontrollieren. Als Massstab gilt hier, die Wirkung von Faktoren über verschiedene Probandengruppen, die in randomisierter Weise entstanden sind, zu kontrollieren.<sup>5</sup> Man spricht von *randomized controlled trials* (RCT). Studien mit Menschen ohne RCT-Design sind empfindlich auf systematische Fehler und daher sind ihre Resultate mit der nötigen Skepsis zu interpretieren. Doppelt blinde RCT-Studien minimieren wie gesagt den Placebo-Effekt.

Viele Studien befassen sich damit, Kennzahlen von Ländern, Bildungssystemen, Hochschulen, von wirtschaftlichen Grössen oder von medizinischen Untersuchungen zu vergleichen. Eine Unmenge von Ranglisten findet sich dazu täglich in Zeitungen und in Fernsehnachrichten. Meistens handelt es sich hier nicht um RCT-Versuche, sondern um Beobachtungen. Die Gruppen sind meist komplex und heterogen. Sie hängen von verschiedenen Kovariablen ab. Bei Ländern beispielsweise hängen sie von der Zusammensetzung der Berufs- und Bildungsschichten, von Immigrationsquoten, von der Verteilung in Dienstleistungs-, Industrie- und Agrarsektor ab. Da die Gruppen nicht durch Randomisierung entstehen und die Faktoren kaum kontrolliert werden, können Resultate unterschiedlich ausfallen, je nachdem, wie man die Gruppen kombiniert. Man nennt dieses Phänomen das *Simpson-Paradoxon*.<sup>6</sup> Hier Beispiele dazu:

**Beispiel 3.10 (Schülerleistungen)** In den USA werden Schülerleistungen via Tests, die von Kontrollorganen durchgeführt werden, ermittelt. Im Rahmen des National Assessment

---

<sup>5</sup> Siehe dazu auch das Beispiel 2.14 zur Chirurgie.

<sup>6</sup> Das Phänomen ist benannt nach E. H. Simpson, der es im Jahr 1951 in [10] beschrieb. Der schottische Statistiker G. U. Yule (siehe [15]) hat es schon 50 Jahre früher entdeckt.

of Educational Progress erhielt man im Jahr 1992 für die Schüler der achten Mittelstufe im Fach Mathematik bei zwei Staaten:

Nebraska	277
New Jersey	272

Der Staat Nebraska scheint damit seine Schülerinnen und Schüler besser auszubilden. Das Bild ist jedoch unvollständig. Ein Faktor, der den Prüfungserfolg mitbestimmt ist, die soziale Herkunft, die in den USA meist durch die Hautfarbe abgebildet werden kann. Aufgelistet nach diesen Untergruppen lauten die Ergebnisse der Studie:

	Weiss	Schwarz	Andere
Nebraska	281	236	259
New Jersey	283	242	260

Bei allen Gruppen – weiss, schwarz und andere Hautfarben – zeigen die Schüler von New Jersey bessere Resultate als jene in Nebraska. Über alle Gruppen hinweg ist aber Nebraska besser! Möglich ist dies, weil in Nebraska der Anteil der weisshäutigen Schülerinnen und Schüler viel höher ist als in New Jersey, nämlich 87 % zu 66 %:

	Weiss	Schwarz	Andere
Nebraska	87 %	5 %	8 %
New Jersey	66 %	15 %	19 %

Der hohe Anteil der weisshäutigen Schülerinnen und Schüler mit hoher Punktzahl führt dazu, dass sie im Gesamtschnitt mehr Gewicht erhalten und damit der Gesamtschnitt in Nebraska höher wird als der in New Jersey. Der Statistiker H. Wainer sagt in [12] zu diesem Beispiel, das das Simpson-Paradoxon darstellt: Das zusammengefasste Ergebnis als auch die Einzelergebnisse können, je nach Sichtweise, relevant sein. Stellt man die Frage, in welchem Staat ein grösserer Anteil von Studierenden zu finden sind, die eine hohe Mathematikpunktzahl haben, so wird man Nebraska nennen. Interessieren sich aber Eltern für die Frage, in welchem Staat die Chance grösser ist, dass ihre Kinder eine bessere Mathematikausbildung haben werden, so ist die Antwort New Jersey. Sowohl das zusammengefasste Ergebnis als auch die Einzelergebnisse können also relevant sein. Adjustierte Resultate können konstruiert werden, wenn man alle Gruppen nach dem gleichen Prinzip wertet. Nimmt man den Staat New Jersey als Referenz, wird man die Punktzahl der weissen Teilgruppe mit 0,66 und die Punktzahl der schwarzen Teilgruppe mit 0,15 multiplizieren:

$$\text{New Jersey} = 283 \cdot 0,66 + 242 \cdot 0,15 + 260 \cdot 0,19 = 272$$

$$\text{Nebraska} = 281 \cdot 0,66 + 236 \cdot 0,15 + 259 \cdot 0,19 = 270$$

□

Das Simpson-Paradoxon ist in der Praxis immer wieder sichtbar. So schien ein Versuch in Schulen in Schottland zu zeigen, dass Milchtrinken schädlich ist. Dabei stellte sich heraus, dass die Lehrpersonen, vor allem kränklichen Schülerinnen und Schülern Milch zu trinken gaben, in der Hoffnung, dass diese gesünder würden (siehe dazu [8]). Der nicht kontrollierte Faktor „kränklich“ verzerrt das Resultat. Man nennt einen solchen Effekt durch eine Kovariable auch einen *Konfundierungseffekt* oder *Vermengungseffekt* (engl. *Confounder*). Beim obigen Beispiel zu den Schülerleistungen in den Staaten Nebraska und New Jersey ist das Resultat verzerrt, weil der Faktor Hautfarbe vermengt ist.

**Beispiel 3.11 (PISA)** PISA ist ein Akronym für „Programme for International Student Assessment“; dabei handelt es sich um eine Schulleistungsstudie, die im Abstand von drei Jahren international durchgeführt wird. Die Studie ist Teil des Indikatorenprogramms INES der Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (OECD), das dazu dient, den OECD-Mitgliedsstaaten Daten über ihre Bildungssysteme zu geben.

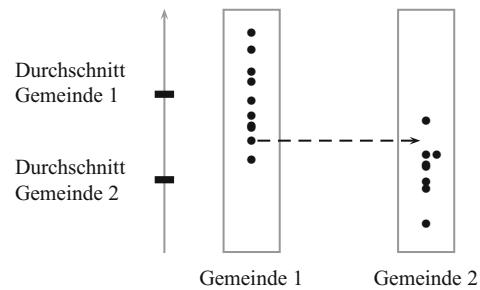
Im Jahr 2000 war ein Forschungsschwerpunkt der PISA-Studie die Leseleistung von 15-jährigen Schülerinnen und Schülern. Folgende Tabelle wurde in [2] publiziert: Finnland 546 Punkte, Kanada 534, Neuseeland 529, ..., Brasilien 396. Die Tabelle kann stark verändert werden, wenn der Vergleich der Länder detailliert über (nicht kontrollierbare) Kovariablen wie die Zusammensetzung der Schulklassen, den Bildungsstand der Eltern, den Migrationshintergrund, die Art der Sprache und Art der Schule gewählt wird. Auch der Faktor Migrationshintergrund kann sinnvoller und weniger sinnvoll umgruppert werden: So stammen Einwanderer nach Deutschland aus niedrigeren sozialen Schichten, Einwanderer nach Neuseeland aus mittleren bis hohen sozialen Schichten. Alle hier möglichen Umgruppierungen erhöhen die Gefahr des Simpson-Paradoxon und machen die Rangliste schwer interpretierbar.

Viele Experten interpretierten die obige Rangliste nicht als Abbild der Leistungen des Schulsystems, sondern als Abbild der Einwanderungs- und Auswanderungsraten, der Homogenität der Klassen, der gesprochenen Dialekte, der sozialen Zusammensetzung oder der Grösse des Landes. Wie üblich ist es schwierig, bei Beobachtungen, wie sie in dieser Studie vorliegen, die Resultate zu interpretieren. □

Verzerrte Resultate entstehen auch, wenn zwischen zwei Untersuchungen zu Gruppen verschiedene Probanden die Gruppen wechseln. So kann, wie Abb. 3.2 zeigt, das durchschnittliche Vermögen von Personen in zwei Gemeinden steigen, nur weil eine Person von der einen Gemeinde zur andern wechselt! Es ist also durchaus möglich, dass ein Medikament scheinbar wirkt, weil Personen während des Versuchs von der Medikamentengruppe zur Placebogruppe wechseln.

In technischen Versuchen können Experimente mit der nötigen Sorgfalt meist geplant und durchgeführt werden. In der Technik lassen sich mit geschickter Kontrolle, nämlich mit *Blockbildung*, zusätzlich die Wirkung von Kovariablen auf eine Messgrösse reduzieren. Das folgende Beispiel zeigt, wie sie funktioniert:

**Abb. 3.2** Wachsendes Durchschnittsvermögen in beiden Gemeinden bei gleich bleibendem Gesamtvermögen



**Beispiel 3.12 (Plattenveredlung)** Ein Firma produziert metallene Platten, die mit Nickel veredelt werden. Die Vernickelung kann mit zwei verschiedenen Verfahren *A* und *B* erfolgen. Erwünscht ist dabei eine möglichst dicke durchschnittliche Nickelschicht.

Ein Versuchsplan, um die Verfahren *A* und *B* zu vergleichen, könnte sein, mit einfacherem zufälligen Ziehen 20 Platten aus der Produktion zu wählen. Aus diesen 20 Platten werden – wieder mit einfacherem zufälligem Ziehen – zwei Gruppen von je 10 Platten gebildet, an denen die Verfahren *A* und *B* getestet werden:

$$\boxed{A} \quad \boxed{B} \quad \boxed{B} \quad \boxed{A} \quad \boxed{B} \quad \dots \quad \boxed{A}$$

Dieser Versuchsplan ist nicht optimal: Die Typ A Unsicherheiten in der Produktion der Platten übertragen sich direkt auf die zwei verschiedenen Veredlungsverfahren.

Eine besserer Versuchsplan besteht darin, mit einfacherem zufälligem Ziehen nur 10 Platten aus der Produktion zu wählen. An den Platten werden beide Verfahren angewendet (beispielsweise indem man jede Platte in zwei Stücke schneidet, und die eine Hälfte dem Verfahren *A*, die andere Hälfte dem Verfahren *B* aussetzt):

$$\boxed{A-B} \quad \boxed{A-B} \quad \boxed{B-A} \quad \dots \quad \boxed{A-B}$$

Diese Blockbildung misst direkt den Effekt zwischen den beiden Verfahren. Sie eliminiert den Mangel der obigen Versuchsanordnung! Die Produktionsstreuung der Platten wird durch den Vergleich der Methoden minimiert.  $\square$

Bei medizinischen und sozialwissenschaftlichen Untersuchungen kann man meist keine Blöcke bilden. Bei der Untersuchung, ob Aspirineinnahme Herzinfarkte reduziert, kann man den Probanden nicht einmal Aspirin und einmal ein Placebo abgegeben.<sup>7</sup> Zusammengefasst hat man:

<sup>7</sup> Dies erhöht nicht nur die Wirkung der Kovariablen, sondern, wie schon oben erwähnt, die Gefahr von fragwürdigen Resultaten. Dies insbesondere dann, wenn die Gruppen nicht durch ein RCT-Design gebildet wurden.

- (1) Experimente – also kontrollierte und bei vorhandener Grundgesamtheit randomisierte Versuche – garantieren, dass statistische Modelle gebraucht werden können, um aus den Daten kausale Schlüsse zu ziehen. In technischen Untersuchungen lassen sich Experimente mit sorgfältigem Aufwand planen.
- (2) Bei Beobachtungen ist es oft schwierig, aus den Daten kausale Schlüsse zu ziehen. Größen können hier von Kovariablen abhängen, die unbekannt sind (sogenannte *sekundäre Variablen*).
- (3) Fehlende Randomisierung, Konfundierungseffekte und Umgruppierungen können systematisch bestimmte Messwerte favorisieren.

### 3.4 Statistische Kontrolle und Vertauschbarkeit

Produktionsprozesse müssen stabil innerhalb gewisser Grenzen ablaufen. Stabil bedeutet, dass der Prozess unter *statistischer Kontrolle* (engl. *statistical control*) ist und die *zeitliche Reihenfolge* der Messresultate unwichtig ist. Viele statistische Verfahren gehen davon aus, dass eine zeitabhängige Folge von Messwerten unter statistischer Kontrolle ist. Dies heisst in der Regel, dass die Daten

- (1) keine zeitlichen Trends oder Zyklen enthalten,
- (2) zufällig von Messwert zu Messwert streuen und
- (3) keine aussergewöhnlich grosse oder kleine Daten vorhanden sind, die auf nicht „normaler“ Streuung der Messgrösse basieren.

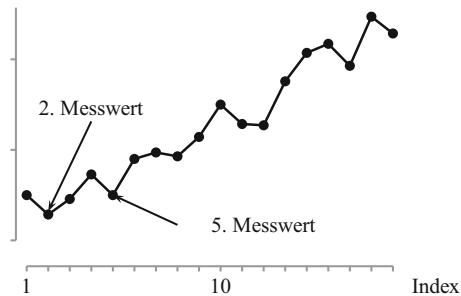
Messwerte heissen *vertauschbar* (engl. *exchangeable*), wenn Rechnungen zu einer Grösse unabhängig von der Reihenfolge der Messungen gemacht werden können. Das Streudiagramm in Abb. 3.3 zeigt 18 Messwerte in der Reihenfolge, in der sie gemessen wurden: Der Versuch ist nicht unter statistischer Kontrolle, da die Werte tendenziell zunehmen. Es wäre sinnlos, den durchschnittlichen Wert zukünftiger Messwerte mit dem arithmetischen Mittel der 18 Messwerte zu bestimmen!

Um aussergewöhnlich grosse oder kleine Messwerte zu markieren, wird im Bereich der statistischen Prozesssteuerung mit der *empirischen Standardabweichung* gearbeitet.<sup>8</sup> Diese versucht, normale Schwankungen der Messwerte zu quantifizieren. Sind  $x_1, x_2, \dots, x_n$  univariate Datenwerte, so ist das *arithmetische Mittel*  $\bar{x}$  (engl. *mean*)

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

<sup>8</sup> In Kap. 11 wird eine zweite Möglichkeit vorgestellt, mit der man aussergewöhnlich grosse oder kleine Datenwerte ausweisen kann.

**Abb. 3.3** Messreihe mit zeitlichem Trend



und die *empirische Standardabweichung*  $s$  der Datenwerte (engl. *estimated standard deviation*  $\widehat{SD}$ ) ist

$$s = \widehat{SD} = \sqrt{\frac{1}{n-1} ((x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2)}$$

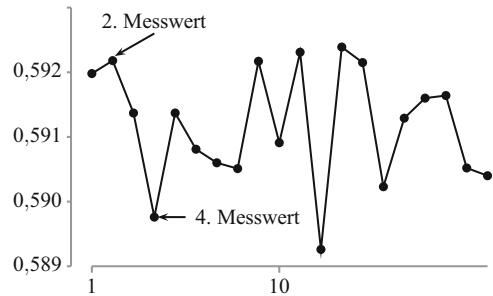
Die obigen Formeln, mit denen man das arithmetische Mittel  $\bar{x}$  und die empirische Standardabweichung  $s$  berechnen kann, sind in den meisten Taschenrechnern und Statistikprogrammen implementiert.

Messwerte, die kleiner als  $\bar{x} - 3 \cdot s$  oder grösser als  $\bar{x} + 3 \cdot s$  sind, werden meist als nicht unter statistischer Kontrolle gewertet. Die beiden berechneten Werte nennt man auch die *geschätzte untere Kontrollgrenze* (engl. *lower control limit*) LCL und die *geschätzte obere Kontrollgrenze* (engl. *upper control limit*) UCL. Es lohnt sich, solche Messwerte besonders zu analysieren. Meist wird diese Formel nur verwendet, wenn mindestens fünfzehn bis zwanzig Messwerte vorliegen.

Um zu beurteilen, ob die Daten unter statistischer Kontrolle sind, benötigt man die Datenwerte in der Reihenfolge, in der sie auftreten. Die so erhaltene Liste nennt man *Urliste*. Sie sollte zur Rekonstruktion eines Experiments aufbewahrt und in einem Laborjournal notiert werden. Man plottet dann die Messwerte aus der Urliste, wie in Abb. 3.3, gegen die Indizes der Messwerte in einem *Streudiagramm* (engl. *scatter plots*):

**Beispiel 3.13 (Druck in einer Vakuumkammer)** Abb. 3.4 visualisiert die zwanzig Messungen zum Unterdruck aus der Vakuumkammer – siehe das Beispiel 1.8 – gegen den Messindex. Im Streudiagramm ist kein zeitlicher Trend zur Zu- oder Abnahme des Unterdrucks sichtbar. Auch Zyklen scheinen nicht vorhanden zu sein. Wie steht es mit aussergewöhnlich tiefen oder hohen Messwerten? Dazu berechnet man die Kontrollgrenzen.

**Abb. 3.4** Streudiagramm der Messwerte aus der Vakuumkammer: erstens keine Trends oder zyklisch wiederkehrende Muster und zweitens zufälliges Streuen um den mittleren Wert



Das arithmetische Mittel  $\bar{p}$  der zwanzig Messwerte beträgt

$$\bar{p} = \frac{0,59198 \text{ bar} + 0,59218 \text{ bar} + \dots + 0,59040 \text{ bar}}{10} = 0,59117 \text{ bar}$$

und die empirische Standardabweichung ist

$$s_p = \sqrt{\frac{1}{20-1} [(0,59198 - 0,59117)^2 + \dots + (0,59040 - 0,59117)^2]} = 0,0009 \text{ bar}$$

Die geschätzte untere Kontrollgrenze lautet somit

$$\text{LCL} = 0,59117 \text{ bar} - 3 \cdot 0,0009 \text{ bar} = 0,5884 \text{ bar}$$

und die geschätzte obere Kontrollgrenze ist

$$\text{UCL} = 0,59117 \text{ bar} + 3 \cdot 0,0009 \text{ bar} = 0,5938 \text{ bar}$$

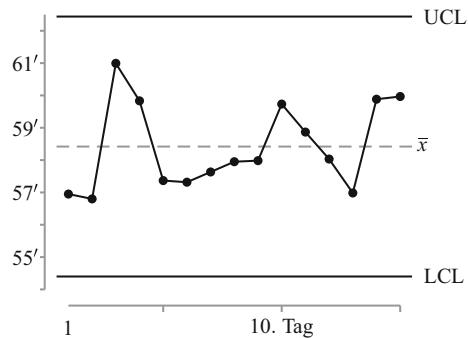
Die Messwerte streuen innerhalb dieser beiden Grenzen. Der Prozess der Vakuumkammer ist damit unter statistischer Kontrolle.  $\square$

**Beispiel 3.14 (Fahrzeiten Intercityzüge von Zürich nach Bern)** Gemäss Fahrplan 2010 der Schweizerischen Bundesbahn beträgt die durchschnittliche Fahrzeit von Intercityzügen von Zürich nach Bern 57 Minuten. Will man zukünftige Fahrzeiten prognostizieren, wie z.B. „Wie gross ist die Wahrscheinlichkeit, dass der morgige Zug mit Nummer IC2345 eine Fahrzeit grösser als 65 Minuten hat?“, benötigt man gemessene Fahrzeiten. Tab. 3.1 zeigt fünfzehn Fahrzeiten zwischen dem 21. Februar 2010 und 7. März 2010. Abb. 3.5 deutet darauf hin, dass die Messwerte erstens keine zeitliche Trends

**Tab. 3.1** Fahrzeiten von Intercityzügen von Zürich nach Bern (Die Datenreihenfolge ist entlang der Spalten.)

56'57"	60'56"	57'22"	57'38"	57'59"	58'52"	56'59"	59'58"
56'48"	59'53"	57'19"	57'57"	59'44"	58'02"	59'53"	

**Abb. 3.5** Streudiagramm der Fahrzeiten: Fahrzeiten unter statistischer Kontrolle



oder zyklisch wiederkehrende Muster aufweisen und zweitens zufällig – wegen vieler Kovariablen, die auf das Zugssystem wirken – um den mittleren Wert streuen. Damit dürfen die Kontrollgrenzen berechnet werden. Die geschätzten Kontrollgrenzen sind

$$\bar{x} \pm 3 \cdot s = 58,42' \pm 3 \cdot 1,34' = 58,42' \pm 4,02'$$

Es sind keine Messwerte ausserhalb der Kontrollgrenzen vorhanden. Der Prozess ist unter statistischer Kontrolle. Dies ist ein gutes Ergebnis für die Schweizerische Bundesbahn. Die Daten eignen sich daher auch gut, um eine Prognose für zukünftige Fahrzeiten zu errechnen.

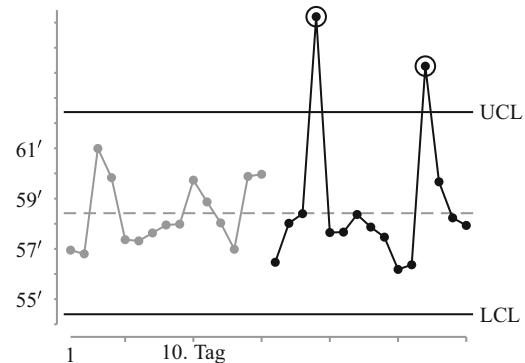
Mit den Kontrollgrenzen lässt sich auch überwachen, ob der Prozess der Fahrten von Intercityzügen von Zürich nach Bern unter statistischer Kontrolle bleibt. Tab. 3.2 zeigt weitere fünfzehn gemessene Fahrzeiten vom 8. März 2010 bis zum 22. März 2010. Die *Kontrollkarte* (engl. *control cart*) in Abb. 3.6 zeigt diese Messwerte. Auf wahrscheinlich spezielle oder sich ändernde Bedingungen weisen die zwei eingekreisten, „aussergewöhnlichen“ Zeiten vom 11. März 2010 und vom 19. März 2010. Dies deutet darauf hin, dass das Zugsystem nicht immer unter statistischer Kontrolle ist. Um die Pünktlichkeit der Züge einzuhalten, sind diese zwei Fahrten zu analysieren und Gegenmassnahmen zu treffen: Die Fahrzeit von 66'14'' war auf starken Schneefall zurückzuführen, diejenige von 64'16'' auf ein technisches Problem an der Lokomotive.

Wie in Kap. 1 erklärt, reagiert ein Qualitätsmanagement nicht auf einzelne Messwerte, sondern handelt nur, wenn die Messreihe nicht mehr unter statistischer Kontrolle ist. Es wäre hier verfehlt, jede Zugfahrt zu analysieren und wegen jeder Verspätung zu reagieren. Erst wenn der Prozess vollständig unter statistischer Kontrolle ist, kann die Qualität (hier die Pünktlichkeit der Züge) verbessert werden. Dazu muss das System als Ganzes

**Tab. 3.2** Weitere gemessene Fahrzeiten von Intercityzügen von Zürich nach Bern (Die Datenreihenfolge ist entlang der Spalten.)

56'28"	58'24"	57'39"	58'22"	57'28"	56'22"	59'40"	57'56"
58'01"	66'14"	57'40"	57'52"	56'11"	64'16"	58'14"	

**Abb. 3.6** LCL und UCL berechnet aus den ersten fünfzehn Beobachtungen mit zwei aussergewöhnlichen Fahrzeiten



**Tab. 3.3** Massen (in Gramm) von 50 Stören aus der Fischzucht der Tropenhaus Frutigen AG (Die Reihenfolge der Messwerte erhält man, wenn man die Messwerte entlang der Spalten liest.)

898	1450	650	746	1344	1710	1388	2214	1208	1066
2050	1516	1206	1428	1734	1538	1480	1194	1300	1500
1198	1452	836	1712	890	1310	1950	1724	1350	1600
644	1196	1290	1064	1240	1384	1422	1500	1145	1276
1294	1200	1380	1134	1458	1506	1612	1540	1228	1336

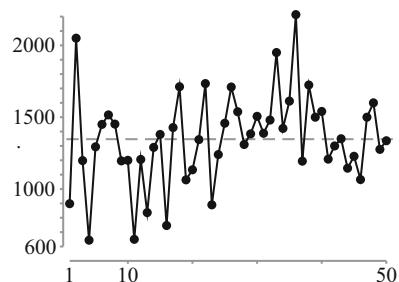
duktionsmethode ändern. Wertvoll ist es auch, das Messverfahren zu analysieren: Sind Messfehler vorhanden? Ist das Messverfahren standardisiert?

3. Ist der Produktionsprozess nicht unter statistischer Kontrolle, so wird der Prozess analysiert. Welche speziellen Faktoren führen zu den fehlerhaften Waren? Auch hier kann es sinnvoll sein, das Messverfahren zu analysieren: Haben die Daten systematische Fehler?

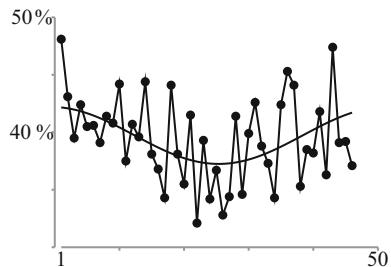
**Beispiel 3.15 (Nicht keimende Blumenzwiebeln)** Ein ähnliches Beispiel, wie die Fahrzeiten von Intercityzügen, wird in Kap. 1 im Beispiel 1.7 diskutiert. Die Kontrollgrenzen wurden aus den ersten fünfzehn Messwerten aus einer Produktion, die unter statistischer Kontrolle war, geschätzt. Man erhielt:  $UCL = \bar{x} + 3 \cdot s = 5,1\% + 3 \cdot 2,5\% = 12,6\%$ . □

**Beispiel 3.16 (Fischzucht)** In Fischzuchten ist es wichtig zu wissen, wie gross die Durchschnittsmasse  $\mu_M$  von Fischen in einem Aufzuchtbettchen ist. Dazu wird mit einem Randomisierungsverfahren eine Stichprobe aus dem Aufzuchtbettchen gezogen und die Massen der untersuchten Fische werden bestimmt. Tab. 3.3 zeigt die Massen von 50 Stören aus der Fischzucht der Tropenhaus Frutigen AG. Abb. 3.8 zeigt die Messwerte in einem Streudiagramm. Sie zeigt, dass kleine Messwerte unter 1000 g nur in der ersten Hälfte der Messreihe auftraten. Die Messungen erfolgten unter freiem Himmel. Im Protokoll der Untersuchung steht, dass die ersten Fische bei starken Windverhältnissen gewogen wurden. Dadurch wurden die abgelesenen Werte der Fischwaage verfälscht. Dies erklärt die im Streudiagramm ersichtliche Tendenz, dass kleine Stichprobenwerte zu Beginn der Messreihe gehäuft auftraten. Das Experiment war nicht unter statistischer Kontrolle. Die 25 zuerst gezogenen Fische wurden daher neu gewogen. □

**Abb. 3.8** Fischmassen: Tendenz zu grösseren Werten (KTI 9029, PFIWI-IW, 2007)



**Abb. 3.9** Streudiagramm der Daten zur Relativdichte mit Jahrestrend



**Beispiel 3.17 (Holzeigenschaften)** In einem Forschungsprojekt der Berner Fachhochschule wurde versucht herauszufinden, wie der Fällzeitpunkt auf die Relativdichte von Holz wirkt. Eine Kovariable, die bekannterweise die Relativdichte beeinflusst, ist die Jahreszeit. Abb. 3.9 zeigt in einem Streudiagramm die 46 Werte von Relativdichten aus [1], die wöchentlich während eines Jahres gemessen wurden. Das Streudiagramm zeigt einen erwarteten zyklischen Jahrestrend. Dieser Faktor kann nicht konstant gehalten werden. Um eine statistische Kontrolle zu erhalten, muss der eingezeichnete Jahreszyklus mit einem mathematischen Instrument aus den Daten entfernt werden. Eine Möglichkeit dies zu tun, ist mit einem *Regressionsmodell* zu arbeiten, wie es in Kap. 12 vorgestellt wird. Kontrollgrenzen können hier nicht mit der Formel  $\bar{x} \pm 3 \cdot s$  berechnet werden. Zuerst muss der „Jahrestrend“ entfernt werden. □

Es sei betont, dass Streudiagramme meist nur grob zeitliche Trends, Zyklen oder Abhängigkeiten zwischen Messwerten aufdecken können. Werkzeuge, mit denen Trends oder Abhängigkeiten feiner abgeschätzt werden können, sind *Autokorrelationskoeffizienten*. Vor allem bei grossen Messreihen sind diese Werkzeuge sehr sinnvoll. Sie werden in Kap. 6 vorgestellt.<sup>9</sup> Man nennt statistische Verfahren *robust*, wenn sie auf Abweichungen von Annahmen unempfindlich reagieren.

**Beispiel 3.18 (Fahrzeiten Intercityzüge von Zürich nach Bern)** Aus den gemessenen Fahrzeiten lässt sich mit statistischen Werkzeugen rechnen, welche durchschnittliche Fahrzeit zukünftige Intercityzüge von Zürich nach Bern haben werden. Das Streudiagramm der Messwerte unterstützt die Annahme, dass die Daten unter statistischer Kontrolle sind. Das Rechenverfahren wäre robust gegenüber dieser Annahme, wenn die Rechnung ihre Gültigkeit behält, auch wenn die Beobachtungen Zyklen enthielten oder aussergewöhnlich grosse oder kleine Messwerte vorhanden wären. □

<sup>9</sup> Im Bereich des SPC werden zusätzliche Werkzeuge – Western Electric Rules – benutzt, um statistische Kontrolle feiner zu beurteilen. Der Prozess ist nicht unter statistischer Kontrolle, wenn (1) Messwerte ausserhalb der geschätzten Kontrollgrenzen UCL oder LCL liegen, (2) zwei von drei aufeinanderfolgenden Messwerten oberhalb  $0,66 \cdot \text{UCL}$  (oder unterhalb  $0,66 \cdot \text{LCL}$ ) liegen, (3) vier von fünf aufeinanderfolgenden Werten oberhalb  $0,33 \cdot \text{UCL}$  (oder unterhalb  $0,33 \cdot \text{LCL}$ ) liegen oder (4) acht aufeinanderfolgende Messwerte auf der gleichen Seite des Mittelwertes liegen.

### 3.5 Qualitätsmanagement: Personen rangieren?

Mit statistischen Methoden können Prozesse begutachtet und gesteuert werden. Streuung und statistische Kontrolle spielen dabei die zentrale Rolle. Insbesondere werden einzelne Messwerte nur ausgeschieden, wenn sie (a) aussergewöhnlich gross oder klein sind oder (b) in Trends oder Zyklen auftreten. Variationen der mittleren Messwerte sind Ausdruck von Zufall und von nicht konstant zu haltenden Kovariablen. Sie werden als Gesamtpaket quantifiziert. Dies ist insbesondere zu bedenken, wenn Produktionsteams oder Personen in Firmen einem Qualitätsmanagement unterworfen werden. Dies betont der bekannte Qualitätsexperte W. E. Deming in [4] auf den Seiten 109–110, wenn er sagt: „Fair rating is impossible.“ So ist es falsch zu denken, dass man Personen auf Grund ihrer Leistungen rangieren kann. Leistungen einer Person hängen von vielen Faktoren oder Kovariablen ab: (a) von der Begabung und vom Leistungswillen der Person selber und (b) von den Personen, die mit ihr arbeiten, von der Arbeitsumgebung (wie auszuführende Arbeit, vorhandene Infrastruktur, Management) und von den Arbeitsbedingungen (Essen im Personalrestaurant, Kaffeepausen, Ferienmöglichkeiten). Diese Faktoren bewirken, dass Leistungen zwischen Personen sehr stark schwanken. Oft ist es auch so, dass die in (b) genannten Faktoren des Arbeitsprozesses die Leistung stärker prägen als die in (a) erwähnten Faktoren der Person selber. Das folgende Beispiel von W. E. Deming in [4] illustriert dies:

**Beispiel 3.19 (Kontrolle von Mitarbeitern)** Sechs Personen mussten an einem einfachen Experiment teilnehmen: Sie erhielten je einen Sack mit total 4000 roten und weissen Perlen. 20 % der Perlen waren rot. Jede Person musste mit *verbundenen Augen* 50 Perlen aus ihrem Sack ziehen. Das Ziel: mit diesem „Produktionsverfahren“ sollen möglichst viele der (schlechten) roten Perlen ausgeschieden werden. Hier das Resultat des Experiments:

	Mike	Peter	Terry	Jack	Louise	Gary
Anzahl rot	9	5	15	4	10	8

Der Erfolg von Terry ist natürlich nur zufällig. Ist aber einer dieser Messwerte aussergewöhnlich, also ausserhalb der geschätzten Kontrollgrenzen, definiert durch  $\bar{x} \pm 3 \cdot s$ ?

Das arithmetische Mittel  $\bar{x}$  der Messwerte ist 8,5 und die empirische Standardabweichung  $s$  beträgt 3,9. Daher ist

$$\text{UCL} = \bar{x} + 3 \cdot s = 8,5 + 3 \cdot 3,9 = 20,2$$

Keine der Testpersonen hat eine aussergewöhnliche Leistung erbracht, da alle Werte kleiner als 20,2 sind. Die statistische Analyse unterstreicht die These, dass es sinnlos ist zu untersuchen, warum Terry 15 rote Perlen und Peter nur 5 rote Perlen gefunden hat!  $\square$

**Tab. 3.4** Anzahl Fehler während einer Kontrollperiode

Name	Anzahl Fehler
Janet	10
Andrew	15
Bill	11
Frank	4
Dick	17
Charlie	23
Alicia	11
Tom	12
Joanne	10

**Beispiel 3.20 (Ford Motorenfabrik)** Ein Beispiel von W. W. Scherkenbach aus der Ford Motorenfabrik illustriert noch einmal die Kontrollgrenzen (siehe [4] auf Seite 112). Als Manager sind Sie verantwortlich für die Leistung einer Gruppe von neun Ingenieurinnen und Ingenieuren. Alle Personen haben in etwa die gleichen Verantwortlichkeiten und haben in etwa die gleiche Wahrscheinlichkeit Fehler – wie Rechen- oder Zeichenfehler, Fehler beim Zusammenstellen von Produkten, … – zu machen. Eine Auflistung der Fehler der Personen findet sich in Tab. 3.4. Sollen Sie Frank für seine ausserordentlich gute Leistung belohnen? Ist es gerecht, Charlie einen Bonus zu verweigern? Die geschätzten unteren und oberen Kontrollgrenzen sind:

$$\left. \begin{array}{l} \text{UCL} \\ \text{LCL} \end{array} \right\} = \bar{x} \pm 3 \cdot s = \bar{x} \pm 3 \cdot 5,31 = 12,55 \pm 15,93$$

Alle Personen liegen mit ihren Fehlern innerhalb der Kontrollgrenzen. Die verschiedenen grossen Fehlerzahlen sind daher auf variierende Bedingungen des Systems, in welchem die Leute arbeiten, zurückzuführen. Alle Leute verdienen damit den gleichen Bonus! Würde man auf Grund der Liste nur Frank belohnen, käme dies einer Lotterie gleich. □

Um die Qualität eines Produktionsteams zu stärken, seien die Vorschläge von W. E. Deming für ein gutes Qualitätsmanagement genannt: (1) Teams sollen in Methoden und Prinzipien der Mitarbeiterführung, der Zusammenarbeit und der Qualitätskontrolle ausgebildet werden, (2) neue müssen Mitarbeiter sorgfältig ausgewählt werden, (3) Mitarbeiter sollen weitergebildet werden, (4) Teamleiter sollen nicht Richter sein, sondern Produktionsteams in stetem Kontakt beraten und von ihnen lernen und (5) Teamleiter sollen fähig sein zu erkennen, ob Mitarbeiter ausserhalb des Produktionsprozesses (auf der guten oder auf der schlechten Seite) oder innerhalb des Produktionsprozesses sind. Dazu können Kontrollkarten, aber es dürfen keine Ranglisten (die unwissenschaftlich sind und schlecht auf Menschen in Produktionsprozessen wirken) benutzt werden. Sie zeigen, dass neben statistischen Werkzeugen, die nur mit Fachwissen eingesetzt werden dürfen, auch soziale Fähigkeiten wichtig sind.

## 3.6 Memorandum zur Datensammlung

Im Folgenden sind die wichtigsten Punkte aus dem Kap. 2 und diesem Kapitel zusammengetragen, die helfen, gutes Datenmaterial zu erhalten. Sie ermöglichen es, aus Daten mit statistischen Werkzeugen effizient zu lernen.<sup>10</sup>

- (1) *Beschreiben Sie das Ziel der Datensammlung:* Nur wenn klar ist, wozu Daten gesammelt werden, kann sinnvoll aus Daten gelernt werden.
- (2) *Definieren Sie die Messgrößen operationell:* Rechnungen zu einer nicht direkt messbaren Grösse oder zu zukünftigen Werten einer unsicheren Grösse hängen davon ab, wie Daten gemessen werden. Die Größen müssen operationell definiert und von allen Beteiligten verstanden werden. Beziehen sich Größen auf Populationen, ist die Grundgesamtheit genau zu definieren.
- (3) *Bestimmen Sie Faktoren und zugehörige Niveaus:* Mit Ursache-Wirkungs-Diagrammen werden Faktoren, die auf eine Messgrösse wirken, bestimmt. Solche Diagramme erlauben es, Datensammlungen kontrolliert ablaufen zu lassen und Überlegungen anzustellen, wie der Effekt von Kovariablen minimiert werden kann. Bestimmt man die Anzahl Faktoren und ihre Niveaus, ergibt sich die Möglichkeit den Aufwand eines Experiments zu beurteilen. Eventuell können Vor- oder Pilotversuche helfen, unbedeutende Faktoren zu eliminieren. Mit dem Bestimmen der Faktoren wird auch der Wirkungsraum des Experiments festgelegt.
- (4) *Wählen Sie bewusst das Design des Versuchs oder der Erhebung:* Die festgelegten Faktoren, die gewünschten Genauigkeiten der Rechnungen, aber auch die Kosten bestimmen, wie ein Experiment oder eine Erhebung durchgeführt wird. Mit faktoriellen Plänen, mit Randomisierung, mit Wiederholungen und eventueller Blockbildung können Versuche meist effizient geplant werden.
- (5) *Sammeln Sie Daten aufmerksam, effizient, kontrolliert und kostenbewusst:* Die Daten, die zum Ziel der Datensammlung führen sollen, können schwierig zu sammeln, zu messen oder zu erfassen sein. Probleme entstehen zum Beispiel bei fehlenden Ressourcen zur Datensammlung oder bei schwierig definierbaren Messgrößen. Geschick, Beharrlichkeit und Phantasie können hier Probleme lösen helfen. Mit einem Managementsystem muss garantiert werden, dass systematische Fehler wegen nicht geeichten oder falsch kalibrierten Geräten oder wegen anderer Ursachen vermieden werden. Um Kosten zu sparen, ist es sinnvoll, die Zeitspanne zur Erfassung der Daten festzulegen.
- (6) *Handeln Sie entsprechend der Daten:* Daten dienen dazu, Probleme zu verstehen, Prozesse zu kontrollieren und aus Daten zu lernen. Da gesammelte Daten Ausschnitte aus der Wirklichkeit darstellen, haben sie eine grosse praktische Relevanz und mehren das Wissen der wissenschaftlich tätigen Personen. Dieses Wissen erlaubt es, gestützt

---

<sup>10</sup> Die aufgeführten Punkte basieren auf drei Methoden von K. Ishikawa (siehe [7]).

auf Modelle und Wahrscheinlichkeitsrechnung wissenschaftlich und vorurteilsfrei zu handeln.

Viele Projekte in Wissenschafts- und Ingenieurgebieten stehen unter Druck, „positive“ Resultate zu liefern. So schildert die Zeitschrift NZZ Folio (siehe [9]) in einem Artikel mit dem Titel *Geschönt, geschlampt, gelogen*, dass ein Forschungsteam aus dem Center for Statistics in Medicine der Universität Oxford 100 Berichte über klinische Versuchsreihen untersucht hat. Dabei stellte das Team fest, dass „ungünstige“ negative Ergebnisse in den publizierten Artikeln nicht erwähnt wurden, um die Chance der Publikation zu erhöhen. Weiter ermittelte es, dass rund die Hälfte der Versuchsreihen grosse Differenzen zwischen den ursprünglichen Zielen der Studie und den berichteten Resultaten hatten. Die deutet darauf hin, dass die Forscher ihre Messwerte einfach nach publizierbarem Material durchkämmt haben. Es ist fragwürdig, erst *nach* der Datenerhebung das Ziel der Datensammlung zu formulieren. Eigentlich lassen sich nämlich in *jeder* Messreihe außergewöhnliche Merkmale finden, wenn man nur lange genug danach sucht. Diese entstehen oft durch zufällige Verteilungen in den Messwerten.

---

### 3.7 Weiterführende Literatur zu Kap. 2 und diesem Kapitel

Ausführliche und vertiefende Informationen zur Sammlung von Daten und zur Qualitätskontrolle finden sich in folgenden Büchern:

1. L. C. Alwan, *Statistical Process Analysis*, (Irwin McGraw-Hill, Boston u. a. 2000)
2. D. R. Cox, *Planning of Experiments* (John Wiley & Sons, New York 1958)
3. C. Daniel, *Applications of Statistics to Industrial Experimentation*, (John Wiley & Sons, New York 1956)
4. W. E. Deming, *Sample Design in Business Research*, (John Wiley & Sons, New York 1960)
5. W. E. Deming, *Out of the Crisis*, (Massachusetts Institute of Technology (MIT) Press Edition, Cambridge Massachusetts 2000)
6. K. Ishikawa, *Guide to Quality Control*, (Asian Productivity Organization, Tokyo 1986)
7. D. C. Montgomery, *Design and Analysis of Experiments*, (John Wiley & Sons, New York 1997)

Nachstehend eine Liste von Büchern, welche unter anderem vertieft Techniken der Versuchsplanung untersuchen:

1. R. A. Fisher, *The Design of Experiments*, 14th ed. (Hafner, New York 1973)
2. T. J. Lorenzen, V. L. Anderson, *Design of Experiments, A No-Name Approach*, Statistics: Textbooks and Monographs (Marcel Decker Inc., New York, Basel, Hongkong 1993)

## Reflexion

### 3.1 Ein Auszug aus der Leica Zeitschrift Nr. 81 von Oktober 2006:

*Der LEICA Courier auf dem Prüfstand*

LEICA wollte es genau wissen: Um zu untersuchen, ob der LEICA Courier den Ansprüchen der (abonnierten) Leserschaft gerecht wird, wurde bei der Fachhochschule Nordwestschweiz eine Studie in Auftrag gegeben. F. B., Studentin in „International Management“, nahm sich der Aufgabe an und verschickte 2800 Fragebögen mit 17 Fragen zur Leserzufriedenheit. „Rund ein Fünftel der Fragebögen wurde zurückgeschickt – pünktlich und exakt ausgefüllt“, freut sich F. B. Nach der ersten Auswertung hat sie das Bild von einer Zeitschrift, die von den Lesern sehr geschätzt und auch gelesen wird. „Rund 95 Prozent der Leser finden den Courier gut bis sehr gut.“

Überrascht Sie das Resultat? Darf sich F. B. freuen? Erläutern Sie im Detail – Definition Grundgesamtheit, Auswahl der Stichprobe – wie Sie diese Studie ausgeführt hätten.

### 3.2 Ein kleineres Unternehmen besteht aus 12 Angestellten, deren Namen

Roman	Franziska	Peter	Urs	Daniel	Rita
Annemarie	Alexander	Silvia	Ursula	Leo	Margrit

lauten. Die Arbeitszufriedenheit soll mit einer Stichprobe aus den 12 Angestellten erforscht werden.

- (a) Bestimmen Sie mit einem Computer sechs Personen, wenn eine einzige solche Befragung mit sechs Personen stattfinden soll.
- (b) Bestimmen Sie mit einem Computer dreimal sechs solcher Personen, wenn drei Befragungen, verteilt auf acht Monate, mit je sechs Personen durchgeführt werden.

### 3.3 Eine Maschine füllt pro Stunde 1000 Flaschen ab, jeweils nummeriert von 1 bis 1000. Sie sollen zur Qualitätskontrolle während 24 Stunden eine Flasche pro Stunde kontrollieren. Bestimmen Sie die Nummern der zu ziehenden Flaschen.

### 3.4 Bei einer Population möchte jemand den Anteil $A$ (in %) der Leute wissen, die ein bestimmtes Merkmal haben. Aus einer Stichprobe mit Stichprobenumfang $n$ kann man $A$ bestimmen. Wenn die Stichprobe durch einfaches zufälliges Ziehen mit Zurücklegen gebildet wurde, hat man mit einer Wahrscheinlichkeit von grob 95 %, dass

$$A = A_{\text{beob}} \pm 1,96 \cdot \frac{\sqrt{A_{\text{beob}} \cdot (100 \% - A_{\text{beob}})}}{\sqrt{n}}$$

ist. Dabei ist  $A_{\text{beob}}$  der beobachtete Anteil in der Stichprobe, der nicht 0 % oder 100 % sein darf. Zeigen Sie, dass die Präzision der Rechnung am kleinsten ist, wenn  $A_{\text{beob}} = 50 \%$  ist. Dies bedeutet, dass ein Anteil am wenigsten präzis bestimmt werden kann, wenn er aus der Stichprobe mit 50 % geschätzt werden muss.

**Tab. 3.5** Rangliste der Gymnasien in der Schweiz, sortiert nach dem Erfolg an der ETH Zürich

Gymnasium	Mittelwert
Gymnasium Immensee ( $n = 32$ )	3,60
Liceo Cantonale Bellinzona ( $n = 79$ )	3,82
Kantonale Maturitätsschule für Erwachsene ( $n = 77$ )	3,83
:	:
Kantonsschule Rämibühl ( $n = 191$ )	4,14
:	:
Kantonsschule Hohe Promenade ( $n = 60$ )	4,55
Gymnasium Liestal ( $n = 56$ )	4,59
Kantonsschule Rychenberg Winterthur ( $n = 61$ )	4,60

**3.5** Hier ein Auszug aus einer Umfrage von CNN/Gallup vom 18. Januar 2005:

*Poll: Americans upbeat about next four years*

Fifty-one percent of respondents said Bush's policies will move the country in the right direction. Fifty-two percent said he will be an outstanding or above average president in his second term. [...] The survey was conducted Friday through Sunday in phone calls to 1,007 adult Americans; it has a margin of error of plus or minus 3 percentage points.

„Überprüfen“ Sie die Fehlerangabe mit der Formel von Aufgabe 3.4. Wäre eine Umfrage mit einem Stichprobenumfang von  $n = 100$  sinnvoll?

**3.6** Viele Untersuchungen – wie etwa [14] – zeigen, dass bei Frauen, die Herpes haben, häufiger Gebärmutterhalskrebs diagnostiziert wird. Ist es sinnvoll, daraus zu schliessen, dass Herpes-Viren das Risiko erhöhen, an Gebärmutterhalskrebs zu erkranken?

**3.7** Die Eidgenössische Technische Hochschule Zürich (ETHZ) untersuchte in der Studie [11], ob der Studienerfolg ihrer Studierenden davon abhängt, an welchem Gymnasium sie die Matura abgeschlossen haben. Sie versuchte dabei die Ausbildungsqualität der Gymnasien aus Prüfungsnoten von an der ETH Studierenden zu berechnen. Das veröffentlichte Resultat mit einer Rangliste findet sich in Tab. 3.5.

- (a) Handelt es sich hier um ein Experiment oder um Beobachtungen?
- (b) Erklären Sie, warum eine Rangliste hier kaum sinnvoll ist. Listen Sie mögliche Kovariablen auf, die auf den „Prüfungserfolg“ wirken könnten.

**3.8** Tab. 3.6 zeigt das Resultat von zwei Elfmeterschützen.

- (a) Wie gross sind die Trefferwahrscheinlichkeiten der beiden Spieler jeweils bei den Heim- und bei den Auswärtsspielen?
- (b) Welcher Spieler hat die höhere Trefferwahrscheinlichkeit?

**Tab. 3.6** Resultat von zwei Elfmeterschützen, bei Heim- und Auswärtsspielen

	Heimspiel		Auswärtsspiel	
	Treffer	kein Treffer	Treffer	kein Treffer
Spieler A	4	1	6	4
Spieler B	14	6	1	1

**3.9** In einer Masterklasse von 20 Studierenden beschliessen fünf Studierende, sich jeweils am Samstagmorgen zu treffen und sich auf die Abschlussprüfung im Fach Mechanik vorzubereiten. Nach der Abschlussprüfung stellt die Klasse fest, dass die fünf Studierenden der Samstag-Morgen-Gruppe im Schnitt 50 % mehr Punkte an der Prüfung erzielen als die anderen 15 Studierenden.

- (a) Weist das Resultat darauf hin, dass Vorbereitungstreffen einen besseren Prüfungserfolg bewirken? Warum oder warum nicht?
- (b) Könnte man mit einem Experiment versuchen, die These zu verifizieren, dass Vorbereitungstreffen einen besseren Prüfungserfolg bewirken? Wenn ja, wie müsste ein solches Experiment aufgebaut werden? Wenn nein, warum nicht?

**3.10** Zwei Ärztegruppen möchten feststellen, ob ein neues Medikament besser heilt als ein bisher benutztes Präparat. Die erste Ärztegruppe, die mit dem alten Medikament sehr zufrieden ist, verabreicht das neue Medikament nur an zehn Prozent der Patienten und erhält das Resultat:

	Medikament neu	Medikament alt
Erfolg	80	810
Kein Erfolg	20	90
Total	100	900

- (a) Wie lauten die Heilungsraten der beiden Medikamente? Ist das neue Medikament wirksamer?
- (b) Die zweite Ärztegruppe ist wagemutiger und probiert das neue Medikament an 90 % der Patienten aus:

	Medikament neu	Medikament alt
Erfolg	855	97
Kein Erfolg	45	3
Total	900	100

Wie lauten die Heilungsraten der beiden Medikamente? Ist das neue Medikament wirksamer?

- (c) Welches Resultat erhalten Sie, wenn Sie beide Studien „naiv“ zusammenführen? Wie könnte man die beiden Studien zusammenfügen, damit das Resultat unverzerrt erscheint?

**3.11** Im Rahmen einer Bachelorarbeit in Maschinentechnik an der Berner Fachhochschule musste die Dimension einer Wasserturbine für ein Pico-Wasserkraftwerk bestimmt werden. Die Dimension der Turbine hängt davon ab, wie gross die Wassermenge des Baches ist, die durch das Kraftwerk fliessen wird. Um die Dimension zu berechnen, wurde während eines Jahres vierzehnmal die Wassermenge des Baches gemessen:

126,85	215,75	182,61	256,53	113,63	177,62	209,62
102,81	409,80	270,08	233,95	59,51	259,35	139,46

Die Reihenfolge der Daten ist entlang der Spalten dargestellt.

- (a) Handelt es sich hier um Beobachtungen oder Werte aus Experimenten?
- (b) Kontrollieren Sie, ob die Messwerte sich nicht gegenseitig „beeinflusst“ haben. Hat es Trends oder zyklische Muster in den Daten? Schätzen Sie die obere und untere Kontrollgrenze, wenn keine Trends oder zyklische Muster vorhanden sind. Sind alle Werte unter Kontrolle?

**3.12** Im Rahmen eines Projekts wurde versucht, die Haftkraft von Klebeetiketten zu optimieren. Tab. 3.7 zeigt zwanzig Messwerte. Streuen die Messwerte unabhängig voneinander? Hat es Trends oder zyklische Muster in den Daten?

**3.13** Die Firma Tillamook-Cheese in Oregon (USA) produziert quaderförmige Frischkäsekörper mit einer mittleren Masse von etwa 19 kg. Die Körper werden gereift und anschliessend für den Handel in kleine Portionen von 500–1000 Gramm zerschnitten. Zur Qualitätskontrolle müssen neben der Zusammensetzung der Körper auch die Durchschnittsmasse der Tagesproduktion bestimmt werden. Tab. 1.13 zeigt 20 Messungen.

- (a) Kontrollieren Sie, ob die Stichprobenwerte keine Trends oder Zyklen zeigen.
- (b) Schätzen Sie die obere und untere Kontrollgrenze: sind ausserordentliche Stichprobenwerte vorhanden?
- (c) Ist es plausibel anzunehmen, dass die Messwerte sich gegenseitig nicht beeinflussen? Deuten die Stichprobenwerte darauf hin, dass die Firma ihre Produktion unter statistischer Kontrolle hat?

**Tab. 3.7** Gemessene Haftkräfte an einer Etikette (Die Datenreihenfolge ist entlang der Spalten dargestellt.)

0,7894478	0,7896109	0,7896697	0,7897264
0,7896139	0,7894955	0,7899264	0,7895403
0,7898179	0,7893841	0,7899313	0,7893542
0,7898388	0,7893035	0,7899085	0,7893552
0,7897761	0,7894020	0,7898478	0,7895980

**Tab. 3.8** Anzahl Fahrradunfälle in der Stadt Bern, 1991–2008

Jahr	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
Unfälle	100	108	120	111	101	104	118	98	91	97
Jahr	2001	2002	2003	2004	2005	2006	2007	2008		
Unfälle	103	97	117	121	129	115	104	84		

**3.14** Die Direktion für Tiefbau, Verkehr und Stadtgrün der Stadt Bern betonte an einem Vortrag von 4. November 2009 bei der IG Velo, dass dank ihrer Arbeit die Fahrradunfälle in den letzten drei Jahren abgenommen habe. Dazu wurden die Anzahl Fahrradunfälle in der Stadt Bern der Jahre 1991–2008 aus Tab. 3.8 benutzt.

- (a) Sie werfen eine Münze mit Seiten  $\oplus$  und  $\ominus$ . Wie gross ist die Wahrscheinlichkeit dreimal nacheinander  $\ominus$  zu werfen?
- (b) Diskutieren Sie die Aussagen in der Graphik mit Ihrem neuen Wissen in Statistik: sind die Fahrradunfälle unter statistischer Kontrolle, sind aussergewöhnliche Streuungen vorhanden, war die Direktion erfolgreich oder hatte sie vielleicht einfach Glück?

**3.15** In Zeitungen finden sich Streudiagramme von Aktienkursen. Sind die Messwerte unter statistischer Kontrolle?

**3.16** Kontrollieren Sie den Versuchsaufbau Ihres letzten Experiments im Physik- oder Ingenieurlabor: Haben Sie die Daten kontrolliert erzeugt? Wenn nein, warum nicht? Haben Sie die Anzahl Messwerte bewusst gewählt? War Ihr Experiment unter statistischer Kontrolle?

---

## Literatur

1. D. Bättig, E. Wyler, *Statistischer Bericht zum Forschungsbericht Fällzeitpunkt und Holzeigenschaften* (Berner Fachhochschule, Burgdorf, 2005)
2. J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, P. Stanat, K.-J. Tillmann, M. Weiss (Hrsg.): *PISA 2000, Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (Leske + Budrich, Opladen, 2001)
3. C. Daniel, *Applications of Statistics to Industrial Experimentation* (John Wiley & Sons, New York, 1976)
4. W. E. Deming, *Out of the Crisis* (MIT Press Edition, 2000)
5. F. Hampel, Is statistics too difficult? The Canadian Journal of Statistics, Vol. **26**, No. 3, 497–513 (1998)
6. A. Huss, A. Spoerri, M. Egger, M. Röösli (for the Swiss National Cohort Study), Residence Near Power Lines and Mortality from Neurodegenerative Diseases: Longitudinal Study of the Swiss Population. American Journal of Epidemiology, Online-Publikation 5.11. (2008)
7. K. Ishikawa, *Guide to Quality Control* (Asian Productivity Organization, 1982)
8. D. V. Lindsley, *Understanding Uncertainty* (J. Wiley & Sons, 2007) S. 122

9. R. Matthews, Geschönt, geschlampt, gelogen. NZZ Folio, 01 (2006)
10. E. H. Simpson, The Interpretation of Interaction in Contingency Tables. Journal of the Royal Statistical Society, Series B, **13**, 238–241 (1951)
11. B. Spicher, *Maturanoten und Studienerfolg* (ETH Zürich, 2008)
12. H. Wainer, *Graphic Discovery. A Trout in the Milk and Other Visual Adventures* (Princeton University Press, 2008)
13. W. A. Wallis, H. V. Roberts, *Methoden der Statistik, ein neuer Weg zu ihrem Verständnis* (rororo, 1969) S. 280
14. E. L. Wynder, J. Cornfield, P. D. Schroff, K. R. Doraiswami, A study of environmental factors in carcinoma of the cervix. American Journal of Obstetrics and Gynecology, Vol. **6**, 1016–52 (1954)
15. G. U. Yale, Notes on the Theory of Association of Attributes in Statistics. Biometrika **2**, 121–134 (1903)

„Ich konnte es auch gar nicht lernen“, sagte die Falsche Suppenschildkröte, „weil ich zu arm war. Ich hatte nur die Pflichtfächer.“  
„Und die waren?“ fragte Alice.

„Also, zunächst einmal das Grosse und das Kleine Nabelweh, natürlich“, antwortete die Falsche Suppenschildkröte, „aber dann auch Deutsch und alle Unterarten – Schönschweifen, Rechtspeibung, Sprachelbeere und Hausversatz.“

Lewis Carroll, *Alice im Wunderland* (Insel Taschenbuch, 1973, S. 99)

## Zusammenfassung

In Kap. 1 wird erklärt, dass Rechnungen aus Daten erst vertrauenswürdig werden, wenn angegeben wird, wie genau und wie plausibel die Resultate sind. Dabei wird Plausibilität wie folgt mit einer Wahrscheinlichkeit beschrieben: „Die mittlere Zeit zwischen zukünftigen, aufeinanderfolgenden starken Erdbeben liegt mit einer Wahrscheinlichkeit von 90 % zwischen 450 und 500 Tagen.“ Was Wahrscheinlichkeiten sind und wie man Aussagen mit Wahrscheinlichkeiten ausdrückt, wird in diesem Kapitel gezeigt. Zudem wird vorgestellt, wie man mit Wahrscheinlichkeiten rechnet. Anschliessend werden Modelle erwähnt, mit denen man beschreiben kann, wie Messwerte streuen. Solche Modelle werden auch mit Wahrscheinlichkeiten formuliert. Zum Schluss des Kapitels wird diskutiert, wie man dank Simulationen, Wahrscheinlichkeiten bei komplizierten Modellen bestimmen kann.

## 4.1 Die drei Rechengesetze

Grössen, wie die zukünftige, durchschnittliche Fahrzeit von Zügen auf der Strecke von Bern nach Zürich, können wegen fehlender Information nur approximativ berechnet werden. Andere Grössen, wie der Druck in einer Vakuumkammer, sind nicht direkt messbar

und daher auch nur approximativ bestimmbar. Wie plausibel solche Rechnungen sind, wird von Wissenschaftlern und Ingenieuren mit Wahrscheinlichkeiten bewertet. Auch Prognosen zu zukünftigen Werten einer unsicheren Grösse formuliert man mit Wahrscheinlichkeiten. So beschreibt das Management der Schweizerischen Bundesbahn im Jahr 2010, dass die im Fahrplan gegebene Fahrzeit von Zügen mit einer Wahrscheinlichkeit von 0,88 nicht mehr als drei Minuten überschritten wird. Ein Wetterdienst beziffert die Wahrscheinlichkeit, dass es am nächsten Tag an einem bestimmten Ort regnen wird, mit 0,9. Was bedeuten Wahrscheinlichkeiten wie 0,88 oder 0,9? In diesem Buch sind Wahrscheinlichkeiten Werte, die quantifizieren, wie *plausibel* eine Aussage ist. Um Plausibilitäten wissenschaftlich nutzen zu können, müssen Regeln dazu aufgestellt werden. Zuerst will man Plausibilitäten vergleichen können. Dazu verlangt man:

- (1) Der Grad der Plausibilität (engl. *believe*) wird durch eine nicht negative Zahl dargestellt.
- (2) Die Plausibilität stimmt mit dem „gesunden Menschenverstand“ überein: Grössere Zahlen bedeuten grössere Plausibilität.

Weiter muss vereinbart werden, auf welcher Basis Plausibilitäten berechnet werden sollen. Man will, dass diese auf Grund von *Informationen*, wie Daten, physikalischen Gesetzen oder Wissen bestimmt werden. Erstrebzt wird, dass verschiedene Personen Aussagen mit derselben Plausibilität werten, wenn sie die gleichen Informationen haben. Dies verhindert, dass Plausibilitäten willkürlich gesetzt werden. Man verlangt daher genauer:

- (3) Auch wenn eine Aussage auf verschiedene Arten dargestellt werden kann, so muss sie die gleiche Plausibilität besitzen.
- (4) Berechnet man die Plausibilität einer Aussage, muss die vorhandene Information benutzt werden.
- (5) Gleches Wissen zu einer Aussage ergibt die gleiche Plausibilität.

R. T. Cox hat im Jahre 1946 in [3] gezeigt, dass diese fünf Regeln dazu führen, Plausibilitäten zu Aussagen mit Wahrscheinlichkeiten zu quantifizieren. Diese müssen zudem drei Rechengesetze erfüllen:<sup>1</sup> die Konvexität, das Additions- und das Multiplikationsgesetz. Um diese vorzustellen, ist es zuerst sinnvoll, relevante Notationen einzuführen. Es ist üblich, Aussagen mit grossen Buchstaben abzukürzen, wie

$$A = \text{„Die Fahrzeit } T \text{ des Zugs IC 8021 von Zürich nach Bern des nächsten Tages wird höchstens 65 Minuten dauern.“}$$

Wie plausibel die Aussage  $A$  ist, wird  $\mathbb{P}(A)$  oder  $\mathbb{P}(T \leq 65')$  geschrieben. Die Plausibilität – eine Zahl zwischen null und eins – hängt davon ab, welches Wissen oder welche Information zur Aussage vorhanden ist. Eine Person aus dem Management der Bahn wird

---

<sup>1</sup> Ausführliche Diskussionen dazu findet man in [11] und [17].

auf Grund von gemessenen Fahrzeiten sagen, dass  $\mathbb{P}(A) = 0,88$  ist. Sie rechnet also mit einer Chance von 88:12, dass die Aussage wahr ist. Mit der Zahl kann sie eine Wette wie folgt abschliessen: „Ist die Aussage richtig, gewinne ich CHF 12.–, ist die Aussage falsch, verliere ich CHF 88.–.“<sup>2</sup> Jemand, der selten mit Zügen fährt, wird vielleicht skeptisch behaupten, dass  $\mathbb{P}(A) = 0,3$  ist. Um dies zu verdeutlichen, werden Wahrscheinlichkeiten auch so geschrieben:

$$\mathbb{P}(A | \mathcal{K}_1) = 0,88 \quad \mathbb{P}(A | \mathcal{K}_2) = 0,3$$

Der vertikale Strich in den Formeln heisst „gegeben“ und  $\mathcal{K}_1$  bzw.  $\mathcal{K}_2$  bezeichnen die vorhandene *Informationen*:  $\mathcal{K}_1$  ist das Wissen der Person aus dem Management auf Grund von Daten und  $\mathcal{K}_2$  ist das Wissen der Person, die selten Zug fährt.<sup>3</sup> Auch Aussagen können als Information genutzt werden. Ist beispielsweise  $B$  die Aussage, dass während der morgigen Fahrt des Intercityzugs IC 8021 das Schienennetz überlastet sein wird, beschreibt  $\mathbb{P}(A | B)$  die Wahrscheinlichkeit, dass  $A$  wahr ist, gegeben die Information, dass  $B$  zutrifft. Der vertikale Strich in der Formel wird wiederum als „gegeben“ gelesen. Analog ist  $\mathbb{P}(A | B, \mathcal{K}_1)$  die Wahrscheinlichkeit, dass  $A$  wahr ist, gegeben die Informationen  $B$  und  $\mathcal{K}_1$ .

Betrachten wir ein weiteres Beispiel. Jemand wirft eine Münze mit Seiten Kopf und Zahl. Die Aussage  $K$  ist: „Die Münze wird auf Kopf fallen.“ Die Person schreibt

$$\mathbb{P}(K | \mathcal{I}) = 0,5$$

Was meint die Person damit? In diesem Buch wird dies so interpretiert: Die Information  $\mathcal{I}$ , die die Person besitzt, erlaubt ihr nicht zu sagen, ob die Münze eher auf Kopf oder eher auf Zahl fallen wird. Die Person sagt also nicht, dass die Münze „zufällig“ auf Kopf oder Zahl fällt. Dies kann ja die Münze auch nicht tun. Physikalische Gesetze, wie das Gesetz von Newton, besagen, dass man berechnen kann, wie der Flug der Münze sein wird. Dazu müsste man die Abwurfhöhe und den Abwurfimpuls der Münze messen. Mit dieser Information  $\mathcal{I}$ , hätte man vielleicht  $\mathbb{P}(K | \mathcal{I}) = 0,9$ . Die Wahrscheinlichkeit beschreibt also, wie plausibel die Aussage für die Person ist. Sie charakterisiert keine Eigenschaft der Münze.

Die drei erwähnten Gesetze zur Wahrscheinlichkeitsrechnung können nun formuliert werden. Sie basieren auf den Regeln (1)–(5) zur Plausibilität. Das erste Gesetz besagt, dass Wahrscheinlichkeiten Zahlen zwischen null und eins sind. Wahre Aussagen haben Wahrscheinlichkeit eins, unwahre Aussagen Wahrscheinlichkeit null:

---

<sup>2</sup> Anders ausgedrückt: Die Plausibilität einer Person misst, wie viel sie bereit ist, bei einer Wette einzusetzen. Daraus kann die Person bei Aussagen zu Unsicherheiten beurteilen, wie gross ihr (monetäres) Risiko ist, falsch zu liegen.

<sup>3</sup> Formuliert man Wahrscheinlichkeiten zu Aussagen, sollte die vorhandene Information genannt werden:  $\mathbb{P}(\text{Aussage} | \text{Information})$ . Manchmal schreibt man auch nur kurz  $\mathbb{P}(\text{Aussage})$ , um Notationen klein zu halten. Es ist aber wichtig, die benutzte Information zu erwähnen. Es besteht sonst die Gefahr, Wahrscheinlichkeiten miteinander zu verrechnen, die auf Grund von verschiedenen Informationen bestimmt wurden. Dies kann zu widersprüchlichen Resultaten führen.

**Theorem 4.1 (Gesetz der Konvexität)**

Die Wahrscheinlichkeit  $\mathbb{P}(A) = \mathbb{P}(A \mid \mathcal{I})$ , wie plausibel  $A$  bei gegebener Information  $\mathcal{I}$  ist, ist eine Zahl zwischen null und eins. Ist die Aussage wahr, so ist die Wahrscheinlichkeit  $\mathbb{P}(A)$  eins. Ist sie unwahr, so ist die Wahrscheinlichkeit  $\mathbb{P}(A)$  null.

Um das zweite Gesetz zu formulieren, braucht man die Negation einer Aussage. Ist  $K$  die oben genannte Aussage „Die Münze wird auf den Kopf fallen“, so ist die Negation von  $K$  die Aussage „Die Münze wird nicht auf den Kopf fallen.“ Die Negation von  $K$  wird in diesem Buch mit  $K^{\text{nicht}}$  bezeichnet. Wenn man die Wahrscheinlichkeit kennt, dass eine Aussage wahr ist, so folgt aus den Regeln zur Plausibilität, dass man auch weiß, wie gross die Wahrscheinlichkeit ist, dass ihre Negation wahr ist. Für diese gilt:

**Theorem 4.2 (Additionsgesetz)**

Die Wahrscheinlichkeit, dass die Aussage  $A$  wahr ist, sei bekannt. Dann lautet die Wahrscheinlichkeit, dass  $A^{\text{nicht}}$  wahr ist,

$$\mathbb{P}(A \mid \mathcal{I}) + \mathbb{P}(A^{\text{nicht}} \mid \mathcal{I}) = 1$$

Ist beispielsweise bei gegebener Information  $\mathbb{P}(A \mid \mathcal{I}) = 0,4$ , so hat die Negation von  $A$ , gegeben die gleiche Information, eine Plausibilität von 0,6.<sup>4</sup>

Das dritte Gesetz ist das Multiplikationsgesetz. Es besagt, wie man die Wahrscheinlichkeit berechnet, dass zwei Aussagen wahr sind. Hier dazu ein Beispiel:  $A$  sei die Aussage „Ütermorgen wird in Bern Schnee liegen“, und  $B$  sei die Aussage „Morgen wird in Bern Schnee liegen.“ Wie gross ist die Wahrscheinlichkeit, dass morgen und übermorgen in Bern Schnee liegen wird? Um dies zu berechnen, ist es sinnvoll zu überlegen, wie wahr die Aussage  $A$  ist. Dies hängt davon ab, ob am Vortag schon Schnee lag. Mit anderen Worten braucht man die Wahrscheinlichkeit  $\mathbb{P}(A \mid B)$ . Hat man diese Wahrscheinlichkeit, wird man als Nächstes überlegen, wie gross die Plausibilität ist, dass auch am ersten Tag Schnee liegt. Dies heißt, man benötigt die Wahrscheinlichkeit, dass  $B$  wahr ist. Das Multiplikationsgesetz sagt, dass aus diesen beiden Wahrscheinlichkeiten die gesuchte Wahrscheinlichkeit  $\mathbb{P}(A \text{ und } B)$  berechnet werden kann:

---

<sup>4</sup> Hat man eine andere Information  $\mathcal{K}$  als  $\mathcal{I}$ , so ist die Plausibilität der Negation von  $A$  nicht notwendigerweise 0,6. Beispielsweise kann  $\mathbb{P}(A^{\text{nicht}} \mid \mathcal{I}) = 0,6$  und  $\mathbb{P}(A^{\text{nicht}} \mid \mathcal{K}) = 0,8$  sein.

**Theorem 4.3 (Multiplikationsgesetz)**

Ist die Wahrscheinlichkeit, dass  $A$  gegeben  $B$  wahr ist, und die Wahrscheinlichkeit, dass  $B$  wahr ist, bekannt, so kennt man die Wahrscheinlichkeit, dass  $A$  und  $B$  wahr sind.<sup>5</sup>

$$\mathbb{P}(A \text{ und } B) = \mathbb{P}(A | B) \cdot \mathbb{P}(B)$$

Die drei Gesetze umfassen die gesamte Wahrscheinlichkeitsrechnung. Es ist also wichtig, mit ihnen rechnen zu können.<sup>6</sup>

**Beispiel 4.1 (Ein häufiger Irrtum)** Im letzten Abschnitt wurden Wahrscheinlichkeiten wie  $\mathbb{P}(A \text{ und } B)$ ,  $\mathbb{P}(A | B)$  und  $\mathbb{P}(B | A)$  eingeführt. Personen, die zum ersten Mal mit Wahrscheinlichkeiten rechnen, verwechseln diese drei Ausdrücke oft und sagen, dass diese Wahrscheinlichkeiten gleich sind. Hier ein Beispiel, das veranschaulicht, dass die drei Wahrscheinlichkeiten sehr verschieden gross sein können. Sie haben als Information, dass sich eine Person in einem Zimmer befindet. Weiter wissen Sie, dass rund 3 % aller Frauen schwanger und rund 50 % aller Menschen Frauen sind. Die Aussage  $A$  sei „Die Person im Zimmer ist schwanger“, und die Aussage  $B$  sei „Die Person im Zimmer ist eine Frau.“ Die Wahrscheinlichkeit, dass die Person eine Frau ist, wenn man weiß, dass die Person schwanger ist, ist einfach:

$$\mathbb{P}(\text{Frau} | \text{schwanger}) = \mathbb{P}(B | A) = 1$$

Die Wahrscheinlichkeit, dass die Person schwanger ist, wenn Sie wissen, dass sie eine Frau ist, lautet

$$\mathbb{P}(\text{schwanger} | \text{Frau}) = \mathbb{P}(A | B) = 0,03$$

Die Wahrscheinlichkeit, dass die Person im Zimmer schwanger und eine Frau ist, also  $\mathbb{P}(A \text{ und } B)$  lautet mit dem Multiplikationsgesetz:

$$\mathbb{P}(\text{Frau und schwanger}) = \mathbb{P}(A | B) \cdot \mathbb{P}(B) = 0,03 \cdot 0,5 = 0,015$$

□

---

<sup>5</sup> Die Multiplikationsregel lässt sich, da  $\mathbb{P}(A \text{ und } B) = \mathbb{P}(B \text{ und } A)$  ist, auch so schreiben:  $\mathbb{P}(A \text{ und } B) = \mathbb{P}(B | A) \cdot \mathbb{P}(A)$ . Genauer müsste man die Information  $\mathcal{I}$  in der Gleichung ausweisen:  $\mathbb{P}(A \text{ und } B | \mathcal{I}) = \mathbb{P}(B | A, \mathcal{I}) \cdot \mathbb{P}(A | \mathcal{I})$ .

<sup>6</sup> Siehe dazu D. V. Lindley in [13] auf Seite 64:

It is a fact that can hardly be emphasized too strongly that these three rules encapsulate everything about probability, and therefore everything about your probability measurement [...].  
... Once you have understood the three rules of probability just stated, you can calculate for yourselves and not read further.

Es folgt eine erste Rechnung mit den Rechengesetzen zu Wahrscheinlichkeiten: Sind  $A$  und  $B$  zwei Aussagen, so hat man nach der Multiplikationsregel

$$\mathbb{P}(B \text{ und } A) = \mathbb{P}(B | A) \cdot \mathbb{P}(A), \quad \mathbb{P}(B^{\text{nicht}} \text{ und } A) = \mathbb{P}(B^{\text{nicht}} | A) \cdot \mathbb{P}(A)$$

Zählt man die beiden Gleichungen zusammen, ergibt sich

$$\mathbb{P}(B \text{ und } A) + \mathbb{P}(B^{\text{nicht}} \text{ und } A) = \{\mathbb{P}(B | A) + \mathbb{P}(B^{\text{nicht}} | A)\} \cdot \mathbb{P}(A)$$

Der Faktor in der eckigen Klammer ist wegen der Konvexität gleich eins. Man erhält damit

$$\mathbb{P}(A) = \mathbb{P}(A \text{ und } B) + \mathbb{P}(A \text{ und } B^{\text{nicht}})$$

Diese Rechnung lässt sich kopieren, wenn man mehrere Aussagen  $B_1, B_2, \dots, B_N$  betrachtet, die die folgende Bedingung erfüllen:

$$\mathbb{P}(B_1 | A) + \mathbb{P}(B_2 | A) + \cdots + \mathbb{P}(B_N | A) = 1$$

Wenn sich die Aussagen  $B_1, B_2, \dots, B_N$  gegenseitig ausschliessen (engl. *mutually exclusive*) und ihre Wahrscheinlichkeiten zu eins aufsummieren, ist dies erfüllt. Dies bedeutet, dass die Aussagen so sind, dass (a) *wenn eine wahr ist, alle andern unwahr sind* und (b) *mindestens eine wahr ist*. Hier ein Beispiel dazu: Von einer nicht direkt messbaren Größe  $\mu$  weiß man, dass sie nur die Werte 3, 6 und 10 annehmen kann. Die Aussagen  $\mu = 3$ ,  $\mu = 6$  und  $\mu = 10$  schließen sich gegenseitig aus. Zudem ist

$$\mathbb{P}(\mu = 3 | \mathcal{I}) + \mathbb{P}(\mu = 6 | \mathcal{I}) + \mathbb{P}(\mu = 10 | \mathcal{I}) = 1$$

für alle Informationen  $\mathcal{I}$ . Die obige Rechnung hat damit gezeigt:

**Theorem 4.4 (Gesetz der totalen Wahrscheinlichkeit)**

Gegeben sind Aussagen  $B_1, B_2, B_3, \dots, B_N$  die sich gegenseitig ausschliessen und deren Wahrscheinlichkeiten, bei gegebener Information  $\mathcal{I}$ , zu eins aufsummieren. Dann gilt

$$\mathbb{P}(A | \mathcal{I}) = \mathbb{P}(A \text{ und } B_1 | \mathcal{I}) + \mathbb{P}(A \text{ und } B_2 | \mathcal{I}) + \cdots + \mathbb{P}(A \text{ und } B_N | \mathcal{I})$$

Es ist nun möglich, die Multiplikationsregel an einem Beispiel zu illustrieren:

**Beispiel 4.2 (OptiMAL)** In Kap. 1 wird in Beispiel 1.1 erwähnt, dass OptiMAL ein schnelles und billiges Verfahren ist, um Personen auf Malaria zu testen. Um zu prüfen, wie

wirksam das Verfahren ist, wurden am Center for Disease Prevention an der University of Miami School of Medicine, 202 Personen untersucht. Die Resultate der Untersuchung sind in einer *Kreuztabelle* zusammengefasst:

	Malaria	ohne Malaria	Total
OptiMAL+	89	2	91
OptiMAL-	7	104	111
Total	96	106	202

Von den 202 untersuchten Personen haben 111 Personen negativ auf OptiMAL reagiert. Malaria hatten 96 Personen. Jemand möchte aus der Gruppe der 202 Personen durch zufälliges Ziehen eine Person auswählen und die Wahrscheinlichkeit bestimmen, dass sie Malaria hat. Es sei  $A$  die Aussage „Die Person hat Malaria.“ Gesucht ist damit  $\mathbb{P}(A)$ . Aus der Kreuztabelle scheint es sinnvoll<sup>7</sup>, da 96 von 202 Personen Malaria haben, die Wahrscheinlichkeit  $\mathbb{P}(A)$  mit 96/202 zu setzen:

$$\mathbb{P}(A) = \mathbb{P}(A \mid \text{Information Tabelle}) = \frac{96}{202} = 0,475$$

Hätte man dies auch mit den Gesetzen zur Wahrscheinlichkeit erhalten? Die Antwort ist ja. Es folgt dazu die Argumentation: Die 202 Personen kann man sich durchnummerniert von 1 bis 202 denken. Die Personen mit Malaria haben die Nummern 1 bis 96. Es bezeichne  $\mathbb{P}(j)$  die Wahrscheinlichkeit, dass die Person  $j$  gezogen wurde. Da man die Person „zufällig“ zieht, besitzt man keine Information, welche Person gewählt wurde. Damit muss  $\mathbb{P}(i) = \mathbb{P}(j)$  für alle  $i$  und  $j$  gelten.<sup>8</sup> Aus der Konvexität folgt damit<sup>9</sup>

$$\mathbb{P}(1) = \mathbb{P}(2) = \dots = \mathbb{P}(202) = \frac{1}{202}$$

Die Wahrscheinlichkeiten summieren sich zu eins. Zudem schliessen sich die Aussagen „Man hat die Person 1 gezogen“, „Man hat die Person 2 gezogen“, ... gegenseitig aus. Mit dem Gesetz der totalen Wahrscheinlichkeit folgt daher:

$$\mathbb{P}(A) = \mathbb{P}(A \text{ und } 1) + \mathbb{P}(A \text{ und } 2) + \dots + \mathbb{P}(A \text{ und } 202)$$

Die einzelnen Summanden kann man mit dem Multiplikationsgesetz bestimmen

$$\mathbb{P}(A \text{ und } j) = \mathbb{P}(A \mid j) \cdot \mathbb{P}(j)$$

---

<sup>7</sup> Im Sinn von gesundem Menschenverstand oder „common sense“.

<sup>8</sup> Wenn dies nicht gelten würde, also wenn beispielsweise  $\mathbb{P}(1) > \mathbb{P}(2)$  wäre, hätte man eine zusätzliche Information. Man müsste erklären, warum die Person 1 plausibler gezogen würde als die Person 2.

<sup>9</sup> Man nennt dies das Prinzip der *Indifferenz*.

Der erste Faktor auf der rechten Seite der Gleichung ist die Wahrscheinlichkeit, dass  $A$  wahr ist, gegeben die Nummer  $j$  der Person. Sie ist eins, wenn  $j = 1, 2, \dots, 96$  ist. Sonst ist sie null. Somit ist

$$\mathbb{P}(A) = \underbrace{1 \cdot \frac{1}{202} + \dots + 1 \cdot \frac{1}{202}}_{96 \times} + \underbrace{0 \cdot \frac{1}{202} + \dots + 0 \cdot \frac{1}{202}}_{106 \times} = \frac{96}{202}$$

Die drei Gesetze zur Wahrscheinlichkeitsrechnung führen zum erwarteten Resultat.

Die Wahrscheinlichkeit, dass  $A$  und die Aussage  $B$  = „Die gezogene Person hat positiv auf OptiMAL reagiert“ wahr sind, lässt sich direkt aus der Tabelle lesen:

$$\mathbb{P}(B \text{ und } A \mid \mathcal{I}) = \frac{89}{202} = 0,44$$

Mit dem Multiplikationsgesetz  $\mathbb{P}(B \text{ und } A) = \mathbb{P}(B \mid A) \cdot \mathbb{P}(A)$  lässt sich dies auch bestimmen. Der Ausdruck  $\mathbb{P}(B \mid A)$  ist die Wahrscheinlichkeit, dass  $B$  wahr ist, gegeben die Person hat Malaria. Gegeben ist also die Aussage  $A$  = „Die Person hat Malaria.“ Dies liest man aus der ersten Spalte der Tabelle:

	<b>Malaria</b>	ohne Malaria	Total
OptiMAL +	<b>89</b>	2	91
OptiMAL -	7	104	111
Total	<b>96</b>	106	202

Man erhält 89/96. Der zweite Faktor  $\mathbb{P}(A)$  ist gleich 96/202:

$$\mathbb{P}(B \text{ und } A) = \mathbb{P}(B \mid A) \cdot \mathbb{P}(A) = \frac{89}{96} \cdot \frac{96}{202} = \frac{89}{202} = 0,44$$

Man erhält das gleiche Resultat. □

Die nächsten Beispiele zeigen, wie man mit den drei Rechenregeln Wahrscheinlichkeiten ausrechnen kann.

**Beispiel 4.3 (Wetterprognose)** Eine Person hört am Radio die Wetternachrichten. Sie schliesst aus den Nachrichten: die Wahrscheinlichkeit ist 0,6, dass morgen an ihrem Wohnsitz starke Windböen auftreten werden (Aussage  $A$ ) und die Wahrscheinlichkeit beträgt 0,5, dass es morgen regnen wird (Aussage  $B$ ). Man hat also<sup>10</sup>

$$\mathbb{P}(A) = 0,6 \quad \mathbb{P}(B) = 0,5$$

<sup>10</sup> Eigentlich müsste man schreiben:  $\mathbb{P}(A \mid \mathcal{K}) = 0,6$  und  $\mathbb{P}(B \mid \mathcal{K}) = 0,5$ . Dabei ist  $\mathcal{K}$  die Information der Person, die die Wetternachrichten gehört hat.

Weiter rechnet die Person mit einer Wahrscheinlichkeit von 0,9, dass es starke Windböen haben wird, falls es regnet:  $\mathbb{P}(A | B) = 0,9$ . Damit ist

$$\mathbb{P}(A \text{ und } B) = \mathbb{P}(A | B) \cdot \mathbb{P}(B) = 0,9 \cdot 0,5 = 0,45$$

Wie gross ist die Wahrscheinlichkeit, dass es regnen wird, wenn man weiss, dass starke Windböen auftreten werden? Dies ist  $\mathbb{P}(B | A)$ . Man nennt dies die zu  $\mathbb{P}(A | B)$  inverse Wahrscheinlichkeit. Mit dem Multiplikationsgesetz erhält man

$$\underbrace{\mathbb{P}(A \text{ und } B)}_{=0,45} = \mathbb{P}(B \text{ und } A) = \mathbb{P}(B | A) \cdot \underbrace{\mathbb{P}(A)}_{=0,6}$$

Also ist  $\mathbb{P}(B | A) = 0,45 / 0,6 = 0,75$ . Die Wahrscheinlichkeiten  $\mathbb{P}(A | B)$ ,  $\mathbb{P}(B | A)$  und  $\mathbb{P}(A \text{ und } B)$  sind alle voneinander verschieden.

Die Wahrscheinlichkeiten kann man auch mit einer Tabelle berechnen:

	$A$	$A^{\text{nicht}}$	Total
$B$			
$B^{\text{nicht}}$			
Total			100

Die Werte  $\mathbb{P}(A) = 0,6$  und  $\mathbb{P}(B) = 0,5$  werden in der untersten Zeile und der Spalte ganz rechts eingetragen:

	$A$	$A^{\text{nicht}}$	Total
$B$			<b>50</b>
$B^{\text{nicht}}$			
Total	<b>60</b>		100

Weil  $\mathbb{P}(A | B) = 0,9$  ist, steht im obersten linken Feld die Zahl  $0,9 \cdot 50 = 45$ . Damit kann die Tabelle ausgefüllt werden:

	$A$	$A^{\text{nicht}}$	Total
$B$	<b>45</b>	5	<b>50</b>
$B^{\text{nicht}}$	15	35	50
Total	<b>60</b>	40	100

Aus der Tabelle liest man ab:  $\mathbb{P}(A \text{ und } B) = 45/100 = 0,45$ . □

**Beispiel 4.4 (Konsum und Arbeitslosigkeit)** Die Chefin der Produktionsabteilung einer Firma geht davon aus, dass die Anzahl der verkauften Konsumgüter im nächsten Geschäftsjahr davon abhängt, wie gross die Arbeitslosenrate ist. Sie rechnet mit einer

Wahrscheinlichkeit von 0,8, dass viele Konsumgüter verkauft werden können, wenn die Arbeitslosenrate tief ist. Die Wahrscheinlichkeit, dass viele Konsumgüter verkauft werden können, wenn die Arbeitslosenrate hoch ist, nennt sie mit 0,4. Weiter nimmt die Chef in an, dass eine Wahrscheinlichkeit von 0,7 besteht, dass die Arbeitslosenrate tief ist. Ist  $vK$  die Aussage „es werden viele Konsumgüter verkauft werden können“ und  $tA$  die Aussage „die Arbeitslosenrate ist tief“, so hat man also

$$\mathbb{P}(vK \mid tA) = 0,8 \quad \mathbb{P}(vK \mid tA^{\text{nicht}}) = 0,4 \quad \mathbb{P}(tA) = 0,7$$

Die Chef in der Produktionsabteilung kann daraus, die Wahrscheinlichkeit berechnen, dass nächstes Jahr viele Konsumgüter verkauft werden können. Die Aussagen  $tA$  und  $tA^{\text{nicht}}$  schliessen sich gegenseitig aus und es ist:

$$\mathbb{P}(tA) + \mathbb{P}(tA^{\text{nicht}}) = 1$$

Mit dem Gesetz der totalen Wahrscheinlichkeit und dem Multiplikationsgesetz ist daher

$$\begin{aligned} \mathbb{P}(vK) &= \mathbb{P}(vK \text{ und } tA) + \mathbb{P}(vK \text{ und } tA^{\text{nicht}}) \\ &= \mathbb{P}(vK \mid tA) \cdot \mathbb{P}(tA) + \mathbb{P}(vK \mid tA^{\text{nicht}}) \cdot \mathbb{P}(tA^{\text{nicht}}) \end{aligned}$$

Damit ist

$$\mathbb{P}(vK) = 0,8 \cdot 0,7 + 0,4 \cdot (1 - 0,7) = 0,68$$

Dieses Resultat kann man, wie bei den obigen Beispielen, auch mit einer Kreuztabelle rechnen:

	$vK$	$vK^{\text{nicht}}$	Total
$tA$			
$tA^{\text{nicht}}$			
Total			100

Weil  $\mathbb{P}(tA) = 0,7$  ist, kann man die letzte Spalte ausfüllen:

	$vK$	$vK^{\text{nicht}}$	Total
$tA$			<b>70</b>
$tA^{\text{nicht}}$			30
Total			100

Jetzt kann die erste Spalte aus den Wahrscheinlichkeiten  $\mathbb{P}(vK \mid tA) = 0,8$  und  $\mathbb{P}(vK \mid tA^{\text{nicht}}) = 0,4$  bestimmt werden:  $0,8 \cdot 70$  ist gleich 56 und  $0,4 \cdot 30$  ist gleich 12:

	$vK$	$vK^{\text{nicht}}$	Total
$tA$	<b>56</b>		70
$tA^{\text{nicht}}$	<b>12</b>		30
Total			100

Die leeren Felder lassen sich nun ausfüllen. Die unterste Zahl  $56 + 12$  in der ersten Spalte liefert die gesuchte Wahrscheinlichkeit:

$$\mathbb{P}(vK) = \frac{56 + 12}{100} = 0,68$$

Rechnungen zu Wahrscheinlichkeiten lassen sich hier also mit den Rechengesetzen oder mit Tabellen durchführen.  $\square$

In der Medizin und in Wettbüros arbeitet man statt mit Wahrscheinlichkeiten auch mit dem Begriff *Chance* (engl. *odd*). Die Chance einer Aussage  $A$  ist der Quotient aus der Wahrscheinlichkeit, dass die Aussage wahr ist, und der Wahrscheinlichkeit, dass sie unwahr ist:

$$\text{Chance von } A = \mathbb{O}(A) = \frac{\mathbb{P}(A)}{\mathbb{P}(A^{\text{nicht}})} = \frac{\mathbb{P}(A)}{1 - \mathbb{P}(A)}$$

Beträgt die Wahrscheinlichkeit 0,75, dass ein Medikament wirkt, so ist seine Heilchance gleich  $0,75/(1 - 0,75) = 3$ , oder 3:1. Analog berechnet man die Wahrscheinlichkeit  $\mathbb{P}(A)$  aus der Chance von  $A$  durch

$$\mathbb{P}(A) = \frac{\mathbb{O}(A)}{1 + \mathbb{O}(A)}$$

Zum Schluss des Abschnitts folgt noch eine Definition:

#### Definition 4.1

Man nennt zwei Aussagen  $A$  und  $B$  – bei gegebener Information  $\mathcal{I}$  – *unabhängig* (engl. *independent*), wenn

$$\mathbb{P}(A \mid B \text{ und } \mathcal{I}) = \mathbb{P}(A \mid \mathcal{I})$$

Die Aussagen  $A$  und  $B$  sind also bei gegebener Information unabhängig, wenn die Plausibilität von  $A$  unabhängig davon ist, ob die Kenntnis von  $B$  vorhanden ist. Zu beachten ist auch, dass zwei Aussagen je bei gegebener Information unabhängig oder abhängig sein können!<sup>11</sup> Bei unabhängigen Aussagen vereinfacht sich das Multiplikationsgesetz zu:

#### Theorem 4.5

*Sind  $A$  und  $B$  bei gegebener Information  $\mathcal{I}$  unabhängig, so hat man*

$$\mathbb{P}(A \text{ und } B \mid \mathcal{I}) = \mathbb{P}(A \mid \mathcal{I}) \cdot \mathbb{P}(B \mid \mathcal{I})$$

---

<sup>11</sup> Wahrscheinlichkeiten drücken aus, wie plausibel Aussagen bei gegebener Information sind. Sie charakterisieren keine Eigenschaften der in den Aussagen vorkommenden Objekten. Vergleichen Sie dazu das Beispiel zum Münzwurf.

Bei unabhängigen Aussagen  $A$  und  $B$  ist also die Wahrscheinlichkeit, dass  $A$  und  $B$  wahr ist, das Produkt von  $\mathbb{P}(A)$  und von  $\mathbb{P}(B)$ .

Im obigen Beispiel zu den Wetterprognosen, sind die Aussagen  $A = \text{„Es werden starke Windböen auftreten“}$  und  $B = \text{„Es wird regnen“}$  bei der vorhandenen Information nicht unabhängig. Man hat nämlich

$$\mathbb{P}(A) = 0,6 \neq \mathbb{P}(A | B) = 0,9$$

Das Wissen der Person, dass es regnen wird, erhöht die Plausibilität von starken Windböen. Zu beachten ist: Sich ausschliessende Aussagen, wie  $A = \text{„Morgen regnet es“}$ , und  $B = \text{„Morgen regnet es nicht“}$ , sind in der Regel nicht unabhängig. Ist etwa  $\mathbb{P}(A) = \mathbb{P}(B) = 0,5$ , so ist  $\mathbb{P}(A) \neq \mathbb{P}(A | B) = 0$ .

## 4.2 Eine andere Interpretation der Wahrscheinlichkeit

In diesem Buch werden *Plausibilitäten* zu Aussagen, gegeben eine Information, mit *Wahrscheinlichkeiten* bezeichnet. Man spricht vom *bayesschen* Wahrscheinlichkeitsbegriff. Verbreitet ist auch eine andere Interpretation: die *frequentistische* Wahrscheinlichkeit. Sie versteht Wahrscheinlichkeiten zu Aussagen als *Anteile* oder *Häufigkeiten* (engl. *frequencies*). Sind  $T$  Fahrzeiten von Zügen, die im nächsten Jahr von Zürich nach Bern fahren, so bedeutet mit der frequentistischen Sichtweise die Wahrscheinlichkeit  $\mathbb{P}(T \leq 65') = 0,88$ , dass 88 % aller Züge von Zürich nach Bern des nächsten Jahres eine Fahrzeit von höchsten 65 Minuten haben werden. Ein zweites Beispiel ist: Jemand wirft eine Münze mit Seiten Kopf und Zahl. Ist  $K$  die Aussage „Die Münze wird auf Kopf fallen“, dann meint die frequentistische Sichtweise

$$\mathbb{P}(K) = 0,5$$

das Folgende: Würde man die Münze sehr oft werfen, würden etwa 50 % der Würfe auf Kopf fallen. Man quantifiziert mit dieser Auffassung also nicht die Information, dass man nicht weiß, ob die Münze auf Kopf oder Zahl fallen wird. Vielmehr beinhaltet diese Interpretation eine viel stärkere Aussage: Wenn man das Experiment unendlich (sic!) mal wiederholen würde, wäre der Anteil Kopf 50 %. Man misst also eine Langzeit-Häufigkeit und eine Eigenschaft der Münze.<sup>12</sup> Mit der frequentistischen Sicht lassen sich

<sup>12</sup> Siehe dazu E. T. Jaynes in [11] auf den Seiten 335–336:

,When I toss a coin, the probability for heads is one-half.‘ What do we mean by this statement? [...] The issue is between the following two interpretations:

- (A) *The available information gives me no reason to expect heads rather than tails, or vice versa – I am completely unable to predict which it will be.*
- (B) *If I toss the coin a very large number of times, in the long run heads will occur about half the time – in other words, the frequency of heads will approach 1/2.*

Wahrscheinlichkeiten nur für Ereignisse formulieren, die (theoretisch) unendlich oft wiederholbar sind.

**Beispiel 4.5 (Narkoserisiko)** Eine Vollnarkose ist nicht ungefährlich. Sie kann zu anhaltenden Schäden wie Atemproblemen oder Gedächtnissstörungen führen. Selten sind auch narkosebedingte Todesfälle möglich. So hat man gemäss [15] für die Wahrscheinlichkeit  $\mathbb{P}(\text{Tod}) \approx 1/200\,000$  bei einer Vollnarkose zu sterben, die in Deutschland durchgeführt wird. Deutet man Wahrscheinlichkeiten frequentistisch, so bedeutet dies: Auf 1 000 000 Vollnarkosen werden ungefähr  $1/200\,000 \times 1\,000\,000 = 5$  Patienten nicht überleben. Für einen einzelnen Patienten  $XY$  liefert dies keine Information. Anders ausgedrückt: Interpretiert man Wahrscheinlichkeiten frequentistisch als Anteile oder Häufigkeiten zu Menschen oder Objekten, so können Vorhersagen für Gruppen von Menschen oder Objekten gemacht werden. Für Einzelpersonen oder einzelne ausgewählte Objekte liefern sie in diesem Sinn keine Information.

Deutet man die Wahrscheinlichkeit als Informationsaussage oder Plausibilität, wie gefährlich eine Vollnarkose ist, so ist die Zahl von 1/200 000 für den Patienten  $XY$  wichtig:

$$\mathbb{P}(XY \text{ stirbt} \mid \text{Wissen } \mathcal{K} \text{ vor Operation}) \approx 1/200\,000$$

Damit kann eine Person abschätzen, ob Sie das Risiko eingehen will, sich einer Vollnarkose zu unterziehen. *Nach* der Operation, also mit zusätzlicher Information  $\mathcal{L}$ , kann die Person  $\mathbb{P}(XY \text{ stirbt} \mid \mathcal{L}) = 0$  setzen.  $\square$

Den Informationsstand zu Aussagen mit Plausibilitäten auszudrücken, ist für Konsumenten und Ingenieurinnen zentral: „Die Wahrscheinlichkeit ist 0,9, dass *mein* gekaufter Fotoapparat mehr als fünf Jahre funktionieren wird“, „Die Wahrscheinlichkeit ist 0,95, dass der Unterdruck in der *konstruierten* Vakuumkammer zwischen 0,589 und 0,593 bar liegt“, oder „Die Wahrscheinlichkeit beträgt 0,95, dass der morgen fahrende IC-Zug 8021 weniger als 30 Minuten Verspätung hat.“ Frequentistische Interpretationen der zwei ersten Aussagen, wie „90 von 100 gleichen Fotoapparate, wie ich gekauft habe, werden mehr als fünf Jahre funktionieren“, oder wie „Auf 1000 Vakuumkammern der absolut gleichen (!) Bauart, wird der Unterdruck in 950 Kammern zwischen 0,589 und 0,593 bar liegen“, scheinen wenig sinnvoll. Die dritte Aussage kann nicht frequentistisch als Langzeitzrate gedeutet werden, fährt doch der IC-Zug 8021 nur einmal. Die Wahrscheinlichkeiten, dass die nächste eintreffende E-Mail Spam enthält oder dass sich der Ozeanspiegel in den folgenden fünfzig Jahren um 50 cm erhöht, lassen sich nicht frequentistisch, sondern nur bayesianisch formulieren.

---

Statement (A) does not describe any property of the coin, but only the *state of knowledge* (or, if you prefer, the state of ignorance). Statement (B) is, at least by implication, asserting something about the coin. Thus (B) is a very much stronger statement than (A).

### 4.3 Wahrscheinlichkeitsmodelle

Im ersten Abschnitt des Kapitels wird die Plausibilität von Aussagen, wie „Die Person hat Malaria“, anhand der gegebenen Information mit einer Wahrscheinlichkeit quantifiziert. In der Wirtschaftswissenschaft und im Ingenieurwesen müssen Rechnungen zu nicht direkt messbaren Größen mit einer Plausibilität versehen werden. So möchte ein Ingenieur Drücke, Temperaturen oder Stromstärken aus gemessenen Werten bestimmen. Eine Ökonomin möchte die Arbeitslosenrate des nächsten Jahres der Schweiz prognostizieren. Dazu braucht man ein Wahrscheinlichkeitsmodell. Es ist sinnvoll, zwei Fälle zu unterscheiden:

1. *Nicht direkt messbare Größen (Parameter) berechnen:* Auf der Basis von beobachteten Zeiten zwischen starken Erdbeben, soll die durchschnittliche Zeit zwischen zukünftigen, starken Erdbeben berechnet werden. Der Druck in einer Kammer soll bestimmt werden. Wegen vorhandener Kovariablen kann er nicht direkt gemessen werden.<sup>13</sup> Solche nicht direkt messbaren Größen kann man aus Daten oder anderen Informationen berechnen. Es lässt sich sagen, wie genau und wie *plausibel* das Resultat ist. Ein Beispiel:

$$\mathbb{P}(450 \text{ Tage} \leq \text{durchschn. Wartezeit} \leq 500 \text{ Tage} \mid \text{Information}) = 0,9$$

Man sagt, dass dies ein Resultat aus der *schliessenden Statistik* (engl. *statistical inference*) ist.

2. *Zukünftige Werte einer unsicheren Größe prognostizieren:* Solche Werte können manchmal mit einem physikalischen Gesetz prognostiziert werden, wenn genügend Information vorhanden ist. Ist dies nicht der Fall, begnügt man sich, Prognosen mit Wahrscheinlichkeiten zu formulieren. Typischerweise trifft man diese Situation in Produktionsprozessen an. So kann man beim Beispiel 1.7 zur Produktion vom Blumenzwiebeln wegen variierender Produktionsbedingungen kaum genau prognostizieren, wie viele nicht keimende Blumenzwiebeln (als *NK* bezeichnet) eine zukünftige Kiste enthalten wird. Vielmehr wird man Aussagen der Form

$$\mathbb{P}(5 \leq NK \leq 10 \mid \text{Information}) = 0,85$$

formulieren. Beschreibt man zukünftige Werte von unsicheren Größen, so spricht man auch von Werten von *Zufallsgrößen* oder *Zufallsvariablen* (engl. *random variable*). Man sagt, dass man eine *Prognose* (engl. *Prediction*) rechnet.

Wahrscheinlichkeitsmodelle werden also benutzt, um zu zeigen, wie plausibel Parameter aus Daten berechnet sind oder um zu prognostizieren, wo zukünftige Werte einer

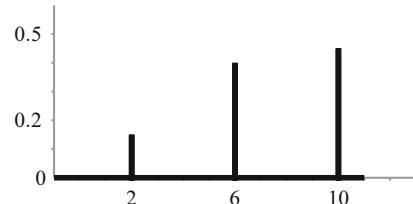
---

<sup>13</sup> Kovariablen sind hier etwa Luftturbulenzen, Inhomogenitäten in der Luft oder Ungenauigkeiten in den Messapparaten. Diese bewirken, dass man Messwerte erhält, die um den gesuchten Druck streuen.

**Abb. 4.1** Graph der Massen-

funktion der Zufallsgrösse

$W_{\text{Schluss}}$ : die *Höhe* der Stäbe ist  
die Wahrscheinlichkeit



Zufallsgrösse liegen werden. Im Folgenden werden zwei Fälle unterschieden: Größen oder Parameter, die nur diskrete Werte, wie ganze Zahlen annehmen, und solche, die kontinuierliche Werte annehmen können.

**Definition 4.2 (Diskretes Wahrscheinlichkeitsmodell)**

Eine unsichere Grösse oder ein Parameter  $G$ , der nur diskrete Werte, wie ganze Zahlen annehmen kann, kann mit einem *diskreten* Wahrscheinlichkeitsmodell beschrieben werden. Dieses gibt an, mit welcher Wahrscheinlichkeit  $p_i$  jeder Wert  $x_i$  angenommen wird:  $\mathbb{P}(G = x_i) = p_i$ . Man nennt dies die *Massenfunktion* (engl. *probability mass function*) des Modells.

Auf Größen des zweiten Falls wird unten eingegangen.

**Beispiel 4.6 (Tagesschlusskurs einer Aktie)** Eine Börsenhändlerin prognostiziert aus ihrer Information  $\mathcal{K}$ , dass der Tagesschlusskurs  $W_{\text{Schluss}}$  einer Aktie CHF 2.– mit Wahrscheinlichkeit 0,15, CHF 6.– mit einer Wahrscheinlichkeit von 0,40 und CHF 10.– mit einer Wahrscheinlichkeit von 0,45 sein wird. Dies ist die Massenfunktion des diskreten Wahrscheinlichkeitsmodells für die zukünftigen Werte des unsicheren Tagesschlusskurses. Ihr Graph ist in Abb. 4.1 gezeichnet. Die Summe aller Höhen der Stäbe beträgt wegen der Konvexität und der Additionsregel eins. Wahrscheinlichkeiten zum Tagesschlusskurs kann man mit der Massenfunktion leicht berechnen, indem man nach der Additionsregel die Höhen der Stäbe aufsummieren. So ist, gegeben das Wissen  $\mathcal{K}$

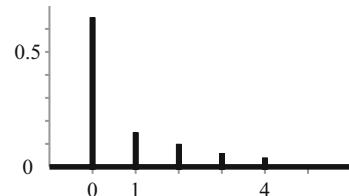
$$\mathbb{P}(W_{\text{Schluss}} = 2 \text{ oder } 6) = \mathbb{P}(W_{\text{Schluss}} = 2) + \mathbb{P}(W_{\text{Schluss}} = 6) = 0,55$$

Analog ist die Wahrscheinlichkeit, dass der Tagesschlusskurs höchstens CHF 3.– ist:

$$\mathbb{P}(W_{\text{Schluss}} \leq 3 \mid \mathcal{K}) = \mathbb{P}(W_{\text{Schluss}} = 2 \mid \mathcal{K}) = 0,15$$

Der grösste Wert der Massenfunktion beträgt 0,45. Er befindet sich beim Wert CHF 10.–. Dies nennt man den wahrscheinlichsten Wert oder den *Modus* (engl. *mode*) des Wahrscheinlichkeitsmodells.  $\square$

**Abb. 4.2** Massenfunktion für die Anzahl defekter Knallkörper per



**Beispiel 4.7 (Anzahl defekter Knallkörper)** Eine Firma produziert Schachteln mit vier Knallkörpern. Eine Person beschreibt aus ihrer Information  $\mathcal{K}$  die Plausibilität zur Anzahl  $N$  defekter Knallkörper in der von ihr gekauften Kiste. Die Anzahl  $N$  ist eine diskrete Grösse und nicht direkt messbar. Sie kann Werte 0, 1, 2, 3 und 4 annehmen. Die Person wählt daher ein diskretes Wahrscheinlichkeitsmodell. Als Massenfunktion nimmt sie:

$$\begin{aligned}\mathbb{P}(N = 0 | \mathcal{K}) &= 0,65, & \mathbb{P}(N = 1 | \mathcal{K}) &= 0,15, & \mathbb{P}(N = 2 | \mathcal{K}) &= 0,10 \\ \mathbb{P}(N = 3 | \mathcal{K}) &= 0,06, & \mathbb{P}(N = 4 | \mathcal{K}) &= 0,04\end{aligned}$$

Abb. 4.2 zeigt den Graphen der Massenfunktion. Die Wahrscheinlichkeit, mehr als zwei defekte Knallkörper in der Schachtel zu finden, beträgt  $0,06 + 0,04 = 0,10$ . Der wahrscheinlichste Wert für  $N$  ist null. Dies ist der Modus des Modells.  $\square$

**Beispiel 4.8 (Roulette und andere Glücksspiele)** Spieler gehen bei vielen Glücksspielen davon aus, dass man nicht weiss, welche Zahl als Nächste gezogen wird. Diese „Indifferenz“ bedeutet, dass Spieler jede zukünftige Zahl  $Z$ , die gezogen werden kann, mit gleicher Wahrscheinlichkeit belegen. Für das Roulette mit Zahlen von 0 bis 36 erhält man die (diskrete) *Gleichverteilung* (engl. *uniform distribution*) mit der Massenfunktion

$$\mathbb{P}(Z = k | \text{Indifferenz}) = 1/37 \quad \text{für } k = 0, 1, 2, \dots, 36$$

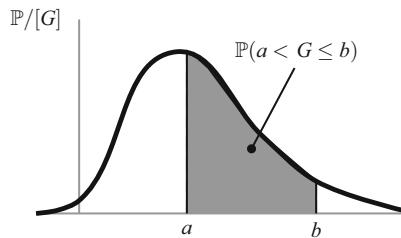
Der Graph der Massenfunktion besteht aus 37 Stäben mit Höhe 1/37.  $\square$

Wahrscheinlichkeiten zu Parametern oder zu Werten von Zufallsgrössen, die kontinuierliche Werte annehmen können, können nicht mit der Massenfunktion angegeben werden. So müssten für den Graph der Massenfunktion über allen diesen unendlich vielen Zahlen Stäbe gebildet werden. Dies ist wenig praktisch. Man geht anders vor:

**Definition 4.3 (Stetiges Wahrscheinlichkeitsmodell)**

Die Plausibilität zu Parametern oder zu Werten von unsicheren Grössen  $G$ , die kontinuierliche Werte annehmen können, wird meist mit einem *stetigen* (engl. *continuous*) Wahrscheinlichkeitsmodell, einer Dichtefunktion, beschrieben. Die

**Abb. 4.3** Ein stetiges Wahrscheinlichkeitsmodell, um die Plausibilität zu einer Grösse  $G$  zu beschreiben



Dichtefunktion (engl. *probability density function* (pdf)) ist eine positive Funktion, so dass die Fläche zwischen ihrem Graphen und der  $x$ -Achse gleich eins ist.<sup>14</sup> Die Wahrscheinlichkeit, dass die Aussage „ $G$  ist zwischen  $a$  und  $b$ “, wahr ist, ist gleich der Fläche unter dem Graphen der Funktion zwischen  $a$  und  $b$ . Abb. 4.3 zeigt die Situation.

Mit der Dichtefunktion  $\text{pdf}(x)$  lässt sich schnell die Wahrscheinlichkeit berechnen, dass die Aussage „ $G$  liegt zwischen  $x$  und  $x + \Delta x$ “, wahr ist. Ist  $\Delta x$  klein, so kann man die Fläche durch eine Rechtecksfläche mit Höhe  $\text{pdf}(x)$  und Breite  $\Delta x$  approximieren:

$$\mathbb{P}(x \leq G \leq x + \Delta x) \approx \text{pdf}(x) \cdot \Delta x \quad (4.1)$$

Mit anderen Worten: Die Wahrscheinlichkeit, dass  $G$  etwa  $x$  ist, ist *proportional* zum Wert  $\text{pdf}(x)$  der Dichtefunktion.

**Beispiel 4.9 (Zerfallszeit von Radon)** Radon ist ein Edelgas und besitzt radioaktive Isotope, die mit der Zeit zerfallen. Die Zerfallszeit  $T$  eines Radonisotops wie  $^{222}\text{Rn}$  ist nicht genau voraussehbar. Sie ist eine kontinuierliche Grösse und kann beliebige Werte grösser als Null annehmen. Daher wird die Plausibilität zu  $T$  mit einem stetigen Wahrscheinlichkeitsmodell beschrieben. Experimente zeigen, dass die zukünftige Zerfallszeit mit der Dichtefunktion

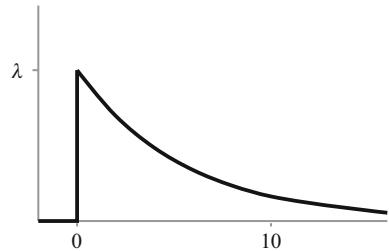
$$\text{pdf}(T = x \mid \lambda) = \lambda \cdot \exp(-\lambda \cdot x) \quad \text{für } x \geq 0$$

mit  $\lambda = 1/5,515 \text{ Tage}^{-1}$  gut modelliert wird. In Abb. 4.4 findet sich ihr Graph. Die Schreibweise  $\text{pdf}(T = x \mid \lambda)$  betont, dass  $\lambda$  bekannt ist:  $\text{pdf}$  ist die Dichtefunktion, *gegeben*  $\lambda$ . Das Modell nennt man *Exponentialverteilung* mit Rate (engl. *rate*)  $\lambda$ . In mathematischer Form wird die Plausibilität zu  $T$  mit der Exponentialverteilung wie folgt geschrieben:

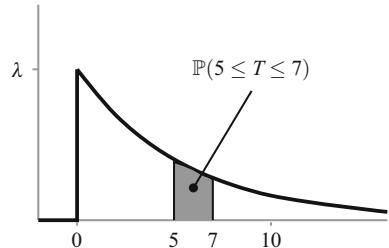
$$T \sim \text{Exponential}(\lambda)$$

<sup>14</sup> Die Einheit der Dichtefunktion ist Wahrscheinlichkeit pro Einheit von  $G$ .

**Abb. 4.4** Graph der Dichtefunktion der Exponentialverteilung



**Abb. 4.5** Wahrscheinlichkeit, dass  $T$  zwischen 5 und 7 Tagen ist



Diese Notation ist auch in Statistikprogrammen verbreitet. Die Wahrscheinlichkeit, dass die Zerfallszeit eines Radon Isotops zwischen fünf und sieben Tagen betragen wird, ist gleich der Fläche unter dem Graphen der Dichtefunktion zwischen fünf und sieben Tagen (siehe Abb. 4.5). Sie lässt sich mit einem Taschenrechner rechnen:

$$\mathbb{P}(5 \text{ Tage} \leq T \leq 7 \text{ Tage} | \lambda) = \int_{5 \text{ Tage}}^{7 \text{ Tage}} \lambda \cdot \exp(-\lambda \cdot x) dx = 0,123$$

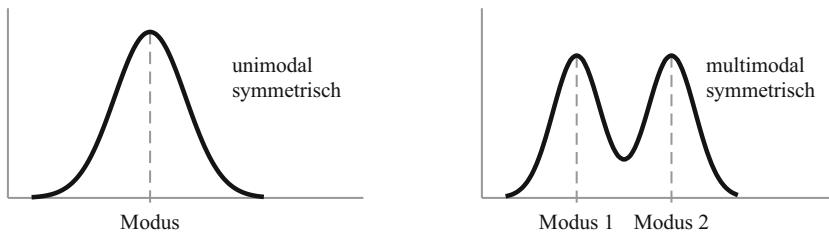
Die Wahrscheinlichkeit, dass die Zerfallszeit grösser als 8 Tage ist, beträgt

$$\mathbb{P}(T \geq 8 \text{ Tage} | \lambda) = \int_{8 \text{ Tage}}^{\infty} \lambda \cdot \exp(-\lambda \cdot x) dx = 0,234$$

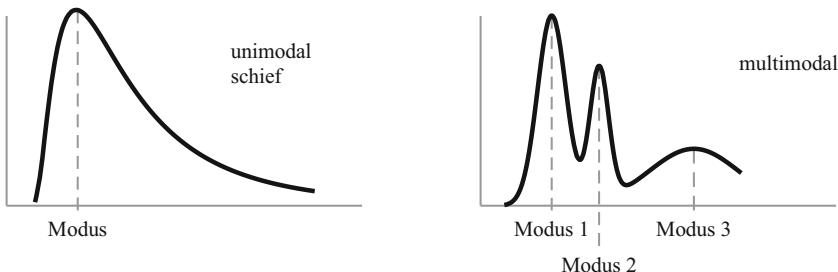
Die Wahrscheinlichkeit, eine Zerfallszeit zwischen 10 und 10,5 Tagen zu messen, ist 0,0141.  $\square$

Anhand des Graphen der Dichtefunktion unterscheidet man verschiedene Typen von Verteilungen: (a) symmetrische Verteilungen (unimodal oder multimodal), wie in Abb. 4.6 und (b) schiefe Verteilungen und multimodale Verteilungen, wie in Abb. 4.7. Stellen, wo die Dichtefunktion lokale Maxima liefert, nennt man *Modalwerte* oder *Moden* (engl. *mode*) des Modells.<sup>15</sup> Bei einem diskreten Modell wird der Wert mit der grössten Wahrscheinlichkeit als Modus bezeichnet.

<sup>15</sup> Sie lassen sich mit einem Computer schnell finden: man zeichnet den Graphen der Dichtefunktion und liest sie aus der Grafik ab.



**Abb. 4.6** Zwei symmetrische Verteilungen



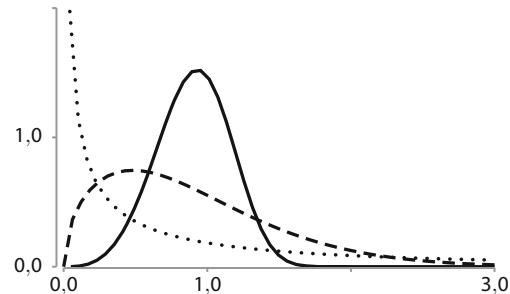
**Abb. 4.7** Zwei schiefe Verteilungen mit ihren Moden

**Beispiel 4.10 (Siebanalyse von Teilchen)** Stoffmengen aus Teilchen, wie Kiesteilchen, gemahlener Zucker oder Kaffeebohnen werden im Bereich der Verfahrenstechnik analysiert. Die Partikel- oder Korngrößen von solchen unregelmässig geformten Objekten wird operativ meist als Durchmesser einer Kugel, die die gleiche Masse wie das Teilchen hat, angegeben. Ein Ingenieur kann mit der vorhandenen Information kaum die Korngrößen der einzelnen Teilchen in der Stoffmenge prognostizieren. Er wird daher Aussagen zu den unsicheren Korngrößen mit Wahrscheinlichkeiten formulieren. Ein Blick in Bücher zur Verfahrenstechnik zeigt, dass die Information zur Teilchengrösse  $G$  meist mit einer RRSB-Verteilung (nach Rosin, Rammler, Sperling und Bennet in [16]) oder Weibull-Verteilung, beschrieben wird. Die Dichtefunktion des Modells lautet:

$$\text{pdf}(G = x \mid d, k) = \frac{k}{d} \cdot (x/d)^{k-1} \cdot \exp(-[x/d]^k) \quad \text{für } x \geq 0$$

Dabei sind  $k$  und  $d$  positive Konstanten. Man nennt  $k$  die Form (engl. *shape*) und  $d$  die Skalierung (engl. *scale*). Die Skalierung  $d$  gibt bei Filtern die Maschenweite an, durch die 63,2 % der Teilchen fallen. Die Zahl  $k$  modelliert die Form der Verteilung. Dies illustrieren die Graphen der Dichtefunktion in Abb. 4.8. Bei der gepunkteten Linie sind  $k = 0,5$ ,  $d = 1$ ; bei der gestrichelten Linie sind  $k = 1,5$ ,  $d = 1$  und bei der ausgezogenen Linie sind  $k = 3$  und  $d = 1$ . Für  $k = 3$  und  $d = 1$  ist die Verteilung unimodal symmetrisch. Der Modus beträgt 0,874. Bei  $k = 1,5$  und  $d = 1$  ist das Modell unimodal schief.  $\square$

**Abb. 4.8** Verschiedene Graphen der Dichte von Weibull-Verteilungen, um Korngrößen von Teilchensystemen zu modellieren



**Beispiel 4.11 (Zeit zwischen starken Erdbeben)** In Kap. 1 in Beispiel 1.3 wird grob skizziert, wie die nicht direkt messbare durchschnittliche Zeit  $\mu$  zwischen aufeinander folgenden, zukünftigen, grossen Erdbeben berechnet werden kann. Die Rechnung basiert auf Daten und Vorinformation  $\mathcal{K}$ . Das Resultat zu  $\mu$  ist daher unsicher. Da dieser Parameter kontinuierliche Werte grösser als Null annehmen könnte, wird die Plausibilität zu  $\mu$  mit einem stetigen Wahrscheinlichkeitsmodell beschrieben. In Kap. 8 wird gezeigt, dass

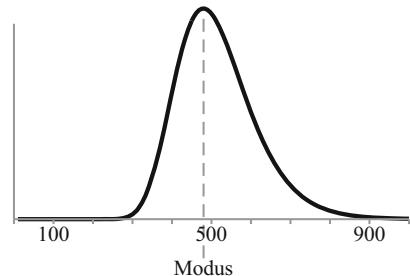
$$\text{pdf}(\mu = x \mid \text{Daten}, \mathcal{K}) = \frac{9,4767 \cdot 10^{87}}{x^{29}} \cdot \exp\{-13\,910,55 \text{ Tage}/x\}$$

eine gute Dichtefunktion ist. Ihr Graph findet sich in Abb. 4.9. Der Modus befindet sich dort, wo der Graph der Dichtefunktion das Maximum hat. An einer Computergrafik abgelesen, beträgt er 479,7 Tage. Man nennt dies auch den „plausibelsten“ Wert für  $\mu$ . Aus der Grafik sieht man, dass  $\mu$  mit hoher Wahrscheinlichkeit zwischen 300 und 900 Tagen ist. Mit der Fläche unter der Grafik kann man berechnen, wie gross die Wahrscheinlichkeit ist, dass  $\mu$  zwischen 400 und 600 Tagen ist:

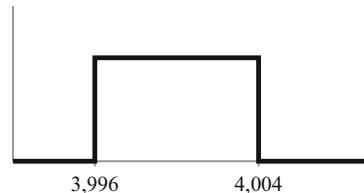
$$\mathbb{P}(400 \leq \mu \leq 600 \mid \text{Daten}, \mathcal{K}) = \int_{400 \text{ Tage}}^{600 \text{ Tage}} \text{pdf}(\mu = x \mid \text{Daten}, \mathcal{K}) dx = 0,72$$

Die Wahrscheinlichkeit ist also 0,72. □

**Abb. 4.9** Plausibilität zur durchschnittlichen Zeit zwischen aufeinanderfolgenden starken Erdbeben



**Abb. 4.10** Dichtefunktion einer Gleichverteilung (Die Fläche des Rechtecks ist eins.)



**Beispiel 4.12 (Messgerät)** Beim Beispiel 1.2 vorgestellten Messgerät zeigt die Gebrauchsleitung in Tab. 1.1, dass gemessene Frequenzen im Bereich von 4 Hz in einer Ordnung von  $\pm 0,10\%$ , also um  $\pm 0,004$  Hz schwanken. Ist  $B$  die Aussage „Die Frequenz des Stromkreises ist 4 Hz“, so beschreibt sie Wahrscheinlichkeiten der Form

$$\mathbb{P}(\text{angezeigte Frequenzen sind im Bereich } XY \mid B)$$

Da angezeigte Frequenzen kontinuierliche Werte annehmen können, beschreibt man sie daher mit einem stetigen Wahrscheinlichkeitsmodell. Man geht davon aus, dass jeder Wert in diesem Bereich *etwa* mit der gleichen Wahrscheinlichkeit auftreten kann. Man modelliert angezeigte Frequenzen  $f_{\text{Messgerät}}$  daher mit einer konstanten Dichtefunktion, wie sie in Abb. 4.10 dargestellt ist. Die Höhe der vertikalen Linie ist so gewählt, dass die Fläche unter dem Graphen gleich eins wird. Wahrscheinlichkeiten kann man einfach berechnen, sind doch die dabei zu berechnenden Flächen Rechtecke. Man nennt dieses Modell eine (stetige) *Gleichverteilung* (engl. *uniform distribution*) auf dem Intervall  $[3,996; 4,004]$ . In Kurzform schreibt man dies:

$$f_{\text{Messgerät}} \sim \text{Uniform}(3,996; 4,004)$$

Die Gleichverteilung ist symmetrisch und hat keinen Modus. □

Zum Schluss dieses Abschnitts eine Bemerkung:

Es ist wichtig sich zu überlegen, ob eine Grösse mit einem *diskreten* oder einem *stetigen* Wahrscheinlichkeitsmodell beschrieben werden soll. Im ersten Fall hat man die *Massenfunktion*, die direkt Wahrscheinlichkeiten ausdrückt. Im zweiten Fall benutzt man eine *Dichtefunktion* und Wahrscheinlichkeiten müssen mit Flächen berechnet werden.

## 4.4 Wie kann man Informationen aus Wahrscheinlichkeitsmodellen zusammenfassen?

Informationen zu nicht direkt messbaren Größen – wie der Druck in einer Vakuumkammer oder die durchschnittlich erwartbare Wartezeit vor einem Schalter einer Bank – werden mit Wahrscheinlichkeitsmodellen dargestellt. Dabei stellt sich die Frage, wie man Plausibilitäten aufschreiben soll. Man kann das Wahrscheinlichkeitsmodell grafisch darstellen. Gerade wenn die Verteilung multimodal ist, ist dies am überzeugendsten. Man kann auch nur den Modus oder die Moden angeben:

$$\text{plausibelster Wert} = \text{Modus der Verteilung}$$

Fasst man die Plausibilität zu einem Parameter mit nur einer Zahl zusammen, so sagt man, dass der gesuchte Parameter durch diesen Wert *geschätzt* (engl. *estimated*) wird. Beliebte Markierungen für Schätzungen eines Parameters  $K$  sind  $K_0$  (Index null) oder  $\hat{K}$  (Hut-Symbol).

**Beispiel 4.13 (Zeit zwischen starken Erdbeben)** Beim obigen Beispiel wird die Plausibilität zur durchschnittlichen Zeit  $\mu$  zwischen zukünftigen, starken Erdbeben mit einem unimodalen und schiefen Wahrscheinlichkeitsmodell beschrieben. Der Modus beträgt 479,7 Tage. Dies schätzt die nicht direkt messbare Größe  $\mu$ .  $\square$

Bei multimodalen Verteilungen wird man die verschiedenen Moden notieren. Eine andere Zahl, um eine Verteilung zu charakterisieren, ist der *Median* oder der *Zentralwert* (engl. *median*). Der Median besagt grob, dass die Wahrscheinlichkeit, dass die Aussagen „Die Größe  $G$  ist größer als der Median“, und „Die Größe  $G$  ist kleiner oder gleich dem Median“, wahr sind, je 0,5 beträgt. Genauer ist der Median die kleinste Zahl, sodass die Aussage „Die Größe  $G$  ist kleiner oder gleich dem Median“, eine Wahrscheinlichkeit von mindestens 0,5 hat.<sup>16</sup> Hier Beispiele dazu:

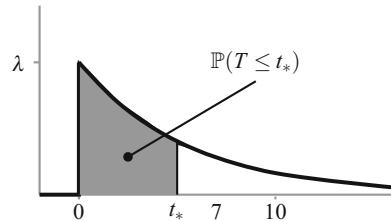
**Beispiel 4.14 (Zerfallszeit von Radon)** Eine zukünftige Zerfallszeit eines  $^{222}\text{Rn}$ -Isotops wird mit der Exponentialverteilung mit Rate  $\lambda = 0,181 \text{ Tage}^{-1}$  prognostiziert:

$$T \sim \text{Exponential}(0,181 \text{ Tage}^{-1})$$

Die Wahrscheinlichkeit, dass sie höchstens  $t_*$  wird, ist gleich der Fläche unter dem Graphen der Dichtefunktion zwischen null und  $t_*$ . Dies illustriert Abb. 4.11. Der Median

<sup>16</sup> Mit einer Formel ausgedrückt hat man: Der Median ist die kleinste Zahl mit  $\mathbb{P}(G \leq \text{Median})$  ist mindestens 0,5.

**Abb. 4.11** Experimentell den Median bestimmen



entspricht dem Wert von  $t_*$ , so dass diese Fläche gleich 0,5 ist. Dazu muss man eine Gleichung für  $t_*$  lösen, die ein Integral enthält:

$$\int_0^{t_*} \text{pdf}(x | \lambda) dx = \int_0^{t_*} \lambda \cdot \exp(-\lambda \cdot x) dx = 0,5$$

Solche Gleichungen können durch Computeralgorithmen, oder indem man *systematisch ausprobiert*, approximativ gelöst werden.<sup>17</sup> Beliebt ist es, den Median mit einer Computersimulation zu berechnen. Dies wird im letzten Abschnitt dieses Kapitels gezeigt. Der hier auftauchende Integrand ist einfach und das Integral kann man analytisch berechnen. Man erhält einen Median von 3,823 Tagen. Physiker nennen diese Zahl auch die Halbwertszeit. Das Modell besagt: Es besteht eine Wahrscheinlichkeit von 0,5, dass die Zerfallszeit eines Radonisotops grösser als 3,823 Tage sein wird. □

**Beispiel 4.15 (Mittelklasse)** Das Institut für ökonomische Forschung in Berlin stuft arbeitende Leute in Deutschland in die *Mittelklasse* ein, wenn sie ein Einkommen besitzen, das zwischen 70 % und 150 % des deutschen Medianeinkommens liegt.<sup>18</sup> Für das Medianeinkommen gilt: bei einer Hälfte aller Arbeitsstellen liegt der Lohn über, bei der anderen Hälfte dagegen unter dem Median. □

Beliebt ist es, die Plausibilität zu einer Grösse, durch zwei Werte  $a$  und  $b$  einzuschachteln. Mit dem Wahrscheinlichkeitsmodell zur Grösse kann man dann beispielsweise sagen:

$$\mathbb{P}(a \leq \text{Grösse} \leq b | \text{Information}) = 0,95$$

Man sagt, dass man ein *Wahrscheinlichkeitsintervall* (engl. *credibility interval*) zum Niveau 0,95 für die Grösse berechnet hat. Man wählt in der Regel die Zahlen  $a$  und  $b$  so, dass die Wahrscheinlichkeit, dass die Grösse grösser als  $b$  und kleiner als  $a$  ist, je 0,025 beträgt.

<sup>17</sup> Ein erster Versuch mit  $t_1 = 4$  Tage gibt  $\int_0^4 \lambda \cdot \exp(-\lambda \cdot x) dx = 0,516$ . Der Wert  $t_1$  ist zu hoch. Eine zweiter Versuch mit  $t_2 = 3,5$  Tagen liefert  $\int_0^{3,5} \lambda \cdot \exp(-\lambda \cdot x) dx = 0,470$ . Der Wert  $t_2$  ist zu tief. Ein guter nächster Versuch  $t_3$  ist das arithmetische Mittel von  $t_1$  und  $t_2$ :  $t_3 = 3,75$  Tage. Mit diesem Verfahren (einer Bisektion) lässt sich der Median „einschachteln“.

<sup>18</sup> The New York Times und Le Monde, 17. Mai 2008.

**Abb. 4.12** Ein Wahrscheinlichkeitsintervall zum Niveau 0,95

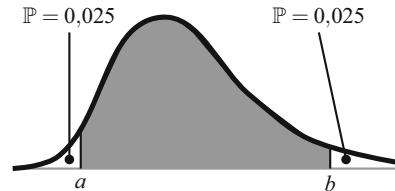


Abb. 4.12 zeigt die Situation. Man sagt, dass  $b$  das 0,975-Quantil und  $a$  das 0,025-Quantil des Modells sind.

Allgemein nennt man den kleinsten Wert  $q_\alpha$ , der mit einer Wahrscheinlichkeit von mindestens  $\alpha$  nicht überschritten wird, das  $\alpha$ -*Quantil* (engl.  $\alpha$ -percentile) des Wahrscheinlichkeitsmodells.

**Beispiel 4.16 (Zeit zwischen starken Erdbeben)** Wie oben erwähnt, kann aus Daten und Vorinformation die durchschnittliche Zeit  $\mu$  zwischen aufeinanderfolgenden, zukünftigen, grossen Erdbeben berechnet werden. Die Plausibilität für  $\mu \geq 0$  wird mit der Dichtefunktion beschrieben:<sup>19</sup>

$$\text{pdf}(\mu = x \mid \text{Daten, Vorinformation}) = \frac{9,4767 \cdot 10^{87}}{x^{29}} \cdot \exp(-13\,910,55 \text{ Tage}/x)$$

Der plausibelste Wert für  $\mu$  ist der Modus  $\mu_0$ . Er beträgt 479,7 Tage. Das 0,25-Quantil  $q_{0,25}$  des Modells berechnet man mit der Gleichung

$$\int_0^{q_{0,25}} \text{pdf}(\mu = x \mid \text{Daten, Vorinformation}) dx = 0,25$$

Die Gleichung kann mit einem Computer oder durch systematisches Ausprobieren gelöst werden. Man erhält  $q_{0,25} = 444$  Tage. Das 0,75-Quantil berechnet man analog. Es ist 572 Tage. Daraus folgt, dass  $\mu$  mit einer Wahrscheinlichkeit von 0,5 zwischen 444 und 572 Tagen liegt. Ein Wahrscheinlichkeitsintervall für  $\mu$  zum Niveau von 0,95, wird durch die Zahlen 354 Tage und 748 Tage gebildet. Die beiden Grenzen sind so gewählt, dass sie das 0,025- und 0,975-Quantil des Modells sind. □

Quantile  $q_\alpha$  mit  $\alpha$  nahe bei 1, wie beispielsweise  $\alpha = 0,99$ , werden im Bereich der Wirtschaftswissenschaft und im Ingenieurwesen beim *Risikomanagement* benutzt. Prognostiziert man einen zukünftigen Wert einer Risikogrösse (wie etwa Schadensummen), so beträgt die Wahrscheinlichkeit, dass dieser das Quantil  $q_\alpha$  überschreitet  $1 - \alpha$ . Man nennt das  $\alpha$ -Quantil in diesem Zusammenhang auch den *VaR*, was *Value at Risk* bedeutet.

<sup>19</sup> Die Rechnung wird in Kap. 9 gezeigt.

**Beispiel 4.17 (Sicherheit in der Informatik)** Für eine Verwaltungsstelle der Schweizerischen Eidgenossenschaft wurden in [12] Schadensummen durch mögliche Angriffe auf hochsensible Daten der Informatikinfrastruktur modelliert. Es ergaben sich die Zahlen:

VaR (95 %) 22 900 CHF

VaR (99 %) 54 400 CHF

Dies heisst, dass zukünftige Schadenfälle mit einer Wahrscheinlichkeit von 0,95 kleiner als 22 900 CHF sein werden. Mit einer Wahrscheinlichkeit von 0,01 werden Schadensummen grösser als 54 400 CHF prognostiziert.  $\square$

Es gibt Quantile, die speziell benannt werden. Das *obere Quartil* (bzw. *untere Quartil*) ist das 0,75-Quantil (bzw. das 0,25-Quantil). Wie der Name sagt, teilen die Quartile die prognostizierten Werte so, dass sie geviertelt werden: 25 % aller prognostizierten Werte liegen über dem oberen Quartil, 25 % liegen unter dem unteren Quartil.

**Beispiel 4.18 (Einkommensverhältnisse)** Tab. 4.1 zeigt die Einkommensentwicklung in den USA, bei der die Einkommensverteilung in fünf Teile mit *Quintilen* zerlegt wurden (aus [18]). Sie besagt, dass im Jahr 2000 eine Wahrscheinlichkeit von 0,80 besteht, dass eine Person in den USA höchstens 141 620 Dollar verdient. Die dargestellten Quintile zeigen, wie sich die Einkommensverhältnisse während der Zeitperioden 1980–1994 und 1994–2000 verschieden entwickelt haben.  $\square$

Die *Fünf-Zahlen-Zusammenfassung* (engl. *Five-Number-Summary*) erlaubt es, die Verteilung eines Wahrscheinlichkeitsmodells mit wenigen Zahlen zu charakterisieren. Die fünf Kennzahlen Median, unteres und oberes Quartil, sowie der kleinste und der grösste mögliche Wert des Modells beschreiben das Modell meist gut.

**Tab. 4.1** Einkommensentwicklung in den USA mit Hilfe von Quintilen (aus [18])

Durchschnittliches Einkommen in realen Dollar des Jahres 2000	1980	1994	2000	1994/1980	2000/1994
5 % Reichste	132 551	210 684	250 146	+ 59 %	+ 18 %
20 % Reichste (oberstes Quintil)	91 634	121 943	141 620	+ 33 %	+ 16 %
4. Quintil	52 169	58 005	65 729	+ 11 %	+ 13 %
3. Quintil	35 431	37 275	42 361	+ 5 %	+ 14 %
2. Quintil	21 527	22 127	25 334	+ 3 %	+ 14 %
20 % Ärmste (1. Quintil)	8920	8934	10 190	+ 0 %	+ 14 %

**Beispiel 4.19 (Zerfallszeit von Radon)** Prognostiziert man Zerfallszeiten von Radon-Isotopen  $^{222}\text{Rn}$  mit der Exponentialverteilung, so lautet die Fünf-Zahlen-Zusammenfassung

min	$q_{0.25}$	med	$q_{0.75}$	max
0 Tage	1,589 Tage	3,823 Tage	7,645 Tage	$\infty$

Die Wahrscheinlichkeit, eine Zerfallszeit zwischen 1,589 und 7,645 Tagen zu messen, beträgt 0,5. Es besteht eine Wahrscheinlichkeit von 0,25 eine Zerfallszeit von mehr als 7,645 Tagen zu erhalten.  $\square$

## 4.5 Monte-Carlo-Simulationen

Um Wahrscheinlichkeitsintervalle und Quantile von stetigen Wahrscheinlichkeitsmodellen zu bestimmen, muss man Integrale berechnen. Diese kann man in der Regel nicht explizit berechnen. Daher werden heute Integrale mit rechenintensiven Computersimulationen – *Monte-Carlo-Simulationen* – approximativ bestimmt. Solche Simulationen sind sehr beliebt, da heute auch persönliche Rechner mit schnellsten Prozessoren ausgestattet sind.

Das Ziel einer Monte-Carlo-Simulation ist das Folgende: Eine Grösse  $x$  sei bei gegebener Information  $\mathcal{I}$  durch ein Wahrscheinlichkeitsmodell beschrieben. Man erzeugt tausend, zehntausend oder hunderttausend Werte  $x_{\text{sim}}$  derart, dass die Wahrscheinlichkeit, dass  $x$  zwischen  $a$  und  $b$  liegt, approximativ

$$\mathbb{P}(a \leq x \leq b \mid \mathcal{I}) \approx \frac{\text{Anzahl erzeugte Werte } x_{\text{sim}} \text{ zwischen } a \text{ und } b}{\text{Gesamtzahl erzeugter Werte } x_{\text{sim}}}$$

ist. Anders ausgedrückt: Ein erzeugter Wert  $x_{\text{sim}}$  tritt proportional zur vorgegebenen Wahrscheinlichkeit  $\mathbb{P}(x = x_{\text{sim}})$  des Modells auf. Auch Quantile sind so einfach berechenbar. Man bildet mit dem Computer die Rangliste der erzeugten  $M$  Werte: kleinster Wert, zweitkleinster Wert, drittkleinster Wert, usw. Der Median entspricht dem Wert in der Mitte dieser Liste. Das 0,2-Quantil ist der  $0,2 \cdot M$ -te Wert und das 0,9-Quantil ist der  $0,9 \cdot M$ -te Wert der Rangliste.

Es gibt viele Methoden solche Werte  $x_1, x_2, x_3, \dots$  zu erzeugen, mit denen nach dem oben Gesagten Wahrscheinlichkeiten, Quantile oder Wahrscheinlichkeitsintervalle eines stetigen Wahrscheinlichkeitsmodells berechnet werden können.

Die hier verwendete Methode, um stetige Wahrscheinlichkeitsmodelle zu simulieren, ist in der Bayes-Statistik weit verbreitet. Es ist die Markov-Kette-Monte-Carlo-Simulation (engl. *Markov Chain Monte Carlo*, auch MCMC-Simulation genannt).

Markov-Kette bedeutet, dass eine Folge von Werten erzeugt wird, bei der  $x_2$  von  $x_1$ ,  $x_3$  von  $x_2$ ,  $x_4$  von  $x_3$  und so weiter abhängen. Hier ein solcher Algorithmus:

Das stetige Wahrscheinlichkeitsmodell sei durch die Dichtefunktion  $\text{pdf}(x \mid \mathcal{I})$  gegeben. Wie kann man eine solche Folge  $x_1, x_2, x_3, \dots$  konstruieren? Hat man schon Werte  $x_1, x_2, \dots, x_m$ , so kann man versuchen einen neu gewählten Punkt  $x^*$  dazuzufügen. Der Vorläuferpunkt  $x_m$  in der erzeugten Folge sollte mit einer Wahrscheinlichkeit auftreten, die proportional zum Wert der Dichtefunktion an dieser Stelle ist:  $\text{pdf}(x_m \mid \mathcal{I})$ . Die Wahrscheinlichkeit, dass der neue Punkt  $x^*$  auftritt, ist proportional zu  $\text{pdf}(x^* \mid \mathcal{I})$ . Ist diese grösser als beim Vorläuferpunkt  $x_m$ , so müsste daher, weil  $x_m$  in der Kette vorkommt,  $x^*$  auch in der Folge auftauchen. In diesem Fall ist der nächste Punkt in der Folge  $x^*$ , also  $x_{m+1} = x^*$ . Ist dies nicht der Fall, so wird man  $x_m$  oder  $x^*$  als nächsten Wert der Folge nehmen. Dabei scheint es sinnvoll,  $x_m$  oder  $x^*$  im Verhältnis ihrer Auftretenswahrscheinlichkeit  $\text{pdf}(x_m \mid \mathcal{I}) : \text{pdf}(x^* \mid \mathcal{I})$  zu wählen! Abb. 4.13 und Abb. 4.14 visualisieren diesen Algorithmus. Er ist im wesentlichen der *Metropolis-Hastings Algorithmus* aus [14] und [9]:

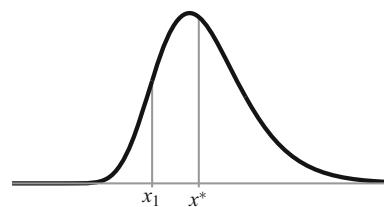
- (1) Wähle einen Startpunkt  $x_1$  und setze  $i = 1$ .
- (2) Springe mit einem Zufallsprozess von  $x_i$  zu einem Punkt  $x^*$ . Ist der Wert  $\text{pdf}(x^* \mid \mathcal{I})$  der Dichtefunktion bei  $x^*$  grösser als bei  $x_i$ , also

$$\text{pdf}(x^* \mid \mathcal{I}) > \text{pdf}(x_i \mid \mathcal{I})$$

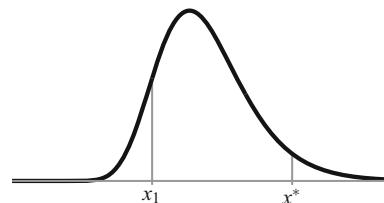
dann ist der nächste Punkt  $x^*$ :  $x_{i+1} = x^*$ . Im anderen Fall berechne das Verhältnis

$$\mathbb{P} = \frac{\text{pdf}(x^* \mid \mathcal{I})}{\text{pdf}(x_i \mid \mathcal{I})}$$

**Abb. 4.13** Fall 1: „Sprung“ zu Vorschlag  $x^*$ : Der nächste Punkt ist  $x_2 = x^*$



**Abb. 4.14** Fall 2: „Sprung“ zu Vorschlag  $x^*$ : das Verhältnis der beiden vertikalen Linien ist 0,3; wähle als nächsten Punkt  $x_2 = x^*$  mit einer Wahrscheinlichkeit von 0,3 oder  $x_2 = x_1$  mit einer Wahrscheinlichkeit von 0,7



Ziehe eine zufällige Zahl  $r$  zwischen Null und Eins. Ist  $r \leq \mathbb{P}$ , so ist  $x_{i+1} = x^*$ , sonst ist  $x_{i+1} = x_i$ .

- (3) Erhöhe den Index  $i$  um Eins und gehe zu Schritt (2).

**Beispiel 4.20 (Ein Beispiel)** Ihre Information  $\mathcal{I}$  zu einer nicht direkt messbaren Größe  $m$ , die beliebige Werte annehmen könnte, beschreibt eine Person mit einem stetigen Wahrscheinlichkeitsmodell. Die dazugehörige Dichtefunktion lautet

$$\text{pdf}(m = x \mid \mathcal{I}) = \frac{\sqrt{2}}{\pi} \cdot \frac{1}{1 + (x - 2)^4}$$

Die Verteilung ist unimodal symmetrisch. Der Modus ist bei  $m_0 = 2$  und Werte von  $m$  liegen mit hoher Plausibilität zwischen  $-1$  und  $5$ .

Um eine MCMC-Kette  $x_1, x_2, x_3, \dots$  nach dem obigen Verfahren zu konstruieren, kann man als Startpunkt  $x_1 = 1$  wählen. Man wähle den Sprungprozess so, dass er durchschnittlich um eine Einheit springt. Als ersten Vorschlagspunkt  $x^*$  habe man 1,6. Es ist  $\text{pdf}(1,6) = 0,439$ . Dieser Wert ist grösser als  $\text{pdf}(1) = 0,225$ . Daher ist  $x_2 = 1,6$ . Man springe nun von  $x_2$  und lande beim nächsten Vorschlagspunkt:  $x^* = 2,7$ . Es ist  $\text{pdf}(2,7) = 0,363$ . Dieser Wert ist kleiner als  $\text{pdf}(1,6)$ . Zu berechnen ist daher der Quotient

$$\mathbb{P} = \frac{\text{pdf}(2,7)}{\text{pdf}(1,6)} = 0,827$$

Man lasse jetzt den Computer eine zufällige Zahl zwischen null und eins ziehen. Diese sei 0,873. Sie ist grösser als 0,827. Damit wird der Vorschlagspunkt abgelehnt. Der dritte Punkt in der Kette ist damit der alte Punkt  $x_2$ . Es ist  $x_3 = 1,6$ . Von  $x_3$  kann man weiter-springen: der nächste Vorschlagspunkt sei  $x^* = 2,1$ . Es ist  $\text{pdf}(2,1) = 0,450$ . Dieser Wert ist grösser als  $\text{pdf}(1,6) = 0,439$ . Somit wird der vierte Punkt  $x_4$  der Kette 2,1 sein.  $\square$

Da die Punkte in der konstruierten Kette nicht unabhängig sind, muss die Kette sehr lang sein, um mit ihr sinnvolle Wahrscheinlichkeiten oder Quantile ausrechnen zu können. Meist braucht man Ketten von 10 000 bis 100 000 Punkten. Hat der Graph der Dichtefunktion eine sehr spitzige Form, sollte man von Punkt zu Punkt wenig springen. Der Algorithmus „findet“ sonst die Spitze der Verteilung kaum. Ist andererseits ihr Graph breit und springt man von Punkt zu Punkt sehr wenig, so braucht es eine riesige Kette, um die ganze Dichtefunktion „abzutasten“.

Der Metropolis-Hastings-Algorithmus findet sich in guten Statistikprogrammen. Er kann daher meist schnell durchgeführt werden. Ein gut implementierter Algorithmus ist sogar in der Lage, den Sprungprozess zu einem Vorschlagspunkt gut zu wählen, wenn der Startwert in der Nähe des Modus liegt und die Verteilung nicht multimodal ist. Empfehlenswert ist es, den Algorithmus mit der *logarithmierten* Dichtefunktion durchzuführen,

um Spitzen zu glätten.<sup>20</sup> Dies liefert meist genauere Resultate. Weiter ist es üblich, den Algorithmus „aufwärmen“ (engl. *burnin* oder *warm-up*) zu lassen: dazu werden die ersten 1–2 % Punkte der Kette weggelassen. Dies verhindert, dass die berechneten Punkte zu stark vom Anfangspunkt abhängen.<sup>21</sup> Hier ein erstes Beispiel dazu:

**Beispiel 4.21 (Zeit zwischen starken Erdbeben)** In Kap. 1 wird in Beispiel 1.3 erläutert, wie die durchschnittliche Zeit  $\mu$  zwischen aufeinanderfolgenden, zukünftigen, grossen Erdbeben aus beobachteten Daten und Vorinformation  $\mathcal{K}$  berechnet werden kann. Die Plausibilität zum Parameter  $\mu$  ist mit der Dichtefunktion

$$\text{pdf}(\mu = x \mid \text{Daten}, \mathcal{K}) = \frac{9,4767 \cdot 10^{87}}{x^{29}} \cdot \exp(-13\,910,55 \text{ Tage}/x)$$

wiedergegeben. Mit dem MCMC-Metropolis-Algorithmus lässt sich eine Kette von 100 100 Punkten des Wahrscheinlichkeitsmodells bestimmen. Beachten Sie, dass der Algorithmus nur die relative Höhe der Dichtefunktion benutzt.<sup>22</sup> Man kann daher auch die Funktion

$$\text{pdf}(\mu = x \mid \text{Daten}, \mathcal{K}) \propto \frac{1}{x^{29}} \cdot \exp(-13\,910,55 \text{ Tage}/x)$$

betrachten. Damit kann auf die für den Computer mühsame Zahl von  $9,4767 \cdot 10^{87}$  verzichtet werden. Numerisch stabiler ist es, den MCMC-Algorithmus mit dem Logarithmus  $L(\mu = x \mid \text{Daten}, \mathcal{K})$  dieser Funktion durchzuführen:

$$L(\mu = x \mid \text{Daten}, \mathcal{K}) \propto -29 \cdot \ln x - \frac{13\,910,55 \text{ Tage}}{x}$$

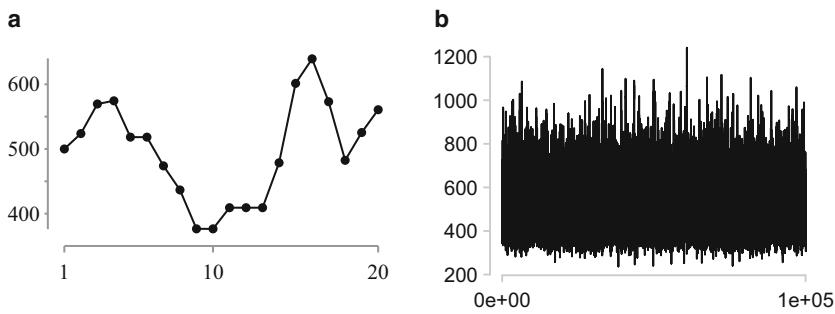
Als Startwert wird  $x_1 = 500$  gewählt. Abb. 4.15a zeigt die ersten zwanzig Punkte und Abb. 4.15b die 100 100 Punkte. Auffallend ist, dass die Punkte nicht in gewissen Bereichen gefangen bleiben, sondern den ganzen Bereich zwischen 300 und 1000 der Dichtefunktion abdecken. Insbesondere im Bereich zwischen 400 und 600, wo die Dichtefunktion hohe Werte hat, sind auch viele Punkte vorhanden. Hier ein Ausschnitt aus der Liste

---

<sup>20</sup> Hat die Verteilung eine schmale Spalte, so sind die Funktionswerte der Dichtefunktion hier meist sehr gross. Mit den logarithmierten Werten kann die Gefahr reduziert werden, Over- oder Underflows zu erhalten. Aus Zahlen wie  $10^{300}$  oder  $10^{-300}$  werden mit dem Logarithmus 300 und  $-300$ .

<sup>21</sup> Ausführliche Informationen, wie man beurteilen kann, ob die gewählte Kette genügend lang ist und der Zufallsprozess gut ist, finden sich in [7].

<sup>22</sup> Man berechnet Quotienten von Funktionswerten von pdf!



**Abb. 4.15** Die ersten zwanzig (a) und die 100 100 Punkte (b) der MCMC-Simulation ( $1 \cdot 10^5$  ist die Gleitkommadarstellung von  $1 \cdot 10^5$  eines Computers)

der simulierten 100 100 Werte:

$$\begin{aligned}x_1 &= 500,0000 & \dots & \dots & x_{100\,095} &= 651,0978 \\x_2 &= 523,8179 & \dots & \dots & x_{100\,096} &= 476,7756 \\x_3 &= 569,5674 & \dots & \dots & x_{100\,097} &= 516,1304 \\x_4 &= 574,4563 & \dots & \dots & x_{100\,098} &= 543,7113 \\x_5 &= 518,3045 & \dots & \dots & x_{100\,099} &= 571,0182 \\x_6 &= 518,3045 & \dots & \dots & x_{100\,100} &= 496,1901\end{aligned}$$

Ersichtlich ist, dass bei den fünf ersten Iterationen, der neue vorgeschlagene Punkt jeweils akzeptiert wurde. Der Punkt  $x_6$  ist gleich dem Punkt  $x_5$ . Hier wurde der neue vorgeschlagene Punkt verworfen. Gesamthaft wurde bei 71,44 % der Rechenschritte der neue Punkt  $x^*$  akzeptiert. Man sagt, dass die *Akzeptanzrate* 71,44 % ist.

Lässt man die ersten 1 % der Kette weg, verbleiben 100 000 Punkte  $x_{1001}, x_{1002}, \dots, x_{100\,100}$ . Mit ihnen kann man Plausibilitäten zum gesuchten Parameter  $\mu$  rechnen. Die Wahrscheinlichkeit, dass die durchschnittliche Zeit  $\mu$  zwischen aufeinander folgenden, zukünftigen, starken Erdbeben grösser als 500 Tage ist, beträgt

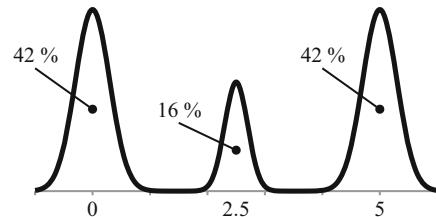
$$\mathbb{P}(\mu > 500 \mid \text{Daten}, \mathcal{K}) \approx \frac{\text{Anzahl } x_i > 500}{100\,000} = \frac{51\,067}{100\,000} = 0,51$$

Der exakte Wert ist ebenfalls 0,51. Ein Wahrscheinlichkeitsintervall zum Niveau 0,5 für  $\mu$  besteht aus den 0,25- und 0,75-Quantilen. Bei einer Kette von 100 000 Punkten entsprechen diese Quantile dem 25 000. und 75 000. Wert in der *Rangliste* der Punkte. Diese betragen 444,0 Tage und 574,6 Tage. Man erhält

$$q_{0,25} \approx 444,0 \text{ Tage}, \quad q_{0,75} \approx 574,6 \text{ Tage}$$

Die exakten Werte sind 443,5 Tage und 572,3 Tage. Analog erhält man aus dem 2500. und 97 500. Wert aus der Rangliste der simulierten Punkte ein Wahrscheinlichkeitsintervall

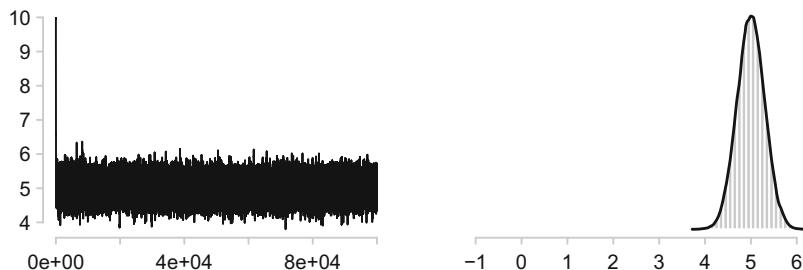
**Abb. 4.16** Eine komplizierte Dichtefunktion mit drei Gipfeln



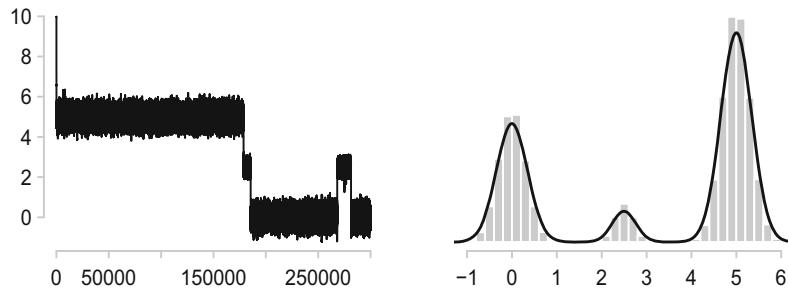
zum Niveau 0,95. Dies sind die Zahlen 353,2 Tage und 745,0 Tage. Das exakte Intervall besteht aus den Grenzen 354,1 Tage und 748,4 Tage.  $\square$

**Beispiel 4.22 (Ein komplexe Verteilung)** Das Beispiel folgt im Wesentlichen den Ausführungen von P. D. Hoff in [10] auf den Seiten 98–104. Es zeigt, dass manchmal sehr lange Ketten konstruiert werden müssen, wenn die Dichtefunktion kompliziert ist. In Abb. 4.16 ist der Graph einer Dichtefunktion mit drei Gipfeln gezeichnet. In den beiden äusseren Spitzen befinden sich je 42 % der Fläche. Die Wahrscheinlichkeit Werte bei der mittleren Spitze zu haben, beträgt 0,16. Abb. 4.17 zeigt den MCMC-Algorithmus mit Startwert  $x_1 = 10$  und einer Kette von 100 000 Iterationen. Beim Algorithmus wurde der neue Punkt in sieben von zehn Fällen akzeptiert. Die Akzeptanzrate des Algorithmus ist damit 70 %. Das linke Bild in Abb. 4.17 zeigt: Die Kette startet beim Punkt  $x = 10$  und verbleibt im Bereich zwischen vier und sechs. Die Kette approximiert die Wahrscheinlichkeit, einen Wert kleiner als vier zu haben, daher mit null. Der richtige Wert ist 0,58. Die 100 000 Iterationspunkte haben somit noch nicht genügt, um die Dichtefunktion zu approximieren. Dies illustriert die rechte Grafik in Abb. 4.17: sie visualisiert die erzeugten Punkte mit einem Histogramm und einer approximierenden Linie. Nur die rechte Spitze der Verteilung ist mit der Kette abgedeckt.

Wählt man eine dreifach so lange Kette, ergibt sich die Situation, die in Abb. 4.18 dargestellt ist. Die Akzeptanzrate der neuen, längeren Kette ist immer noch rund 70 %. Die Kette deckt die Bereiche aller drei Spitzen der Dichtefunktion ab. Die zwei linken



**Abb. 4.17** Dichtefunktion mit drei Gipfeln: MCMC-Kette mit 100 000 Punkten und Akzeptanzrate von 70 %

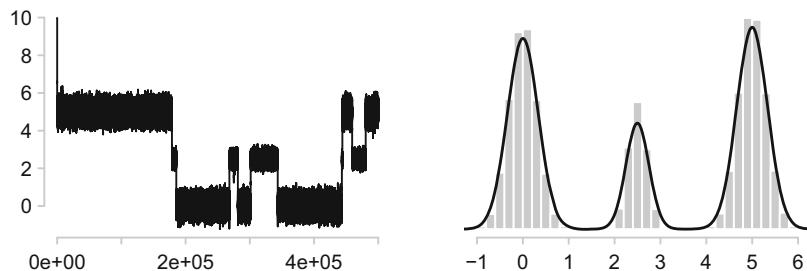


**Abb. 4.18** Dichtefunktion mit drei Gipfeln: MCMC-Kette mit 300 000 Punkten und Akzeptanzrate von 70 %

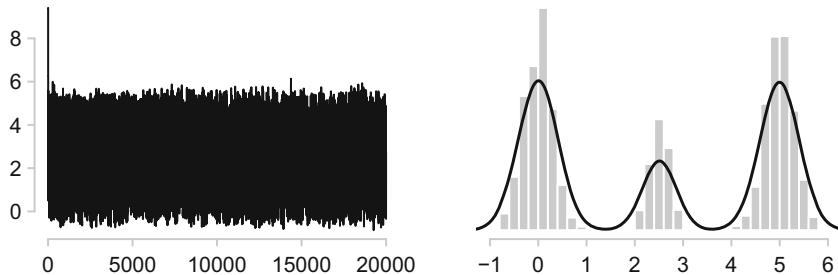
Spitzen der Verteilung sind noch zu klein. Daher wird die Kette Wahrscheinlichkeiten schlecht approximieren. So ist

$$\mathbb{P}(\text{Wert} < 3) \approx \frac{101\,252}{300\,000} = 0,338$$

Der richtige Wert dieser Wahrscheinlichkeit beträgt 0,58. In Abb. 4.19 ist eine Kette von 500 000 Punkten gezeigt. Auch sie hat eine Akzeptanzrate von 70 %. Um also hier Wahrscheinlichkeiten und Quantile berechnen zu können, braucht man eine sehr lange Kette. Man sagt auch, dass die Kette sehr langsam *konvergiert* oder *stationär* geworden ist. Warum so eine langsame Konvergenz? Der Grund liegt darin, dass der Sprungprozess, um Punkte in der Kette zu generieren, eine zu kleine Streuung hat. Die erzeugten Punkte bewegen sich so sehr langsam. Sie verbleiben meist lange in engen Bereichen um die drei Spitzen. Dies sieht am auch an der hohen Akzeptanzrate von 70 %. Ist nämlich der vorgeschlagene Punkt  $x^*$  nahe beim letzten Punkt  $x_m$ , so sind die entsprechenden Funktionswerte der Dichtefunktion auch etwa gleich gross. Somit ist die Wahrscheinlichkeit gross, dass der vorgeschlagene Punkt akzeptiert wird.



**Abb. 4.19** Dichtefunktion mit drei Gipfeln: MCMC-Kette mit 500 000 Punkten und Akzeptanzrate von 70 %



**Abb. 4.20** Dichtefunktion mit drei Gipfeln: MCMC-Kette mit 20 000 Punkten und Akzeptanzrate von 25 %

Abb. 4.20 zeigt noch einmal eine Kette von „nur“ 20 000 Punkten mit Startwert  $x_1 = 10$ , bei dem der Sprungprozess so gewählt wurde, dass er viel stärker streut. Hier ist

$$\mathbb{P}(\text{Wert} < 1) \approx \frac{8415}{20\,000} = 0,42$$

Dies entspricht dem exakten Wert von 0,42. Die Akzeptanzrate 25 % der Kette ist auch viel tiefer. Eine noch grössere Streuung des Sprungprozesses wäre wiederum schlecht. Viele vorgeschlagene Punkte würden in die Region der Dichtefunktion fallen, wo diese sehr kleine Werte hat. Die Kette bliebe damit an einem Punkt „gefangen“. Als Folge hat man eine kleine Akzeptanzrate.  $\square$

Das obige Beispiel zeigt, dass man nicht immer klar beurteilen kann, ob die erzeugte Kette genügend lang ist. Dies gilt vor allem bei multimodalen Verteilungen. Es ist daher empfehlenswert, mehrere Ketten mit Sprungprozessen zu konstruieren, die verschieden gross streuen. Als *Faustregel* gilt: Wähle die Streuung so, dass die Akzeptanzrate des MCMC-Algorithmus zwischen 20 und 50 % liegt (siehe [10]). Man kann auch anders beurteilen, ob eine Kette stationär ist (und damit das Wahrscheinlichkeitsmodell gut approximiert). Dazu misst man, wie stark die erzeugten Punkte voneinander abhängen. Je stärker sie dies tun, um so länger muss in der Regel die Kette gewählt werden. Eine Kette von 100 000 Punkten mit starker Abhängigkeit, hat vielleicht die Aussagekraft einer Kette von 1000 unabhängigen Punkten. Beim obigen Beispiel hat die sehr lange Kette von 500 000 Punkten eine Aussagekraft von einer aus unabhängigen Punkten bestehenden Kette mit rund 20 Punkten! Die Kette mit 20 000 Punkten mit einem Sprungprozess, der gross streut, eine solche, die rund 1300 Punkten entspricht.<sup>23</sup> Beliebt ist es zu messen, ob die verschiedenen Ketten nur Teile der Verteilung abtasten (wovon abzuraten ist) oder ob alle Ketten die gesamte Verteilung erfassen (was zu empfehlen ist). Hat man  $M$  Ketten mit zufällig gewählten Startpunkten, so charakterisieren die empirischen Standardabwei-

<sup>23</sup> Wie dies gerechnet wird, siehe [10] auf den Seiten 102–103.

chungen  $s_1, s_2, \dots, s_M$  der Punkte der Ketten, wie ausgedehnt die einzelnen Ketten sind. Die mittlere quadratische Ausdehnung  $s_{\text{Ketten}}^2$  der Ketten ist

$$s_{\text{Ketten}}^2 = \frac{s_1^2 + s_2^2 + \dots + s_M^2}{M}$$

Mit der empirischen Standardabweichung  $s_{\text{Schwerpunkt}}$  der Mittelwerte der  $M$  Ketten kann man messen, wie die einzelnen Ketten zueinander versetzt sind. Das Verhältnis

$$\widehat{R}^2 = 1 + \frac{s_{\text{Schwerpunkt}}^2}{s_{\text{Kette}}^2}$$

nennt man nach Gelmann und Rubin ([5]) das Quadrat des *potential scale reduction factors* oder der *Gelman-Rubin-Diagnostik*. Tasten die verschiedenen Ketten die gesamte Verteilung ab, so ist  $\widehat{R} = 1$ , da alle Ketten den gleichen Schwerpunkt haben. Decken aber die einzelnen Ketten nur verschiedene Teile der Verteilung ab, so wird  $\widehat{R} > 1$ .

Ausführlich informieren die Statistiker A. Gelman und K. Shirley in [6], wie Markov-Kette-Monte-Carlo-Simulationen qualitativ beurteilt werden können. Man findet in dieser Publikation auch eine ausführliche Literaturliste zu Markov-Kette-Monte-Carlo-Simulationen.

## 4.6 Weiterführende Literatur zu diesem Kapitel

Ausführliche und weiterführende Informationen, weshalb man Wahrscheinlichkeiten als Plausibilitäten interpretieren soll und zu den Rechenregeln für Wahrscheinlichkeiten findet man in:

1. W. Bolstad, *Introduction to Bayesian Statistics* (Hoboken, NJ, John Wiley & Sons, 2004)
2. E. T. Jaynes, *Probability Theory. The Logic of Science* (Cambridge University Press, 2010)
3. D. V. Lindley, *Understanding Uncertainty* (Hoboken, NJ, John Wiley & Sons, 2007)
4. D. S. Sivia, J. Skilling, *Data Analysis, a Bayesian Tutorial*, 2nd edition (Oxford Science Publications 2010)

## Reflexion

**4.1** Es sind Aussagen  $A$  und  $B$  gegeben, die unabhängig sind, mit  $\mathbb{P}(A) = 0,6$  und  $\mathbb{P}(B) = 0,2$ .

- (a) Berechnen Sie die folgenden Wahrscheinlichkeiten:

$$\mathbb{P}(A | B), \quad \mathbb{P}(B | A), \quad \mathbb{P}(A \text{ und } B), \quad \mathbb{P}(A^{\text{nicht}}), \quad \mathbb{P}(A^{\text{nicht}} \text{ und } B^{\text{nicht}})$$

Arbeiten Sie mit einer Tabelle oder mit den Rechenregeln.

- (b) Bestimmen Sie die Chancen von  $A$  und von  $B$ .

**4.2** Zu zwei Aussagen  $A$  und  $B$  kennt man die Wahrscheinlichkeiten  $\mathbb{P}(A) = 0,3$ ,  $\mathbb{P}(A | B) = 0,6$  und  $\mathbb{P}(B) = 0,4$ .

- (a) Berechnen Sie die Wahrscheinlichkeiten  $\mathbb{P}(A \text{ und } B)$ ,  $\mathbb{P}(B | A)$ ,  $\mathbb{P}(A^{\text{nicht}} | B)$ , und  $\mathbb{P}(A^{\text{nicht}})$ . Arbeiten Sie dazu mit einer Tabelle oder mit den Rechenregeln.  
(b) Wie lauten die Chancen (1) von  $A$ , (2) von  $B$  und (3) von  $A^{\text{nicht}}$  gegeben  $B$ ?

**4.3** Zu zwei Aussagen  $A$  und  $B$  hat man

$$\mathbb{P}(A) = 0,55, \quad \mathbb{P}(B) = 0,25, \quad \mathbb{P}(A \text{ und } B) = 0,12$$

Berechnen Sie die Wahrscheinlichkeiten  $\mathbb{P}(A | B)$  und  $\mathbb{P}(B | A)$ . Sind  $A$  und  $B$  unabhängig?

**4.4** Gruppe B Streptokokus (GBS) gehört zu den häufigsten lebensbedrohenden Infektionen bei Neugeborenen. Gemäss [4] sind zwischen 5 % und 30 % aller schwangeren Frauen Trägerinnen des GBS. Studien haben gezeigt, dass die Behandlung von infizierten Frauen während der Schwangerschaft die Häufigkeit durch GBS verursachten Blutvergiftungen verringert. Daher ist der GBS-Nachweis wichtig. Der ‚Strep B Test‘ ist ein schnelles Verfahren für den Nachweis des GBS Antigens aus Vagina- oder Zervixabstrichen. Um die Sensitivität und Spezifität des Verfahrens zu berechnen, wurden in einer klinischen Studie 101 Personen untersucht.<sup>24</sup> Die Resultate sind: (1) 50 Frauen sind mit GBS Bakterien infiziert, (2) positiver Strep B Test bei 51 Personen und (3) von diesen 51 Personen sind 46 mit GBS Bakterien infiziert.

---

<sup>24</sup> Strep B Test 014B255, ulti med Products (Deutschland) GmbH, Reeshoop 1, 22926 Ahrensburg.

- (a) Stellen Sie in der folgenden Kreuztabelle die Daten übersichtlich dar.

	infiziert	nicht infiziert	Total
Strep B+			
Strep B-			
Total			

- (b) Wie lauten die Sensitivität  $\mathbb{P}(\text{Strep B}+ \mid \text{infiziert})$  und die Spezifität  $\mathbb{P}(\text{Strep B}- \mid \text{nicht infiziert})$  des Strep B Tests?

**4.5** Ein Manager denkt, dass er am nächsten Tag einen Verkaufsvertrag mit einer Wahrscheinlichkeit von 0,75 erfolgreich abschliessen kann (Aussage A) und eine anschliessend geplante Sitzung mit einer Wahrscheinlichkeit von 0,35 kürzer als eine Stunde dauern wird (Aussage B). Weiter geht er davon aus, dass bei erfolgreichem Verkaufsvertrag, eine Wahrscheinlichkeit von 0,4 besteht, dass die Sitzung kürzer als eine Stunde dauert.

- (a) Notieren Sie die gegebenen Wahrscheinlichkeiten mit mathematischen Symbolen.  
 (b) Gesucht ist die Wahrscheinlichkeit, dass der Vertrag erfolgreich abgeschlossen wird und die Sitzung kurz dauert. Berechnen Sie die Wahrscheinlichkeit, indem Sie die folgende Kreuztabelle ausfüllen:

	Vertrag ok	Vertrag nicht ok	Total
kurze Sitzung			
lange Sitzung			
Total			100

- (c) Können Sie die Wahrscheinlichkeit in (b) auch direkt mit den Rechenregeln bestimmen?

**4.6** Ein Hausmakler geht davon aus, dass die Häuserpreise im nächsten Jahr davon abhängen, wie hoch die Hypothekenzinsen sein werden. Die Wahrscheinlichkeit von höheren Häuserpreisen setzt er 0,9, wenn die Hypothekenzinsen nicht erhöht werden, und er setzt sie 0,1, wenn diese erhöht werden. Zudem geht er davon aus, dass mit einer Wahrscheinlichkeit von 0,6 die Hypothekenzinsen nicht erhöht werden.

- (a) Formulieren Sie die obigen Wahrscheinlichkeiten mit mathematischen Symbolen.  
 (b) Berechnen Sie die Wahrscheinlichkeit, dass die Häuserpreise steigen werden. Führen Sie die Rechnung einmal mit einer Tabelle und einmal mit den Rechenregeln zur Wahrscheinlichkeit durch.

**4.7** Bei einem Aids-Test betrage die Sensitivität  $\mathbb{P}(\text{Test +} \mid \text{Aids}) = 0,998$  und die Spezifität  $\mathbb{P}(\text{Test -} \mid \text{nicht Aids})$  sei 0,992. Füllen Sie mit diesen Angaben die folgende Kreuztabelle aus:

	HIV infiziert	nicht HIV infiziert	Total
Test positiv			
Test negativ			
Total	16 000		4 000 000

Vergleichen Sie Ihre Rechnungen mit dem Kommentar aus [2]:

Selbst kleinste Unsicherheiten im einzelnen Test können bei grossen Serien zu riesigen Fehlermengen führen.

Verlangt man beispielsweise von einem Aids-Test eine Sensitivität von 99,8 Prozent, bedeutet dies, dass der Test unter 1000 HIV-Infizierten 998 als solche erkennt. Und die Spezifität eines Tests sagt, wie viele gesunde Menschen er als gesund ausweist. Weist ein Aids-Test eine Spezifität von 99,2 Prozent auf, ergeben sich bei 1000 HIV-freien Untersuchten 992 richtige Resultate. Obwohl die Fehlerquote dieses Tests im Promille-Bereich liegt, würde eine breit angelegte Untersuchung zu einem Debakel führen.

In der Schweiz zählen etwa 4 Millionen zur Gruppe der sexuell Aktiven. Man schätzt, dass 0,4 Prozent von ihnen mit Aids-Viren angesteckt sind. Landesweit muss man also von 16 000 Infizierten ausgehen, die eine vorsorgliche Untersuchung ausfindig machen müsste. Bei einem Test mit der genannten Sensitivität und Spezifität käme Folgendes heraus: Von den 16 000 Virusträgern würden 15 968 als positiv erkannt und 32 verfehlt. Für die Gruppe der 3 984 000 Nichträger gäbe es 3 952 128 korrekte Negativresultate, aber 31 872 Personen erhielten vorerst den falschen, höchst unangenehmen Befund „HIV-positiv“. Die insgesamt 47 840 Personen mit positivem Testergebnis müssten nun noch weitere aufwändige Tests über sich ergehen lassen, bis 33 Prozent tatsächlich Infizierten unter ihnen gefunden wären.

Deshalb sind landesweite Untersuchungen bei medizinischen Problemen, die nur einen kleinen Teil der Bevölkerung betreffen, aus statistischer Sicht ganz allgemein wenig sinnvoll. Mit der bestehenden Praxis, einen Test nur den Angehörigen einer Risikogruppe zu empfehlen, fährt man eindeutig besser.

**4.8** Ein System besteht aus zwei unabhängigen Komponenten, die je mit einer Wahrscheinlichkeit von 0,95 funktionieren. Berechnen Sie die Wahrscheinlichkeit, dass das System ausfällt, (a) wenn die Komponenten in Serie und (b) wenn sie parallel geschaltet sind. (*Ein Tipp:* Stellen Sie wiederum eine Tabelle auf!)

**4.9** Zu zwei Aussagen  $A$  und  $B$  kennen Sie die Wahrscheinlichkeiten  $\mathbb{P}(A) = 0,25$ ,  $\mathbb{P}(B) = 0,15$  und  $\mathbb{P}(A \mid B) = 0,8$ . Berechnen Sie die zu  $\mathbb{P}(A \mid B)$  inverse Wahrscheinlichkeit  $\mathbb{P}(B \mid A)$ . (*Ein Tipp:* Bestimmen Sie die Wahrscheinlichkeit  $\mathbb{P}(A \text{ und } B)$  mit der Multiplikationsregel auf zwei Arten.)

**4.10** Eine Firma, die Papier produziert, kauft drei Holzsorten  $X$ ,  $Y$  und  $Z$  ein. Je nach Holzsorte, wird die Papierqualität gut oder ausgezeichnet sein. Die Wahrscheinlichkeit,

dass die Papierqualität ausgezeichnet ist, beträgt 0,7, wenn die Sorte  $X$  benutzt wird, sie beträgt 0,8, wenn die Sorte  $Y$  gewählt wird und 0,9, wenn die Sorte  $Z$  verwendet wird. Die Produktionschefin nimmt an, dass während des nächsten Monats mit einer Wahrscheinlichkeit von 0,5 die Holzsorte  $X$ , mit einer Wahrscheinlichkeit von 0,4 die Sorte  $Y$  und mit einer Wahrscheinlichkeit von 0,1 die Sorte  $Z$  benutzt wird.

- (a) Notieren Sie sorgfältig die gegebenen Wahrscheinlichkeiten mit mathematischen Symbolen.
- (b) Bestimmen Sie die Wahrscheinlichkeit, dass das während des nächsten Monats produzierte Papier ausgezeichnet sein wird und dabei die Holzsorte  $Y$  verwendet wird.
- (c) Wie gross ist die Wahrscheinlichkeit, dass die Papierqualität während des nächsten Monats ausgezeichnet sein wird? (*Tipp:* Benutzen Sie das Gesetz der totalen Wahrscheinlichkeit.)

**4.11** Eine Aktie besitze einen Wert von CHF 50.– zu Beginn eines Handelstages. Ein Börsenhändler geht davon aus, dass der Tagesschlusskurs  $W_{\text{Schluss}}$  die Werte CHF 60.– mit einer Wahrscheinlichkeit von 0,6, CHF 53.– mit einer Wahrscheinlichkeit von 0,1 und CHF 40.– mit einer Wahrscheinlichkeit von 0,3 annehmen wird.

- (a) Zeichnen Sie den Graphen der Massenfunktion für den Tagesschlusskurs  $W_{\text{Schluss}}$ . Wie lautet der Modus der Verteilung?
- (b) Bestimmen Sie mit der Massenfunktion die Wahrscheinlichkeiten  $\mathbb{P}(W_{\text{Schluss}} \leq 55)$ ,  $\mathbb{P}(W_{\text{Schluss}} = 40)$  und  $\mathbb{P}(38 < W_{\text{Schluss}} < 53)$ .

**4.12** Eine Person will ihre Plausibilität zu einer Grösse  $r$  mit Wahrscheinlichkeiten formulieren. Sie weiss, dass die Grösse nur die Werte 0, 1, 5 und 10 haben kann. Sie fasst ihre Information  $\mathcal{K}$  zur Grösse mit dem folgenden diskreten Wahrscheinlichkeitsmodell zusammen:

$$\begin{aligned}\mathbb{P}(r = 0 \mid \mathcal{K}) &= 0,3 & \mathbb{P}(r = 1 \mid \mathcal{K}) &= 0,1 \\ \mathbb{P}(r = 5 \mid \mathcal{K}) &= 0,2 & \mathbb{P}(r = 10 \mid \mathcal{K}) &= 0,4\end{aligned}$$

- (a) Zeichnen Sie den Graphen der Massenfunktion des Modells.
- (b) Berechnen Sie, bei gegebenem Wissen  $\mathcal{K}$ , die Wahrscheinlichkeiten  $\mathbb{P}(r \leq 1)$ ,  $\mathbb{P}(r \leq 5,5)$ ,  $\mathbb{P}(2 < r \leq 4)$  und  $\mathbb{P}(r > 9)$ .

**4.13** Aussagen zu Werten einer Grösse  $Y$ , die Werte  $0, 1, 2, 3, \dots$ , annehmen kann, seien mit dem diskreten Wahrscheinlichkeitsmodell mit Massenfunktion

$$\mathbb{P}(Y = k \mid p) = (1 - p)^k \cdot p \quad \text{mit } k = 0, 1, 2, 3, \dots$$

beschreibbar. Die Zahl  $p$  ist eine Zahl zwischen 0 und 1. Man nennt diese Verteilung die *geometrische Verteilung*.

- (a) Zeichnen Sie den Graphen der Massenfunktion für  $p = 0,1$  und für  $p = 0,5$ . Wie ist das Modell verteilt?
- (b) Berechnen Sie, für  $p = 0,4$  die folgenden Wahrscheinlichkeiten: (1) dass ein Wert von  $Y$  gleich drei, (2) gleich vier, fünf oder acht, (3) kleiner zehn und (4) grösser als sechs ist.

**4.14** Jemand möchte Aussagen zu einem Parameter  $V$  machen. Die Person weiss, dass sie eine beliebige Zahl zwischen  $-\infty$  und  $+\infty$  sein könnte. Sie beschreibt daher ihr Wissen zu  $V$  mit einem stetigen Wahrscheinlichkeitsmodell. Dieses hat die Dichtefunktion

$$\text{pdf}(V = x) = \text{pdf}(x) = a \cdot \frac{1}{1 + x^4}$$

- (a) Wie muss  $a$  gewählt werden, dass  $\text{pdf}(x)$  eine Dichtefunktion ist?
- (b) Zeichnen Sie den Graphen der Dichtefunktion. Wie ist das Modell verteilt? Wie lautet der Modus?
- (c) Bestimmen Sie mit der Dichtefunktion die Wahrscheinlichkeit, dass  $V$  zwischen  $-3$  und  $1$  liegt. Wie gross ist die Wahrscheinlichkeit, dass  $V$  grösser als  $2,5$  ist? Wie lautet die Wahrscheinlichkeit, dass  $V$  zwischen  $2,3$  und  $2,35$  liegt?
- (d) Berechnen Sie mit – systematischem Ausprobieren – die  $0,025$ -,  $0,25$ -,  $0,75$ - und  $0,975$ -Quantile des Wahrscheinlichkeitsmodells. Wie lautet ein Wahrscheinlichkeitsintervall für  $V$  zum Niveau  $0,95$ ?
- (e) Lösen Sie die Aufgaben (c) und (d) auch mit einer Markov-Kette-Monte-Carlo Simulation. Erhalten Sie dabei ähnliche Resultate?

**4.15** Eine Person beschreibt die Plausibilität zu einer Grösse  $M \geq 0$  mit einem stetigen Wahrscheinlichkeitsmodell, gegeben mit der *halben Cauchyverteilung*:

$$\text{pdf}(M = x \mid \sigma) = \frac{2}{\pi \cdot \sigma} \cdot \frac{1}{1 + x^2/\sigma^2} \quad \text{für } x \geq 0$$

Man nennt  $\sigma$  den Skalierungsparameter der halben Cauchyverteilung.

- (a) Zeichnen Sie den Graphen der Dichtefunktion für  $\sigma = 1$ ,  $\sigma = 3$  und  $\sigma = 10$ . Welchen Effekt hat der Parameter  $\sigma$  auf den Graphen der Dichtefunktion?
- (b) Bestimmen Sie die folgenden Wahrscheinlichkeiten:

$$\mathbb{P}(1 \leq M \leq 2 \mid \sigma = 1), \quad \mathbb{P}(M \leq 0,5 \mid \sigma = 1), \quad \mathbb{P}(M > 6 \mid \sigma = 1)$$

- (c) Berechnen Sie die  $0,5$ - und  $0,9$ -Quantile des Modells, wenn  $\sigma = 1$  ist.

**4.16** Sein Wissen zu einer nicht negativen, stetigen Grösse  $T$  beschreibt jemand mit einer Exponentialverteilung mit Rate  $\lambda = 0,1$ :  $T \sim \text{Exponential}(0,1)$ . Berechnen Sie mit einem Taschenrechner oder einem Statistikprogramm die folgenden Wahrscheinlichkeiten:

$$\mathbb{P}(3 \leq T \leq 15 | \lambda), \quad \mathbb{P}(T < 10 | \lambda), \quad \mathbb{P}(T \geq 7 | \lambda)$$

Informieren Sie sich, wie Sie mit einem Statistikprogramm Quantile der Exponentialverteilung berechnen können. Wie lautet das 0,8-Quantil von  $T$ ?

**4.17** Bei gegebener Information  $\mathcal{K}$  macht eine Person Aussagen zu einer Grösse  $m$ , die zwischen 10 und 20 liegt, mit einem stetigen Wahrscheinlichkeitsmodell. Die Dichtefunktion des Modells ist

$$\text{pdf}(m = x | \mathcal{K}) = a \cdot \frac{(x - 10)^3 \cdot (20 - x)^4}{1000}$$

- (a) Wie muss  $a$  gewählt werden, dass  $\text{pdf}(x | \mathcal{K})$  eine Dichtefunktion ist?
- (b) Zeichnen Sie den Graphen der Dichtefunktion. Wie ist das Modell verteilt? Wie lautet der Modus?
- (c) Bestimmen Sie die Wahrscheinlichkeit, dass  $m$  zwischen 15 und 16 liegt. Wie gross ist die Wahrscheinlichkeit, dass  $m$  grösser als 12,5 ist? Wie lautet die Wahrscheinlichkeit, dass  $m$  zwischen 12 und 15 liegt?
- (d) Bestimmen Sie mit Ihrem Taschenrechner (und durch systematisches Ausprobieren) den Median, das 0,1-Quantil, das 0,6-Quantile und die beiden Quartile.
- (e) Wie lautet die Fünfzahlen-Zusammenfassung des Modells?
- (f) Bilden Sie eine MCMC-Kette, um die Dichtefunktion zu approximieren. Beurteilen Sie mit einem Histogramme der gebildeten Kette, ob die Approximation gut ist. Bestimmen Sie daraus Wahrscheinlichkeitsintervalle für  $m$  zum Niveau 0,68 und 0,95.

**4.18** Ihr Wissen aus Daten zu einem Parameter  $K > 0$  beschreibt eine Person mit einem stetigen Wahrscheinlichkeitsmodell mit Dichtefunktion

$$\text{pdf}(K = x | \text{Daten}) = 0,00135 \cdot x^3 \cdot \exp(-0,3 \cdot x) \quad \text{für } x \geq 0$$

- (a) Zeichnen Sie den Graphen der Dichtefunktion.
- (b) Wie lauten der Median, das 0,3-Quantil und die Quartile des Modells?
- (c) Bestimmen Sie Werte  $a$  und  $b$  so, dass  $K$  mit einer Wahrscheinlichkeit von 0,9 zwischen diesen Werten liegt. (*Tipp:* Berechnen Sie geeignete Quantile!)
- (d) Wie lautet der natürliche Logarithmus der Dichtefunktion? Vereinfachen Sie den erhaltenen Ausdruck so weit als möglich.
- (e) Bilden Sie eine MCMC-Kette, um die Dichtefunktion zu approximieren. Benutzen Sie dabei den Logarithmus der Dichtefunktion. Beurteilen Sie mit einem Histogramm der gebildeten Kette, ob die Approximation gut ist. Bestimmen Sie daraus Wahrscheinlichkeitsintervalle für  $K$  zum Niveau 0,5, 0,68 und 0,95.

**4.19** Bei der Firma Roche AG in Sisseln wurde eine Charge von gemahlenem  $\beta$ -Karotin Pulver mit einem Sieb analysiert. Das dabei gewonnene Information zur Teilchengrösse  $G$  (in  $\mu\text{m}$ ) wurde in [8] mit einer Weibull-Verteilung beschrieben. Dies ist ein stetiges Wahrscheinlichkeitsmodell. Es besitzt – für  $G > 0$  – die Dichtefunktion

$$\text{pdf}(G = x) = \frac{3,40}{158,53 \mu\text{m}} \cdot \left( \frac{x}{158,53 \mu\text{m}} \right)^{2,40} \exp \left( - \left( \frac{x}{158,53 \mu\text{m}} \right)^{3,40} \right) \quad \text{für } x > 0$$

- (a) Zeichnen Sie den Graphen der Dichtefunktion. Wie ist das Modell verteilt? Was sind die Modalwerte?
- (b) Berechnen Sie die Wahrscheinlichkeit, dass eine gemessene Teilchengrösse zwischen 100  $\mu\text{m}$  und 200  $\mu\text{m}$  liegen wird.
- (c) Wie gross ist die Wahrscheinlichkeit, dass eine gemessene Teilchengrösse grösser als 150  $\mu\text{m}$  sein wird?
- (d) Berechnen Sie mit Ihrem Taschenrechner (und mit geschicktem Ausprobieren) den Median, die 0,1-, 0,25- und 0,75-Quantile der Verteilung. Wie lautet die Fünf-Zahlen-Zusammenfassung des Wahrscheinlichkeitsmodells?
- (e) Wie lautet der natürliche Logarithmus der Dichtefunktion? Vereinfachen Sie den erhaltenen Ausdruck so weit als möglich.
- (f) Bilden Sie eine MCMC-Kette, um die Dichtefunktion zu approximieren. Beurteilen Sie mit einem Histogramm der gebildeten Kette, ob die Approximation gut ist. Lösen Sie mit der Simulation die Aufgaben (b)–(d). Bestimmen Sie daraus auch Wahrscheinlichkeitsintervalle für die Teilchengrösse zum Niveau 0,68 und 0,95.

**4.20** Anhand von Schadenereignissen über 200 000 CHF hat eine Versicherung in der Schweiz versucht, mit einem stetigen Wahrscheinlichkeitsmodell solche zukünftigen, grossen Schadensummen  $S$  zu prognostizieren. Sie erhielt dabei in [1] die Dichtefunktion

$$\text{pdf}(S = x \mid \text{Daten}) = \frac{1}{231\,408,2} \left( 1 + \frac{0,4 \cdot (x - 200\,000 \text{ CHF})}{231\,408,2} \right)^{-3,5}$$

Dabei ist  $x$  grösser als 200 000 CHF.

- (a) Zeichnen Sie den Graphen des Modells. Wie ist das Modell verteilt?
- (b) Wie plausibel sind die folgenden Aussagen? (1)  $S$  wird grösser als eine Million CHF sein, (2)  $S$  wird zwischen 300 000 und 500 000 CHF sein und (3)  $S$  wird kleiner als 600 000 CHF sein.
- (c) Wie lautet der Medianschaden? Berechnen Sie durch systematisches Ausprobieren die 0,25- und 0,75-Quantile des Modells. Wie lautet die Fünf-Zahlen-Zusammenfassung des Modells?
- (d) Konstruieren Sie eine MCMC-Kette, um die Dichtefunktion des Modells für  $S$  zu approximieren. Beurteilen Sie mit einem Histogramm der gebildeten Kette, ob die Approximation gut ist. Lösen Sie die Aufgaben (b) und (c) mit dieser Simulation. Bestimmen Sie daraus auch die 0,25- und 0,75-Quantile des Modells.

**4.21** Finden Sie beim Statistik-Programm Ihres Rechners, wie Sie die Dichtefunktion und wie Sie Quantile der Exponentialverteilung aufrufen können.

---

## Literatur

1. D. Bättig, A. Leu, *Einfluss von Lothar auf die Risikostruktur der XX Versicherung* (Berner Fachhochschule, Burgdorf, 2003)
2. H. Cerruti, Neue Zürcher Zeitung, Folio **04** (1999)
3. R. T. Cox, Probability, frequency and reasonable expectation. Am. J. Phys. **14**, 1–13 (1946)
4. R. G. Finch, G. L. French, I. Phillips, Group B streptococci in the female genital tract. Br. Med. J. **1**, 1245–1247 (1976)
5. A. Gelman, D. B. Rubin, Inference from iterative simulation using multiple sequences (with discussion). Statistical Science **7**, 457–511 (1992)
6. A. Gelman, K. Shirley: Inference from Simulations and Monitoring Convergence. In: Brooks, S., Gelman, A., Jones, G. L., Meng, X.-L. (Hrsg.) *Handbook of Markov Chain Monte Carlo*, Kapitel 6. (Chapman and Hall/CRC, 2011)
7. C. J. Geyser, Pratical Markov Chain Monte Carlo. Statistical Science **7**(4), 473–511 (1992)
8. K. Graf, Verfahrenstechnik. Manuskript (Berner Fachhochschule, Burgdorf, 2008)
9. W. K. Hastings, Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57**, 97–109 (1970)
10. P. D. Hoff, *A First Course in Bayesian Statistical Methods*, Springer Text in Statistics (Springer Verlag, 2009)
11. E. T. Jaynes, *Probability Theory, The Logic of Science* (Cambridge University Press, 2003)
12. A. Leu, E. Wyler: JAR – Return on IT Security, Bericht Institut für Risiko- und Extremwertanalyse, Berner Fachhochschule (2006)
13. D. V. Lindley, *Understanding Uncertainty* (John Wiley & Sons, 2007)
14. N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller, Equation of state calculations by fast computing machines. J. Chem. Phys. **2**, 1007–1092 (1953)
15. B. A. Orser, Schlummern ohne Risiko. Spektrum der Wissenschaft, **2** (2008)
16. P. Rosin, E. Rammler, The Laws Governing the Fineness of Powdered Coal. Journal of the Institute of Fuel **7**, 29–36 (1933)
17. D. S. Sivia, J. Skilling, *Data Analysis, a Bayesian Tutorial* (Oxford University Press, 2006)
18. E. Todd, *Apres l'empire, Essai sur la décomposition du système américain* (Gallimard, 2002)

*In diesem Augenblick rief Fünf, der schon dauernd ängstlich über den Garten hingeblickt hatte: „ Die Königin! Die Königin!“, und sogleich warfen sich die drei Gärtnner flach auf die Erde. Das Geräusch von vielen Schritten wurde vernehmbar, und Alice wandte sich gespannt um.*

*Lewis Carroll, Alice im Wunderland (Insel Taschenbuch, 1973, S. 81)*

## Zusammenfassung

Ein Ziel dieses Kapitels ist es zu zeigen, wie eine nicht direkt messbare Grösse aus Daten berechnet werden kann. Dazu braucht man, wie schon in Kap. 1 erklärt, ein Datenmodell, das beschreibt, wie die Daten streuen, und Vorinformation. Mit der Regel von Bayes kann daraus der Parameter bestimmt werden. Thomas Bayes, ein englischer Priester, hat sie um 1746 hergeleitet. Sie wurde im Jahr 1763 von Richard Price in [1] nach dem Tod Bayes veröffentlicht. Pierre-Simon Laplace entdeckte sie im Jahr 1812 in [4] neu und bemerkte, dass er mit Wahrscheinlichkeiten und der erwähnten Regel Plausibilitäten zu nicht direkt messbaren Größen aus Astronomie, Natur- und Sozialwissenschaften berechnen konnte. Er benutzte sie unter anderem, um das Verhältnis zwischen Jungen- und Mädchen geburten mit einer Genauigkeitsangabe und einer Plausibilität zu bestimmen. (S. B. McGrawe schildert im Buch [5] detailliert, wie die Regel von Bayes entstanden ist und wie die Gültigkeit und Anwendbarkeit der Regel während 150 Jahren oft in Frage gestellt wurden. In der heutigen Wissenschaft spielt sie aber eine nicht mehr wegzudenkende Rolle.)

## 5.1 Die Regel von Bayes

Oft hat man die Situation, dass man die Wahrscheinlichkeit zu einer ersten Aussage, wenn eine zweite Aussage als wahr angenommen wird, einfach berechnen kann. Ist etwa  $A$  die Aussage „Morgen regnet es in Bern heftig“, und  $B$  die Aussage „Die Marktgasse in

Bern wird morgen nass sein“, so ist  $\mathbb{P}(B \mid A) = 1$ . Schwieriger ist es, die zu dieser Wahrscheinlichkeit inverse Wahrscheinlichkeit  $\mathbb{P}(A \mid B)$  zu berechnen: Wie gross ist die Wahrscheinlichkeit, dass es regnet, wenn die Marktgasse nass ist? Ohne zusätzliche Information ist dies nicht möglich. Hier ein anderes Beispiel, das schon in Kap. 1 vorgestellt ist:

**Beispiel 5.1 (OptiMAL)** In Beispiel 1.1 wird erwähnt, dass OptiMAL ein Verfahren ist, um Personen auf Malaria zu testen. Mittels Probanden kann man ermitteln, wie zuverlässig dieses Verfahren ist. Es interessiert, ob es bei einer kranken Person auch wirklich positiv reagiert. Dies nennt man die *Sensitivität* oder die *Richtigpositiv-Rate*. Mathematisch geschrieben ist dies  $\mathbb{P}(\text{OptiMAL} + \mid \text{Patient hat Malaria})$ . Eine Ärztin möchte aber wissen, ob ein Patient Malaria hat, wenn OptiMAL positiv ausfällt. Dies ist die zur Sensitivität inverse Wahrscheinlichkeit.  $\square$

Wie man inverse Wahrscheinlichkeiten berechnet, sagt die Regel von Bayes. Sie wird nun hergeleitet. Mit dem Multiplikationsgesetz zu Wahrscheinlichkeiten ist einerseits

$$\mathbb{P}(A \text{ und } B) = \mathbb{P}(A \mid B) \cdot \mathbb{P}(B)$$

und andererseits hat man

$$\mathbb{P}(B \text{ und } A) = \mathbb{P}(B \mid A) \cdot \mathbb{P}(A)$$

Die linken Seiten der beiden Gleichungen sind identisch. Daher ist

$$\mathbb{P}(A \mid B) \cdot \mathbb{P}(B) = \mathbb{P}(B \mid A) \cdot \mathbb{P}(A)$$

Wenn  $\mathbb{P}(A) \neq 0$  ist, kann man dies nach  $\mathbb{P}(B \mid A)$  auflösen. Man schliesst daraus:<sup>1</sup>

---

<sup>1</sup> Die Regel von Bayes und Laplace wird im Folgenden wie üblich Regel von Bayes genannt. Es war aber Laplace, der die Formel benutzte, um aus Daten zu lernen. Hier dazu ein Auszug aus [5]:

Weil er [Laplace mit der Regel von Bayes] es geschafft hatte, die Spieltheorie in praktische Mathematik zu übersetzen, dominierte sein Werk ein ganzes Jahrhundert lang die Wahrscheinlichkeitstheorie und die Statistik. Glenn Shafer von der Rutgers University Newark, New Jersey, formulierte es so: „Meiner Meinung nach geht alles auf die Arbeit von Laplace zurück. Und wir interpretieren das nur in Thomas Bayes hinein. Laplace hat es modern formuliert. In gewissem Sinn ist alles Laplace.“

**Theorem 5.1 (Regel von Bayes und Laplace)**

Sind  $A$  und  $B$  Aussagen mit  $\mathbb{P}(A) \neq 0$ , so lautet die zu  $\mathbb{P}(A | B)$  inverse Wahrscheinlichkeit:

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

Dies lässt sich auch anders schreiben. Weiss man, wie plausibel die Aussage  $A$  ist, so ist der Nenner in der Formel eine feste Zahl. Er hängt nicht von der Art der Aussage  $B$  ab. Dies bedeutet:

**Theorem 5.2 (Lernen aus Information)**

$A$  sei eine Aussage mit  $\mathbb{P}(A) \neq 0$ . Dann gilt für alle Aussagen  $B$ :

$$\mathbb{P}(B | A) \text{ ist proportional zu } \mathbb{P}(A | B) \cdot \mathbb{P}(B)$$

Dabei hängt die Proportionalitätskonstante nicht von der Aussage  $B$  ab.

Arbeitet man mit stetigen Wahrscheinlichkeitsmodellen, um die Plausibilität zu  $B$  auszudrücken, so hat man Dichtefunktionen. In diesem Fall gilt analog

$$\text{pdf}(B | A) \text{ ist proportional zu } \mathbb{P}(A | B) \cdot \text{pdf}(B)$$

Die obige Form des „Lernen aus Information“ wird später immer wieder verwendet. Sie sagt, wie man die Plausibilität zu  $B$  neu berechnen kann, wenn man die zusätzliche Information  $A$  hat. Beim Beispiel zu OptiMAL ist  $B$  die Aussage „Der Patient hat Malaria.“ Die Wahrscheinlichkeit  $\mathbb{P}(B)$  muss von der Ärztin vorgegeben werden. Daraus kann man die Wahrscheinlichkeit  $\mathbb{P}(B | A)$  berechnen: die Wahrscheinlichkeit, dass der Patient Malaria hat, *gegeben* die zusätzliche Information  $A$  aus dem medizinischen Test. In Kap. 1 wird diese Wahrscheinlichkeit mit einer Chance formuliert. Man erhält mit der Regel von Bayes:<sup>2</sup>

$$\mathbb{O}(\text{krank} | \text{Test+}) = \frac{\text{Sensitivität}}{1 - \text{Spezifität}} \cdot \mathbb{O}(\text{krank})$$

---

<sup>2</sup> Es ist  $\mathbb{P}(B | A)$  proportional zu  $\mathbb{P}(A | B) \cdot \mathbb{P}(B)$  und für die Negation von  $B$  ist analog  $\mathbb{P}(B^{\text{nicht}} | A)$  proportional zu  $\mathbb{P}(A | B^{\text{nicht}}) \cdot \mathbb{P}(B^{\text{nicht}})$ . Bei beiden Gleichungen ist der Proportionalitätsfaktor gleich. Dividiert man daher die erste durch die zweite Gleichung, erhält man für die Chance  $\mathbb{O}(B | A)$

$$\mathbb{O}(B | A) = \frac{\mathbb{P}(A | B)}{\mathbb{P}(A | B^{\text{nicht}})} \cdot \mathbb{O}(B)$$

Hier ein weiteres Beispiel dazu:

**Beispiel 5.2 (Zeit zwischen starken Erdbeben)** Beim Beispiel 1.3 zu den Zeiten zwischen starken Erdbeben von Kap. 1 will man die durchschnittliche Zeit  $\mu$  zwischen zukünftigen, aufeinanderfolgenden starken Erdbeben berechnen. Die Zahl  $\mu$  ist eine nicht direkt messbare Größe. Deshalb wird man  $\mu$  aus den beobachteten 28 Wartezeiten bestimmen und angeben, wie plausibel beispielsweise Aussagen der Form  $\mu = (420 \pm 10)$  Tage oder  $\mu = (500 \pm 200)$  Tage sind. Mathematisch geschrieben sucht man also  $\mathbb{P}(\mu = \dots | \text{Daten})$ . Die Regel von Bayes besagt, wie man dies tun kann:

$$\mathbb{P}(\mu = \dots | \text{Daten}) = \frac{\mathbb{P}(\text{Daten} | \mu = \dots)}{\mathbb{P}(\text{Daten})} \cdot \mathbb{P}(\mu = \dots)$$

Der Ausdruck  $\mathbb{P}(\text{Daten} | \mu = \dots)$  im Zähler des Bruches ist die inverse Wahrscheinlichkeit zur gesuchten Wahrscheinlichkeit. Man nennt ihn die *Likelihood-Funktion* oder die *Likelihood*. Er sagt, wie plausibel Werte von Zeiten zwischen starken Erdbeben sind, wenn man die durchschnittliche Wartezeit  $\mu$  kennt. Um ihn zu bestimmen, muss man wissen, wie Zeiten zwischen starken Erdbeben um  $\mu$  streuen. Dazu braucht man ein *Streu-* oder *Datenmodell*. Solche Modelle lassen sich mit im Buch später vorgestellten Werkzeugen aufstellen. Den Ausdruck  $\mathbb{P}(\text{Daten})$  im Nenner der obigen Gleichung nennt man die *Evidenz* oder die *marginale Likelihood*. Er hängt nicht von  $\mu$  ab. Also hat man wieder die Formel zum Lernen aus Information:

$$\mathbb{P}(\mu = \dots | \text{Daten}) \propto \mathbb{P}(\text{Daten} | \mu = \dots) \cdot \mathbb{P}(\mu = \dots)$$

Das Zeichen  $\propto$  ist schon im Kap. 4 eingeführt worden. Es bedeutet „ist proportional.“ Der Faktor  $\mathbb{P}(\mu = \dots)$  ganz rechts besagt, wie plausibel Aussagen zu  $\mu = \dots$  sind, bevor die Daten betrachtet werden. Dies nennt man die *A priori*-Wahrscheinlichkeit (engl. *prior-probability*) oder den *Prior* zu  $\mu$ . Dazu benötigt man Informationen, sei es Vorwissen oder seien es Rechnungen aus anderen Datensätzen.

Die Regel von Bayes aktualisiert also den Prior oder das (ungenaue) Vorwissen zu  $\mu$ . Sie führt dazu die Daten (kodiert in der Likelihood) und den Prior (das Vorwissen) zusammen. Man nennt  $\mathbb{P}(\mu = \dots | \text{Daten})$  die *A posteriori*-Wahrscheinlichkeit (engl. *posterior-probability*) zu  $\mu$ .  $\square$

Wie die einzelnen Bestandteile in der Regel von Bayes berechnet werden, um nicht direkt messbare Größen zu bestimmen, zeigt exemplarisch der nächste Abschnitt.

---

Der Nenner im ersten Faktor lässt sich mit dem Additionsgebot umschreiben. Dies führt zu

$$\mathbb{O}(B | A) = \frac{\mathbb{P}(A | B)}{1 - \mathbb{P}(A^{\text{nicht}} | B^{\text{nicht}})} \cdot \mathbb{O}(B)$$

Diese Formel wird beim medizinischen Test zu OptiMAL benutzt.

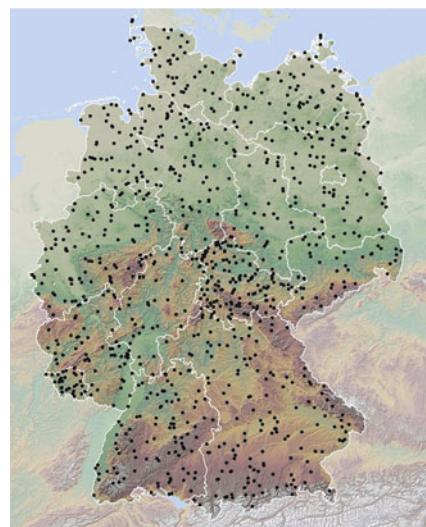
## 5.2 Berechnen eines Anteils

Oft sind Objekte aus einer Grundgesamtheit in zwei Kategorien aufteilbar. So hat man bei einer Produktionsserie Geräte, die defekt oder nicht defekt sind. Der Anteil der defekten Geräte einer Produktionsserie ist eine oft gesuchte Zahl. Patienten können ein bestimmtes Medikament vertragen oder nicht vertragen. Eine medizinische Kontrollstelle möchte dann berechnen, wie gross der Anteil der Personen ist, die das Medikament nicht vertragen.

**Beispiel 5.3 (HNV-Indikator)** Die Europäische Union verfolgt mit vielen Indikatoren, wie sich die Landwirtschaft entwickelt. Der High Nature Value Ackerland Indikator (HNV-Indikator) besagt, ob eine intensiv genutzte landwirtschaftliche Fläche einen Naturschutzwert hat. Mit diesem Indikator fallen Flächen in zwei Kategorien: sie haben einen Naturschutzwert oder sie haben keinen. Im Jahr 2009 fand eine erste Erfassung des HNV-Indikators statt. Man erhielt: Mit einer Wahrscheinlichkeit von 0,68 ist der Anteil der HNV Ackerland-Bestände an der landwirtschaftlichen Fläche in Deutschland 13,0 %  $\pm$  0,4 %.

Um den HNV-Indikator zu kennen, wird die Fläche von Deutschland in  $1 \text{ km}^2$  grosse, quadratische Flächen zerlegt. Mit einer zufällig ausgewählten Stichprobe der Flächen wird dann versucht, den HNV-Indikator zu berechnen. Abb. 5.1 zeigt die dazu benutzten Stichprobenflächen. Stichprobenflächen mit weniger als 5 % Anteil an Landwirtschaftsfläche werden hingegen nicht untersucht. Um den HNV-Indikator trotzdem zu bestimmen, ist es nötig, den Anteil  $A$  solcher Flächen in Deutschland zu kennen. Wir wollen solche Flächen Klasse 1-Flächen nennen, die anderen Klasse 0-Flächen. Der Anteil  $A$  ist eine Zahl zwischen null und eins. Ist  $A = 1$ , so sind alle  $1 \text{ km}^2$ -Flächen in Deutschland in Klasse

**Abb. 5.1** Punkte zeigen die zufällig gewählten Probe- flächen im Jahr 2009 (aus Bundesamt für Naturschutz 2013, Statistisches Bundes- amt 2004, Geobasisdaten: ©GeoBasis-DE/BKG)



eins. Falls  $A$  bekannt ist, ist es einfach die Wahrscheinlichkeit auszurechnen, dass eine in randomisierter Art gezogene Fläche  $F_i$  in der ersten oder nullten Klasse ist:

$$\mathbb{P}(F_i \text{ in Klasse } 1 | A) = A, \quad \mathbb{P}(F_i \text{ in Klasse } 0 | A) = 1 - A$$

Dies ist das *Datenmodell* oder das *Streumodell*. Es handelt sich um die *Bernoulli-Verteilung*. Dies schreibt man als

$$\text{Datenmodell: } i\text{-te Fläche } F_i \sim \text{Bernoulli}(A)$$

Der Anteil  $A$  ist nicht bekannt und muss aus einer Stichprobe berechnet werden. Hier das Resultat aus acht von einer Biologin ausgewählten Stichprobenflächen:

Klasse	1	1	0	0	0	1	0	0
--------	---	---	---	---	---	---	---	---

Die Frage ist: Wie lautet – gegeben die acht Messwerte – der Anteil  $A$ ? Beachten Sie: Der Parameter  $A$  bezieht sich auf alle  $1 \text{ km}^2$ -Flächen in Deutschland und nicht auf die Stichprobe mit acht Flächen. Man sieht, dass drei von acht Stichprobenflächen in Klasse eins sind. Ist es daher vernünftig,  $A$  mit ungefähr  $3/8 = 0,375$  anzugeben? Wie genau und wie plausibel ist dieses Resultat? Gemäß den Angaben in Kap. 4 sollte die Biologin ihr Wissen zu  $A$  daher mit einer Wahrscheinlichkeit ausdrücken. Gesucht ist also

$$\mathbb{P}(A | \text{Daten})$$

Dies ist die zu den obigen Ausdrücken inverse Wahrscheinlichkeit. Mit der Regel von Bayes kann sie berechnet werden:

$$\mathbb{P}(A | \text{Daten}) = \frac{\mathbb{P}(\text{Daten} | A)}{\mathbb{P}(\text{Daten})} \cdot \mathbb{P}(A)$$

Der Nenner des Bruchs, die Evidenz, hängt nicht vom Parameter  $A$  ab. Man hat somit die im vorigen Abschnitt schon vorgestellte „Lernformel“

$$\mathbb{P}(A | \text{Daten}) \propto \mathbb{P}(\text{Daten} | A) \cdot \mathbb{P}(A) \tag{5.1}$$

Der Faktor ganz rechts quantifiziert die Plausibilität zu  $A$  aus möglicher Vorinformation. Er ist der Prior oder die *A priori-Verteilung*. Da für  $A$  kontinuierliche Werte zwischen null und eins in Frage kommen, kann man die Plausibilität zu  $A$  mit einem *stetigen* Wahrscheinlichkeitsmodell formulieren. Die Biologin habe keine Vorinformation und halte jeden Wert von  $A$  für gleich plausibel, zum Beispiel:

$$\mathbb{P}(A = 0,5) = \mathbb{P}(A = 0,4) = \mathbb{P}(A = 0,2) = \dots$$

**Abb. 5.2** Information zum Anteil  $A$  vor den Messungen:  
eine flache Verteilung

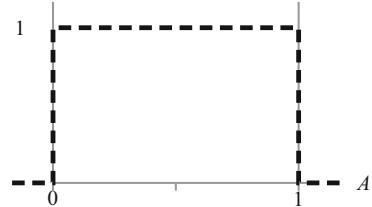


Abb. 5.2 visualisiert diese minimale Vorinformation: eine *flache* Verteilung, die Gleichverteilung auf dem Intervall  $[0; 1]$ :

$$\text{Prior: } A \sim \text{Uniform}(0; 1)$$

Die Dichtefunktion ist eins für  $A$  zwischen null und eins. Sie ist proportional zu  $\mathbb{P}(A)$ . Damit folgt

$$\mathbb{P}(A) = \mathbb{P}(A \mid \text{min. Vorinformation}) \propto 1$$

Der verbleibende Faktor  $\mathbb{P}(\text{Daten} \mid A)$  in Gleichung (5.1) sagt, wie plausibel die gemessenen Werte sind, falls  $A$  bekannt wäre. Man nennt ihn die *Likelihood* oder die *Likelihood-Funktion*. Sie ist mit dem Datenmodell bestimmbar. Nimmt man an, dass die Messwerte der Stichprobenflächen unabhängig sind,<sup>3</sup> so ist nach dem Multiplikationsgesetz für unabhängige Aussagen:

$$\begin{aligned} \mathbb{P}(\text{Daten} \mid A) &= \mathbb{P}(1 \text{ und } 1 \text{ und } 0 \text{ und } \dots \text{ und } 0 \mid A) \\ &= \mathbb{P}(1 \mid A) \cdot \mathbb{P}(1 \mid A) \cdot \mathbb{P}(0 \mid A) \cdot \dots \cdot \mathbb{P}(0 \mid A) \end{aligned}$$

Mit dem Datenmodell ergibt sich damit für die Likelihood-Funktion:

$$\mathbb{P}(\text{Daten} \mid A) = A \cdot A \cdot (1 - A) \cdot \dots \cdot (1 - A) = A^3 \cdot (1 - A)^5$$

Die Likelihood ist also ein Term, der vom gesuchten Parameter  $A$  abhängt. Setzt man ihn in die obige Gleichung ein, ergibt sich für die Plausibilität zu  $A$ :

$$\mathbb{P}(A \mid \text{Daten, min. Vorinformation}) \propto \mathbb{P}(\text{Daten} \mid A) \cdot 1 = A^3 \cdot (1 - A)^5 \cdot 1$$

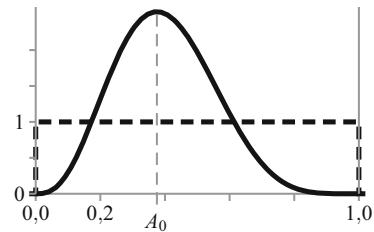
Dies ist die Dichtefunktion, genauer die *A posteriori-Verteilung*, des gesuchten Anteils  $A$ . Der Graph der Verteilung ist in Abb. 5.3 gezeichnet. Man sieht, wo  $A$  mit hoher Wahrscheinlichkeit liegt. Der plausibelste Wert von  $A$  ist der Modus  $A_0$  der A posteriori-Verteilung. Er lässt sich mit einer Wertetabelle und einem Computer schnell finden. Er beträgt – nicht überraschend –  $3/8 = 0,375$ . Man kann den gesuchten Anteil durch diesen Wert *schätzen*:

$$A \approx A_0 = 0,375$$

---

<sup>3</sup> Dies ist bei den Stichprobenflächen sicher zu hinterfragen. Weil die Stichprobenflächen randomisiert gewählt wurden, ist die Unabhängigkeit wahrscheinlich.

**Abb. 5.3** Plausibilität des Anteils  $A$ , gegeben die Information aus 8 Messungen, davon 3 in Klasse 1; gestrichelt die Plausibilität zu  $A$  vor der Datensammlung



Man nennt dies die *MAP-Schätzung* (aus Maximum A Posterior) für den Anteil  $A$ . Aus der Grafik der A posteriori-Verteilung ist ersichtlich, dass der Anteil  $A$  nicht sehr präzis lokalisiert ist: Mit hoher Wahrscheinlichkeit dürfte  $A$  zwischen 0,1 und 0,9 sein.

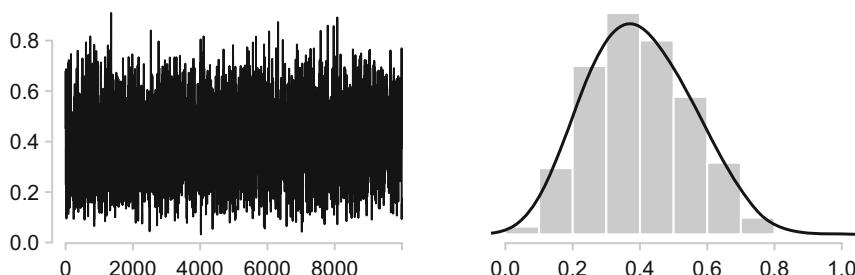
Mit einer MCMC-Simulation kann man Wahrscheinlichkeitsintervalle für den Anteil  $A$  berechnen. Die fehlende Proportionalitätskonstante im Ausdruck für die Dichtefunktion der A posteriori-Verteilung braucht man dazu nicht. Dies ist praktisch! Bei Statistikprogrammen, die mit der Bayes-Regel arbeiten, genügt es, die Daten, das Datenmodell und den Prior für den Parameter  $A$  zu nennen. Das Datenmodell und der Prior werden in einer mathematischen Schreibweise angegeben:

$$\text{Datenmodell: } i\text{-ter Messwert} \sim \text{Bernoulli}(A), \quad \text{Prior: } A \sim \text{Uniform}(0 ; 1)$$

Daraus erstellen Statistikprogramme mit dem Produkt Likelihood  $\times$  Prior die A posteriori-Verteilung des Parameters  $A$  und tasten diese mit MCMC-Simulationen ab. Abb. 5.4 zeigt eine solche Kette von 10 000 Punkten. Das 0,25-Quantil besteht aus dem 2500. kleinsten Punkt der Kette und beträgt 0,29. Das 0,75-Quantil, bestimmt mit dem 2500. grössten Punkt der Kette, ist 0,51. Man erhält: Es besteht eine Wahrscheinlichkeit von 0,5, dass

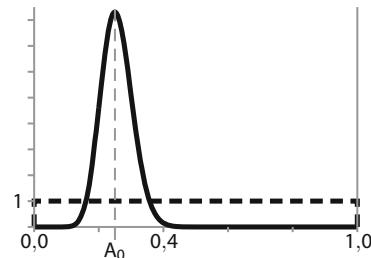
$$0,29 \leq A \leq 0,50$$

Ein Wahrscheinlichkeitsintervall zum Niveau 0,95 für  $A$  ist durch die Zahlen 0,14 und 0,70 gegeben. Die Fünf-Zahlen-Zusammenfassung der A posteriori-Verteilung des Anteils  $A$



**Abb. 5.4** MCMC-Kette für die A posteriori-Verteilung des Anteils  $A$

**Abb. 5.5** Plausibilität des Anteils  $A$ , gegeben 80 Messungen, davon 20 in Klasse 1



lautet:

Plausibilität zu $A$ :	min	$q_{0,25}$	med	$q_{0,75}$	max
	0,0	0,29	0,39	0,50	1,0

Man kann die Wahrscheinlichkeitsintervalle auch mit Integralen berechnen. Die fehlende Konstante für die  $A$  posteriori-Dichtefunktion muss dazu bestimmt werden. Die Fläche unter dem Graphen der  $A$  posteriori-Dichte muss eins sein. Man erhält daraus die Dichtefunktion

$$\text{pdf}(A \mid \text{Daten, min. Vorinformation}) = 504 \cdot A^3 \cdot (1 - A)^5$$

Mit systematischem Ausprobieren lassen sich Quantile und damit Wahrscheinlichkeitsintervalle ausrechnen. Man erhält die gleichen Resultate wie mit der Simulation.

Die Biologin hätte statt acht Flächen, 80 untersuchen können. Davon seien 20 in Klasse eins. Die  $A$  posteriori-Dichtefunktion des Anteils  $A$  ist dann (vgl. Abb. 5.5):

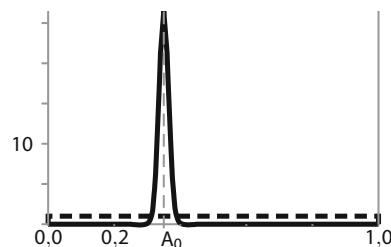
$$\text{pdf}(A \mid \text{Daten, min. Vorinformation}) \propto A^{20} \cdot (1 - A)^{60} \cdot 1$$

Der plausibelste Wert von  $A$  ist gleich 0,25. Der Graph der Dichtefunktion zeigt eine schmale Spitze. Der Anteil  $A$  ist damit präziser als bei acht Messungen bestimmt. So hat man mit einer Wahrscheinlichkeit von 0,5, dass  $A$  zwischen 0,22 und 0,29 ist. Weiter beträgt die Wahrscheinlichkeit 0,95, dass  $A$  zwischen 0,17 und 0,36 liegt.

Abb. 5.6 zeigt, wie der Anteil noch präziser bestimmt werden könnte, wenn der Stichprobenumfang erhöht wird. Dargestellt ist die  $A$  posteriori-Verteilung von  $A$ , wenn 1000 Flächen untersucht würden, davon 350 in der Klasse eins.  $\square$

Das Beispiel zu den Stichprobenflächen für den HNV-Indikator zeigt:

**Abb. 5.6** Plausibilität des Anteils  $A$ , gegeben 1000 Messungen, davon 350 in Klasse 1



**Theorem 5.3 (Bestimmen eines Anteils)**

In einer Grundgesamtheit befinden sich Objekte in zwei Kategorien 0 und 1. Gesucht ist der Anteil  $A$  an Objekten in der Grundgesamtheit, die in Kategorie 1 sind. Eine Stichprobe mit  $n$  Messungen, die unabhängig sind, liefert das Resultat:

$$\text{Anzahl in Kategorie 1: } \alpha, \quad \text{Anzahl in Kategorie 0: } n - \alpha$$

Aus Vorinformation  $\mathcal{I}$  kann mit den Daten die Plausibilität zu  $A$  aktualisiert werden:

$$\underbrace{\text{pdf}(A \mid \text{Daten}, \mathcal{I})}_{\text{Posterior}} \propto \underbrace{A^\alpha \cdot (1 - A)^{n-\alpha}}_{\text{Likelihood aus Datenmodell}} \cdot \underbrace{\text{pdf}(A \mid \mathcal{I})}_{\text{Prior}}$$

Die  $A$  posteriori-Verteilung des Anteils  $A$  der Objekte in Kategorie 1 ist damit bis auf eine Konstante bestimmt. Wahrscheinlichkeitsintervalle zu  $A$  lassen sich deshalb mit einer MCMC-Simulation einfach bestimmen. Ist nur minimale Vorinformation  $\mathcal{I}$  vorhanden, so ist der Prior zu  $A$  flach:  $\text{pdf}(A \mid \mathcal{I}) = 1$ . Man hat dann

$$\text{Datenmodell: } i\text{-ter Messwert} \sim \text{Bernoulli}(A), \quad \text{Prior: } A \sim \text{Uniform}(0; 1)$$

Man nennt den erhaltenen Posterior für  $A$  die *Beta-Verteilung* mit Kennzahlen  $\alpha + 1$  und  $n - \alpha + 1$ . Beim obigen Beispiel mit  $\text{pdf}(A \mid \text{Daten}) \propto A^3 \cdot (1 - A)^5$  handelt es sich um die Beta-Verteilung mit Kennzahlen 4 und 6. Die Beta-Verteilung ist in den meisten Statistikprogrammen verfügbar.<sup>4</sup> Damit sind hier zugehörige Wahrscheinlichkeitsintervalle schnell abrufbar.

Auch mit nur einer Messung kann man berechnen, wo  $A$  mit hoher Wahrscheinlichkeit ist. Ist das Merkmal nicht beobachtet worden, so ist

$$\text{pdf}(A \mid \text{Daten, min. Vorinformation}) \propto A^0 \cdot (1 - A)^{1-0} \cdot 1 = 1 - A$$

Dies ist eine Beta-Verteilung mit Kennzahlen eins und zwei. Der plausibelste Wert von  $A$  ist null. Der Median der Verteilung beträgt 0,29. Es besteht daher eine Wahrscheinlichkeit von 0,5, dass der gesuchte Anteil  $A$  zwischen 0 und 0,29 ist.

---

<sup>4</sup> Die Beta-Verteilung mit Kennzahlen  $a$  und  $b$  hat die Dichtefunktion

$$\text{pdf}(x \mid a, b) = \frac{(a + b - 1)!}{(a - 1)! \cdot (b - 1)!} \cdot x^{a-1} \cdot (1 - x)^{b-1} \quad \text{für } 0 \leq x \leq 1$$

Der Quotient mit den Fakultäten bewirkt, dass die Fläche unter der Dichtefunktion eins ist. Er wurde schon von Bayes berechnet. Dabei ist  $k!$  das Produkt der ersten  $k$  natürlichen Zahlen:  $1! = 1$ ,  $2! = 1 \cdot 2$ ,  $3! = 1 \cdot 2 \cdot 3$ . Man setzt  $0! = 1$ .

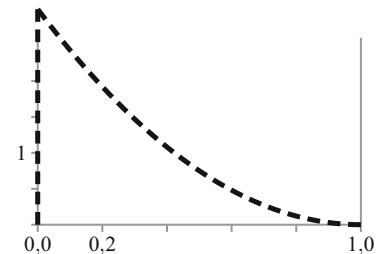
### 5.3 Wie hängt das Resultat von der Vorinformation ab?

Man kann untersuchen, wie die vorherigen Resultate vom Prior abhängen. So habe ein Biologe als Information  $\mathcal{K}$ , dass der Anteil der Flächen in Deutschland, die in Klasse eins sind, tendenziell klein ist, etwa  $\text{pdf}(A | \mathcal{K}) \propto A^0 \cdot (1 - A)^2$ . Dies ist eine Beta-Verteilung mit Kennzahlen 1 und 3. Es könnte sein, dass er dieses Wissen durch Resultate aus einer früheren Untersuchung hat – zwei Stichprobenflächen, alle in Klasse null. Die Dichtefunktion, die diese Information spiegelt, ist in Abb. 5.7 gezeichnet. Wiederum seien acht Stichprobenflächen untersucht worden. Davon seien drei in Klasse eins. Die A posteriori-Verteilung von  $A$  ist dann

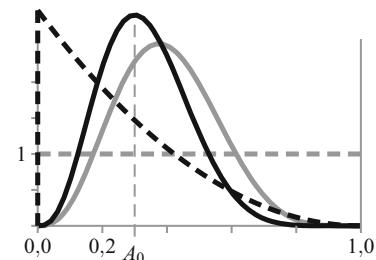
$$\underbrace{\text{pdf}(A | \text{Daten}, \mathcal{K})}_{\text{A posteriori}} \propto \underbrace{A^3 \cdot (1 - A)^5}_{\text{Likelihood}} \cdot \underbrace{\text{pdf}(A | \mathcal{K})}_{\text{A priori}} \propto A^3 \cdot (1 - A)^5 \cdot (1 - A)^2$$

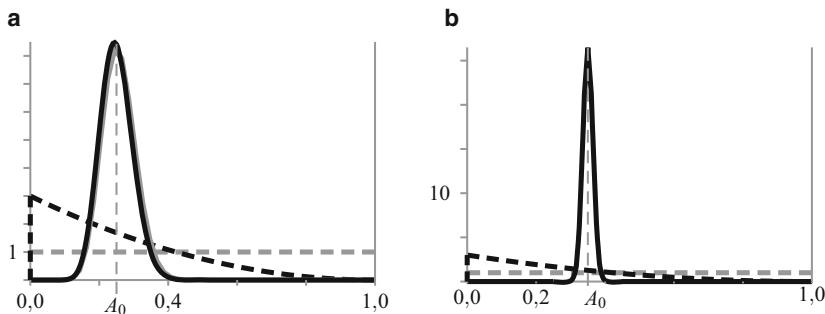
Dies ist eine Beta-Verteilung mit Kennzahlen 4 und 8. Abb. 5.8 zeigt ihren Graphen. In grauer Farbe sind zusätzlich – gestrichelt – die A priori- und – ausgezogen – die A posteriori-Verteilung bei flachem Prior zu  $A$  gezeichnet. Der plausibelste Wert von  $A$  ist nicht mehr 0,375, sondern er ist 0,3! Die zusätzlich benutzte Information  $\mathcal{K}$  des Biologen spiegelt sich stark im Resultat. Dies ist durchaus wünschbar. Man will nämlich die beste Schätzung für  $A$  haben, gegeben die gesamte Information. Wegen der Vorinformation  $\mathcal{K}$  ist die A posteriori-Verteilung schmäler geworden. Der gesuchte Parameter ist präziser als bei Unwissen lokalisiert. So besteht eine Wahrscheinlichkeit von 0,5, dass  $A$  zwischen 0,23 und 0,42 liegt. Bei minimaler Vorinformation zu  $A$  sind die entsprechenden Grenzen des Wahrscheinlichkeitsintervalls 0,29 und 0,50. Dies ist nicht erstaunlich: Je mehr Infor-

**Abb. 5.7** Plausibilität des Anteils  $A$  vor der Datenerhebung (Prior)



**Abb. 5.8** In schwarzer Farbe: Gestrichelt der Prior zu  $A$ , ausgezogen der Posterior; In grauer Farbe: gestrichelt flacher Prior, ausgezogen zugehöriger Posterior



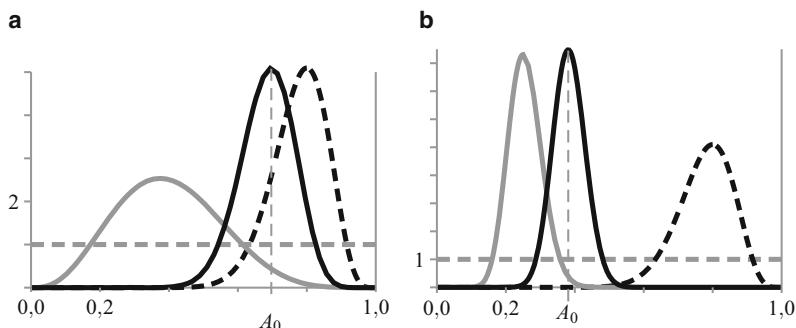


**Abb. 5.9** Plausibilität zum Anteil  $A$  bei 80 Messungen (a) und 1000 Messungen (b)

mation in die Regel von Bayes gesteckt wird, um so präziser werden gesuchte Parameter in der Regel bestimmt.

Abb. 5.9 zeigt die  $A$  posteriori-Verteilung mit den gleichen Daten wie im obigen Abschnitt, mit 80 Messungen und mit 1000 Messungen. Der Prior von  $A$  ist gestrichelt gezeichnet. Die grauen Linien stellen – zum Vergleich mit den obigen Rechnungen – die Rechnung zu  $A$  bei flachem Prior dar. Je grösser die Datenmenge wird, desto kleiner wirkt die Vorinformation  $\mathcal{K}$  des Biologen auf das Resultat. Bei 1000 Messungen unterscheiden sich die beiden  $A$  posteriori-Verteilungen von  $A$  kaum voneinander.

Abb. 5.10 zeigt, wie sich die  $A$  posteriori-Verteilung von  $A$  verhält, wenn der Biologe eine Vorinformation  $\mathcal{J}$  hat, die der Grundgesamtheit kaum entspricht. Der Biologe geht davon aus, dass mit hoher Wahrscheinlichkeit der Anteil  $A$  um 0,8 liegen müsste.<sup>5</sup> Da-



**Abb. 5.10** Plausibilität zum Anteil  $A$  bei acht Messungen (a) und 80 Messungen (b); Graph der  $A$  priori-Verteilung schwarz gestrichelt, in grauer Farbe die  $A$  posteriori-Plausibilität zum Anteil  $A$ , wenn der Prior flach ist

<sup>5</sup> Beispielsweise hat eine Voruntersuchung ergeben, dass bei zehn Flächen acht in Kategorie 1 waren. Die  $A$  priori-Verteilung von  $A$  ist dann  $\text{pdf}(A) \propto A^2 \cdot (1 - A)^8$ .

bei ist die Vorinformation wiederum mit einer gestrichelten Linie gezeichnet. In grauen Linien sind die Rechnungen bei minimaler Vorinformation zu  $A$  dargestellt: Bei wenigen Messungen beeinflusst die Vorinformation des Biologen die A posteriori-Verteilung von  $A$  stark. Die plausibelste Schätzung ist  $A_0 = 0,697$ , statt 0,375 beim flachen Prior zu  $A$ . Die Präzision der Schätzung ist schon gross. Sie ist vor allem ein Ausdruck der Vorinformation  $J$ . Die A posteriori-Verteilung bei minimaler Vorinformation zu  $A$  (graue Linie) und die neu gewählte A priori-Verteilung überlagern sich nur wenig. Dies kann ein Zeichen sein, dass die Information  $J$  des Biologen schlecht ist oder die Daten systematische Fehler haben. Bei 80 Messungen (Grafik rechts) spielt  $J$  eine weniger wichtige Rolle. Hätte man 1000 Messungen, wären die graue und schwarze Linie fast deckungsgleich. Die vielen Daten spielen die dominante Rolle und „korrigieren“ die falsche Vorinformation.

**A priori, Likelihood, A posteriori:** Die A posteriori-Plausibilität aktualisiert die Plausibilität zu einer Grösse aus dem A priori-Wissen mit Hilfe der Likelihood. Die Likelihood wird vom Datenmodell und den Daten bestimmt. Das A priori-Wissen setzt sich aus Vorinformation zusammen. Je mehr Messungen man hat, umso dominanter wird die Likelihood und umso weniger spielt das A priori-Wissen eine Rolle.<sup>6</sup>

Hat man viele Datenwerte, so ist die Likelihood dominant. In diesem Fall lassen sich Wahrscheinlichkeitsintervalle approximativ mit einfachen Formeln und mit dem Standardfehler SE angeben.<sup>7</sup> Ist der aus den Daten beobachtete Anteil  $A_0$  nicht nahe bei null oder nahe bei eins, so ist

$$A = A_0 \pm 1,96 \cdot \text{SE}(A) = A_0 \pm 1,96 \cdot \sqrt{\frac{A_0 \cdot (1 - A_0)}{n}}$$

ein Wahrscheinlichkeitsintervall für  $A$  zum Niveau von *etwa* 0,95 (also einem Fehler 1. Art von etwa 5 %).

Beim oben besprochenen Beispiel zur Entwicklung der Landwirtschaft ist der Anteil  $A$  der Flächen in Klasse eins gesucht. Dreissig von achtzig Stichprobenflächen seien in der Klasse eins. Mit der Formel lautet ein Wahrscheinlichkeitsintervall für  $A$  zum Niveau von etwa 0,95:

$$A = \frac{30}{80} \pm 1,96 \cdot \sqrt{\frac{30/80 \cdot (1 - 30/80)}{80}} = 0,375 \pm 0,106$$

---

<sup>6</sup> Man kann untersuchen, wie stark ein Resultat von der A priori-Verteilung abhängt. Dazu kann man verschiedene A priori-Verteilungen in die Regel von Bayes einsetzen. Man spricht auch von einer *Sensitivitätsanalyse*.

<sup>7</sup> Was genau der Standardfehler SE ist, wird in Kap. 14 erklärt. Die Approximationsformel folgt aus der Methode von Laplace, die jenem Kapitel vorgestellt wird. Obwohl sie sehr beliebt ist, können die Resultate schlecht sein. Auch erhält man bessere Resultate, wenn man zuerst den Posterior von  $A$  mit der logit-Transformation umwandelt, und anschliessend die Laplace-Methode benutzt.

Bei diesem Resultat ist der Fehler 1. Art 5 % und der Fehler 2. Art beträgt  $\pm 0,106$ . Der exakte Fehler 1. Art lautet 4,6 %.

**Beispiel 5.4 (OptiMAL)** In Beispiel 1.1 berechnet eine Ärztin die A posteriori-Wahrscheinlichkeit  $\mathbb{P}(\text{Malaria} | \text{OptiMAL+}, \text{Vorinformation})$ , dass ihr Patient Malaria hat, aus einer einzigen Messung. Diese ist der positive OptiMAL-Test. Der Prior (das Vorwissen) beeinflusst damit das Resultat wesentlich.  $\square$

**Beispiel 5.5 (Versicherungspolicen)** Eine Versicherungsfirma, die Schäden an Häusern infolge Feuer, Wassereinbrüchen oder Stürmen versichert, will neu auch Policen in einer Region anbieten, in der sie bisher nicht tätig war. Dazu muss die Versicherungsprämie für solche Policen festgelegt werden. Der in der Firma tätige Aktuar braucht dafür unter anderem den Anteil  $A$  der Policen, die im nächsten Jahr Schadensfälle melden werden. Wie soll der Aktuar den Anteil  $A$  berechnen, wenn keine Daten zur neuen Region vorhanden sind? Wählt er minimale Vorinformation zu  $A$ , so ist jeder Anteil  $A$  gleich wahrscheinlich. Der Anteil  $A$  ist damit nicht festgelegt und die Versicherungsprämie kann nicht bestimmt werden.

Der Aktuar hat Vorwissen aus den Policen der bisherigen Regionen: von 100 Kontrakten meldeten fünf Schadensfälle. Der Aktuar kann daher versuchen, mit dieser Information  $A$  zu beschreiben:

$$\text{pdf}_{\text{neue Region}}(A | \text{Information}) \propto A^5 \cdot (1 - A)^{95}$$

Der plausibelste Wert  $A_0$  von  $A$  ist  $5/100 = 0,05$  und mit einer Wahrscheinlichkeit von 0,95 ist

$$0,02 \leq A \leq 0,11$$

Eine Versicherungsprämie kann damit berechnet werden. Nach einem Jahr hat die Firma sechs Policen in der neuen Region abgeschlossen. Der Inhaber einer Police meldet einen Schaden. Wie lautet nun  $A$ ? Mit dem obigen Vorwissen ist

$$\text{pdf}_{\text{neue Region}}(A | \text{Daten, Information}) \propto \underbrace{A^1 \cdot (1 - A)^5}_{\text{Likelihood}} \cdot \underbrace{A^5 \cdot (1 - A)^{95}}_{\text{Prior} = \text{Information}}$$

Die Information aus der alten Region ist hier entscheidend. Sie ermöglicht dem Aktuar, den Anteil  $A$  mit genügender Präzision zu berechnen! Der plausibelste Wert  $A_0$  von  $A$  und ein Wahrscheinlichkeitsintervall zum Niveau 0,95 sind

$$A_0 = \frac{6}{106} = 0,06 \quad \text{und} \quad 0,03 \leq A \leq 0,12$$

Würde der Aktuar nur die sechs Policen der neuen Region betrachten, hätte er  $A_0 = 1/6 = 0,17$  erhalten. Ein Wahrscheinlichkeitsintervall zum Niveau 0,95 wäre  $0,04 \leq$

$A \leq 0,58$ . Wegen der wenigen Policen wäre das Resultat daher sehr unsicher und kaum brauchbar.

Das vorgestellte Verfahren lässt sich allgemein durchführen. Es ist im Versicherungswesen beliebt und wurde 1918 von A. W. Whitney in [6] entwickelt. Gegeben ist:

	Policen mit Schäden	Policen ohne Schäden	Total
Alte Region	$a$	$b$	$N$
Neue Region	$\alpha$	$\beta$	$n$
Total	$a + \alpha$	$b + \beta$	$N + n$

In der Tabelle ist  $n$  meist viel kleiner als  $N$ . Die Zahl  $n$  kann sogar null sein. Die A posteriori-Plausibilität zum Anteil  $A$  an Policen mit Schadensmeldungen für die neue Region ist

$$\text{pdf}_{\text{neue Reg.}}(A | \text{Daten, Inf.}) \propto \underbrace{A^\alpha \cdot (1 - A)^\beta}_{\text{Likelihood}} \cdot \underbrace{A^a \cdot (1 - A)^b}_{\text{Inf. alte Region}} = A^{\alpha+a} \cdot (1 - A)^{\beta+b}$$

Dies ist eine Beta-Verteilung mit Kennzahlen  $\alpha + a + 1$  und  $\beta + b + 1$ . Der plausibelste Wert  $A_0$  von  $A$  beträgt

$$A_0 = \frac{\alpha + a}{n + N} = \frac{\alpha}{n} \cdot z + \frac{a}{N} \cdot (1 - z) \quad \text{mit } z = \frac{n}{n + N}$$

Der plausibelste Wert  $A_0$  ist daher das Mittel aus dem mit  $z$  gewichteten plausibelsten, unsicheren Wert  $\alpha/n$  der neuen Region (den Daten) und dem mit  $1-z$  gewichteten plausibelsten, präziseren Wert  $a/N$  der alten Region (der Vorinformation oder dem Vorwissen). Die Zahl  $z$  nennt man die *Kredibilität* (engl. *Credibility*) der Vertragspolicen für die neue Region. Ist  $n$  klein (wenige oder keine Daten aus der neuen Region), so ist die Kredibilität klein. Je mehr Daten zur neuen Region vorhanden sind, umso grösser wird  $n$ . Ist  $n$  gross, so ist  $z \approx 1$ . Der plausibelste Wert  $A_0$  ist dann etwa  $\alpha/n$  und die Information aus der alten Region spielt kaum mehr eine Rolle.  $\square$

## 5.4 Versuchsplanung und Unabhängigkeit

Mit der A posteriori-Verteilung  $\text{pdf}(A | \text{Daten})$  des Anteils  $A$  an Objekten in einer Grundgesamtheit, die in Kategorie eins sind, kann man Wahrscheinlichkeitsintervalle zu  $A$  berechnen. Etwa in der Form: Mit einer Wahrscheinlichkeit von 0,95 ist

$$A = A_0 \pm \delta A = 0,42 \pm 0,05$$

Dabei ist hier  $\pm \delta A = \pm 0,05$  die Präzision oder der Fehler 2. Art des Resultats. Die Präzision hängt von der Anzahl Messungen  $n$  ab. Lässt sich aus der A posteriori-Verteilung von  $A$  bestimmen, wie gross  $n$  gewählt werden muss, um eine verlangte Präzision zu erreichen? Es stellt sich heraus, dass die Antwort nicht ganz einfach ist.

Sind die Messwerte unabhängig, so besagt die Formel mit dem Standardfehler, dass mit einer Wahrscheinlichkeit von etwa 0,95

$$A = A_0 \pm \delta A \approx A_0 \pm 1,96 \cdot \sqrt{\frac{A_0 \cdot (1 - A_0)}{n}}$$

Dabei ist  $A_0$  der plausibelste Wert von  $A$ . Die verlangte Präzision von  $A$  hängt damit nicht allein von  $n$  ab. Wesentlich ist auch zu wissen, wo etwa  $A$  (und damit  $A_0$ ) ist. Stellt man die obige Formel nach der Stichprobengröße  $n$  um, erhält man

$$n \approx \left( \frac{1,96}{\delta A} \right)^2 \cdot A_0 \cdot (1 - A_0)$$

Der Stichprobenumfang muss umso grösser gewählt werden, je kleiner  $\delta A$  und umso präziser das Resultat für  $A$  gewünscht wird. Der Faktor  $A_0 \cdot (1 - A_0)$  ist klein für  $A_0$  in der Nähe von null oder eins. Er ist am grössten, wenn  $A_0$  von  $A$  bei 0,5 liegt. Der Stichprobenumfang, muss also umso höher gewählt werden, je näher der Anteil  $A$  tendenziell bei 0,5 liegt.

*Eine kritische Bemerkung:* In der obigen Formel, um einen Anteil  $A$  zu berechnen, wird angenommen, dass die Messwerte unabhängig sind. In der Praxis ist dies kaum überprüfbar. So führt ein kleiner Wirkungsraum im Experiment meist zu Abhängigkeiten der Messwerte. Daher sollte man gegenüber der berechneten Präzision von  $A$  skeptisch sein.

Wie die Regel von Bayes sagt, hängt die A posteriori-Verteilung von  $A$  von der Likelihood ab:

$$\mathbb{P}(A \mid \text{Daten}, \mathcal{I}) \propto \underbrace{\mathbb{P}(\text{Daten} \mid A)}_{\text{Likelihood}} \cdot \mathbb{P}(A \mid \mathcal{I})$$

Die Likelihood-Funktion ist einfach zu berechnen, wenn die Messwerte unabhängig sind. Das folgende Beispiel illustriert, wie die Likelihood schwieriger zu bestimmen ist, wenn die Messwerte nicht unabhängig sind:

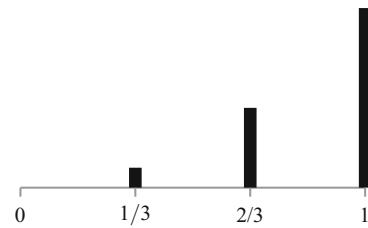
**Beispiel 5.6 (Eine Grundgesamtheit mit drei Objekten)** Eine Grundgesamtheit bestehe aus der Objekten, zwei davon in Klasse 1 und eines in Klasse 0. Der Anteil  $A$  der Objekte in Klasse 1 ist also  $2/3 = 0,667$ .

Eine Person habe nur als Information  $\mathcal{I}$ , dass die Grundgesamtheit drei Objekte in den Klassen 1 und 0 hat. Der Anteil  $A$  kann daher 0, 1/3, 2/3 oder 1 sein. Der Prior für den Anteil  $A$  ist für die Person somit:

$$\mathbb{P}(A = 0 \mid \mathcal{I}) = \mathbb{P}\left(A = \frac{1}{3} \mid \mathcal{I}\right) = \mathbb{P}\left(A = \frac{2}{3} \mid \mathcal{I}\right) = \mathbb{P}(A = 1 \mid \mathcal{I}) = 0,25$$

Eine Stichprobe soll helfen,  $A$  genauer zu bestimmen. Die Person zieht dazu zufällig zwei Objekte aus der Grundgesamtheit. Ihr Resultat: zweimal Klasse 1. Sie hat damit für die A

**Abb. 5.11** Posterior zum Anteil  $A$  bei den Daten, gezogen mit Zurücklegen



posteriori-Plausibilität des Anteils  $A$ :

$$\mathbb{P}(A \mid \text{Daten}, \mathcal{I}) \propto \underbrace{\mathbb{P}(\text{1 und 1} \mid A)}_{\text{Likelihood}} \cdot \underbrace{\mathbb{P}(A \mid \mathcal{I})}_{\text{A priori-Plausibilität}}$$

Sind die Messwerte durch Ziehen *mit Zurücklegen* ermittelt, beeinflussen sie sich gegenseitig nicht. Sie sind unabhängig. Daher ist die Likelihood

$$\mathbb{P}(\text{1 und 1} \mid A) = \mathbb{P}(\text{erste 1} \mid A) \cdot \mathbb{P}(\text{zweite 1} \mid A) = A \cdot A = A^2$$

Abb. 5.11 zeigt die entstandene A posteriori-Plausibilität. So ist die Wahrscheinlichkeit, dass  $A = 0$  nun null. Der plausibelste Wert von  $A$  ist eins. Man hat

$$\mathbb{P}(A = 1 \mid \text{Daten Version 1}, \mathcal{I}) = 0,643$$

Was passiert aber, wenn die Person die Messwerte durch Ziehen *ohne Zurücklegen* bestimmt hat? Die erste Messung ist eine Eins. In der Grundgesamtheit verbleiben damit noch zwei Objekte: eine Eins und eine Null. Die Wahrscheinlichkeit  $\mathbb{P}(\text{zweite 1} \mid A)$  ist damit nicht mehr  $A = 2/3$ , sondern sie ist tiefer: 0,5! Die Messwerte sind damit nicht mehr unabhängig. Mit der Multiplikationsregel der Wahrscheinlichkeitsrechnung erhält man

$$\begin{aligned} \mathbb{P}(\text{1 und 1} \mid A) &= \mathbb{P}(\text{zweite 1} \mid A \text{ und erste 1}) \cdot \mathbb{P}(\text{erste 1} \mid A) \\ &= \frac{\text{Restanzahl Objekte in Klasse 1}}{3 - 1} \cdot A = \frac{3 - A}{2} \cdot A \end{aligned}$$

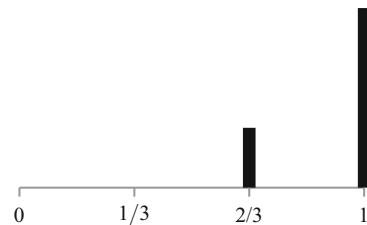
In Abb. 5.12 ist die berechnete A posteriori-Plausibilität zu  $A$  dargestellt. So ist nun auch die Wahrscheinlichkeit, dass  $A = 1/3$  ist, gleich null. Der plausibelste Wert von  $A$  ist wiederum eins. Man hat

$$\mathbb{P}(A = 1 \mid \text{Daten Version 2}, \mathcal{I}) = 0,750$$

Die beiden A posteriori-Verteilungen zu  $A$  sind also verschieden. □

Die Rechnung im obigen Beispiel lässt sich allgemein durchführen. Ist der Stichprobenumfang  $n$  gross, so wird wiederum der Likelihood dominant und der gesuchte Anteil

**Abb. 5.12** Posterior zum Anteil  $A$  bei den Daten, gezogen ohne Zurücklegen



$A$  von Objekten in Klasse 1 lässt sich mit einer einfachen Formel bestimmen. Wenn die Datenwerte durch zufälliges Ziehen *ohne Zurücklegen* entstanden sind, ist

$$A = A_0 \pm 1,96 \cdot \sqrt{\frac{A_0 \cdot (1 - A_0)}{n}} \cdot \sqrt{1 - \frac{n-1}{N-1}}$$

ein Wahrscheinlichkeitsintervall für  $A$  zum Niveau von *etwa* 0,95. Dabei ist  $A_0$  der aus der Stichprobe beobachtete Anteil und  $N$  ist die Anzahl Elemente in der Grundgesamtheit.

### Reflexion

**5.1**  $A$  sei eine Aussage mit  $\mathbb{P}(A) = 0,7$ . Berechnen Sie mit der Regel von Bayes – oder mit Tabellen wie in Kap. 4 – die zu  $\mathbb{P}(A | B)$  inverse Wahrscheinlichkeit, wenn  $\mathbb{P}(A | B) = 0,8$  und  $\mathbb{P}(B) = 0,4$  sind. Führen Sie die Rechnung noch einmal durch, wenn  $\mathbb{P}(A | B) = 0,8$  und  $\mathbb{P}(B) = 0,8$  sind.

**5.2** In einem Tiergehege mit vielen kleinen Tieren soll der Anteil der weiblichen Tiere berechnet werden. Dazu wurden zwölf Tiere nacheinander gefangen und ihr Geschlecht festgestellt. Gezogen wurden die Tiere durch Ziehen mit Zurücklegen, um zu garantieren, dass die Messwerte unabhängig sind. Hier das Resultat:

Gefangene Tiere: 12, weiblich: 10, männlich: 2

Es ist keine weitere Information zum Anteil an weiblichen Tieren im Gehege vorhanden.

- Wie lautet die Plausibilität  $\text{pdf}(A | \text{Daten})$  zum Anteil  $A$  der weiblichen Tiere im Gehege? Bestimmen Sie den plausibelsten Wert für  $A$ .
- Bestimmen Sie Wahrscheinlichkeitsintervalle zum Niveau 0,5 und 0,95 für  $A$  aus der A posteriori-Verteilung von  $A$ . Wie lautet die Fünf-Zahlen-Zusammenfassung zu  $A$ ? Führen Sie die Rechnungen auf zwei Arten aus: einmal mit Integralen, einmal mit einer MCMC-Simulation.
- Eine Person geht davon aus, dass der Anteil  $A$  der weiblichen Tiere eher hoch ist. Sie nimmt dazu als a priori-Verteilung für  $A$  die Dichtefunktion  $\text{pdf}(A) \propto A^3$ . Zeichnen

Sie den Graphen dieser Verteilung. Wie lautet die A posteriori-Verteilung von  $A$ ? Zeichnen Sie ihren Graphen.

- (d) Bestimmen Sie Wahrscheinlichkeitsintervalle zum Niveau 0,5 und 0,95 für  $A$  aus der A posteriori-Verteilung von  $A$  von Aufgabe (c). Haben sich die Intervalle im Vergleich zur Aufgabe (b) geändert?
- (e) Führen Sie die Rechnung von (a) noch einmal durch, diesmal mit einer Stichprobe von 36 Tieren, bei der 30 weiblich waren. Vergleichen Sie die beiden A posteriori-Verteilungen.

**5.3** Die Leitung einer Firma, die 3000 Personen angestellt hat, möchte erfahren, ob die Angestellten mit ihrer Arbeit zufrieden sind. Dazu werden mit zufälligem Ziehen zwanzig Angestellte ausgewählt und befragt. Die Antworten treten in den Kategorien „zufrieden“ und „nicht zufrieden“ auf. Hier das Resultat:

Befragte Personen: 20, Antwort „zufrieden“: 18

Gehen Sie im Folgenden davon aus, dass die Antworten unabhängig sind. Weitere Informationen zur Arbeitszufriedenheit haben Sie nicht.

- (a) Berechnen Sie den Anteil  $A$  der Personen in der Firma, die mit ihrer Arbeit zufrieden sind. Geben Sie das Folgende als Resultat an: die A posteriori-Verteilung, den plausibelsten Wert und Wahrscheinlichkeitsintervalle zum Niveau 0,5 und 0,95. Bestimmen Sie mit der „Standardfehler-Formel“ auch ein Wahrscheinlichkeitsintervall für  $A$  zum Niveau von etwa 0,95.
- (b) Geben Sie die Fünf-Zahlen-Zusammenfassung des Resultats aus der Aufgabe (a) an.

**5.4** Mit einer Umfrage möchte eine Firma wissen, wie gross der Anteil  $A$  der erwachsenen Personen in der Schweiz ihr Produkt „Supergut!“ kennen. Dazu werden 1000 Personen befragt. Als Antworten waren erlaubt: „Kenne ich!“, oder „Kenne ich nicht!“. 604 Personen geben an, das Produkt zu kennen. Als (un)plausible? Annahme gilt: die Antworten besitzen keine systematischen Fehler und sind unabhängig.

- (a) Berechnen Sie den Anteil  $A$  der erwachsenen Personen in der Schweiz, die das Produkt der Firma kennen. Geben Sie als Resultat an: die A posteriori-Verteilung, den plausibelsten Wert und Wahrscheinlichkeitsintervalle zum Niveau 0,5 und 0,95. Bestimmen Sie mit der „Standardfehler-Formel“ auch ein Wahrscheinlichkeitsintervall für  $A$  zum Niveau von ungefähr 0,95.
- (b) Geben Sie die Fünf-Zahlen-Zusammenfassung des Resultats aus der Aufgabe (a) an.

**5.5** In Publikationen werden Ergebnisse zur Wirkung von Medikamenten in verschiedensten Formen dargestellt. In einem Kurs mussten 119 Ärzte, die in der Schweiz praktizieren, vier solcher Ergebnisse lesen (siehe [2]). Dabei interpretierten 13 Ärzte die Ergebnisse richtig und die anderen 106 Ärzte falsch.

- (a) Berechnen Sie aus der gegebenen Information den Anteil  $A$  aller in der Schweiz praktizierenden Ärzte, die die vier Ergebnisse richtig interpretieren würden. Geben Sie als Resultat an: die A posteriori-Verteilung von  $A$ , den plausibelsten Wert und Wahrscheinlichkeitsintervalle zum Niveau 0,5 und 0,95. Bestimmen Sie mit der „Standardfehler-Formel“ auch ein Wahrscheinlichkeitsintervall für  $A$  zum Niveau von ungefähr 0,95.
- (b) Geben Sie die Fünf-Zahlen-Zusammenfassung des Resultats aus der Aufgabe (a) an.

**5.6** Eine Produktionsserie eines teuren Geräts umfasst nur 20 Stück. Eine Person besitzt acht dieser Geräte. Sie stellt fest, dass drei davon Mängel haben. Berechnen Sie ein Wahrscheinlichkeitsintervall zum Niveau von etwa 0,95 für den Anteil  $A$  aller Geräte der Produktionsserie, die Mängel haben. (*Tipp:* Gehen Sie davon aus, dass die acht Geräte durch Ziehen ohne Zurücklegen aus der Produktionsserie gewählt wurden. Benutzen Sie die Formel am Schluss des Kapitels.)

---

## Literatur

1. T. Bayes, R. Price: An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes. Phil. Trans., Roy. Soc. B, **53**, 370–418 (1763)
2. Ch. Damur, J. Steurer, Beurteilen Ärzte Therapieergebnisse anders als Studenten? Schweiz. Med. Wochenschrift **130**, 171–176 (2000)
3. H. Jeffreys, *The Theory of Probability*, (Oxford University Press, 1961)
4. P. S. de Laplace, *Théorie analytique des probabilités* (Courcier Imprimeur, Paris, 1812)
5. S. B. McGrawne, *Die Theorie, die nicht sterben wollte (Wie der englische Pastor Thomas Bayes eine Regel entdeckte, die nach 150 Jahren voller Kontroversen heute aus Wissenschaft, Technik und Gesellschaft nicht mehr wegzudenken ist)* (Springer Spektrum, Springer Verlag Berlin Heidelberg, 2014)
6. A. W. Whitney, The theory of experience rating, Proceedings if the Casualty Actuarial Society **4**, 274–292 (1918)

„Ahem!“ sagte die Maus mit gewaltiger Stimme. „Seid ihr alle bereit? Es folgt nun das Allertrockenste, was mir bekannt ist. Darf ich um allgemeine Ruhe bitten! ,Frühzeitig schon hatte Napoleon sich um die süddeutschen Fürsten bemüht. Dagegen waren die Unterhandlungen mit Sachsen, Braunschweig und Sachsen-Wei— —“  
„Hm!“ sagte der Weih.  
„Brr!“ sagte der Marabu und fröstelte.

Lewis Carroll, Alice im Wunderland (Insel Taschenbuch, 1973, S. 28f.)

## Zusammenfassung

In Untersuchungen müssen oft mehrere Größen berechnet werden. So interessieren bei der medizinischen Studie in Beispiel 2.15 die Heilraten von Aspirin bei Herzinfarkten und Schlaganfällen. In Beispiel 2.1 zur Biodiversität interessieren die durchschnittlichen Artenzahlen von verschiedenen Pflanzenarten. Um die Plausibilität zu solchen Größen zu beschreiben, braucht man ein gemeinsames Wahrscheinlichkeitsmodell. Wie damit gerechnet wird, wird in diesem Kapitel erklärt. Dabei wird das Gesetz der Marginalisierung vorgestellt. Es wird später verwendet, um zukünftige Werte von unsicheren Größen zu prognostizieren.

Oft wird optimistisch angenommen, dass die Messwerte oder Beobachtungen unabhängig sind. Dies hat zur Folge, dass nicht direkt messbare Größen mit zunehmender Anzahl Messungen  $n$  immer präziser berechnet werden können. Solchen Rechnungen sollte man mit der nötigen Skepsis begegnen. In diesem Kapitel wird ein Werkzeug vorgestellt, das zu beurteilen hilft, ob in einer zeitabhängige Folge von Messwerten oder Beobachtungen Trends oder Abhängigkeiten vorhanden sind.

## 6.1 Das gemeinsame Modell

Wie beschreibt man die Plausibilität von mehreren Größen? Zuerst muss man entscheiden, ob ein diskretes oder stetiges Modell benutzen will. Im ersten Fall hat man eine Massenfunktion, im zweiten Fall eine Dichtefunktion. Können  $G$  und  $H$  nur diskrete Werte annehmen, so beschreibt die Massenfunktion  $\mathbb{P}(G = g \text{ und } H = h)$  die Wahrscheinlichkeit, dass  $G$  den Wert  $g$  und  $H$  den Wert  $h$  annimmt. Man nennt dies die *gemeinsame* Massenfunktion (engl. *joint mass distribution function*) von  $G$  und  $H$ . Das Multiplikationsgesetz sagt

$$\mathbb{P}(G = g \text{ und } H = h) = \mathbb{P}(G = g \mid H = h) \cdot \mathbb{P}(H = h)$$

Hängt der erste Faktor für alle möglichen Werte  $g$  nicht von  $h$  ab, so sagt man, dass  $G$  und  $H$  unabhängig sind. Man hat dann

$$\mathbb{P}(G = g \text{ und } H = h) = \mathbb{P}(G = g) \cdot \mathbb{P}(H = h)$$

Dies bedeutet: Sind  $G$  und  $H$  unabhängig, so ist die gemeinsame Massenfunktion also gleich dem *Produkt* der einzelnen Massenfunktionen.

**Beispiel 6.1 (Ein Zahlenbeispiel)** Eine Person wisse aus ihrer Information  $\mathcal{I}$ , dass die Größe  $X$  nur Werte 1, 4, 5 und 6 und die Größe  $Y$  nur die Werte 5, 6 und 7 annehmen kann. Mit der Massenfunktion, gegeben durch Tab. 6.1, kann sie angeben, wie plausibel diese Werte aus ihrer vorhandenen Information sind.

Die Wahrscheinlichkeit ist 0,01, dass  $X = 1$  und  $Y = 6$  sind. Weiter liest man aus Tab. 6.1:  $\mathbb{P}(X = 4 \text{ und } Y = 7) = 0,13$ , sowie  $\mathbb{P}(X = 6 \text{ und } Y = 6) = 0,16$ .

Mit dem Gesetz der totalen Wahrscheinlichkeit lässt sich bestimmen, wie plausibel  $X = 4$  ist. Die Aussagen  $Y = 5$ ,  $Y = 6$  und  $Y = 7$  schliessen sich gegenseitig aus und ihre Wahrscheinlichkeiten summieren sich zu eins. Damit ist

$$\mathbb{P}(X = 4) = \mathbb{P}(X = 4 \text{ und } Y = 5) + \mathbb{P}(X = 4 \text{ und } Y = 6) + \mathbb{P}(X = 4 \text{ und } Y = 7)$$

Man erhält  $\mathbb{P}(X = 4) = 0,10 + 0,08 + 0,13 = 0,31$ . Die Massenfunktion von  $X$  kann man also berechnen, indem man die Spalten in Tab. 6.1 aufsummiert:<sup>1</sup>

$$\mathbb{P}(X = 1) = 0,18 \quad \mathbb{P}(X = 5) = 0,29 \quad \mathbb{P}(X = 6) = 0,22$$

Analog erhält man die Massenfunktion von  $Y$ , wenn man die Zeilen in Tab. 6.1 aufsummiert. Aus der gemeinsamen Verteilung kann man damit die einzelnen Verteilungen, die sogenannten *Rand-* oder *Marginalverteilungen* (engl. *marginal distribution*), von  $X$  und

---

<sup>1</sup> Anders ausgedrückt: Die Massenfunktion von  $X$  berechnet man, indem man die gemeinsame Massenfunktion über die Werte der zweiten Variable  $Y$  aufsummiert.

**Tab. 6.1** Gemeinsames Wahrscheinlichkeitsmodell für  $X$  und  $Y$

	$X = 1$	$X = 4$	$X = 5$	$X = 6$
$Y = 5$	0,05	0,10	0,14	0,03
$Y = 6$	0,01	0,08	0,02	0,16
$Y = 7$	0,12	0,13	0,13	0,03

**Tab. 6.2** Gemeinsames Modell für  $X$  und  $Y$  mit den beiden Randverteilungen

	$X = 1$	$X = 4$	$X = 5$	$X = 6$	total
$Y = 5$	0,05	0,10	0,14	0,03	0,32
$Y = 6$	0,01	0,08	0,02	0,16	0,27
$Y = 7$	0,12	0,13	0,13	0,03	0,41
total	0,18	0,31	0,29	0,22	

$Y$  berechnen. Das Resultat ist in Tab. 6.2 dargestellt. Die beiden Grössen sind bei der gegebenen Information  $I$  nicht unabhängig modelliert. So ist  $\mathbb{P}(X = 4 \text{ und } Y = 6) = 0,08$ . Aber man hat  $\mathbb{P}(X = 4) \cdot \mathbb{P}(Y = 6) = 0,31 \cdot 0,27 = 0,0837$ .

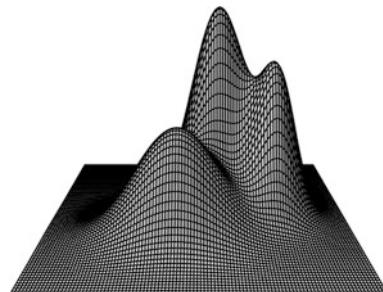
Der wahrscheinlichste Wert der gemeinsamen Verteilung von  $X$  und  $Y$  ist  $X = 6$  und  $Y = 6$ . Er kann für eine Person, die sich für beide Grössen interessiert, wichtig sein. Ist für eine Person nur die Grösse  $X$  bedeutend, so wird sie den wahrscheinlichsten Wert von  $X$  wissen wollen. Er beträgt  $X = 4$ .  $\square$

Ein Wahrscheinlichkeitsmodell für  $X$  und  $Y$  mit kontinuierlichen Werten ist durch eine Dichtefunktion  $\text{pdf}(x, y)$  gegeben. Der Graph der Dichtefunktion ist eine Fläche im Raum, wie in Abb. 6.1 gezeigt. Er wird meist nicht dargestellt, da er sehr komplex ist. Wahrscheinlichkeiten zu  $X$  und  $Y$  sind gleich dem Volumen unter dem Graphen der Dichtefunktion. Man kann die Dichtefunktion statt durch ihren Graphen auch durch ihre Höhenangabe, wie bei Karten, darstellen. Abb. 6.2 zeigt die Höhenlinien der obigen Dichtefunktion.

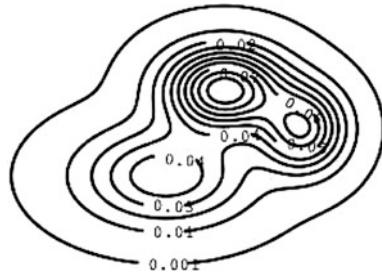
Mit dem Gesetz der totalen Wahrscheinlichkeit kann man die Dichtefunktionen  $\text{pdf}_X(x)$  und  $\text{pdf}_Y(y)$  der Randverteilungen von  $X$  und  $Y$  berechnen. Man hat

$$\mathbb{P}(x \leq X \leq x + \Delta x) = \sum_y \mathbb{P}(x \leq X \leq x + \Delta x \text{ und } y \leq Y \leq y + \Delta y)$$

**Abb. 6.1** Graph einer Dichtefunktion, die die Plausibilität zu zwei Grössen beschreibt



**Abb. 6.2** Niveaulinien der Dichtefunktion aus Abb. 6.1



Die linke Seite der Gleichung ist gleich  $\text{pdf}_X(x) \cdot \Delta x$ . Die rechte Seite ist  $\text{pdf}_{XY}(x, y) \cdot \Delta x \cdot \Delta y$ . Dividiert man die Gleichung durch  $\Delta x$ , ergibt sich

$$\text{pdf}_X(x) = \sum_y \text{pdf}_{XY}(x, y) \cdot \Delta y$$

Wählt man  $\Delta y$  infinitesimal klein, erhält man ein Integral:

**Theorem 6.1 (Gesetz der Marginalisierung)**

*Wird die Information  $\mathcal{I}$  zu zwei Größen  $X$  und  $Y$  mit der gemeinsamen Dichtefunktion  $\text{pdf}(x, y \mid \mathcal{I})$  beschrieben, so lautet die Dichtefunktion  $\text{pdf}(x \mid \mathcal{I})$  der Randverteilung der Größe  $X$ :*

$$\text{pdf}(x \mid \mathcal{I}) = \int \text{pdf}(x, y \mid \mathcal{I}) \, dy$$

Beliebt ist es, dies in Kurzform so zu schreiben:

$$\mathbb{P}(X \mid \mathcal{I}) = \int \mathbb{P}(X \text{ und } Y \mid \mathcal{I}) \cdot dY$$

Man kann dieses Gesetz auch so lesen: Hat man die gemeinsame Plausibilität zu zwei Größen  $X$  und  $Y$ , so kann man daraus die Plausibilität zu  $X$  isolieren: *Man summiert über alle möglichen Werte von  $Y$  auf.* Man sagt auch, dass man die Größe  $Y$  ausintegriert.

## 6.2 Randverteilungen mit Monte-Carlo-Simulationen rechnen

Um die Randverteilungen von gemeinsamen stetigen Wahrscheinlichkeitsmodellen zu rechnen, müssen wie oben angegeben, Integrale bestimmt werden. Meist können solche Integrale nicht explizit ausgerechnet werden. Üblich ist es daher, mit Monte-Carlo-Simulationen, meist einer MCMC-Simulation, zu arbeiten. Sind  $X$  und  $Y$  zwei Größen mit einer gemeinsamen Dichtefunktion  $\text{pdf}(x, y \mid \mathcal{I})$ , so kann man nach dem gleichen Re-

**Abb. 6.3** MCMC-Simulation:  
Kette von 150 Punkten mit der  
Dichtefunktion aus Abb. 6.1



zept wie bei eindimensionalen Modellen eine Kette von Punkten  $(x_1, y_1), (x_2, y_2), \dots, (x_L, y_L)$  bestimmen, die dem gemeinsamen Modell von  $X$  und  $Y$  folgen. Abb. 6.3 zeigt die Dichtefunktion von Abb. 6.2 mit einer Kette von 150 Punkten. Dabei ist es wichtig, den Sprungprozess der MCMC-Simulation gut an die Form der Verteilung anzupassen, um eine ausgewogene Akzeptanzrate zu erhalten.

Wahrscheinlichkeiten für die Randverteilung von  $X$  berechnen sich aus der Kette, indem man die  $x$ -Koordinaten der Kette betrachtet. Die  $x$ -Koordinaten der Kette stellen nämlich die Randverteilung von  $X$  dar. Damit ist

$$\mathbb{P}(a \leq X \leq b) \approx \frac{\text{Anzahl Punkte mit } x\text{-Koordinate zwischen } a \text{ und } b}{\text{Anzahl simulierte Punkte}}$$

Auch Wahrscheinlichkeitsintervalle für  $X$  sind berechenbar. Wählt man den 5 % kleinsten und den 5 % grössten  $x$ -Wert aus den simulierten Punkten, hat man ein Wahrscheinlichkeitsintervall für  $X$  zum Niveau 0,9.

### 6.3 Verbundene Größen und Korrelation

Gemeinsame Wahrscheinlichkeitsmodelle sind besonders einfach, wenn die damit beschriebenen Objekte unabhängig beschrieben werden. Man erhält das gemeinsame Wahrscheinlichkeitsmodell, indem man die Wahrscheinlichkeitsmodelle für die einzelnen Größen wählt und diese miteinander multipliziert. Mit Beobachtungen oder Messwerten kann man versuchen zu beurteilen, ob Größen unabhängig modelliert werden können. Hier ein Beispiel dazu:

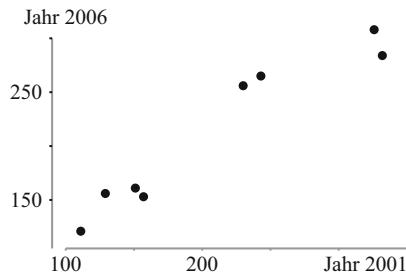
**Beispiel 6.2 (Biodiversität)** Im Rahmen der Untersuchung der Biodiversität, das in Beispiel 2.1 vorgestellt ist, will man die durchschnittliche Artenzahlen pro Parzelle  $\mu_{2001}$  und  $\mu_{2006}$  in den Jahren 2001 und 2006 in der Schweiz berechnen. Tab. 6.3 zeigt die Messresultate der Artenzahlen in den beiden Jahren aus acht Parzellen. Aus den Daten kann man Plausibilitäten zu den beiden nicht direkt messbaren Größen  $\mu_{2001}$  und  $\mu_{2006}$  berechnen:

$$\mathbb{P}(\mu_{2001} | \text{Daten}) \quad \text{und} \quad \mathbb{P}(\mu_{2006} | \text{Daten})$$

**Tab. 6.3** Artenzahlen in acht Parzellen aus einer Untersuchung zur Biodiversität in der Schweiz in den Jahren 2001 und 2006 (Biodiversitätsmonitoring Schweiz BDM, Hintermann & Weber AG, Reinach (Basel), Herbst 2007)

	Parzelle 1	Parzelle 2	Parzelle 3	Parzelle 4	Parzelle 5	Parzelle 6	Parzelle 7	Parzelle 8
2001	326	111	332	129	151	230	243	157
2006	308	121	284	156	161	256	265	153

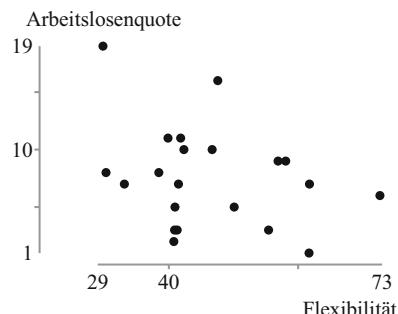
**Abb. 6.4** Streudiagramm der Artenzahlen aus Tab. 6.3



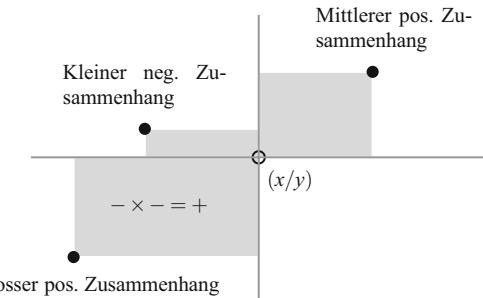
Aussagen zu  $\mu_{2001}$  und  $\mu_{2006}$  dürften jedoch kaum unabhängig sein. Dies zeigt sich darin, dass Parzellen mit hohen Artenzahlen im Jahr 2001 tendenziell auch hohe Artenzahlen im Jahr 2006 haben. Die Messwerte sind *positiv verbunden*. In Abb. 6.4 ist dies in einem Streudiagramm visualisiert. Das Diagramm entsteht, indem jeweils die beiden betreffenden, beobachteten Artenzahlen (z. B. die Zahlen 326 und 308 für die Parzelle 1) als Punkt mit Koordinaten einträgt. Will man aus den Daten Aussagen zu beiden Parametern formulieren, braucht man daher die gemeinsame Verteilung von  $\mu_{2001}$  und  $\mu_{2006}$ . □

**Beispiel 6.3 (Arbeitslosigkeit und Flexibilität)** Abb. 6.5 zeigt in einem Streudiagramm die Merkmale Flexibilität des Arbeitsmarkts und Arbeitslosigkeit von 21 Industrieländern in den Jahren 1984–1990. Die beiden Größen könnten *negativ verbunden* sein: Eine tendenziell höhere Flexibilität scheint mit einer tendenziell tieferen Arbeitslosigkeit verbunden. Man kann versuchen, mit einem Parameter zu messen, wie stark diese Verbundenheit ist. Dazu braucht man ein gemeinsames Wahrscheinlichkeitsmodell für die Flexibilität und die Arbeitslosenrate. □

**Abb. 6.5** Streudiagramm der Arbeitslosenquote (in %) und der Flexibilität (Di Tella und MacCulloch, aus The Economist und aus NZZ, 2000)



**Abb. 6.6** Berechnung des empirischen Korrelationskoeffizienten



Mit dem *empirischen Korrelationskoeffizienten nach Pearson* (engl. *Pearson correlation*) kann man messen, wie stark Messwerte oder Beobachtungen bivariater Merkmale  $(x_1/y_1), (x_2/y_2), \dots, (x_n/y_n)$  einer Grösse miteinander verbunden sind. Sind  $\bar{x}$  und  $\bar{y}$  die arithmetischen Mittel der  $x$ - und  $y$ -Werte, so ist der Schwerpunkt der Messpunkte, wenn man sie in einem Streudiagramm darstellt, beim Punkt  $(\bar{x}/\bar{y})$ . Das Produkt

$$(x_i - \bar{x}) \cdot (y_i - \bar{y})$$

besitzt das Vorzeichen +, wenn der Messpunkt  $(x_i/y_i)$  oben rechts oder unten links des Schwerpunkts  $(\bar{x}/\bar{y})$  liegt. Man spricht von einem positiven Zusammenhang. Es ist um so grösser, je weiter der Messpunkt vom Schwerpunkt  $(\bar{x}/\bar{y})$  entfernt ist. Dies illustriert Abb. 6.6. Man mittelt dieses Produkt über alle Messwerte und dividiert durch die empirische Standardabweichung  $s_x$  und  $s_y$  der  $x$ - und  $y$ -Werte, um eine dimensionslose Zahl zu erhalten:

$$\rho_{\text{emp}} = \frac{\frac{1}{n-1} [(x_1 - \bar{x}) \cdot (y_1 - \bar{y}) + \dots + (x_n - \bar{x}) \cdot (y_n - \bar{y})]}{s_x \cdot s_y}$$

Diese Zahl – der empirische Korrelationskoeffizient nach Pearson – liegt immer zwischen –1 und 1. Eine beliebte *Faustregel* ist:

Ist  $\rho_{\text{emp}} > \frac{2}{\sqrt{n}}$  oder  $\rho_{\text{emp}} < -\frac{2}{\sqrt{n}}$ , so sollten die beiden Merkmale einer bivariaten Grösse nicht unabhängig voneinander modelliert werden.

**Beispiel 6.4 (Arbeitslosigkeit und Flexibilität)** Mit den Daten der oben erwähnten Studie kann man versuchen, das Wissen zur Flexibilität des Arbeitsmarkts und zur Arbeitslosigkeit in Industrieländern zu beschreiben. Können Plausibilitäten, gegeben die Daten, zu den beiden Grössen unabhängig voneinander formuliert werden? Die Daten sind in Tab. 6.4 aufgelistet. Die arithmetischen Mittel und die empirischen Standardabweichun-

**Tab. 6.4** Flexibilität und Arbeitslosenrate in % aus verschiedenen Ländern

Flexibilität	38,5	41,3	41,8	56,9	61,8	50,1	42,3	41,5	47,6	39,9
Arbeitslosenrate	8	3	11	9	7	5	10	7	16	11
Flexibilität	55,4	46,7	40,9	41,0	29,8	40,8	61,7	58,1	72,7	30,3
Arbeitslosenrate	3	10	3	5	19	2	1	9	6	8

gen der beiden Messgrößen sind

$$\overline{\text{Flexibilität}} = 46,72, \quad s_{\text{Flex.}} = 11,14, \\ \overline{\text{Arbeitslosenrate}} = 7,62, \quad s_{\text{Arbeits.}} = 4,44$$

Der empirische Pearson Korrelationskoeffizient lautet

$$\rho_{\text{emp}} = \frac{\frac{1}{20} [(38,5 - 46,72) \cdot (8 - 7,62) + \dots + (33,1 - 46,72) \cdot (7 - 7,62)]}{11,14 \cdot 4,44} \\ = -0,33$$

Die Daten sind also negativ verbunden. Der Wert  $-0,33$  ist aber nicht kleiner als  $-2/\sqrt{21} = -0,44$ . Man kann es daher wagen, Flexibilität und Arbeitslosigkeit, als unabhängig zu betrachten. Mit anderen Worten:

$$\mathbb{P}(\text{Arbeitslosigkeit} \mid \text{Flexibilität, Daten}) = \mathbb{P}(\text{Arbeitslosigkeit} \mid \text{Daten})$$

Anders ausgedrückt: Es dürfte schwierig sein, mit den vorliegenden Daten Aussagen zur Arbeitslosigkeit in Industrieländern auf Grund der Flexibilität des Arbeitsmarkts zu formulieren.  $\square$

**Beispiel 6.5 (Biodiversität)** Beim obigen Beispiel zu den acht Parzellen beträgt der empirische Korrelationskoeffizient 0,96. Die Daten sind stark positiv verbunden. Der Wert ist grösser als  $2/\sqrt{8} = 0,71$ . Die Plausibilität zu den beiden Artenzahlen in den Jahren 2001 und 2006 sollte daher mit einem gemeinsamen Wahrscheinlichkeitsmodell beschrieben werden.  $\square$

Zwei Zufallsgrößen  $X$  und  $Y$ , wie beim besprochenen Beispiel die Flexibilität und die Arbeitslosigkeit, können korreliert sein. Wenn Beobachtungen vorliegen, bedeutet dies aber noch nicht, dass die beiden Größen auch kausal zusammenhängen. So kann es sein, dass die Variable  $Y$ , scheinbar von  $X$  abhängt, in Realität das Resultat einer Kovariablen  $Z$  ist. In einem solchen Fall spricht man von einem *Konfundierungs-* oder *Vermengungseffekt* (engl. *confounded variables*).

## 6.4 Autokorrelation und Unabhängigkeit

Oft geht man bei Untersuchungen davon aus, dass eine *zeitabhängige* Folge von Messwerten unter statistischer Kontrolle ist und die Messwerte nicht voneinander abhängen. Damit können gesuchte, nicht direkt messbare Größen mit einer Präzision bestimmt werden, die proportional zu  $1/\sqrt{n}$  ist. Dabei ist  $n$  die Anzahl Messungen. Wie schon in Kap. 3 gesagt, zeigen Erfahrungen, dass Abhängigkeiten zwischen nacheinander gemessenen Werten kaum vermeidbar sind. Sie sind vor allem dann spürbar, wenn der Wirkungsraum des Experiments klein ist: beispielsweise, wenn die Messungen durch die gleiche Person oder das gleiche Labor ausgeführt werden. Als Konsequenz hat man dann: die Präzision verhält sich nicht mehr wie  $1/\sqrt{n}$ , sondern sie kann grösser (in der Praxis selten) oder viel kleiner sein. Es ist daher sinnvoll zu beurteilen, ob eine zeitliche Folge von Messwerten oder Beobachtungen Trends oder Abhängigkeiten hat. Ein Werkzeug dazu ist die *Autokorrelationsfunktion* (engl. *autocorrelationfunction acf*). Wie sie definiert ist, wird im folgenden Beispiel gezeigt:

**Beispiel 6.6 (Chloridgehalt)** Man kann den Chloridgehalt einer Kalium-Chlorid-Lösung indirekt messen, indem man die Chloridionen mit Silberionen ausfällt. Dieses Verfahren wurde siebenmal nacheinander an einer Lösung durchgeführt. Im Labor Chemie der Berner Fachhochschule erhielt man die Werte (in mol/m<sup>3</sup>):

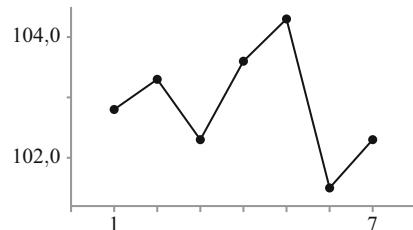
102,8 103,3 102,3 103,6 104,3 101,5 102,1

Abb. 6.7 veranschaulicht, dass die Messwerte unter statistischer Kontrolle sind. Es gibt keine Trends nach höheren oder tieferen Messwerten und aussergewöhnlich hohe oder tiefe Messwerte sind nicht vorhanden.

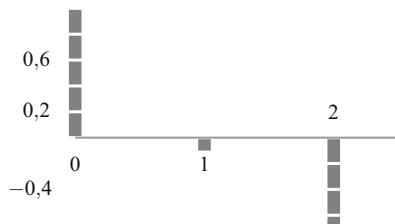
Wie kann man beurteilen, ob die Messwerte voneinander abhängen? Als Erstes kann man messen, wie die einzelnen Messwerte mit sich selber verbunden sind. Man verdoppelt dazu die Liste der Messwerte und erhält sieben Punktpaare mit gleichen Koordinaten:

X	102,8	103,3	102,3	103,6	104,3	101,5	102,1
Y	102,8	103,3	102,3	103,6	104,3	101,5	102,1

**Abb. 6.7** Streudiagramm der sechs Chloridgehalte



**Abb. 6.8** Verschiedene Korrelationen in der Messreihe der Chloridgehalte



Die Punkte liegen auf einer Geraden, die eine Steigung von  $45^\circ$  hat. Daher ist der empirische Korrelationskoeffizient  $\rho_{\text{emp}}$  nach Pearson gleich eins. Die Messwerte sind natürlich stark mit sich selber verbunden. Interessanter ist es zu messen, wie der erste Messwert mit dem zweiten, der zweite mit dem dritten usw. verbunden ist. Dazu muss man die erste Zeile bei der obigen Tabelle um eine Einheit nach rechts schieben:

X	-	102,8	103,3	102,3	103,6	104,3	101,5	102,1
Y	102,8	103,3	102,3	103,6	104,3	101,5	102,1	-

Der empirische Korrelationsfaktor der in der Tabelle verbleibenden sechs Punktpaare misst, wie stark diese miteinander verbunden sind. Man erhält  $\rho_{\text{emp}} = -0,10$ . Tendenziell folgt also auf einen grösseren Messwert ein kleinerer Messwert. Diese Zahl ist nicht kleiner als  $-2/\sqrt{6} = -0,82$ , d. h. benachbarte Messwerte können als unabhängig betrachtet werden.

Man kann als Nächstes messen, wie stark der erste Messwert mit dem dritten, der zweite mit dem vierten und so weiter verbunden ist:

X	-	-	102,8	103,3	102,3	103,6	104,3	101,5	101,5
Y	102,8	103,3	102,3	103,6	104,3	101,5	102,1	-	-

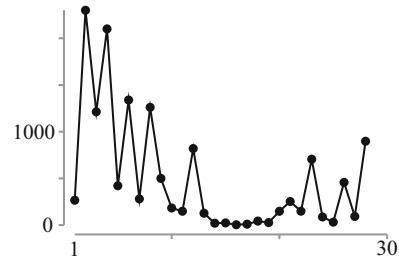
Aus den fünf Paaren erhält einen empirischen Korrelationskoeffizienten von  $-0,67$ . Abb. 6.8 visualisiert die berechneten Korrelationskoeffizienten.  $\square$

Im obigen Beispiel wird gezeigt, wie Abhängigkeiten in einer Datenreihe quantifiziert werden können. Der empirische Korrelationskoeffizient  $\rho_{\text{emp},r}$  zwischen den um  $r$  versetzten Datenwerten misst dies. Dazu müssen jedesmal neue arithmetische Mittel und empirische Standardabweichungen berechnet werden. Um die Rechnungen zu vereinfachen, wird der empirische Korrelationskoeffizient  $\rho_{\text{emp},r}$  mit einer einfacheren Formel approximiert:

$$\rho_{\text{emp},r} \approx \text{acf}(r) = \frac{\frac{1}{n-1} [(x_1 - \bar{x}) \cdot (x_{r+1} - \bar{x}) + \dots + (x_{n-r} - \bar{x}) \cdot (x_n - \bar{x})]}{s_x \cdot s_x}$$

Man nennt die Funktion, die von der Versetzung  $r$  (engl. *lag*) abhängt, die Autokorrelationsfunktion der Messreihe. Sie hat den Wert eins, wenn der Lag null ist:  $\text{acf}(0) = 1$ .

**Abb. 6.9** Streudiagramm der 28 Zeiten zwischen aufeinanderfolgenden, starken Erdbeben

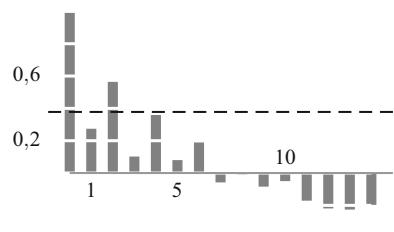


Statistikprogramme können die Werte dieser Funktion schnell ausrechnen. Zudem stellen sie meist ihren Graphen wie beim obigen Beispiel mit Stäbchen dar. Auch zeichnen sie automatisch zwei waagrechte Linien auf den Höhen  $2/\sqrt{n}$  und  $-2/\sqrt{n}$  ein. Als Regel dazu gilt (siehe [2]):

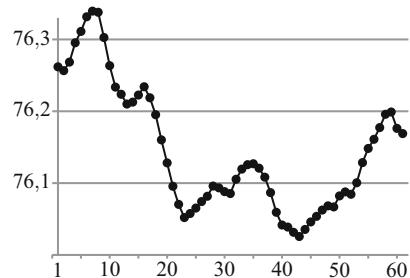
Sind Werte der Autokorrelationsfunktion in *systematischer Art* oder *gruppenweise* grösser  $2/\sqrt{n}$  oder kleiner als  $-2/\sqrt{n}$ , so sollten die Datenwerte nicht als trendfrei oder als unabhängig betrachtet werden.

**Beispiel 6.7 (Zeit zwischen starken Erdbeben)** Beim Beispiel 1.3 zu den  $n = 28$  Zeiten zwischen aufeinanderfolgenden, starken Erdbeben interessiert die durchschnittliche Zeit  $\mu$  zwischen aufeinanderfolgenden, zukünftigen, starken Erdbeben. Am einfachsten kann man  $\mu$  berechnen, wenn man annimmt, dass die Wartezeiten unabhängig sind. Abb. 6.9 zeigt die Wartezeiten in einem Streudiagramm. Abb. 6.10 visualisiert den Graphen der Autokorrelationsfunktion. Für den Lag  $r = 0$  ist der Wert der Autokorrelationsfunktion eins. Ein Wert ist ausserhalb des Bandes, das durch die Zahlen  $-2/\sqrt{28}$  und  $2/\sqrt{28}$  begrenzt wird. Es ist die Korrelation für den Lag  $r = 2$ . Dies liegt daran, dass die beobachteten Zeiten hin und her pendeln: auf eine lange Wartezeit folgt eine kurze, auf eine kurze ein lange. Es finden sich aber keine Werte in systematischer Weise ausserhalb der gestrichelten Linien. Daher ist es sinnvoll, die beobachteten Zeiten als trendfrei und unabhängig zu betrachten. Skeptisch sollte man trotzdem bleiben.  $\square$

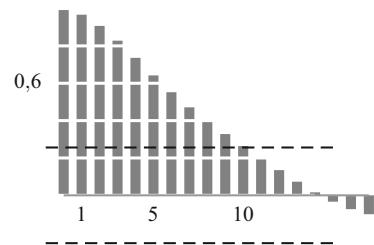
**Abb. 6.10** Graph der Autokorrelationsfunktion für die 28 Zeiten zwischen aufeinanderfolgenden, starken Erdbeben



**Abb. 6.11** Haftkraft (in N) auf einem Klebestreifen, gemessen an 61 Stellen



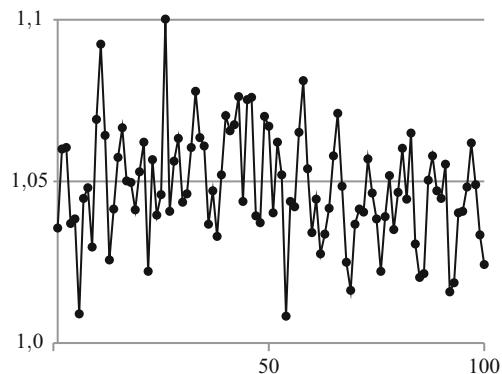
**Abb. 6.12** Graph der Autokorrelationsfunktion der Haftkräfte (mit Bändern  $\pm 2/\sqrt{61}$ )



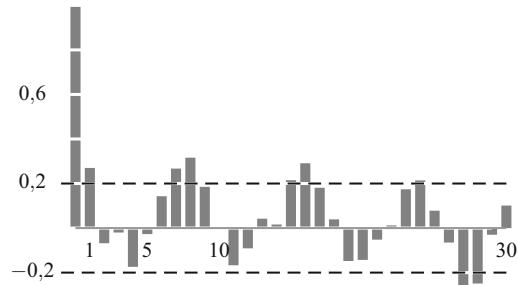
**Beispiel 6.8 (Klebestreifen)** Im Rahmen eines Projekts an der Berner Fachhochschule in Burgdorf wurde versucht, die Haftkraft von Klebestreifen zu optimieren. Gemessen wurde dabei die Haftkraft entlang von 61 aneinandergereihten Punkten entlang des Klebestreifens. Abb. 6.11 zeigt das Streudiagramm der Messwerte. Abb. 6.12 visualisiert die ersten siebzehn Werte der Autokorrelationsfunktion. Schon das Streudiagramm zeigt, dass nacheinander gemessene Werte stark verbunden sind. Der Graph der Autokorrelation verdeutlicht dies. In der Gruppe von Lag 1 bis Lag 10 sind die Messwerte positiv korreliert. Auf einen hohen (bzw. tiefen) Messwert folgt ein hoher (bzw. tiefer) Messwert. Die Messwerte sind also nicht unabhängig. □

**Beispiel 6.9 (Motordrehzahlen)** Im Rahmen eines Projekts an der Berner Fachhochschule aus dem Jahr 2009 wurde versucht, die Drehzahl  $D$  eines Motors zu berechnen. Weil die Experimentierumgebung nicht konstant gehalten werden konnte, streuen gemessene Drehzahlen um  $D$ . Daher muss die Drehzahl  $D$  mit einer Plausibilität, also einer Wahrscheinlichkeit, aus Daten bestimmt werden. Das Streudiagramm in Abb. 6.13 zeigt 100 Messwerte, die während vier Sekunden erhalten wurden. Die Messwerte scheinen mit Blick auf das Streudiagramm trendfrei und unabhängig zu sein. Die Autokorrelationsfunktion – siehe Abb. 6.14 – zeigt aber systematisch Werte ausserhalb der gestrichelten Linien: in den Bereichen Lag 6–10 und Lag 14–16 und so weiter. Die Daten haben ein zyklisches Verhalten. Dies ist wegen der Eigenschwingung des Motors, die als Kovariablen auf gemessene Drehzahlen wirkt. Will man die Drehzahl  $D$  berechnen, ist es daher empfehlenswert, den zyklischen Trend aus den Daten zu entfernen. □

**Abb. 6.13** Streudiagramm der gemessenen Motordrehzahlen (in 1000 Umdrehungen pro Minute)



**Abb. 6.14** Graph der Autokorrelationsfunktion für die Motordrehzahlen



**Beispiel 6.10 (Asparaginsäure)** W. S. Gosset hat im Jahr 1927 in [4] den Stickstoffgehalt  $S$  in einer Asparaginsäure untersucht. Gesucht war der Stickstoffgehalt  $N$  in der Säure. Dazu benutzte er 135 Messungen. Nimmt man an, dass die Messungen unabhängig sind, so hat man für die Präzision des gesuchten Stickstoffgehalts

$$\text{Erreichbare Präzision für } N \propto \frac{1}{\sqrt{n}} = \frac{1}{\sqrt{135}} = \frac{1}{135^{0.5}} = \frac{1}{11,62}$$

Untersuchungen von H. Graf, F. Hampel und J.-D. Tacier in [3] haben aber gezeigt, dass die Daten voneinander abhängen. Korrekterweise gilt hier mit einem Modell, das diese Abhängigkeiten beschreibt

$$\text{Erreichbare Präzision für } N \propto \frac{1}{n^{0,17}} = \frac{1}{135^{0,17}} = \frac{1}{2,30}$$

Damit ist

$$\frac{\text{falsche erreichbare Präzision}}{\text{tatsächlich erreichbare Präzision}} = \frac{11,62}{2,30} = 5,05$$

Rechnet man also damit, dass die Messwerte unabhängig sind, so erhält man ein Resultat, dass fünfmal zu optimistisch ist! Dies muss vermieden werden.  $\square$

Wie man Trends oder Abhängigkeiten zwischen Daten modellieren kann, findet man in einschlägigen Büchern zur Statistik. In ein solchen Modellen fliesst meist die acf-Funktion ein. Beim obigen Beispiel zu den Klebeetiketten etwa durch eine mit wachsendem Lag  $r$  geometrisch abnehmende Funktion  $\text{acf}(r) = \rho^r$ . Dabei ist  $\rho$  ein unbekannter Parameter. Die Diskussion solcher Modelle würde den Inhalt eines einführenden Statistikkurses sprengen. Deshalb wird hier darauf verzichtet.

## Reflexion

**6.1** An einer Kreuzung befinden sich zwei Ampeln, die die Farben rot, gelb und grün anzeigen. Eine Person beschreibt die Plausibilität, dass die Ampeln diese Farben anzeigen werden, durch die folgende Massenfunktion:

	rot 1	gelb 1	grün 1
rot 2	0,10	0,05	0,25
gelb 2	0,06	0,01	0,20
grün 2	0,30	0,03	0,00

- (a) Wie lauten die Wahrscheinlichkeiten, dass (1) beide Ampeln rot anzeigen, (2) die erste Ampel grün und die zweite Ampel rot anzeigt? Sind die beiden Ampeln durch die Person unabhängig modelliert?
- (b) Bestimmen Sie die Randverteilungen des Wahrscheinlichkeitsmodells für die Ampeln. Wie gross sind die Wahrscheinlichkeiten, dass (1) die Ampel eins grün anzeigt, (2) die Ampel zwei rot anzeigt?
- (c) Wie lautet der wahrscheinlichste Zustand der beiden Ampeln? Was ist der wahrscheinlichste Zustand der Ampel zwei?

**6.2** Eine Person beschreibt ihre Plausibilität – gegeben eine Information  $\mathcal{I}$  – zu zwei diskreten Grössen  $A$  (mit Werten 2, 5, 6, und 10) und  $B$  (mit Werten 0, 1 und 3) mit dem gemeinsamen Wahrscheinlichkeitsmodell:

	$A = 2$	$A = 5$	$A = 6$	$A = 10$
$B = 0$	0,09	0,10	0,06	0,08
$B = 1$	0,16	0,02	0,01	0,09
$B = 3$	0,13	0,07	0,06	0,13

- (a) Wie lautet der Modus der gemeinsamen Verteilung?
- (b) Bestimmen Sie die Randverteilung der Grösse  $B$ . Zeichnen Sie den Graphen der Massenfunktion dieser Randverteilung. Wie lautet der plausibelste Wert von  $B$ ?

- (c) Integrieren Sie die Grösse  $B$  aus, um die Randverteilung der Grösse  $A$  zu erhalten. Zeichnen Sie den Graphen der Massenfunktion der Randverteilung. Wie lautet der plausibelste Wert von  $A$ ?

**6.3** Gegeben sind zwei Größen  $X$  und  $Y$ , die je Werte zwischen null und eins annehmen können. Jemand formuliert sein Wissen zu diesen beiden Größen mit einem gemeinsamen stetigen Wahrscheinlichkeitsmodell:

$$\text{pdf}_{XY}(x, y \mid \mathcal{I}) = 0,54745 \cdot [2 - (x - 0,5)^2 - (y - 0,4)^2]$$

- (a) Benutzen Sie einen Computer, um die Dichtefunktion zu visualisieren. Lassen Sie den Computer entweder den Graphen oder die Höhenlinien der Dichtefunktion zeichnen.  
 (b) Wie lautet der Modus der gemeinsamen Verteilung?  
 (c) Die Randverteilung von  $X$  lautet

$$\text{pdf}_X(x \mid \mathcal{I}) = \int_0^1 \text{pdf}_{XY}(x, y \mid \mathcal{I}) dy$$

Bestimmen Sie dieses Integral mit einem Taschenrechner. Zeichnen Sie den Graphen der erhaltenen Dichtefunktion. Was ist der Modus der Randverteilung von  $X$ ?

- (d) Integrieren Sie auch  $X$  aus, um die Dichtefunktion  $Y$  zu erhalten.

**6.4** Eine Person formuliert ihr Wissen zu zwei Größen  $U$  und  $V$ , die beliebige Werte annehmen können mit einem stetigen Wahrscheinlichkeitsmodell. Die dazugehörige gemeinsame Dichtefunktion lautet

$$\text{pdf}(u, v \mid \mathcal{I}) \propto \exp\left(-\frac{1}{1,28} \left[ \frac{(u-1)^2}{4} + \frac{(v-4)^2}{9} - \frac{(u-1)(v-4)}{5} \right]\right)$$

- (a) Benutzen Sie einen Computer, um die Dichtefunktion zu visualisieren. Lassen Sie den Computer entweder den Graphen oder Höhenlinien zeichnen.  
 (b) Wie lautet der Modus der gemeinsamen Verteilung?  
 (c) Führen Sie mit dem Logarithmus von  $\text{pdf}(u, v \mid \mathcal{I})$  eine MCMC-Simulation durch, um die Wahrscheinlichkeit zu berechnen, dass  $U$  zwischen  $-0,5$  und  $2,5$  liegt.  
 (d) Zeichnen Sie mit Hilfe der MCMC-Simulation den Graphen der Dichtefunktion der Randverteilung von  $U$ .  
 (e) Bestimmen Sie aus (c) die Wahrscheinlichkeit, dass  $U$  zwischen  $-1,0$  und  $3,0$  und  $V$  zwischen  $1,5$  und  $7,5$  ist.

**6.5** Die Beratungsfirma Heidrick & Struggles hat untersucht, ob die Lohnhöhe von Verwaltungsratsmitgliedern mit der Qualität ihrer Arbeit, basierend auf 41 Kriterien, zusammenhängt. Die Resultate von 13 Ländern in Europa sind in Aufgabe 1.10 in Kap. 1 erwähnt und in Tab. 1.10 dargestellt.

**Tab. 6.5** Noten aus zwei Prüfungen von zwölf Studierenden (Daten aus Berner Fachhochschule, Burgdorf)

Studierender	1	2	3	4	5	6	7	8	9	10	11	12
Prüfung 1	4,2	4,1	5,3	5,5	5,2	5,7	4,1	4,3	4,4	4,0	4,7	5,1
Prüfung 2	4,0	4,0	5,8	5,4	5,9	5,5	4,5	4,8	3,6	4,9	5,6	5,1

- (a) Stellen Sie die Messpunkte in einem Streudiagramm dar und beurteilen sie grob, ob die Beobachtungsgrößen positiv, negativ oder nicht korreliert sind.
- (b) Bestimmen Sie den empirischen Korrelationskoeffizienten der Beobachtungspaare. Ist es sinnvoll, die Plausibilität zu den beiden Messgrößen mit einem gemeinsamen Wahrscheinlichkeitsmodell zu beschreiben?

**6.6** Eine Lehrperson untersucht Noten aus dem Modul Analysis in Maschinentechnik an der Berner Fachhochschule. Sie möchte wissen, ob sie aus der Note  $N_1$  der ersten Prüfung auf die Note  $N_2$  der zweiten Prüfung hochrechnen kann. Dazu müssten die beiden Messgrößen  $N_1$  und  $N_2$  korreliert sein. Tab. 6.5 zeigt die Noten aus einer Stichprobe von zwölf Studierenden.

- (a) Stellen Sie die Beobachtungen in einem Streudiagramm dar. Beurteilen Sie grob, ob die Beobachtungen der Noten 1 und Noten 2 positiv, negativ oder nicht korreliert sind.
- (b) Berechnen Sie den empirischen Korrelationskoeffizienten zwischen den Noten. Ist es sinnvoll anzunehmen, dass die beiden Messgrößen  $N_1$  und  $N_2$  voneinander abhängen?

**6.7** Mit einer Untersuchung zu Rucksäcken versucht eine Person von der Masse von Rucksäcken auf ihr Volumen zu rechnen. Dies geht nur, wenn die beiden Größen abhängig sind. Um dies festzustellen, benutzt die Person die Messungen von 24 Rucksäcken, die in Tab. 6.6 dargestellt sind.

- (a) Es ist nicht einfach, das Volumen eines Rucksacks operationell zu messen. Wie würden Sie das tun?
- (b) Der ASTM-Standard gibt an, wie man Volumen von Rucksäcken operationell messen kann. Viele Produzenten halten sich daran. Finden Sie den ASTM-Standard im Internet.<sup>2</sup>

<sup>2</sup> Ein Zitat dazu aus [1]: Trying to calculate the pack volume is hopeless: none of the shapes are either square or round. About the only reliable way to assess the volume of the main bag is to fill it up with something like sawdust or polystyrene foam 'packing peanuts', and then to measure the amount used in a nice rectangular box which can be accurately measured. In fact there is an ASTM Standard for this, using 20 mm hollow plastic spheres (essentially ping pong balls).

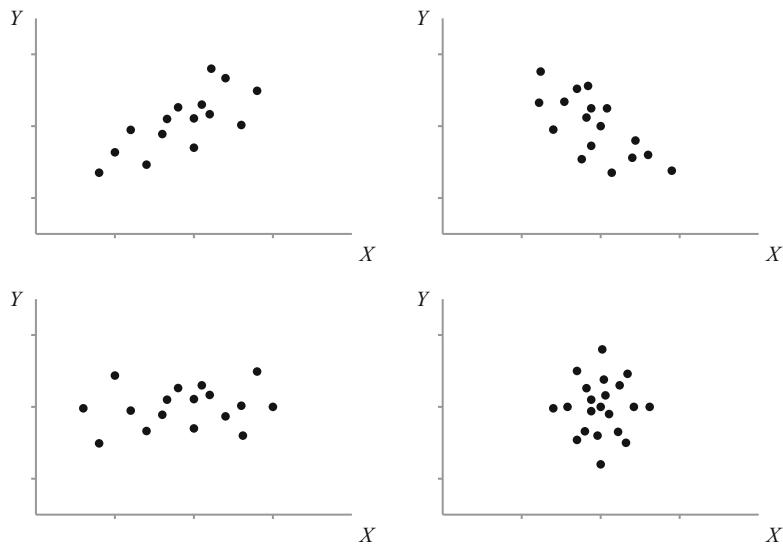
**Tab. 6.6** Masse und Volumen von 24 leichten Rucksäcken (aus [1])

Marke	Modell	Masse (in kg)	Volumen (in L)
Crux	AK47	1,17	45
Crux	AK57	1,30	52
Elem. Horizons	Northern Lite	1,23	52
GoLite	Quest (M)	1,42	57
GoLite	Quest (W)	1,29	53
GoLite	Odyssey (M)	1,59	72
GoLite	Odyssey (W)	1,42	64
Granite Gear	Escape AC 60	1,48	49
Granite Gear	Vapor Flash Ki	1,38	37
Granite Gear	Nimbus Ozone	1,44	53
JanSport	Big Bear 63	1,63	64
Lightwave	UltraHike 60	1,20	55
Lightwave	Fastpack 50	1,19	48
Lightwave	Wildtrek 55	1,46	49
Lowe Alpine	Nanon 50:60	1,42	53
Lowe Alpine	Zepton ND50	1,08	49
Mont-Bell	Versalite 50	1,33	50
Mont-Bell	Versalite 50	1,13	46
One Planet	Shadow	1,51	53
One Planet	Shadow (W)	1,45	51
Osprey	Exos 46	1,05	40
Osprey	Exos 58 U	1,19	50
REI	Flash 65	1,35	50
REI	Flash 50	1,18	43
ULA	Circuit	1,16	48
ULA	Catalyst	1,49	46
ULA	Camino	1,45	36

- (c) Stellen Sie die Messpunkte in einem Streudiagramm dar. Beurteilen Sie daraus grob, ob die beiden Messgrößen positiv, negativ oder nicht korreliert sind.
- (d) Berechnen Sie den empirischen Korrelationskoeffizienten der Messwertpaare. Ist es sinnvoll anzunehmen, dass die beiden Messgrößen Volumen und Masse voneinander abhängen?

**6.8** Abb. 6.15 zeigt vier Grafiken mit Messwerten von Untersuchungen zu zwei Größen  $X$  und  $Y$ . In welchen Fällen sind  $X$  und  $Y$  eher positiv korreliert, negativ korreliert oder nicht korreliert?

**6.9** In einer Vakuumkammer müssen gemäss vorgegebenen den Spezifikationen Messwerte des Unterdrucks möglichst zwischen 0,590 bar und 0,593 bar liegen. Um dies zu



**Abb. 6.15** Vier verschiedene Streudiagramme zu gemessenen Werten von  $X$  und  $Y$

kontrollieren wurden 20 Messungen durchgeführt. Sie sind in Tab. 1.7 beim Beispiel 1.8 dargestellt.

- (a) Überprüfen Sie mit einem Streudiagramm, ob Trends in den Daten vorliegen. Wenn nicht, hat es Messwerte, die nicht unter statistischer Kontrolle sind?
- (b) Ist es sinnvoll anzunehmen, dass die Messwerte trendfrei und unabhängig modelliert werden können? Beantworten Sie die Frage mit Hilfe der Autokorrelationsfunktion der Messreihe.

**6.10** Die Firma Tillamook-Cheese in Oregon (USA) produziert Frischkäsekörper mit einer mittleren Masse von ungefähr 19 kg. Die Körper werden anschliessend für den Handel in kleine Portionen von 500–1000 Gramm zerschnitten. Tab. 1.13 in Kap. 1 zeigt 20 Messungen von Massen der Körper aus einem Teil der Tagesproduktion.

- (a) Überprüfen Sie, ob die Messwerte keine Trends zeigen. Sind Messwerte ausserhalb der Kontrollgrenzen vorhanden?
- (b) Ist es plausibel anzunehmen, dass die Messwerte trendfrei und unabhängig modelliert werden können? Beantworten Sie die Frage mit Hilfe der Autokorrelationsfunktion der Messreihe.

## Literatur

1. R. Caffin, Lightweight Internal Frame Packs: a State of the Market Report – Part 1A: Testing Overview and Lists of Packs Tested. BackpackingLight.com, ISSN 1537-0364 (2010)
2. C. Chatfield, *The analysis of time series, an introduction*, 6th ed. (Chapman & Hall/CRC, New York, 2004)
3. H. Graf, F. R. Hampel, J.-D. Tacier: The problem of unsuspected serial correlations, In Robust and Nonlinear Time Series Analysis. J. Franke, W. Härdle, R. D. Martin (eds.), Lecture Notes in Statistics 26. Springer, New York, 127–145 (1984)
4. Student, Errors of routine analysis. *Biometrika* **19**, 151–164 (1927)

„Wie lautet euer Urteil?“ fragte der König die Schöffen.  
„Halt, noch nicht!“ rief das Weisse Kaninchen dazwischen, „vor dem Urteil kommt noch allerlei anderes!“  
„Ruft den ersten Zeugen“, sagte der König; und das Weisse Kaninchen stiess dreimal in seine Fanfare und rief: „Erster Zeuge!“  
Lewis Carroll, Alice im Wunderland (Insel Taschenbuch, 1973, S. 112)

## Zusammenfassung

Mit statistischen Werkzeugen lassen sich nicht direkt messbare Größen bestimmen. Dies ist exemplarisch in Kap. 5 gezeigt. Mit den Gesetzen zur Wahrscheinlichkeitsrechnung lassen sich auch zukünftige gemessene oder beobachtete Werte einer Größe prognostizieren. Ein solcher Blick in die Zukunft ist wegen fehlender Information meist mit Unsicherheit verbunden. Daher wird angegeben, wo solche Messwerte mit welcher Wahrscheinlichkeit liegen werden. Für diese Aufgabe braucht man einerseits ein Datenmodell für mögliche Messwerte und andererseits die Plausibilität zu den Parametern des Datenmodells. Zukünftige Messwerte oder Beobachtungen kann man daraus mit dem Gesetz der Marginalisierung, das im vorigen Kapitel erklärt ist, prognostizieren. Dazu muss man Integrale ausrechnen. Sie sind kaum explizit berechenbar. Daher wird ein Verfahren vorgestellt, das auf einer Computersimulation aufbaut.

## 7.1 Objekte, die in zwei Kategorien auftreten

Oft hat man eine Grundgesamtheit von Objekten, die in zwei Kategorien null oder eins eingeteilt werden können. So hat man bei einer Produktionsserie Geräte, die defekt oder nicht defekt sind. Der Anteil der defekten Geräte einer Produktionsserie ist eine mögliche, gesuchte Zahl. Patienten können ein bestimmtes Medikament vertragen oder nicht vertragen. Eine medizinische Kontrollstelle möchte daher wissen, wie gross der Anteil

der Personen ist, die das Medikament nicht vertragen. Wie man aus einer Stichprobe und Vorinformation solche Anteile berechnen kann, ist in Kap. 5 gezeigt. Wie gross ist aber die Wahrscheinlichkeit, dass das nächste untersuchte Gerät defekt ist oder die nächste untersuchte Person das Medikament nicht verträgt? Das folgende Beispiel illustriert die dazu nötige Rechnung.

**Beispiel 7.1 (HNV-Indikator)** Um den HNV-Indikator von Deutschland zu kennen (siehe das Beispiel 5.3), wird die Fläche von Deutschland in  $1 \text{ km}^2$  grosse Flächen zerlegt. Von diesen Flächen wird eine Stichprobe analysiert. Stichprobenflächen mit weniger als 5 % Anteil an Landwirtschaftsfläche werden nicht untersucht. Wir wollen solche Flächen Klasse 1-Flächen nennen, die anderen Klasse 0-Flächen. Es interessiert der Anteil  $A$  der  $1 \text{ km}^2$ -Flächen in Deutschland, die in der Klasse eins sind. Eine Stichprobe von acht Flächen, drei davon in Klasse eins, liefert Information zu  $A$ . Da kein Zensus vorliegt, ist das Wissen zu  $A$  unsicher. Es sollte daher mit einer Wahrscheinlichkeit ausgedrückt werden. Die Messwerte seien unabhängig. Weitere Information  $\mathcal{I}$  zu  $A$  ist nicht vorhanden. Daher ist

$$\text{Datenmodell: } i\text{-ter Messwert } \sim \text{Bernoulli}(A), \quad \text{Prior: } A \sim \text{Uniform}(0; 1)$$

Die  $A$  posteriori-Verteilung des Anteils  $A$  ist nach Theorem 5.3

$$\text{pdf}(A \mid \text{Daten}, \mathcal{I}) \propto \mathbb{P}(\text{Daten} \mid A) \cdot \text{pdf}(A \mid \mathcal{I}) \propto \underbrace{A^3 \cdot (1 - A)^5}_{\text{Likelihood}} \cdot \underbrace{\frac{1}{A}}_{\text{Prior}}$$

Die fehlende Konstante kann man berechnen: Die Fläche unterhalb des Graphen der Funktion muss eins sein. Man erhält

$$\text{pdf}(A \mid \text{Daten}, \mathcal{I}) = 504 \cdot A^3 \cdot (1 - A)^5$$

Der plausibelste Wert für  $A$  ist der Modus  $A_0 = 3/8 = 0,375$ .

Eine Biologin wählt nun eine zusätzliche Fläche aus. Wie gross ist die Wahrscheinlichkeit, dass die Fläche in Klasse eins liegt? Wenn die Biologin den Anteil  $A$  genau kennen würde, wäre die Antwort einfach. Das Datenmodell besagt nämlich, dass

$$\mathbb{P}(\text{Klasse 1} \mid A) = A \tag{7.1}$$

Leider kennt die Biologin  $A$  nicht. Wie soll man also vorgehen? Hier zwei Verfahren dazu:

- (1) Man könnte in Gleichung (7.1) den plausibelsten Wert von 0,375 für  $A$  einsetzen und behaupten: Die Wahrscheinlichkeit beträgt ungefähr 0,375, dass die nächste untersuchte Fläche in Klasse 1 liegt.
- (2) Eine zweite Vorgehensweise ist vorsichtiger: Viele Anteile  $A$  sind gemäss der  $A$  posteriori-Verteilung möglich. Zum Beispiel kann  $A$  der Wert 0,375 oder der Wert 0,5

sein. Der Wert 0,375 ist aber wahrscheinlicher als der Wert 0,5:

$$\frac{\text{pdf}(A = 0,375 \mid \text{Daten}, \mathcal{I})}{\text{pdf}(A = 0,5 \mid \text{Daten}, \mathcal{I})} = \frac{0,375^3 \cdot (1 - 0,375)^5}{0,5^3 \cdot (1 - 0,5)^5} = \frac{2,535}{1,969} = 1,287$$

Es scheint daher vernünftig, die gesuchte Wahrscheinlichkeit aus den „zwei“ Datenmodellen  $\mathbb{P}(\text{Klasse } 1 \mid A = 0,375)$  und  $\mathbb{P}(\text{Klasse } 1 \mid A = 0,5)$  zu mitteln. Dabei sollte der Wert 0,375 um 1,287-fach gewichtet werden als der Wert 0,5. Also

$$\mathbb{P}(\text{nächste Fläche in Klasse } 1) \approx \frac{0,375 \cdot 2,535 + 0,5 \cdot 1,969}{2,535 + 1,969} = 0,430$$

Das zweite Verfahren lässt sich weiter ausbauen. Man kann alle möglichen Werte für den Anteil  $A$  nehmen, diese gewichten mit der Wahrscheinlichkeit, dass der Anteil  $A$  diese Werte annimmt, und aufsummieren. Mit einer Formel ausgedrückt, hat man:

$$\mathbb{P}(\text{Klasse } 1 \mid \text{Daten}, \mathcal{I}) \approx \int_0^1 \underbrace{A}_{\text{Anteil}} \cdot \underbrace{504 \cdot A^3 \cdot (1 - A)^5}_{\text{Plausibilität}} \, dA = 0,4$$

Es zeigt sich, dass man damit die gesuchte Wahrscheinlichkeit erhält. Hier die Rechnung dazu: Gesucht ist die Wahrscheinlichkeit, dass die nächste untersuchte Fläche in Klasse eins liegt, gegeben das Wissen aus den acht Stichprobenflächen. Mathematisch geschrieben ist dies

$$\mathbb{P}(\text{Klasse } 1 \mid \text{Daten}, \mathcal{I})$$

Die Biologin kennt aber andere Wahrscheinlichkeiten. Zuerst hat sie das Datenmodell. Dieses sagt, wie gross die Wahrscheinlichkeit ist, dass die nächste Fläche in Klasse eins ist, wenn der Anteil  $A$  bekannt wäre:  $\mathbb{P}(\text{Klasse } 1 \mid A) = A$ . Welche Werte von  $A$  am plausibelsten sind, kennt die Biologin aus der  $A$  posteriori-Verteilung. Dies ist  $\mathbb{P}(A) = \mathbb{P}(A \mid \text{Daten}, \mathcal{I})$ . Bildet man das Produkt der beiden Wahrscheinlichkeiten, erhält man aus dem Multiplikationsgesetz das gemeinsame Wahrscheinlichkeitsmodell von  $A$  und einer Messung „Klasse 1“:

$$\mathbb{P}(\text{Klasse } 1 \mid A) \cdot \mathbb{P}(A) = \mathbb{P}(\text{Klasse } 1 \text{ und } A)$$

Aus diesem gemeinsamen Modell für den Parameter  $A$  und für die Aussage „Klasse 1“ lässt sich mit der Marginalisierung aus Theorem 6.1 der Parameter  $A$  eliminieren. Man erhält die gesuchte Wahrscheinlichkeit, indem man über alle möglichen Anteile  $A$  zwischen null und eins aufsummiert:

$$\mathbb{P}(\text{Klasse } 1 \mid \text{Daten}, \mathcal{I}) = \int_0^1 \mathbb{P}(\text{Klasse } 1 \text{ und } A \mid \text{Daten}, \mathcal{I}) \, dA$$

Explizit ergibt sich damit die oben erwähnte Formel:

$$\begin{aligned}\mathbb{P}(\text{Klasse } 1 \mid \text{Daten}, \mathcal{I}) &= \int_0^1 \mathbb{P}(\text{Klasse } 1 \mid A) \cdot \mathbb{P}(A \mid \text{Daten}, \mathcal{I}) dA \\ &= \int_0^1 A \cdot [504 \cdot A^3 \cdot (1 - A)^5] dA = 0,4\end{aligned}$$

Die Biologin prognostiziert also, dass die nächste Stichprobenfläche mit einer Wahrscheinlichkeit von 0,4 in Klasse eins und mit einer Wahrscheinlichkeit von 0,6 in Klasse null sein wird. Man nennt dies auch das *Prognosemodell* für den nächsten Messwert.

Der Wert 0,4 entspricht nicht dem plausibelsten Wert von  $A$ , der 0,375 lautet. Warum ist dies so? Der Grund liegt darin, dass die A posteriori-Verteilung von  $A$  leicht schief ist. Die Wahrscheinlichkeit, dass  $A$  grösser als der plausibelste Wert 0,375 ist, beträgt

$$\mathbb{P}(A > 0,375 \mid \text{Daten}, \mathcal{I}) = \int_{0,375}^1 504 \cdot A^3 \cdot (1 - A)^5 dA = 0,55$$

Tendenziell ist  $A$  somit grösser als der plausibelste Wert 0,375. Das Gesetz der Marginalisierung berücksichtigt dies. Es gewichtet das Datenmodell  $\mathbb{P}(\text{Klasse } 1 \mid A)$  mit der A posteriori-Verteilung von  $A$  und summiert über alle möglichen Anteile  $A$  auf. Die Wahrscheinlichkeit, dass die nächste untersuchte Fläche in Klasse 1 ist, ist daher nicht 0,375, sondern sie liegt leicht höher.  $\square$

Die obige Rechnung zeigt, wie man eine Prognose berechnen kann, dass ein Objekt aus einer Grundgesamtheit in eine von zwei Kategorien 0 und 1 sein wird. Mit dem Gesetz der Marginalisierung ist

$$\mathbb{P}(\text{Objekt in } 1 \mid \text{Daten}) = \int_0^1 \text{pdf}(\text{Objekt in } 1 \text{ und Anteil ist } A) dA$$

Der Integrand wird aus dem Produkt des Datenmodells und des Posteriors von  $A$  berechnet:

$$\text{pdf}(\text{Objekt in } 1 \text{ und Anteil ist } A) = \underbrace{A}_{\text{Datenmodell}} \cdot \underbrace{\text{pdf}(A \mid \text{Daten})}_{\text{Posterior}}$$

Bei minimaler Vorinformation kann man das Integral explizit berechnen. Man hat:

**Theorem 7.1 (Prognose zu zwei Kategorien bei minimaler Vorinformation)**

Objekte können in zwei Kategorien 0 und 1 auftreten. Bei  $n$  unabhängigen Messungen sind  $\alpha$  der Objekte in der Kategorie 1. Weitere Information ist nicht vorhanden.

*Die Wahrscheinlichkeit, dass das nächste gemessene Objekt in Kategorie 1 ist, beträgt*

$$\mathbb{P}(\text{Objekt ist in Kategorie 1} \mid \text{Daten, min. Vorinformation}) = \frac{\alpha + 1}{n + 2}$$

Man nennt dies auch die *Sukzessionsregel von Laplace* (engl. *Laplace's law of succession*). Die Regel funktioniert auch, wenn keine Messungen vorhanden sind: Ist  $n = \alpha = 0$ , so ist die Wahrscheinlichkeit 0,5, dass das nächste zufällig gewählte Objekt in Kategorie 1 liegt. Dies entspricht Indifferenz zum Anteil.

Die Sukzessionsregel von Laplace lässt sich verallgemeinern. So gibt es entsprechende Formeln, wenn Zusatzinformationen und nicht minimale Vorinformation vorhanden ist. Eine andere Variante ist: Man hat Objekte, die in  $B$  verschiedenen Kategorien sind.<sup>1</sup> Weiter habe man vor der Datensammlung keine weitere Information und beobachte  $\alpha$  mal die Kategorie  $K$  aus  $n$  unabhängigen Messwerten. Dann ist:

$$\mathbb{P}(\text{nächstes Objekt ist in Kategorie } K \mid \text{Daten, min. Vorinformation}) = \frac{\alpha + 1}{n + B}$$

Die Prognose hängt also davon ab, wie viele Kategorien auftreten können. Ein Beispiel dazu sind Objekte, die in vier Kategorien sein können. Bei sechs untersuchten, unabhängigen Objekten liegen zwei in Kategorie  $X$ . Dann ist die Wahrscheinlichkeit, dass das nächste gezogene Objekt in Kategorie  $X$  ist

$$\mathbb{P}(\text{nächstes Objekt ist in Kategorie } X \mid \text{Daten, min. Vorinformation}) = \frac{2 + 1}{6 + 4} = 0,3$$

Können die Objekte nur in zwei Kategorien auftreten, so ist

$$\mathbb{P}(\text{nächstes Objekt ist in Kategorie } X \mid \text{Daten, min. Vor.}) = \frac{2 + 1}{6 + 2} = 0,375$$

Die Wahrscheinlichkeit ist grösser geworden, da das Objekt in weniger Kategorien sein kann.

**Beispiel 7.2 (HNV-Indikator)** Beim obigen Beispiel werden acht Flächen untersucht, die sich in zwei Klassen aufteilen. Drei davon finden sich in der Klasse eins. Die Wahr-

---

<sup>1</sup> Man hat Anteile  $A_1, A_2, \dots, A_B$  der Kategorien  $1, 2, \dots, B$ . Die gemeinsame A-posteriori-Verteilung, um die Plausibilität zu den Anteilen bei einem flachen Prior zu beschreiben, ist eine Dirichlet-Verteilung. Die Randverteilungen der Anteile sind wieder Beta-Verteilungen.

scheinlichkeit, dass die nächste, zufällig gezogene Probefläche in Klasse eins fällt, ist

$$\mathbb{P}(\text{Klasse eins} \mid \text{Daten, min. Vorinformation}) = \frac{3+1}{8+2} = 0,4$$

Für einen Biologen ist es auch interessant zu wissen, wie gross der Anteil  $A_{\text{Militär}}$  aller  $1 \text{ km}^2$ -Flächen in Deutschland ist, die in militärischen Übungsplätzen liegen. Solche Flächen können nämlich nicht besichtigt werden. Damit hat man zwei Kategorien: militärische und nicht-militärische Übungsplätze. Der Biologe wählt den flachen Prior zu diesem Anteil. Alle acht Stichprobenflächen sind nicht solche Plätze. Daher ist der Posterior von  $A_{\text{Militär}}$  gemäss Theorem 5.3:

$$\text{pdf}(A_{\text{Militär}} \mid \text{Daten, min. Vorinformation}) \propto A_{\text{Militär}}^0 \cdot (1 - A_{\text{Militär}})^8$$

Dies ist eine Beta-Verteilung mit Kennzahlen 1 und 9. Der plausibelste Wert von  $A_{\text{Militär}}$  ist 0, was nicht erstaunlich ist. Mit einer Wahrscheinlichkeit von 0,5 ist  $A_{\text{Militär}}$  kleiner als 0,14. Mit einer Wahrscheinlichkeit von 0,95 ist  $A_{\text{Militär}} \leq 0,34$ .

Mit der Sukzessionsregel lautet die Wahrscheinlichkeit, dass die folgende untersuchte Fläche auf einem militärischen Übungsplatz liegt:

$$\mathbb{P}(\text{Fläche ist Übungsplatz} \mid \text{Daten, min. Vorinformation}) = \frac{0+1}{8+2} = 0,1$$

Sie beträgt also 0,1 und nicht – wie vielleicht erwartet – null.  $\square$

## 7.2 Ein Verfahren, um Messwerte zu prognostizieren

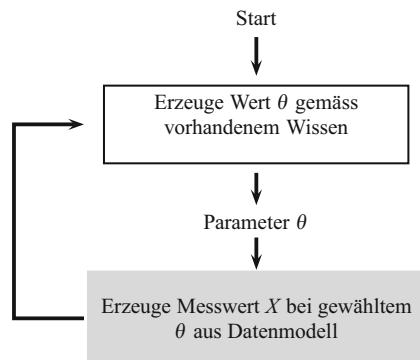
Die oben berechneten Prognosen folgen aus dem Gesetz der Marginalisierung. Benutzt wird dabei ein Datenmodell, das von Parametern abhängt. Dieses sagt, wie Messwerte streuen. Die Plausibilität zu den Parametern des Datenmodells wird mit einem Wahrscheinlichkeitsmodell angegeben. Das Vorgehen, um weitere Messwerte zu prognostizieren, kann man folgt formulieren:

### Theorem 7.2 (Prognose von Messwerten)

*Ein Messwert  $X$  einer Zufallsgrösse, gegeben eine Information  $\mathcal{I}$ , kann wie folgt prognostiziert werden: Man braucht (1) ein Datenmodell in Funktion von Parametern  $\theta$ , das besagt, wie Messwerte streuen, und (2) die Plausibilität zu den Parametern  $\theta$  des Datenmodells. Dann hat man*

$$\mathbb{P}(X \mid \mathcal{I}) = \int \mathbb{P}(X \text{ und } \theta \mid \mathcal{I}) d\theta = \int \underbrace{\mathbb{P}(X \mid \theta \text{ und } \mathcal{I})}_{\text{Datenmodell für } X} \cdot \underbrace{\mathbb{P}(\theta \mid \mathcal{I})}_{\text{Plausibilität zu } \theta} d\theta$$

**Abb. 7.1** Monte-Carlo-Simulation um Messwerte oder Beobachtungen von Zufallsgrößen zu prognostizieren



In Worten: *Die Wahrscheinlichkeit eines Werts  $X$  der Zufallsgröße berechnet man, indem in das Datenmodell alle möglichen Werte der Modellparameter eingesetzt werden, gewichtet mit der Wahrscheinlichkeit, dass die Modellparameter diese Werte annehmen. Dann wird aufsummiert.*

Das Verfahren ist ein Spezialfall des Bayes'schen Mittelns von Modellen (engl. Bayes Model Averaging): für die Prognose  $\mathbb{P}(X)$  eines Messwerts mittelt man über die „verschiedenen“ Datenmodelle  $\mathbb{P}(X | \theta)$  für alle möglichen Parameterwerte  $\theta$  aus.

Das erste obige Integral eliminiert durch Integration den Parameter  $\theta$  aus dem gemeinsamen Wahrscheinlichkeitsmodell für Messwerte und Parameter. Das zweite Integral folgt aus dem Gesetz der Multiplikation. Bei den obigen Beispielen ist der Anteil  $A$  der einzige Parameter,  $\mathbb{P}(\text{Messwert} = 1 | A) = A$  das Datenmodell und die Plausibilität zu  $A$  erhält man aus der A posteriori-Verteilung von  $A$ .

Die oben erwähnten Integrale können in der Regel nicht explizit berechnet werden. Mit der folgenden Monte-Carlo-Simulation lassen sie sich aber approximativ bestimmen:

### Theorem 7.3 (Simulation des Prognosemodells)

Eine Simulation, um Werte einer Zufallsgröße zu prognostizieren, ist:

- (1) Erzeuge einen Wert  $\theta$  nach der Plausibilität  $pdf(\theta | I)$  für den Parameter  $\theta$ .
- (2) Bilde damit einen Messwert  $X$  nach dem Datenmodell  $\mathbb{P}(X | \theta)$ .
- (3) Wiederhole das Verfahren von (1) und (2) mehrere tausend Mal.

Die Simulation ist in Abb. 7.1 dargestellt. Die Wahrscheinlichkeit, dass die Zufallsgrösse einen Wert zwischen  $a$  und  $b$  haben wird, berechnet man aus der Simulation. Es ist:

$$\mathbb{P}(a \leq X \leq b | I) \approx \frac{\text{Anzahl erzeugte Werte von } X \text{ zwischen } a \text{ und } b}{\text{Anzahl Wiederholungen des Verfahrens}}$$

Dies wird im Folgenden illustriert:

**Beispiel 7.3 (Qualität eines Expertensystems)** Die Autobahnen in der Schweiz sind stark befahren. Verkehrsüberlastungen oder Unfälle führen daher zu Staus und Verzögerungen. Fahrbahnen, die mit Schnee und Eis bedeckt sind, erhöhen die Unfallgefahr besonders stark. Deshalb werden im Winter automatische Expertensysteme eingesetzt, die vor gefährlichen Strassenzuständen warnen (siehe [1]). Sie helfen dem Unterhaltspersonal, die Strassen rechtzeitig zu salzen und sie von Schnee und Eis freizuhalten.

Alarne des Expertensystems treten in zwei Kategorien auf: korrekte und falsche Alarne. Eine Ingenieurin möchte wissen, wie gross der Anteil  $A_{\text{falsch}}$  der falschen Alarne ist. Bei 20 Warnungen des Systems findet sie zwei falsche Alarne. Weiter hat sie keine andere Information zu  $A_{\text{falsch}}$ . Die A posteriori-Verteilung von  $A_{\text{falsch}}$  ist damit eine Beta-Verteilung mit Kennzahlen 3 und 19:

$$\text{pdf}(A_{\text{falsch}} | \text{Daten, min. Vorinformation}) \propto A_{\text{falsch}}^2 \cdot (1 - A_{\text{falsch}})^{18}$$

Der plausibelste Wert für  $A_{\text{falsch}}$  ist 0,10. Wahrscheinlichkeitsintervalle für diesen Parameter lassen sich mit einer MCMC-Kette, bestehend aus 10 000 Punkten, bestimmen.

Wie gross ist die Wahrscheinlichkeit, dass der nächste Alarm des Expertensystems falsch ist? Nach der Sukzessionsregel mit zwei Merkmalen von Theorem 7.1 ist sie

$$\mathbb{P}(\text{falscher Alarm} | \text{Daten, min. Vorinformation}) = \frac{2 + 1}{20 + 2} = 0,14$$

Mit einer nach dem obigen Verfahren konstruierten Monte-Carlo-Simulation kann man dies ebenfalls berechnen. Mögliche Anteile  $A_{\text{falsch}}$  lassen sich einfach aus der Kette der MCMC-Simulation ablesen. Mit diesen Werten lassen falsche oder korrekte Alarne simulieren. Hier ein Pseudocode, der dies umsetzt:

```

for i = 1 to 10000 do
    Afalsch[i] = i-ter Punkt aus der MCMC-Simulation
                für Afalsch
    Messwert[i] = falsch (1) bzw. korrekt (0),
                 mit Wahrscheinlichkeit Afalsch[i]
                 bzw. 1-Afalsch[i]
end

```

Die folgende Tabelle zeigt die ersten sechs Werte der Simulation, bei der zehntausend Alarme erzeugt wurden:

$A_{\text{falsch}}$ aus der MCMC-Kette	Daraus simuliert: falsch (1) / korrekt (0)
0,2186398	0
0,2835980	0
0,1798409	1
0,1740201	0
0,2415998	1
...	...

Von den 10 000 erzeugten Null- und Eins-Werten waren 1425 eins. Damit ist

$$\mathbb{P}(\text{falscher Alarm} \mid \text{Daten, min. Vorinformation}) \approx \frac{1425}{10\,000} = 0,1425$$

Das Resultat stimmt gut mit dem exakten Wert von 0,14 überein.  $\square$

## Reflexion

**7.1** In einem Tiergehege mit vielen kleinen Tieren soll der Anteil der weiblichen Tiere berechnet werden. Dazu wurden zwölf Tiere nacheinander gefangen und ihr Geschlecht festgestellt. Gezogen wurden die Tiere durch Ziehen mit Zurücklegen, um zu garantieren, dass die Messwerte unabhängig sind. Hier das Resultat:

Gefangene Tiere: 12, weiblich: 10, männlich: 2

Weitere Information zum Anteil an weiblichen Tieren ist nicht vorhanden.

- (a) Prognostizieren Sie: Wie gross ist die Wahrscheinlichkeit – gegeben die Daten –, dass das nächste, zufällig gewählte Tier weiblich ist? Bestimmen Sie diese Wahrscheinlichkeit auch mit einer Monte-Carlo-Simulation.
- (b) Führen Sie die Rechnung von (a) noch einmal durch, diesmal mit einer Stichprobe von 20 Tieren, bei der alle weiblich waren.

**7.2** Die Leitung einer Firma, die 3000 Personen angestellt hat, möchte erfahren, ob die Angestellten mit ihrer Arbeit zufrieden sind. Dazu werden mit zufälligem Ziehen zwanzig Angestellte ausgewählt und befragt. Die Antworten treten in Kategorien „zufrieden“ und „nicht zufrieden“ auf. Hier das Resultat:

Befragte Personen: 20, Antwort „zufrieden“: 18

Hat man keine weiteren Informationen zur Arbeitszufriedenheit, so lautet der Posterior zum Anteil  $A$  aller Personen in der Firma, die mit ihrer Arbeit zufrieden sind:

$$\text{pdf}(A \mid \text{Daten, min. Vorinformation}) \propto A^{18} \cdot (1 - A)^2$$

- (a) Berechnen Sie die Wahrscheinlichkeit, dass die nächste, zufällig ausgewählte Person sagt, dass sie mit der Arbeit zufrieden ist.
- (b) Bestimmen Sie die Wahrscheinlichkeit in (a) auch mit einer Monte-Carlo-Simulation.
- (c) Wie lautet die Wahrscheinlichkeit in (a), wenn der Fragebogen die Antworten „zufrieden“, „nicht zufrieden“ und „weiss nicht“ zulässt?

**7.3** Mit einer Umfrage möchte eine Firma wissen, wie gross der Anteil  $A$  der erwachsenen Personen in der Schweiz ihr Produkt „Supergut!“ kennen. Dazu werden 1000 Personen befragt. Als Antworten sind erlaubt: „Kenne ich!“ oder „Kenne ich nicht!“. 604 Personen geben an, das Produkt zu kennen. Als (unplausible?) Annahme gilt: die Antworten besitzen keine systematischen Fehler und sind unabhängig.

- (a) Berechnen Sie die Wahrscheinlichkeit, dass die nächste, zufällig ausgewählte Person sagt, dass sie das Produkt „Supergut!“ kennt.
- (b) Wie lautet die Antwort zu (a), wenn der Fragebogen neben den oben erwähnten Antworten auch die Antwort „Ich bin mir nicht sicher“, zulässt?

**7.4** In Publikationen werden Ergebnisse zur Wirkung von Medikamenten in verschiedenen Formen dargestellt. In einem Kurs mussten 119 Ärzte, die in der Schweiz praktizieren, vier solcher Ergebnisse lesen. Dabei interpretierten nach [2] 13 Ärzte die Ergebnisse richtig und die anderen 106 Ärzte falsch.

- (a) Berechnen Sie aus der gegebenen Information den Anteil  $A$  aller in der Schweiz praktizierenden Ärzte, die die vier Ergebnisse richtig interpretieren würden. Geben Sie als Resultat an: die  $A$  posteriori-Verteilung von  $A$ , den plausibelsten Wert und Wahrscheinlichkeitsintervalle zum Niveau 0,5 und 0,95. Bestimmen Sie mit den „Standardfehler-Formeln“ auch Wahrscheinlichkeitsintervalle für  $A$  zum Niveau von ungefähr 0,5 und ungefähr 0,95.
- (b) Geben Sie die Fünf-Zahlen-Zusammenfassung des Resultats aus der Aufgabe (a) an.
- (c) Berechnen Sie aus der vorhandenen Information die Wahrscheinlichkeit, dass der nächste befragte Arzt die Ergebnisse richtig interpretiert.

## Literatur

1. D. Bättig, S. Eggimann, O. Mermoud, U. Mori, S. Stankowski, Strassenglätte-Prognosesystem. Forschungsauftrag ASTRA 2008/002, OBF (2010)
2. Ch. Damur, J. Steurer, Beurteilen Ärzte Therapieergebnisse anders als Studenten? Schweiz. Med. Wochenschrift, **130**, 171–176 (2000)

„Was weisst du von dieser Angelegenheit?“ fragte der König Alice.  
„Nichts“, sagte Alice.  
„Nicht das geringste?“ forschte der König weiter.  
„Nicht das geringste“, sagte Alice.  
„Das ist sehr wichtig“, sagte der König, zu den Schöffen gewandt.  
Die wollten sich das gerade aufschreiben, als das Weisse Kaninchen einfiel: „Unwichtig meinen Euer Majestät natürlich“, sagte es sehr unterwürfig, doch runzelte es dabei die Stirn und schnitt allerlei Gesichter.  
*Lewis Carroll, Alice im Wunderland (Insel Taschenbuch, 1973, S. 120.)*

## Zusammenfassung

Nicht direkt messbare Größen, wie der Anteil von Objekten in einer Grundgesamtheit, die in einer gewissen Kategorie sind, kann man aus Daten und zusätzlichen Informationen berechnen. In Kap. 5 ist gezeigt, dass die Regel von Bayes die Vorinformation (den Prior) zur nicht direkt messbaren Größe aktualisiert. Man erhält eine Genauigkeit und eine Plausibilität zur gesuchten Größe. Für die Regel von Bayes braucht man ein Datenmodell, das besagt, wie Messwerte streuen. Damit glaubwürdig wird, was gerechnet wird, müssen das Datenmodell und der Prior erklärt werden. Die in diesem Kapitel vorgestellten Argumente, um Modelle zu wählen, sind Skalierungs- und Informationsregeln. Am Schluss des Kapitels wird eine wichtige Kennzahl eines Wahrscheinlichkeitsmodells definiert. Es ist der Erwartungswert oder der durchschnittlich erwartbare Wert. Hat man Information dazu, kann dies helfen, ein Wahrscheinlichkeitsmodell auszuwählen.

## 8.1 Das Problem: Modell und Vorwissen

Mit der Regel von Bayes lassen sich nicht direkt messbare Größen, auch Parameter genannt, bestimmen. Bei technischen oder physikalischen Problemen sind solche Parameter etwa die Masse eines Körpers, die Temperatur einer Substanz, die Konzentration einer Lösung oder der Druck in einer Kammer. In der Medizin kann dies die Heilrate eines Medikaments sein. Dabei variieren die Messungen oder Beobachtungen, weil die Experimentier- oder die Versuchsumgebung nicht konstant gehalten werden kann oder Messungen von Proband zu Proband verschieden ausfallen. So werden angezeigte Körpergewichte variieren, weil das Luftvolumen im Körper ändert und Personen nicht immer gleich ruhig auf der Waage stehen. Auch die Messinstrumente besitzen Messunsicherheiten und bewirken, dass die Messwerte streuen. Theorem 5.2 zeigt, wie die Plausibilität zu einem Parameter berechnet werden kann:

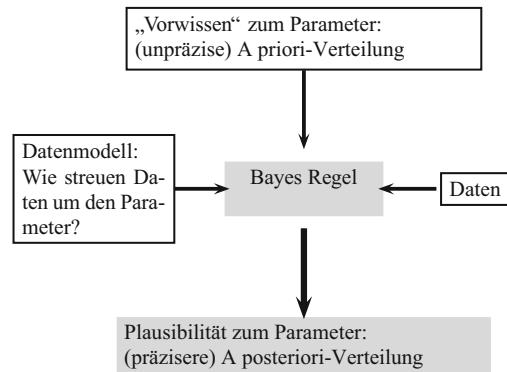
Wie präzis und wie plausibel eine nicht direkt messbare Größe oder ein Parameter  $\theta$  berechnet ist, beschreibt die A posteriori-Plausibilität:

$$\underbrace{\mathbb{P}(\theta \mid \text{Daten}, \mathcal{I})}_{\text{A posteriori-Plausibilität}} \propto \underbrace{\mathbb{P}(\text{Daten} \mid \theta)}_{\text{Likelihood}} \cdot \underbrace{\mathbb{P}(\theta \mid \mathcal{I})}_{\text{A priori-Plausibilität}} \quad (8.1)$$

Die A priori-Plausibilität (der Prior) kodiert die Plausibilität zu  $\theta$  aus der Vorinformation  $\mathcal{I}$ . Die Likelihood  $\mathbb{P}(\text{Daten} \mid \theta)$  beschreibt, wie die Messungen streuen. Sie wird aus dem Daten- oder Streumodell berechnet.

In Abb. 8.1 ist das Vorgehen schematisch dargestellt. Wie soll man den Prior und das Datenmodell wählen? Wie lautet insbesondere der Prior bei minimaler Vorinformation zur gesuchten Größe? Die nächsten zwei Abschnitte stellen zwei Methoden vor, mit denen solche Fragen beantwortet werden können.

**Abb. 8.1** Wie man mit der Regel von Bayes die Plausibilität zu einem Parameter berechnet oder aktualisiert



## 8.2 Transformation und minimale Vorinformation

Was bedeutet minimale Vorinformation zu einer nicht direkt messbaren Grösse? Wie wählt man ein Wahrscheinlichkeitsmodell, dass diese widerspiegelt? Um dies auszuleuchten, ist es sinnvoll, zuerst ein anderes Problem zu analysieren. Manchmal müssen Parameter verrechnet werden. So kennt man den Anteil defekter Geräte in einer Gesamtproduktion und daraus müssen Kosten wegen Garantieansprüche berechnet werden. Wie kann man hier die Plausibilität zur zweiten Grösse direkt aus der ersten Grösse berechnen?

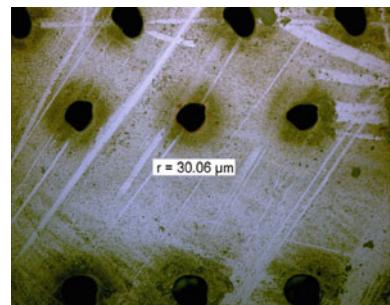
**Beispiel 8.1 (Bohrlöcher mit Laserimpulsen)** Ultrakurze Laserpulse mit einer Dauer von wenigen Femtosekunden können benutzt werden, um kleine Öffnungen in Halbleiter zu bohren. Das Resultat zeigt Abb. 8.2. In einem Labor in Burgdorf wurden Versuche angestellt, um die Möglichkeiten und Grenzen des Bohrens mit Laserimpulsen zu erforschen. Gesucht waren die Fläche  $F$  und der Durchmesser  $D$  eines Kreislochs, die ein ps-Laser Lithium Niobat Platten brennt. Die beiden Größen können wegen Kovariablen (nicht konstant zu haltende Experimentierfaktoren) und Unreinheiten in den Platten nicht direkt gemessen werden. Messungen variieren um  $F$  und  $D$ . Die Information zu  $F$  und  $D$  ist daher mit Wahrscheinlichkeiten zu beschreiben. So hat man für  $F$  das Wahrscheinlichkeitsintervall

$$\mathbb{P}(150 \mu\text{m}^2 \leq F \leq 180 \mu\text{m}^2 \mid \text{Daten, Vorwissen}) = 0,95$$

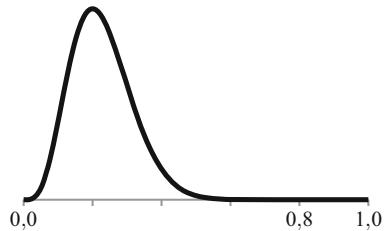
Da man den Durchmesser  $D$  aus der Kreisfläche  $F$  berechnen kann, müsste eigentlich ein entsprechendes Wahrscheinlichkeitsintervall für  $D$  daraus bestimmbar sein. Wie macht man dies?  $\square$

**Beispiel 8.2 (Kunden und Kosten)** Eine Firma mit einem Kundenkreis von rund 50 000 Personen will ein neues Produkt herstellen. Mit einer Umfrage versucht sie den Anteil  $A$  der 50 000 Personen zu berechnen, die das neue Produkt kaufen würden. Von zwanzig durch Randomisierung ausgewählten Personen, würden vier das Produkt kaufen. Nach Theorem 5.3 ist – bei minimaler Vorinformation zu  $A$  – die A posteriori-Verteilung von  $A$

**Abb. 8.2** Mit Laserimpulsen gebohrte Löcher (aus [5])



**Abb. 8.3** Plausibilität zum Anteil  $A$ , bestimmt aus der Umfrage



eine Beta-Verteilung mit Kennzahlen 5 und 17:

$$\text{pdf}(A \mid \text{Daten, min. Vorinformation}) \propto A^4 \cdot (1 - A)^{16} \cdot 1$$

Der Graph der Verteilung findet sich in Abb. 8.3. Der plausibelste Wert für  $A$  ist  $A_0 = 4/20 = 0,2$ . Die Fünf-Zahlen-Zusammenfassung der Plausibilität zu  $A$  lautet:

	Minimum	$q_{0,25}$	Median	$q_{0,75}$	Maximum
Anteil A	0	0,163	0,219	0,283	1

Für die Firma ist es auch wichtig zu wissen, wie hoch die Gesamtkosten  $K$  für die Produktion des neuen Produkts sein werden. Diese sind proportional zur Wurzel der produzierten Geräte (oder dem Anteil der Kunden, die das Gerät kaufen werden):

$$K = 2 \cdot \sqrt{A}$$

Was ist der plausibelste Wert von  $K$ ? Wie pflanzt sich die Plausibilität zu  $A$  auf diejenige zu  $K$  fort?  $\square$

Die folgenden Rechnungen zeigen, wie die Plausibilität zu einer Grösse auf eine andere Grösse umgerechnet werden kann. Eine Person beschreibt, bei gegebener Information  $I$ , ihr Wissen zu einer Grösse  $X$  mit einem Wahrscheinlichkeitsmodell. Die Person betrachtet eine zweite Grösse  $Y$ , die funktionell von  $X$  abhängt:  $Y = g(X)$ . Beispielsweise kann dies  $Y = X^3$  oder  $Y$  der Kehrwert  $1/X$  sein. Dabei hängen  $X$  und  $Y$  in eindeutiger Weise voneinander ab:  $X$  kann man aus  $Y$  und  $Y$  aus  $X$  berechnen. Damit liegt  $X$  zwischen zwei Zahlen  $a$  und  $b$ , genau dann, wenn  $Y = g(X)$  zwischen  $g(a)$  und  $g(b)$  liegt. Ist dabei  $g$  monoton wachsend – d. h. für  $a < b$  ist auch  $g(a) < g(b)$  –, so ist

$$\mathbb{P}(a \leq X \leq b) = \mathbb{P}(g(a) \leq Y \leq g(b))$$

Ist andererseits die Transformation  $g$  monoton fallend – d. h. für  $a < b$  ist  $g(a) > g(b)$  –, hat man

$$\mathbb{P}(a \leq X \leq b) = \mathbb{P}(g(b) \leq Y \leq g(a))$$

Deshalb ist:

**Theorem 8.1 (Übertragen von Wahrscheinlichkeitsintervallen)**

Wahrscheinlichkeitsintervalle (und damit Fünf-Zahlen-Zusammenfassungen) lassen sich bei eindeutigen Transformationen schnell übertragen. Man transformiert einfach die entsprechenden Zahlen und ordnet die Zahlen der Grösse nach um.

**Beispiel 8.3 (Kunden und Kosten)** Beim Beispiel 8.2 zu den Kunden und Kosten hat man das folgende Wissen zu den Kosten  $K = 2 \cdot \sqrt{A}$ :

	Minimum	$q_{0.25}$	Median	$q_{0.75}$	Maximum
Anteil $A$	0	0,163	0,219	0,283	1
Kosten $K$	$0 (= 2 \cdot \sqrt{0})$	$0,807 (= 2 \cdot \sqrt{0,163})$	0,936	1,064	2

Es besteht eine Wahrscheinlichkeit von 0,9, dass  $A$  zwischen 0,10 und 0,38 ist. Mit einer Wahrscheinlichkeit von 0,9 ist deshalb  $K$  zwischen  $2 \cdot \sqrt{0,10} = 0,63$  und  $2 \cdot \sqrt{0,38} = 1,23$ .

Je mehr Geräte des neuen Typs verkauft werden, um so kleiner sind die Kosten  $L$  für Werbeaktionen des Geräts. Es gelte  $L = 3/A$ . Diese Transformation ist monoton fallend. Überträgt man die Fünf-Zahlen-Zusammenfassung von  $A$  auf  $L$ , wird aus dem Minimum von  $A$  das Maximum von  $L$ , dem 0,25-Quantil von  $A$  das 0,75-Quantil von  $L$ . Man erhält:

	Minimum	0,25-Quantil	Median	0,75-Quantil	Maximum
Kosten $L$	$3 (= 3/1)$	$10,60 (= 3/0,283)$	13,70	18,40	$3/0 = \infty$

Es besteht eine Wahrscheinlichkeit von 0,9, dass die Werbekosten  $L$  zwischen  $3/0,18 = 16,7$  und  $3/0,10 = 30$  sind.  $\square$

Wie lautet die Dichtefunktion, um die Plausibilität zur transformierten Grösse  $Y = g(X)$  zu charakterisieren? Man hat für kleine  $\Delta x$

$$\mathbb{P}(x \leq X \leq x + \Delta x \mid \mathcal{I}) = \int_x^{x+\Delta x} \text{pdf}_X(x \mid \mathcal{I}) dx \approx \text{pdf}_X(x \mid \mathcal{I}) \cdot |\Delta x|$$

Dabei ist  $\text{pdf}_X(x \mid \mathcal{I})$  die Dichtefunktion des Wahrscheinlichkeitsmodells für die Grösse  $X$ . Wenn  $X$  zwischen  $x$  und  $x + \Delta x$  liegt, so ist  $Y$  zwischen  $y = g(x)$  und  $y + \Delta y$ . Also ist:

$$\text{pdf}_X(x \mid \mathcal{I}) \cdot |\Delta x| \approx \mathbb{P}(Y \text{ zwischen } y \text{ und } y + \Delta y \mid \mathcal{I}) \approx \text{pdf}_Y(y \mid \mathcal{I}) \cdot |\Delta y|$$

Hier ist  $\text{pdf}_Y(y \mid \mathcal{I})$  die Dichtefunktion zum Wahrscheinlichkeitsmodell, das die Grösse  $Y$  beschreibt. Mit anderen Worten ist

$$\text{pdf}_Y(y \mid \mathcal{I}) \cdot |\Delta y| \approx \text{pdf}_X(x \mid \mathcal{I}) \cdot |\Delta x|$$

Da  $x$  und  $y$  über  $y = g(x)$  zusammenhängen, sind auch  $\Delta x$  und  $\Delta y$  miteinander verknüpft. Für kleine  $\Delta y$  und  $\Delta x$  ist der Quotient der beiden Ausdrücke gleich der Ableitung von  $g$ :

$$\frac{\Delta y}{\Delta x} \approx \frac{dg}{dx} = g'(x)$$

Daraus folgt:

**Theorem 8.2 (Transformation von stetigen Wahrscheinlichkeitsmodellen)**

Die Plausibilität zu einer Grösse  $Y$  sei durch das stetige Wahrscheinlichkeitsmodell mit Dichtefunktion  $pdf_Y(y)$  gegeben. Hängt eine zweite Grösse  $X$  eindeutig von  $Y$  durch eine Gleichung  $Y = g(X)$  ab, so berechnet man ihre Dichtefunktion wie folgt: Man substituiert im Produkt  $pdf_Y(y) \cdot dy$  die Variable  $y$  durch die Transformation  $g(x)$  und  $dy$  durch das Differenzial  $dy = g'(x) \cdot dx$ . Der erhaltene Faktor ohne  $dx$  ist dann – bis auf das Vorzeichen – die Dichtefunktion  $pdf_X(x)$  von  $X$ .

**Beispiel 8.4 (Kunden und Kosten)** Beim obigen Beispiel zu den Kunden und Kosten interessiert der Anteil  $A$  der Personen aus dem Kundenkreis, die das neue Produkt kaufen wollen. Die A posteriori-Verteilung des Anteils ist

$$pdf_A(A \mid \text{Daten, min. Vorinformation}) = 101\,745 \cdot A^4 \cdot (1 - A)^{16} \cdot 1$$

Die Konstante 101 745 ist so gewählt, dass die Fläche unter dem Graphen der Dichtefunktion eins wird. Der Modus ist  $4/20 = 0,2$ . Die Kosten und der Anteil hängen in eindeutiger Weise zusammen:  $K = 2 \cdot \sqrt{A}$  und  $A = 0,25 \cdot K^2$ . Um die A posteriori-Verteilung von  $K$  zu bestimmen, müssen im Produkt

$$pdf_A(A \mid \text{Daten, min. Vorinformation}) \cdot dA = 101745 \cdot A^4 \cdot (1 - A)^{16} \cdot dA$$

alle Ausdrücke mit  $A$  durch  $K$  substituiert werden. Einerseits ist  $A = 0,25 \cdot K^2$ . Andererseits ist

$$dA = A'(K) \cdot dK = 0,5 \cdot K \cdot dK$$

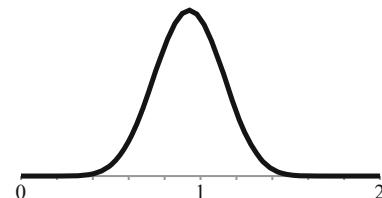
Diese Terme setzt man in die Dichtefunktion für  $A$  ein. Deshalb erhält man

$$pdf_A(A \mid \text{Daten, min. Vor.}) \cdot dA = 101754 \cdot (0,25 \cdot K^2)^4 \cdot (1 - 0,25 \cdot K^2)^{16} \cdot 0,5 \cdot K \cdot dK$$

Die A posteriori-Verteilung von  $K$  lautet damit:

$$pdf_K(K \mid \text{Daten, min. Vorinformation}) = 198,7207 \cdot K^9 \cdot (1 - 0,25 \cdot K^2)^{16}$$

**Abb. 8.4** Plausibilität zu den Kosten  $K$ , ermittelt aus einer Umfrage



**Abb. 8.5** Flacher Prior zu einem Lageparameter

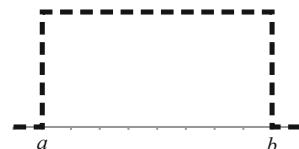


Abb. 8.4 visualisiert den Graphen der Dichtefunktion. Der plausibelste Wert für  $K$  ist der Modus  $K_0 = 0,937$ . Er ist nicht gleich dem Wert  $2 \cdot \sqrt{0,2} = 0,894$  (dem plausibelsten Wert von  $A$  in die Gleichung für  $K$  eingesetzt).  $\square$

Der Schluss dieses Abschnitts beschäftigt sich mit der Frage, was minimale Vorinformation zu Parametern bedeutet. Wie soll man die A priori-Verteilung zu einer nicht direkt messbaren Grösse wählen, wenn keine Information zu ihr vorhanden ist? Dazu ist es sinnvoll, zwischen *Lageparametern* (engl. *location parameter*) und *Skalierungsparametern* (engl. *scale parameter*) zu unterscheiden.

Ein Lageparameter  $\alpha$  kann ein gesuchter Zeitpunkt oder eine gesuchte Koordinate sein. Er liege zwischen zwei Zahlen  $a$  und  $b$ . Weitere Information habe man nicht. Ist  $z$  eine beliebige Zahl, so bedeutet dies, dass es keine Rolle spielen sollte, ob  $\alpha$  oder  $\beta = \alpha + z$  plausibler ist. Aus dem Transformationsgesetz ist wegen  $d\beta = d\alpha$  deshalb

$$\text{pdf}(\alpha \mid \text{min. Vor.}) \cdot d\alpha = \text{pdf}_\beta(\beta \mid \text{min. Vor.}) \cdot d\beta = \text{pdf}(\alpha + z \mid \text{min. Vor.}) \cdot d\alpha$$

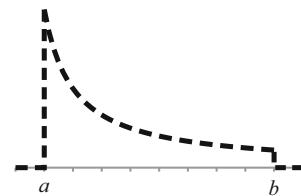
Dies heisst, dass  $\text{pdf}(\alpha \mid \text{min. Vor.}) = \text{pdf}(\alpha + z \mid \text{min. Vor.})$  ist. Die Dichtefunktion muss also konstant sein. Man schreibt  $\text{pdf}(\alpha \mid \text{min. Vorinformation}) \propto 1$ . Abb. 8.5 zeigt ihren Graphen: eine Gleichverteilung.<sup>1</sup>

Etwas anders verhält es sich, wenn man ausdrücken will, wo ein positiver Parameter  $\alpha$ , wie eine durchschnittliche Wartezeit, sei es die vor einem Schalter oder die zwischen starken Erdbeben, liegt. Ist diese durchschnittliche Wartezeit vor einem Schalter 4, 8 oder 20 Minuten? Ist sie so sogar doppelt so gross? Oder ist sie 4, 8 oder 20 Sekunden? Hat man keine weitere Information dazu, so dürfte es nicht relevant sein, ob die Wartezeit ein paar Sekunden oder ein paar Minuten ist. Mit anderen Worten:  $\alpha$  oder  $\beta = z \cdot \alpha$  sind gleich plausibel. Deshalb ist

$$\text{pdf}(\alpha \mid \text{min. Vor.}) \cdot d\alpha = \text{pdf}_\beta(\beta \mid \text{min. Vor.}) \cdot d\beta = \text{pdf}(z \cdot \alpha \mid \text{min. Vor.}) \cdot z \cdot d\alpha$$

<sup>1</sup> Man spricht auch von einer flachen Verteilung.

**Abb. 8.6** Jeffreys' Prior zu einem Skalierungsparameter



Also ist  $\text{pdf}(\alpha \mid \text{min. Vor.}) = \text{pdf}(z \cdot \alpha \mid \text{min. Vor.}) \cdot z$ . Setzt man  $\alpha = 1$ , ergibt sich  $\text{pdf}(z \mid \text{min. Vor.}) = \text{pdf}(1 \mid \text{min. Vor.})/z \propto 1/z$ . Abb. 8.6 zeigt den Graphen dieser Dichtefunktion. Man nennt diese Verteilung nach [3] auch *Jeffreys' A priori-Verteilung* (engl. *Jeffreys's prior*).

Damit hat man zusammengefasst:

**Theorem 8.3 (Minimale Vorinformation oder Indifferenz zu Parametern)**

Ist  $\mu$  ein Parameter in einem bestimmten endlichen Bereich, so beschreibt

- (1) die Dichtefunktion  $\text{pdf}(\mu) \propto 1$  minimale Vorinformation zu  $\mu$  bei einem Lageparameter.
- (2) die Dichtefunktion  $\text{pdf}(\mu) \propto 1/\mu$  minimale Vorinformation zu  $\mu$  bei einem Skalierungsparameter.

Die Verteilung mit Dichtefunktion  $1/\mu$  für  $\mu$  bedeutet, dass ihr Logarithmus  $z = \log \mu$  einer Gleichverteilung folgt. In der Tat ist  $\mu = \exp(z)$  und  $d\mu = \exp(z) \cdot dz$  und somit

$$\frac{1}{\mu} \cdot d\mu = \frac{1}{\exp(z)} \cdot \exp(z) \cdot dz = 1 \cdot dz$$

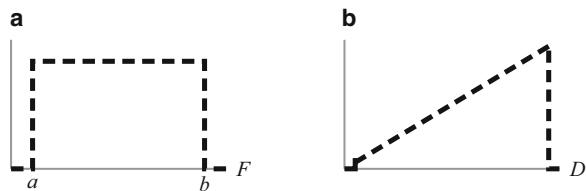
Minimale Vorinformation zu Skalierungsparametern, wie durchschnittlichen Wartezeiten, Längen, Flächen und Volumen oder Amplituden von Schwingungen heisst also, dass alle Werte der *logarithmierten* Grösse gleich plausibel sind. So sind logarithmische Skalen zu solchen Parametern, wie Dezibel für Schallintensitäten und Magnituden für Erdbebenstärken, auch weit verbreitet.

**Beispiel 8.5 (Bohrlöcher mit Laserimpulsen)** Beim Beispiel 8.1 suchen die Ingenieure die Fläche  $F$  und den Durchmesser  $D$  eines Bohrlochs. Jemand modelliert sein Wissen zur Fläche  $F$  vor den Messungen als Lageparameter:

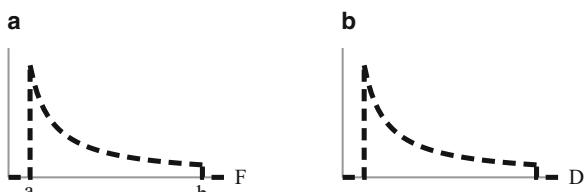
$$\text{pdf}_F(F \mid \text{min. Vorinformation}) \propto 1 \quad \text{für } a \leq F \leq b$$

Die beiden Zahlen  $F$  und  $D$  hängen in eindeutiger Weise voneinander ab. So ist  $F = 0,25\pi \cdot D^2$  und  $D = 2 \cdot \sqrt{F/\pi}$ . Um die Dichtefunktion der Verteilung von  $D$  zu berechnen,

**Abb. 8.7** Minimale Vorinformation zu  $F$  (a) und zu  $D$  (b), wenn  $F$  als Lageparameter betrachtet wird



**Abb. 8.8** Minimale Vorinformation zu  $F$  (a) und zu  $D$  (b), wenn  $F$  als Skalierungsparameter betrachtet wird



muss man in

$$\text{pdf}_F(F \mid \text{min. Vorinformation}) \cdot dF \propto 1 \cdot dF$$

den Ausdruck  $dF$  ersetzen. Man hat  $dF = F'(D) \cdot dD = 0,25\pi \cdot 2 \cdot D \cdot dD$ . Daher ist

$$1 \cdot dF \propto 0,25\pi \cdot 2 \cdot D \cdot dD$$

Die Dichtefunktion von  $D$  ist damit  $\text{pdf}_D(D \mid \text{min. Vorinformation}) \propto D$ . Abb. 8.7 zeigt das Resultat. Größere Werte des Durchmessers  $D$  sind also plausibler als kleine Werte von  $D$ ! Dies kann störend wirken.

Sinnvoller ist es, die Fläche  $F$  als skalierbare Größe zu betrachten. Dann ist

$$\text{pdf}_F(F \mid \text{min. Vorinformation}) \propto 1/F \quad \text{für } a \leq F \leq b$$

Die Dichtefunktion, um die Plausibilität zu  $D$  zu beschreiben, ist

$$\frac{1}{F} \cdot dF \propto \frac{1}{0,25\pi \cdot D^2} \cdot 0,25\pi \cdot 2 \cdot D \cdot dD \propto \frac{1}{D} \cdot dD$$

Abb. 8.8 zeigt die erhaltenen Dichtefunktionen. Die Plausibilität zu den beiden Größen  $F$  und  $D$  wird so mit dem gleichen Modell beschrieben!<sup>2</sup>  $\square$

### 8.3 Unordnung und relative Entropie

Wie in Kap. 4 erklärt, beschreiben in diesem Buch Wahrscheinlichkeitsmodelle die Plausibilität von Aussagen in Funktion der gegebenen Information. So sagt „Die Wahrscheinlichkeit, dass eine Münze, die man wirft, auf Kopf fallen wird, ist 0,5“, nichts über die

<sup>2</sup> Jeffreys' Prior  $\text{pdf}(x) \propto 1/x$  ist invariant, wenn man  $x$  mit einer Potenz  $y = x^n$  transformiert.

Münze aus. Die Aussage besagt, dass die vorhandene Information nicht erlaubt zu prognostizieren, ob die Münze eher auf Kopf oder auf Zahl fällt. Eine Person, die die Plausibilität zu einer Grösse formuliert, sollte also die zur Verfügung stehende Information in das Modell stecken. Das Prinzip der *maximalen Entropie* (engl. *maximum entropy*) versucht dies zu tun. Entropie ist ein Begriff aus der Physik und misst, welche Unordnung (oder das Gegenteil davon, welche Information) ein System hat. Bekannt ist dieser Begriff auch im Bereich der Bilderkennung: hier lässt sich damit bestimmen, wie informativ ein verrauchtes oder teilzerstörtes Bild ist. Bei Wahrscheinlichkeitsmodellen kann mit der Entropie gemessen werden, wie gross die Unordnung oder Streuung gegenüber dem Modell bei minimaler Vorinformation ist. Man definiert Entropie wie folgt:

Eine Grösse könne nur die Werte  $x_1, x_2, \dots$  annehmen. Besteht minimale Vorinformation, so sei die Wahrscheinlichkeit  $m_i \neq 0$ , dass der Wert  $x_i$  angenommen wird. Bei zusätzlicher Information formuliert eine Person ihr Wissen neu: der Wert  $x_i$  hat Wahrscheinlichkeit  $p_i$ . Die *relative Entropie*  $S$  des neuen Wahrscheinlichkeitsmodells ist dann

$$S = -p_1 \cdot \ln \frac{p_1}{m_1} - p_2 \cdot \ln \frac{p_2}{m_2} - p_3 \cdot \ln \frac{p_3}{m_3} - \dots$$

Dabei wird, wenn ein Summand mit  $p_i = 0$  auftaucht,  $0 \cdot \ln 0 = 0$  gesetzt. Man nennt  $S$  nach [4] auch die *Kullback-Leibler-Information* oder *KL-Divergenz*. Wenn alle  $m_i$ 's gleich sind, so spricht man nach [6] von der *Shannon-Entropie*. Sie lautet:<sup>3</sup>

$$S_{\text{Shannon}} = -p_1 \cdot \ln p_1 - p_2 \cdot \ln p_2 - p_3 \cdot \ln p_3 - \dots$$

Die Shannon-Entropie ist ein Mass für die Streuung eines Wahrscheinlichkeitsmodells. Je grösser die Shannon-Entropie eines Wahrscheinlichkeitsmodells ist, umso „breiter“ ist ihre Massenfunktion, und umso unschärfer werden damit mit dem Modell gemachte Aussagen zu einem Parameter oder zu Werten von Zufallsgrössen.

---

<sup>3</sup> Warum misst man Entropie oder Information mit einer logarithmischen Funktion? Betrachten Sie dazu eine Informationstafel, die aus drei Feldern besteht. Die Felder können mit Farben weiss und schwarz belegt werden, um ein „Signal“ zu senden:



Damit kann eine Person  $N = 2^3$  verschiedene Signale bilden. Mit zwei Informationstafeln können doppelt so viele Signale gesendet werden:

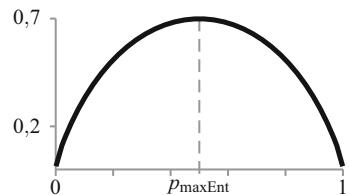


Man kann auf  $M = N \cdot N = 2^3 \cdot 2^3$  mögliche Arten die Felder mit schwarzer oder weisser Farbe belegen. Logarithmiert man diesen Wert, so hat man

$$\ln(N \cdot N) = \ln(N) + \ln(N) = 2 \cdot \ln(N)$$

Der Logarithmus  $\ln(M)$  misst somit, wie viel Information mit der Tafel gesendet werden kann.

**Abb. 8.9** Graph der Entropie  $S(p)$  beim Tagesschlusskurs einer Aktie



Es folgen zwei Beispiele, die illustrieren, wie Entropie berechnet wird und benutzt werden kann, um Wahrscheinlichkeiten zu setzen:

**Beispiel 8.6 (Tagesschlusskurs einer Aktie)** Eine Börsenhändlerin prognostiziert, basierend auf ihrem Wissen  $\mathcal{K}$ , dass der Tagesschlusskurs  $W_{\text{Schluss}}$  einer Aktie mit Wahrscheinlichkeit  $p$  CHF 2,00 oder mit Wahrscheinlichkeit  $1 - p$  CHF 2,20 sein wird. Die Shannon-Entropie  $S$  des Wahrscheinlichkeitsmodells ist

$$S(p) = -p \cdot \ln p - (1 - p) \cdot \ln(1 - p)$$

Der Graph von  $S$  findet sich in Abb. 8.9. Die Entropie ist maximal, wenn  $p = 0,5$ , also wenn das Wissen  $\mathcal{K}$  minimale Vorinformation ist. Man kann das Resultat auch anders interpretieren: Setzt die Börsenhändlerin  $p$  nicht 0,5, so besitzt sie zusätzliche Information zum Börsenschlusskurs.  $\square$

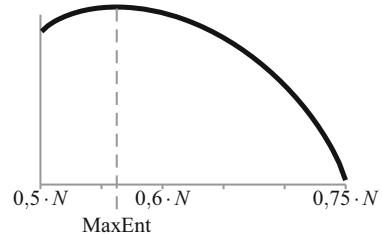
**Beispiel 8.7 (Kängurus)** Dieses Beispiel, von Gull und Skilling in [1] vorgestellt, handelt davon, wie man Wahrscheinlichkeiten zu erfundenen Eigenschaften von Kängurus zuordnen kann. Es zeigt, dass die Entropie hilft, fehlende Information zu ergänzen. Gegeben sei die folgende Information  $\mathcal{I}$ :

Drei Viertel aller Kängurus in Australien trinken Foster's und drei Viertel der Kängurus sind linkshändig.

Daraus soll die Plausibilität  $p$  berechnet werden, dass ein Känguru Foster's trinkt und linkshändig ist. Es sind vier Aussagen zu den Kängurus möglich: (1) Ist linkshändig und trinkt Foster's, (2) ist linkshändig und trinkt kein Foster's, (3) ist rechtshändig und trinkt Foster's und (4) ist rechtshändig und trinkt kein Foster's. Bei minimaler Vorinformation ist es sinnvoll, alle vier Aussagen mit einer Wahrscheinlichkeit von  $m_1 = m_2 = m_3 = m_4 = 0,25$  zu belegen. Geht man davon aus, dass  $N$  Kängurus in Australien leben, so können aus der Information  $\mathcal{I}$  die unterste Zeile und die rechte Spalte des folgenden Schemas berechnet werden:

	linkshändig	rechtshändig	Total
Foster's			$0,75 \cdot N$
Foster's <sup>nicht</sup>			$0,25 \cdot N$
Total	$0,75 \cdot N$	$0,25 \cdot N$	$N$

**Abb. 8.10** Entropie in Funktion von  $n$  beim Beispiel zu den Kängurus



Weitere Zahlen lassen sich in der Tabelle in Funktion einer einzigen (unbekannten) Zahl  $n$  bestimmen:

	linkshändig	rechtshändig	Total
Foster's	$n$	$0,75 \cdot N - n$	$0,75 \cdot N$
Foster's nicht	$0,75 \cdot N - n$	$n - 0,5 \cdot N$	$0,25 \cdot N$
Total	$0,75 \cdot N$	$0,25 \cdot N$	$N$

Wie soll man aber  $n$  wählen? Sicher ist  $n$  zwischen  $0,5 \cdot N$  und  $0,75 \cdot N$ , damit die Zahlen in der Tabelle nicht negativ werden. Man kann  $n$  so wählen, dass die Entropie des Wahrscheinlichkeitsmodells maximal wird. Setzt man

$$p_1 = \frac{n}{N}, \quad p_2 = p_3 = \frac{0,75 \cdot N - n}{N}, \quad p_4 = \frac{n - 0,5 \cdot N}{N}$$

so lautet die relative Entropie  $S$

$$S = -p_1 \cdot \ln \frac{p_1}{0,25} - p_2 \cdot \ln \frac{p_2}{0,25} - p_3 \cdot \ln \frac{p_3}{0,25} - p_4 \cdot \ln \frac{p_4}{0,25}$$

Abb. 8.10 zeigt den Graphen der Entropie in Funktion von  $n$ . Das Maximum befindet sich bei

$$n = 0,75^2 \cdot N = 0,5625 \cdot N$$

Dies bedeutet, dass die gesuchte Wahrscheinlichkeit  $p$ , dass ein Känguru linkshändig ist und Foster's trinkt, aus der vorhandenen Information gleich  $0,75^2 = 0,5625$  gesetzt wird. Das Resultat besagt, dass die Aussagen „Känguru ist linkshändig“ und „Känguru trinkt Foster's“ aus der vorgegebenen Information unabhängig modelliert werden:

$$\begin{aligned} \mathbb{P}(\text{Foster's und linkshändig} \mid \mathcal{I}) &= \mathbb{P}(\text{Foster's} \mid \mathcal{I}) \cdot \mathbb{P}(\text{linkshändig} \mid \mathcal{I}) \\ &= 0,75 \cdot 0,75 = 0,5625 \end{aligned}$$

Dies illustriert, wie sinnvoll das Prinzip der maximalen Entropie Wahrscheinlichkeiten setzt. Beschreibt man die gesuchte Wahrscheinlichkeit  $p$  mit einem anderen Wert als 0,5625, ist zusätzliche Information vorhanden. Diese besagt, dass linkshändig und

Foster's trinken voneinander abhängig sind. Beispielsweise, warum linkshändige Kängurus eher Foster's trinken als rechtshändige Kängurus.<sup>4</sup> □

Die beiden vorgestellten Beispiele zeigen, wie aus der vorhandenen Information mit einem automatischen Verfahren Wahrscheinlichkeitsmodelle konstruiert werden können:

**Theorem 8.4 (Prinzip der maximalen Entropie (MaxEnt))**

*Man will die Plausibilität zu Grösse mit einem Wahrscheinlichkeitsmodell beschreiben. Führt die vorhandene Information dazu, dass das Modell nicht eindeutig wählbar ist, wählt man dasjenige aus, das die grösste (relative) Entropie hat.*

Weiss man etwa bei gegebener Information  $\mathcal{I}$  nur, dass ein Parameter nur  $N$  endliche Wert  $x_i$  annehmen kann, so können die Wahrscheinlichkeiten  $p_i$  für  $x_i$  mit MaxEnt besetzt werden. Man kann zeigen, dass dann alle  $p_i$  gleich gross werden:  $\mathbb{P}(\text{Parameter} = x_i \mid \mathcal{I}) = p_i = 1/N$ . Würde eine Person einen der Werte plausibler als andere Werte setzen, müsste sie ihre zusätzliche Information deklarieren. Dies kann Zusatzwissen aus Daten sein.

Für ein stetiges Wahrscheinlichkeitsmodell definiert man die relative Entropie  $S$  mit der Dichtefunktion  $\text{pdf}(x)$ :

$$S = - \int_{-\infty}^{\infty} \text{pdf}(x) \cdot \ln \left( \frac{\text{pdf}(x)}{m(x)} \right) dx$$

Wiederum ist  $m(x)$  die Dichtefunktion, die minimale Vorinformation widerspiegelt. Ist  $m(x) \propto 1$ , so spricht man von der Shannon-Entropie.

**Beispiel 8.8 (Siebanalyse von Teilchen)** Die Plausibilität zur Korngrösse  $G$  von Teilchen, wie Kiessteinchen oder von gemahlenem Zucker, wird meistens mit einer Weibull-Verteilung beschrieben (siehe das Beispiel 4.10). Die Dichtefunktion des stetigen Wahrscheinlichkeitsmodells ist:

$$\text{pdf}(G = x \mid d, k) = \frac{k}{d} \cdot (x/d)^{k-1} \cdot \exp(-[x/d]^k) \quad \text{für } x \geq 0$$

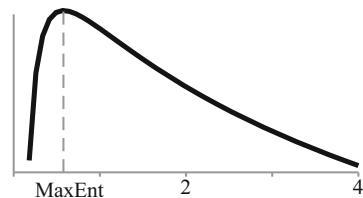
Dabei sind die Form  $k$  und die Skalierung  $d$  positive Konstanten. Die Skalierung  $d$  gibt bei Filtern die Maschenweite an, durch die 63,2 % der Teilchen fallen. Eine Ingenieurin weiss aus einem Experiment, dass bei den untersuchten Teilchen  $d = 1 \text{ cm}$  ist. Damit hat sie:

$$\text{pdf}(G = x \mid k, d = 1 \text{ cm}) = k \cdot x^{k-1} \cdot \exp(-x^k) \quad \text{für } x \geq 0$$

---

<sup>4</sup> Eine ausführliche und weiterführende Diskussion findet sich in [2].

**Abb. 8.11** Graph der Entropie  $S(k)$



Wie soll sie den Parameter  $k$  wählen, wenn sie keine zusätzliche Information zu den Teilchengrößen hat? Mit dem Prinzip der maximalen Entropie kann sie  $k$  so setzen, dass das Wahrscheinlichkeitsmodell eine maximale Shannon-Entropie hat. Die Shannon-Entropie  $S = S(k)$  lautet

$$S(k) = - \int_0^{\infty} k \cdot x^{k-1} \cdot \exp(-x^k) \cdot \ln [k \cdot x^{k-1} \cdot \exp(-x^k)] \, dx$$

Das Integral lässt sich vereinfachen. Eine Rechnung ergibt

$$S(k) = -\ln k + \left(1 - \frac{1}{k}\right) \cdot \gamma + 1$$

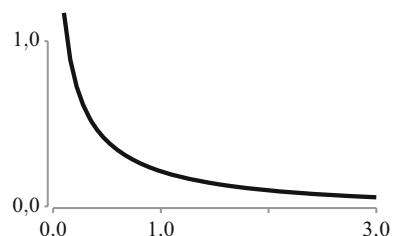
Dabei ist  $\gamma = 0,5772$ . Der Graph der Shannon-Entropie ist in Abb. 8.11 dargestellt. Die Entropie – und damit die „Unordnung“ des Wahrscheinlichkeitsmodells – wird maximal, wenn  $k = \gamma = 0,5772$  ist. Nach MaxEnt ist das Modell, um Teilchengrößen zu beschreiben, deshalb:

$$\text{pdf}(G = x \mid \text{Information}) = \gamma \cdot x^{\gamma-1} \cdot \exp(-x^\gamma) \quad \text{für } x \geq 0$$

Der Graph des Modells findet sich in Abb. 8.12. Die Verteilung ist schief.  $\square$

Wahrscheinlichkeitsmodelle lassen sich konstruieren, wenn man das Transformationsgesetz oder das Prinzip der maximalen Entropie anwendet. Hat man minimale Vorinformation zu einem Lageparameter, der zwischen zwei Werten liegt, so folgt aus dem Transformationsgesetz, dass die Plausibilität zu diesem mit einer Gleichverteilung modelliert werden sollte. Zum gleichen Resultat führt MaxEnt. Unter allen stetigen Verteilungen, mit Werten in einem endlichen Bereich, hat die Gleichverteilung die grösste Entropie.<sup>5</sup>

**Abb. 8.12** Weibullverteilung mit  $d = 1$  und maximaler Entropie ( $k = 0,5772$ ), um die Teilchengröße zu beschreiben



<sup>5</sup> Einen Beweis findet man in [7].

## 8.4 Der Erwartungswert als Information

Die Breite (oder die Entropie) eines Wahrscheinlichkeitsmodells ist wichtig. Aber auch die „Lage“ des Modells spielt oft eine Rolle. So nimmt man im Ingenieurwesen an, dass Messwerte um einen „mittleren Wert“ variieren. Ein solcher mittlerer Wert meint den durchschnittlich erwartbaren Wert oder *Erwartungswert* (engl. *expected value*) des Datenmodells. Er ist wie folgt definiert:

### Definition 8.1

Ein Wahrscheinlichkeitsmodell, um die Plausibilität zu einer Grösse  $G$  zu beschreiben, sei diskret. Die Grösse  $G$  kann dabei die Werte  $x_1, x_2, \dots$  mit Wahrscheinlichkeiten  $\mathbb{P}(G = x_1), \mathbb{P}(G = x_2), \dots$  annehmen. Der Erwartungswert  $\mu$  des Modells ist das gewichtete Mittel dieser Werte

$$\mu = x_1 \cdot \mathbb{P}(G = x_1) + x_2 \cdot \mathbb{P}(G = x_2) + \dots$$

Bei einem stetigen Wahrscheinlichkeitsmodell mit Dichtefunktion  $\text{pdf}(x)$  erhält man den Erwartungswert, indem man  $x \cdot \mathbb{P}(G \approx x)$  aufsummiert. Mit der Gleichung (4.1) erhält man ein Integral:

$$\mu = \sum x \cdot \mathbb{P}(G \approx x) = \sum x \cdot \text{pdf}(x) \cdot \Delta x = \int_{-\infty}^{\infty} x \cdot \text{pdf}(x) \, dx$$

**Beispiel 8.9 (Tagesschlusskurs einer Aktie)** Beim Beispiel 4.6 zum Tagesschlusskurs  $W_{\text{Schluss}}$  einer Aktie weiss eine Börsenhändlerin, dass  $W_{\text{Schluss}}$  nur die Werte CHF 2.–, CHF 6.– und CHF 10.– annehmen kann. Genauer beschreibt sie die Plausibilität dazu mit einem diskreten Wahrscheinlichkeitsmodell:

$W_{\text{Schluss}}$	2,00	6,00	10,00
Wahrscheinlichkeit	0,15	0,40	0,45

Der Erwartungswert  $\mu$  des Wahrscheinlichkeitsmodells lautet

$$\begin{aligned} \mu &= 2 \cdot \mathbb{P}(W_{\text{Schluss}} = 2) + 6 \cdot \mathbb{P}(W_{\text{Schluss}} = 6) + 10 \cdot \mathbb{P}(W_{\text{Schluss}} = 10) \\ &= 2 \cdot 0,15 + 6 \cdot 0,40 + 10 \cdot 0,45 = 7,20 \text{ CHF} \end{aligned}$$

Man nennt diesen Wert auch den „fairen“ Preis der Aktie bei gegebenem Wahrscheinlichkeitsmodell. □

**Beispiel 8.10 (Zerfallszeit von Radon)** Die Dichtefunktion für die Zerfallszeit  $T$  von  $^{225}\text{Radon}$ , die in Beispiel 4.9 vorgestellt ist, lautet

$$\text{pdf}(T = x \mid \lambda) = \lambda \cdot \exp(-\lambda \cdot x) \quad \text{für } x \geq 0$$

Dabei ist  $\lambda = 1/5,515 \text{ Tage}^{-1}$  die Rate der Exponentialverteilung. Der Erwartungswert  $\mu$  des Modells – die durchschnittlich erwartbare Zerfallszeit – ist:

$$\mu = \int_{-\infty}^{\infty} x \cdot \text{pdf}(T = x \mid \lambda) dx = \int_0^{\infty} x \cdot \lambda \cdot \exp(-\lambda \cdot x) dx = \frac{1}{\lambda}$$

Der Erwartungswert der Exponentialverteilung ist also der Kehrwert der Rate  $\lambda$ . Die im Schnitt erwartbare Zerfallszeit für  $^{225}\text{Rn}$  ist mit diesem Modell daher 5,515 Tage. Er ist deutlich höher als der Median, der 3,823 Tage beträgt.  $\square$

**Beispiel 8.11 (Siebanalyse von Teilchen)** Eine Person beschreibt ihr Wissen zur Korngrösse  $G$  von gemahlenen Teilchen mit einer Weibull-Verteilung mit Parametern  $d = 1 \text{ cm}$  und  $k = 1,5$ :

$$\text{pdf}(G = x) = \frac{1,5}{1 \text{ cm}} \cdot (x/(1 \text{ cm}))^{0,5} \cdot \exp(-[x/(1 \text{ cm})]^{1,5}) \quad \text{für } x \geq 0$$

Der Erwartungswert  $\mu$  des Modells – also die durchschnittlich erwartbare Korngrösse – lautet

$$\mu = \int_0^{\infty} x \cdot \frac{1,5}{1 \text{ cm}} \cdot (x/(1 \text{ cm}))^{0,5} \cdot \exp(-[x/(1 \text{ cm})]^{1,5}) dx = 0,903 \text{ cm}$$

Mit dem Wahrscheinlichkeitsmodell sagt die Person, dass die durchschnittliche Korngrösse 0,903 cm ist. Diese unterscheidet sich vom Modus 0,481 cm – der „wahrscheinlichsten“ Korngrösse –, da die Verteilung schief ist.  $\square$

Obwohl der Erwartungswert als durchschnittlicher Wert eines Wahrscheinlichkeitsmodells gut interpretierbar ist, besitzt er auch Nachteile. So gibt es Modelle, die *keinen endlichen Erwartungswert* haben!<sup>6</sup> Die Information, dass bei einem Wahrscheinlichkeitsmodell der Erwartungswert vorgegeben ist, kann verarbeitet werden. Hier ein Beispiel:

**Beispiel 8.12 (Tagesschlusskurs einer Aktie)** Ein Börsenhändler prognostiziert, basierend auf seiner Information  $\mathcal{K}$ , dass der Tagesschlusskurs einer Aktie mit Wahrscheinlichkeit  $p$  CHF 2.– und mit Wahrscheinlichkeit  $1 - p$  CHF 6.– sein wird. Er weiss weiter,

---

<sup>6</sup> Modelle ohne endlichen Erwartungswert haben die Eigenschaft, dass sehr grosse Werte mit hoher Wahrscheinlichkeit prognostiziert werden. Es gibt Grössen, die mit einem solchen Modell beschrieben werden.

dass der durchschnittlich erwartbare Preis der Aktie CHF 5.– ist. Es gilt also

$$2 \cdot p + 6 \cdot (1 - p) = 5$$

Daraus folgt, dass der Börsenhändler  $p = 0,25$  setzen sollte.  $\square$

## Reflexion

**8.1** Ein Ökonom formuliert seine Plausibilität zu einem Parameter  $K$ , der kontinuierliche Werte annehmen könnte, mit der Dichtefunktion:

$$\text{pdf}_K(K = x \mid \mathcal{I}) = \frac{1}{\pi} \cdot \frac{1}{1 + (x - 2)^2}$$

- (a) Was ist der plausibelste Wert  $K_0$  von  $K$ ?
- (b) Berechnen Sie Wahrscheinlichkeitsintervalle zum Niveau 0,5 und 0,95 für den Parameter  $K$ .
- (c) Bestimmen aus (b) Wahrscheinlichkeitsintervalle zum Niveau 0,5 und 0,95 für die Grösse  $L = \exp(K)$ .
- (d) Wie lauten die Dichtefunktion, um die Plausibilität zur Grösse  $L$  zu formulieren?
- (e) Bestimmen Sie aus der Rechnung von (d) den „wahrscheinlichsten“ Wert von  $L$ .

**8.2** Jemand formuliert aus seiner vorhandenen Information  $\mathcal{I}$ , wo zukünftige Werte einer unsicheren Grösse  $X \geq 0$  liegen werden, mit einem stetigen Wahrscheinlichkeitsmodell:

$$\text{pdf}_X(X \mid \mathcal{I}) = \frac{3}{(X + 1)^4}$$

- (a) Wie lautet der plausibelste Wert von  $X$ ? Bestimmen Sie die Fünf-Zahlen-Zusammenfassung des Wahrscheinlichkeitsmodells.
- (b) Eine Ingenieurin möchte wissen, wo zukünftige Werte der Grösse  $Y = \sqrt{X}$  liegen. Wie lautet die Fünf-Zahlen-Zusammenfassung zu  $Y$ ? Wie lautet die Dichtefunktion, um  $Y$  zu beschreiben? Was ist der plausibelste Wert von  $Y$ ?
- (c) Lösen Sie die gleichen Aufgaben wie in (b), diesmal für die Grösse  $Z = \ln X$ .
- (d) Beantworten Sie die gleichen Fragen wie in (b) für die transformierte Grösse  $B = 1/(1 + X)$ .

**8.3** Eine Ingenieurin formuliert aus ihrer vorhandenen Information  $\mathcal{I}$ , wie gross die Längen  $L$  von maschinell produzierten Würfeln mit Masse 100 g sind. Sie benutzt dazu die Gleichverteilung

$$\text{pdf}_L(L \mid \mathcal{I}) = \frac{1}{10} \quad \text{für} \quad 10 \text{ cm} \leq L \leq 20 \text{ cm}$$

- (a) Wie lautet der plausibelste Wert von  $L$ ? Bestimmen Sie die Fünf-Zahlen-Zusammenfassung des Wahrscheinlichkeitsmodells.
- (b) Die Ingenieurin möchte Informationen zum Volumen  $V = L^3$  und zur Dichte  $\rho = 100 \text{ g}/L^3$  der Würfel haben. Wie lauten die Fünf-Zahlen-Zusammenfassungen von  $V$  und  $\rho$ ? Wie lauten die Wahrscheinlichkeitsmodelle, um  $V$  und  $\rho$  zu beschreiben? Was sind die plausibelsten Werte von  $V$  und  $\rho$ ?
- (c) Wie gross ist die Wahrscheinlichkeit, dass das Volumen  $V$  eines Würfels zwischen  $2000 \text{ cm}^3$  und  $5000 \text{ cm}^3$  ist?

**8.4** Ein Meteorologe möchte wissen, wie gross der Anteil  $A$  (eine Zahl zwischen null und eins) der korrekten Wettervorhersagen seines Wetterdiensts ist. Aus Daten formuliert er sein Wissen dazu mit einem stetigen Wahrscheinlichkeitsmodell:

$$\text{pdf}_A(A \mid \text{Daten}) = 858 \cdot A^{10} \cdot (1 - A)^2$$

- (a) Zeichnen Sie den Graphen der Dichtefunktion. Was ist der plausibelste Wert  $A_0$  von  $A$ ?
- (b) Berechnen Sie Wahrscheinlichkeitsintervalle zum Niveau 0,5 und 0,95 für den Anteil  $A$ .
- (c) Der Parameter  $A$ , der zwischen null und eins liegt, lässt sich mit der *logit*-Transformation in eine Grösse  $K$  zwischen  $-\infty$  und  $+\infty$  umwandeln:

$$K = \text{logit}(A) = \ln\left(\frac{A}{1 - A}\right)$$

Die Transformation ist eindeutig. So ist  $A = 1/(1 + \exp(-K))$ . Bestimmen Sie aus Aufgabe (b) Wahrscheinlichkeitsintervalle zum Niveau 0,5 und 0,95 für die transformierte Grösse  $K$ .

- (d) Wie lautet die Dichtefunktion, um die Plausibilität zur logit-transformierten Grösse  $K$  zu formulieren? Zeichnen Sie den Graphen dieser Dichtefunktion.
- (e) Bestimmen Sie aus den Rechnungen von (d) den Modus von  $K$ .

**8.5** Eine Person weiss von einer Grösse  $h$ , dass  $h$  nur 1, 2, 3, 4, 5 und 6 sein kann. Sie formuliert die Plausibilität der Werte mit zwei diskreten Wahrscheinlichkeitsmodellen:

	1	2	3	4	5	6
$\mathbb{P}$ , Modell 1	0,05	0,10	0,40	0,30	0,10	0,05
$\mathbb{P}$ , Modell 2	0,10	0,15	0,25	0,25	0,10	0,15

- (a) Zeichnen Sie die Graphen der beiden Massenfunktionen. Was vermuten Sie: Welches Modell besitzt eine grössere Shannon-Entropie?
- (b) Berechnen Sie die Shannon-Entropien der beiden Wahrscheinlichkeitsmodelle.

**8.6** Eine Person formuliert ihr Wissen zu einer positiven Grösse  $X \geq 0$  mit einer Exponentialverteilung:  $X \sim \text{Exponential}(\lambda)$ . Es ist also:

$$\text{pdf}(X = x | \lambda) = \lambda \cdot \exp(-\lambda \cdot x)$$

- (a) Zeichnen Sie die Graphen der Dichtefunktion für  $\lambda = 0,5$ ,  $\lambda = 2$  und  $\lambda = 5$ . Was vermuten Sie: Welches der drei Modelle besitzt die grösste Shannon-Entropie?
- (b) Berechnen Sie die Shannon-Entropie der drei Modelle aus (a).

**8.7** Eine Versicherungsgesellschaft geht davon aus, dass bei ihren Haushaltversicherungen (pro Police) zwischen 0 und 5 Schadensfälle pro Jahr auftreten werden. Die Anzahl der Schadensfälle des nächsten Jahrs modelliert die Versicherungsgesellschaft mit dem diskreten Wahrscheinlichkeitsmodell

Anzahl Schadensfälle	0	1	2	3	4	5
Wahrscheinlichkeit	0,80	0,10	0,05	0,03	0,01	0,01

Zeichnen Sie den Graphen der Massenfunktion des Modells. Berechnen Sie die im Schnitt erwartbare Anzahl  $\mu$  Schadensfälle für das nächste Jahr.

**8.8** Eine Ingenieurin formuliert aus ihrer vorhandenen Information  $\mathcal{I}$ , wo eine positive nicht direkt messbare Grösse  $b \geq 0$  liegt, mit einem stetigen Wahrscheinlichkeitsmodell:

$$\text{pdf}(b = x | \mathcal{I}) = x \cdot \exp(-x)$$

- (a) Zeichnen Sie den Graphen der Dichtefunktion. Wie ist das Modell verteilt?
- (b) Wie lautet der plausibelste Wert von  $b$ ? Bestimmen Sie den Median des Modells. Berechnen Sie den Erwartungswert des Modells.
- (c) Welche der berechneten Kennzahlen – plausibelster Wert, Median oder Erwartungswert – beschreibt den „Schwerpunkt“ der Verteilung am besten?

**8.9** Bei der Firma Roche AG in Sisseln wurde eine Charge von gemahlenem  $\beta$ -Karotin Pulver mit einem Sieb analysiert. Mit der dabei gewonnenen Information  $\mathcal{W}$  wurde die Teilchengrösse mit einer Weibull-Verteilung beschrieben. Die Dichtefunktion für die Teilchengrösse in  $\mu\text{m}$  ist für  $x > 0$ :

$$\text{pdf}(x | \mathcal{W}) = \frac{3,40}{158,53 \mu\text{m}} \cdot \left( \frac{x}{158,53 \mu\text{m}} \right)^{2,40} \cdot \exp \left( - \left( \frac{x}{158,53 \mu\text{m}} \right)^{3,40} \right)$$

- (a) Zeichnen Sie den Graphen der Dichtefunktion. Wie ist das Modell verteilt? Was ist der Median und der plausibelste Wert der Modells?
- (b) Berechnen Sie den Erwartungswert des Modells.

- (c) Was bedeuten hier die Zahlen Median, plausibelster Wert und Erwartungswert?
- (d) Erklären Sie, warum hier die drei Kennzahlen Median, plausibelster Wert und Erwartungswert etwa gleich sind.
- (e) Bestimmen Sie durch systematisches Ausprobieren mit dem Taschenrechner oder mit einer MCMC-Simulation Wahrscheinlichkeitsintervalle zum Niveau 0,5 und 0,95 für die Teilchengrösse.

**8.10** Bei einem Spiel kann man mit einer Wahrscheinlichkeit von  $p$  den Betrag CHF 10.– gewinnen oder mit einer Wahrscheinlichkeit von  $1 - p$  den Betrag CHF 2.– verlieren. Eine Person weiss, dass der Spielbetreiber pro Spiel durchschnittlich CHF 0,10 gewinnt. Wie gross ist die Wahrscheinlichkeit  $p$ ?

**8.11** Bei einem Spiel kann man CHF 5.–, CHF 20.– gewinnen oder CHF 10.– verlieren:

Gewinn in CHF	5,00	20,00	-10,00
Wahrscheinlichkeit	$p$	$q$	$1 - p - q$

Eine Person weiss zudem, dass der Spielbetreiber pro Spiel durchschnittlich CHF 0,50 gewinnt.

- (a) Bestimmen Sie  $q$  in Funktion von  $p$  aus dieser Information.
- (b) Wie lautet die Shannon-Entropie des Wahrscheinlichkeitsmodells in Funktion von  $p$ ? Wie gross sollte die Person die Wahrscheinlichkeit  $p$  setzen, wenn sie mit dem Prinzip der maximalen Entropie arbeitet?

## Literatur

1. S. F. Gull, J. Skilling, Maximum entropy image reconstruction, IEE Proc., 131F, 646–659 (1984)
2. E. T. Jaynes, Monkeys, Kangaroos, and N. Fourth Annual Workshop on Bayesian/Maximum Entropy Methods, University of Calgary (1984). Auch in: Proceedings Volume: Maximum Entropy and Bayesian Methods in Applied Statistics, James H. Justice, Editor (Cambridge University Press, 1986)
3. H. Jeffreys, *Theory of Probability* (Clarendon Press, Oxford, 1939)
4. S. Kullback, R. A. Leibler, On information and sufficiency, Annals of Mathematical Statistics **22**(19), 79–86 (1951)
5. R. Schor, Bohren von Lithium Niobat mit ps-Impulsen, Bachelorarbeit Maschinentechnik, Berner Fachhochschule (2007)
6. C. E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. **27**, 379–423 und 623–656 (1948)
7. D. S. Sivia, J. Skilling, *Data Analysis, A Bayesian Tutorial*, 2nd ed. (Oxford University Press, 2010)

*Diesmal war kein Schildchen daran mit der Anschrift ‚Trink mich‘, aber sie entkorkte es trotzdem und führte es an die Lippen. „Irgend etwas Interessantes passiert ja immer“, sagte sie sich, „sobald ich etwas esse oder trinke; ich will doch einmal sehen, wie diese Flasche hier wirkt. Hoffentlich lässt sie mich wieder grösser werden, denn langsam bin ich es wirklich leid, so winzig klein herumzulaufen!“  
Lewis Carroll, Alice im Wunderland (Insel Taschenbuch, 1973, S. 37.)*

## Zusammenfassung

In vielen Untersuchungen werden nicht negative, kontinuierliche Größen, wie Wartezeiten vor Flugschaltern, Lebensdauern von Geräten, Zerfallszeiten von radioaktiven Elementen oder Schadenssummen bei Unwettern, betrachtet. Außerdem werden auch positive diskrete Größen studiert, wie die Anzahl Unfälle während eines Jahres, wie die monatlichen Schadensfälle bei einer Versicherung, oder die Anzahl Zerfälle von  $\alpha$ -Teilchen während einer Stunde oder die Anzahl Löcher in produzierten porösen Membranen. Wie Parameter von solchen Größen berechnet werden können, wird in diesem Kapitel gezeigt. Die Resultate hängen dabei vom Datenmodell ab, das besagt, wie Messwerte der Größen streuen. Es ist daher sinnvoll, das Datenmodell zu beurteilen. Wie dies gemacht werden kann, wird in diesem Kapitel ebenfalls diskutiert.

## 9.1 Die Exponentialverteilung

Positive Größen, die *kontinuierliche* Werte annehmen können, sind, wie oben schon erwähnt, z. B. Wartezeiten vor Bahn- und Flugschaltern, Wartezeiten zwischen besonderen Ereignissen, Lebensdauern von Beleuchtungskörpern oder von Computerfestplatten, Zerfallszeiten von radioaktiven Elementen oder Schadenssummen bei Unwettern. Wie kann man die vorhandene Information zu solchen Größen mit einem Wahrscheinlichkeitsmo-

dell beschreiben? Man kann als Information annehmen, dass das Wahrscheinlichkeitsmodell einen endlichen Erwartungswert hat. So wird die durchschnittlich erwartbare Wartezeit vor Bahn- und Flugschaltern oder die durchschnittlich erwartbare Lebensdauer der produzierten Beleuchtungskörper endlich sein. Diese Information lässt sich mit dem Prinzip der maximalen Entropie in das Modell einbauen:<sup>1</sup>

**Theorem 9.1 (MaxEnt für die Exponentialverteilung)**

*Die Plausibilität zu einer kontinuierlichen Größe  $X$  soll mit einem stetigen Wahrscheinlichkeitsmodell beschrieben werden. Als Information  $I$  hat man: (1)  $X$  kann nicht negative Werte annehmen und (2) das Modell hat einen endlichen Erwartungswert  $\mu$ . Das Modell, das dies mit der grössten Shannon Entropie umsetzt, ist die Exponentialverteilung mit Rate  $\lambda = 1/\mu$ :*

$$X \sim \text{Exponential}(1/\mu) : \quad \text{pdf}(X = x \mid \mu, I) = \frac{1}{\mu} \cdot \exp(-x/\mu) \quad \text{für } x \geq 0$$

Die Exponentialverteilung ist ein Wahrscheinlichkeitsmodell mit einem Parameter, der Rate  $\lambda$ . Den Graphen der Dichtefunktion der Exponentialverteilung findet man in Abb. 4.4. Hierzu folgt ein Beispiel:

**Beispiel 9.1 (Zeit zwischen starken Erdbeben)** Beim Beispiel 1.3 von Kap. 1 zu den Zeiten zwischen aufeinanderfolgenden Erdbeben mit einer Stärke von mindestens 8 will man

- (a) die durchschnittliche Zeit  $\mu$  zwischen zwei zukünftigen, aufeinanderfolgenden, starken Erdbeben berechnen und
- (b) die (einzelne) Zeit bis zum nächsten starken Erdbeben prognostizieren.

Als Daten hat man die 28 Zeiten zwischen aufeinanderfolgenden Erdbeben (auch Wartezeiten genannt) zwischen dem 1. Januar 1969 und dem 31. Dezember 2007, die in Tab. 1.2 aufgelistet sind.

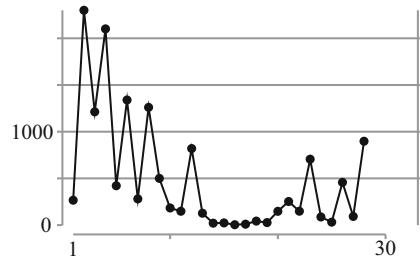
Abb. 9.1 und Abb. 9.2 zeigen das Streudiagramm der Beobachtungen und die Autokorrelationsfunktion. Es sind kaum Tendenzen zu längeren oder kürzeren Wartezeiten sichtbar. Auffallend sind jedoch ein paar ausserordentlich lange Wartezeiten. Die Beobachtungen scheinen ausserdem unabhängig zu sein.<sup>2</sup>

Bei der Frage (a) scheint es „vernünftig“, die durchschnittliche Zeit  $\mu$  zwischen aufeinanderfolgenden, zukünftigen, starken Erdbeben durch das arithmetische Mittel der be-

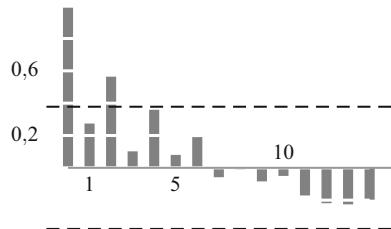
<sup>1</sup> Für einen Beweis, siehe [3]

<sup>2</sup> Dies darf hier durchaus kritisch hinterfragt werden.

**Abb. 9.1** Streudiagramm der 28 Zeiten zwischen aufeinanderfolgenden, starken Erdbeben



**Abb. 9.2** Der Graph der Autokorrelationsfunktion für die 28 Zeiten zwischen aufeinanderfolgenden, starken Erdbeben



obachteten Zeiten zu schätzen. Dieses beträgt

$$\mu = \frac{261,64 + 2300,14 + 1212,66 + \dots + 897,79}{28} \text{ Tage} = 496,81 \text{ Tage}$$

Ist dies wirklich der plausibelste Wert für  $\mu$ ? Wenn ja, wie genau ist das Resultat? Um diese Fragen zu beantworten, braucht man ein Wahrscheinlichkeitsmodell für  $\mu$ . Da  $\mu$  kontinuierliche Werte annehmen kann, ist dies die Dichtefunktion  $\text{pdf}(\mu | \text{Daten})$ . Der plausibelste Wert für  $\mu$  ist dann der Modus dieser Verteilung.

Mit der Regel von Bayes ist die Plausibilität zu  $\mu$  gleich dem Produkt der Likelihood-Funktion (bestimmt aus den Daten) und dem Prior (gegeben durch Vorinformation):

$$\underbrace{\text{pdf}(\mu | \text{Daten, Vorinformation})}_{\text{A posteriori-Verteilung}} \propto \underbrace{\mathbb{P}(\text{Daten} | \mu)}_{\text{Likelihood}} \cdot \underbrace{\text{pdf}(\mu | \text{Vorinformation})}_{\text{Prior}}$$

Die Plausibilität zu  $\mu$  wird also aus der Vorinformation mit Hilfe der Daten aktualisiert. Wir wollen das Folgende annehmen: Man hat nur die Vorinformation, dass  $\mu$  zwischen 100 und 2000 Tagen liegt. Der Parameter  $\mu$  ist ein Skalierungsparameter. Daher kann man  $\mu$  mit der Jeffreys A priori-Verteilung charakterisieren:

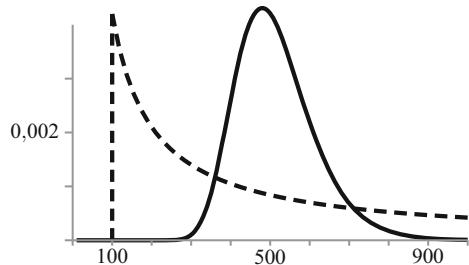
$$\text{pdf}(\mu | \text{Vorinformation}) = \min. \text{Vorinformation} \propto 1/\mu \quad \text{für } 100 \leq \mu \leq 2000$$

Dies bedeutet, dass die Plausibilität für den Logarithmus  $\ln \mu$  mit einer Gleichverteilung formuliert wird:

$$\text{Prior: } \ln \mu \sim \text{Uniform}(\ln 100 ; \ln 2000)$$

Die Likelihood-Funktion  $\mathbb{P}(\text{Daten} | \mu)$  wird aus dem Datenmodell berechnet. Dieses besagt, wie Zeiten zwischen aufeinanderfolgenden Erdbeben streuen. Angenommen wird als

**Abb. 9.3** Plausibilität zu  $\mu$ :  
gestrichelt vor den Daten,  
ausgezogen die A posteriori-Verteilung aus den Daten



einige Information: die Zeiten sind positiv und streuen um den durchschnittlichen endlichen Wert  $\mu$ . Weiter sind Zeiten kontinuierliche, positive Werte. Benutzt man MaxEnt von Theorem 9.1, so wird eine Wartezeit daher mit einer Exponentialverteilung modelliert:

$$\text{Datenmodell: } i\text{-te Wartezeit} \sim \text{Exponential}(1/\mu)$$

Es ist also:

$$\text{pdf}(i\text{-te Wartezeit} = x \mid \mu) = \frac{1}{\mu} \cdot \exp(-x/\mu)$$

Nimmt man an, dass die Beobachtungen unabhängig sind, ist die Likelihood-Funktion nach dem Multiplikationsgesetz ein Produkt solcher Faktoren:

$$\mathbb{P}(\text{Daten} \mid \mu) \propto \frac{1}{\mu} \exp\left(-\frac{261,64}{\mu}\right) \cdot \frac{1}{\mu} \exp\left(-\frac{2300,14}{\mu}\right) \cdots \cdot \frac{1}{\mu} \exp\left(-\frac{897,79}{\mu}\right)$$

Damit sind die Bestandteile der Regel von Bayes bekannt. Es ist

$$\text{pdf}(\mu \mid \text{Daten, min. Vorinformation}) \propto \frac{1}{\mu^{28}} \cdot \exp(-13910,55 \text{ Tage}/\mu) \cdot \frac{1}{\mu}$$

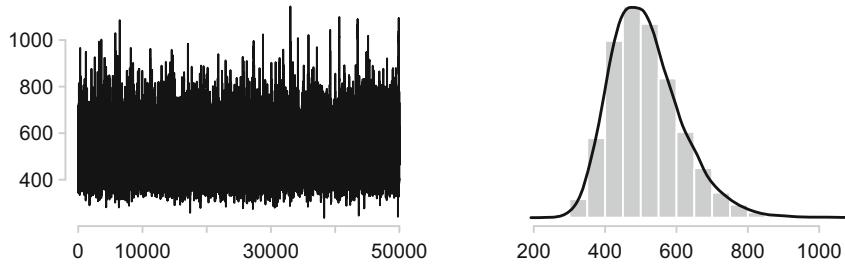
Dies ist die A posteriori-Dichtefunktion von  $\mu$ . Ihr Graph in Abb. 9.3 zeigt eine unimodale schiefe Verteilung. Der plausibelste Wert ist der Modus der A posteriori-Verteilung. Am schnellsten findet man ihn mit einer Wertetabelle und einem Computer. Man kann auch analytisch rechnen, indem man die A posteriori-Dichtefunktion ableitet und null setzt. Man erhält 479,7 Tage. Der plausibelste Wert für  $\mu$  ist nicht gleich dem arithmetischen Mittel von 496,8 Tagen der 28 beobachteten Zeiten. Ein vielleicht überraschendes Resultat.

Wahrscheinlichkeitsintervalle für  $\mu$  lassen sich mit einer MCMC-Simulation bestimmen. Mit Statistikprogrammen erzeugt man die MCMC-Simulation in einfacher Weise. Man gibt die Daten ein und wählt das Datenmodell sowie den Prior für den Parameter  $\mu$ :

$$\text{Datenmodell: } i\text{-te Wartezeit} \sim \text{Exponential}(1/\mu)$$

$$\text{Prior: } \ln \mu \sim \text{Uniform}(\ln 100 ; \ln 2000)$$

Ein Statistikprogramm berechnet daraus mit der Formel Likelihood  $\times$  Prior die A posteriori-Verteilung von  $\mu$  und tastet diese Verteilung mit einer MCMC-Simulation ab. Abb. 9.4



**Abb. 9.4** MCMC-Kette der A posteriori-Verteilung für die durchschnittliche Zeit  $\mu$  zwischen zukünftigen starken Erdbeben

zeigt eine Kette von 50 000 Punkten mit einer Akzeptanzrate von 41 %. 25 % der simulierten Punkte der Kette sind kleiner als 444 Tage. Das 0,25-Quantil der A posteriori-Verteilung von  $\mu$  ist also 444 Tage. Aus der Kette lässt sich analog die Fünf-Zahlen-Zusammenfassung bestimmen, um die Plausibilität zu  $\mu$  zu formulieren:

min	$q_{0,25}$	med	$q_{0,75}$	max
0 Tage	444 Tage	504 Tage	574 Tage	$\infty$

Es besteht eine Wahrscheinlichkeit von 0,5, dass  $\mu$  zwischen 444 und 574 Tagen und eine Wahrscheinlichkeit von 95 %, dass  $\mu$  zwischen 356 und 744 Tagen liegt.

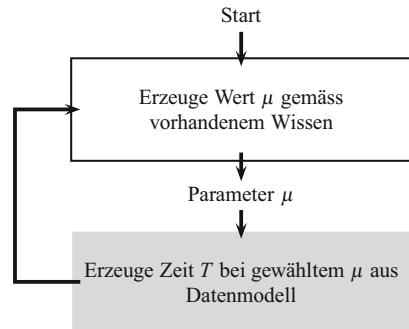
Mit der A posteriori-Verteilung von  $\mu$  ist die Frage (a) – Wie lautet die durchschnittliche Zeit zwischen aufeinanderfolgenden, zukünftigen, starken Erdbeben? – beantwortet. Es ist nun auch möglich, die Frage (b) – Wie lange geht es bis zum nächsten starken Erdbeben? – zu beantworten. Dies kann mit Theorem 7.2 getan werden. Man braucht dazu (1) das Datenmodell der Wartezeiten und (2) die Plausibilität zu  $\mu$ . In das Datenmodell setzt man alle möglichen Werte von  $\mu$  ein, gewichtet diese mit der Wahrscheinlichkeit, dass  $\mu$  diese Werte annimmt und summiert dann auf. Mit einer Formel ausgedrückt ist die Dichtefunktion für die Zeit  $T_{\text{zuk}}$  bis zum nächsten starken Erdbeben:

$$\text{pdf}(T_{\text{zuk}} = x \mid \text{Daten, min. Vor.}) \propto \underbrace{\int \frac{1}{\mu} \cdot \exp\left(-\frac{x}{\mu}\right) \cdot \frac{1}{\mu^{29}} \cdot \exp\left(-\frac{13\,910,55 \text{ Tage}}{\mu}\right) d\mu}_{\begin{array}{l} \text{Datenmodell:} \\ \text{pdf}(T_{\text{Zukunft}}=x \mid \mu) \end{array}} \underbrace{\exp\left(-\frac{13\,910,55 \text{ Tage}}{\mu}\right)}_{\begin{array}{l} \text{Plausibilität zu } \mu: \\ \text{pdf}(\mu \mid \text{Daten, min. Vorinformation}) \end{array}}$$

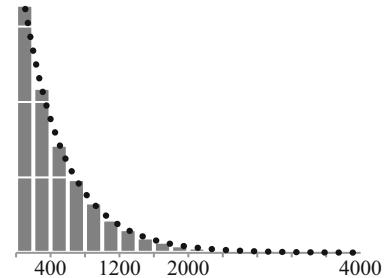
Dieses Integral kann man explizit berechnen.<sup>3</sup> Einfacher ist es, das Prognosemodell mit der vorgestellten Monte-Carlo-Simulation von Theorem 7.3 auszurechnen. Wie dies gemacht wird, fasst Abb. 9.5 zusammen. 50 000 Werte des Parameters  $\mu$  lassen sich aus der

<sup>3</sup> Am einfachsten ist es,  $\mu$  durch  $\mu = 1/\lambda$  zu substituieren. Das Integral lässt sich anschliessend mit einem Taschenrechner oder von Hand mit partieller Integration bestimmen.

**Abb. 9.5** Simulation für das Prognosemodell



**Abb. 9.6** Prognosemodell für die Zeit zwischen aufeinanderfolgender Erdbeben



Kette der MCMC-Simulation für die A posteriori-Verteilung von  $\mu$  ablesen. Mit diesen Werten lassen sich anschliessend Wartezeiten mit dem Datenmodell der Exponentialverteilung simulieren. Hier ein Pseudo-Code, der dies umsetzt:

```

for i = 1 to 50000 do
    mu[i] = i-ter Punkt aus der MCMC-Simulation für mu
    Zeit[i] = erzeugt aus Exponentialverteilung,
              die eine Rate 1/mu[i] hat
end
  
```

Das Histogramm in Abb. 9.6 visualisiert 50 000 simulierte Zeiten. Die Wahrscheinlichkeit, dass die Zeit  $T_{\text{zuk}}$  bis zum nächsten starken Erdbeben grösser als ein halbes Jahr (180 Tage) ist, beträgt

$$\mathbb{P}(T_{\text{zuk}} > 180 \mid \text{Daten, m. Vor.}) \approx \frac{\text{Anzahl simulierte Zeiten} > 180}{\text{Anzahl simulierte Werte}} = \frac{34\,869}{50\,000} = 0,697$$

Eine exakte Rechnung mit dem obigen Integral liefert das gleiche Resultat. Man berechnet in analoger Weise: Es besteht weiter eine Wahrscheinlichkeit von 0,52, dass die Zeit bis zum nächsten starken Erdbeben höchstens 365 Tage beträgt.

Seit einem halben Jahr habe kein starkes Erdbeben stattgefunden. Wie gross ist Wahrscheinlichkeit  $\mathbb{P}$ , dass das nächste starke Erdbeben sich frühestens in einem halben Jahr

ereignet? Mathematisch notiert ist dies

$$\mathbb{P} = \mathbb{P}(T_{\text{zuk}} > 360 \text{ Tage} \mid T_{\text{zuk}} > 180 \text{ Tage})$$

Mit der Simulation erhält man

$$\mathbb{P} \approx \frac{\text{Anzahl simulierte Zeiten} > 360}{\text{Anzahl simulierte Zeiten} > 180} = \frac{24\,424}{34\,869} = 0,700$$

Die Wahrscheinlichkeit ist leicht höher als diejenige, die oben berechnet ist. Hat sich also in letzter Zeit kein starkes Erdbeben ereignet, so wird es nicht wahrscheinlicher, dass das Erdbeben bald auftritt!  $\square$

Aus dem Beispiel zu den Erdbeben folgt allgemein:

**Theorem 9.2 (Das Exponentialmodell: Parameter berechnen und Messwerte prognostizieren)**

Gegeben seien  $x_1, x_2, \dots, x_n$  unabhängig modellierbare Datenwerte, die unter statistischer Kontrolle sind, die exponentialverteilt um den Parameter  $\mu > 0$  streuen:

$$\text{Datenmodell: } i\text{-ter Messwert} \sim \text{Exponential}(1/\mu)$$

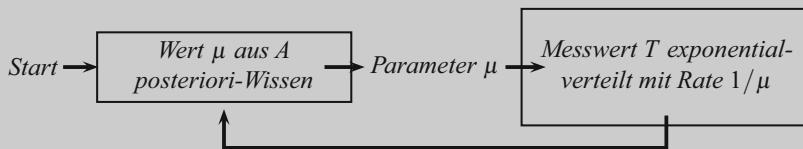
(A) Aus der Vorinformation  $\mathcal{I}$  zu  $\mu$  lässt sich mit den Daten die Plausibilität zu  $\mu$  aktualisieren:

$$\underbrace{\text{pdf}(\mu \mid \text{Daten}, \mathcal{I})}_{\text{A posteriori-Wissen}} \propto \underbrace{\frac{1}{\mu^n} \cdot \exp\left(-\frac{x_1 + x_2 + \dots + x_n}{\mu}\right)}_{\text{Likelihood}} \cdot \underbrace{\text{pdf}(\mu \mid \mathcal{I})}_{\text{A priori-Wissen}}$$

(B) Prognosen für Messwerte  $T$  berechnet man mit dem Gesetz der Marginalisierung

$$\text{pdf}(T = x \mid \text{Daten}) = \int_0^{\infty} \underbrace{\frac{1}{\mu} \cdot \exp(-x/\mu)}_{\text{Datenmodell}} \cdot \underbrace{\text{pdf}(\mu \mid \text{Daten}, \mathcal{I})}_{\text{A posteriori-Wissen zu } \mu} \, d\mu$$

oder man arbeitet mit einer Simulation:



Ist die Vorinformation zum Parameter  $\mu$  minimal, so ist  $\text{pdf}(\mu) \propto 1/\mu$ . Die A posteriori-Verteilung für  $\mu$  und die Dichtefunktion für das Prognosemodell kann man dann explizit berechnen und Wahrscheinlichkeitsintervalle mit Formeln bestimmen. Man findet die Rechnung in Anhang A in Abschn. A.1.

Die Likelihood im Ausdruck für den Posterior von  $\mu$  ist dominant, wenn man viele Datenwerte hat. Dies ist in Abschn. 5.3 erwähnt. In diesem Fall ist der Einfluss des Priors klein und Wahrscheinlichkeitsintervalle für  $\mu$  lassen sich approximativ mit einfachen Formeln und mit dem Standardfehler SE angeben.<sup>4</sup> Man hat mit einer Wahrscheinlichkeit von etwa 0,95:

$$\mu = \bar{x} \pm 1,96 \cdot \text{SE}(\exp) = \bar{x} \pm 1,96 \cdot \frac{\bar{x}}{\sqrt{n}} \quad (9.1)$$

Dabei ist  $\bar{x}$  das arithmetische Mittel der  $n$  unabhängigen Datenwerte.

**Beispiel 9.2 (Zeit zwischen starken Erdbeben)** Beim Beispiel 9.1 hat man 28 Beobachtungen. Das arithmetische Mittel der Datenwerte ist 496,81 Tage. Ein Wahrscheinlichkeitsintervall zum Niveau von ungefähr 0,95 – also einem Fehler 1. Art von 5 % – für die durchschnittliche Zeit  $\mu$  zwischen zukünftigen, aufeinanderfolgenden, starken Erdbeben ist mit der obigen Formel

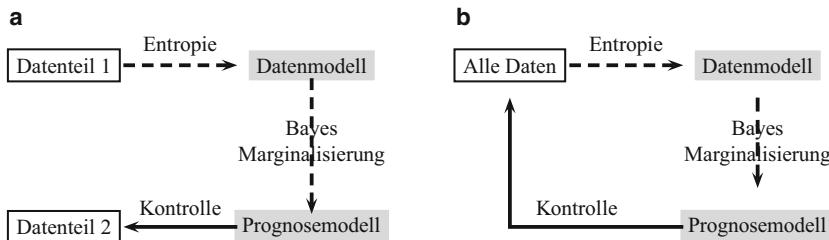
$$\mu = 496,81 \text{ Tage} \pm 1,96 \cdot \frac{496,81 \text{ Tage}}{\sqrt{28}} = (497 \pm 184) \text{ Tage}$$

Der exakte Fehler 1. Art beim Resultat ist 7 %. Das Resultat sollte skeptisch betrachtet werden. Angenommen wird, dass die Beobachtungen unabhängig sind. Insbesondere ist der Faktor  $1/\sqrt{n}$  zweifelhaft. □

Das Resultat zur mittleren Zeit zwischen aufeinanderfolgenden, zukünftigen starken Erdbeben hängt vom Datenmodell – der Exponentialverteilung – ab. Es ist daher sinnvoll, das Datenmodell zu überprüfen. Dazu gibt es verschiedene Verfahren. Ein beliebtes Verfahren ist die *Kreuzvalidierung* (engl. *cross-validation* oder *outsample prediction*). Man berechnet dazu die Parameter des Datenmodells aus einem Teil der Daten. Mit der Marginalisierung erhält man daraus das Prognosemodell. Ein Prognosemodell sollte dann die restlichen Daten gut prognostizieren können. Leider benötigt man dazu meistens zwei grosse Datenmengen: eine, um die Parameter des Datenmodells präzis zu bestimmen, und eine, um die Prognose an möglichst vielen Messpunkten zu testen. Ein weniger zuverlässiges Verfahren ist: Man verwertet alle Daten, um das Prognosemodell zu rechnen und schaut anschliessend, ob das Modell die schon benutzten Daten gut „prognostiziert“. Man spricht dann von *insample prediction*. Abb. 9.7 zeigt beide Verfahren. Wie die insample prediction gemacht wird, zeigt das folgende Beispiel:

---

<sup>4</sup> Was der Standardfehler ist, wird in Kap. 14 erklärt.



**Abb. 9.7** Outsample prediction (a) und insample prediction (b)

**Beispiel 9.3 (Zeit zwischen starken Erdbeben)** Hier die der Grösse nach geordneten 28 beobachteten Zeiten, die kleinste zuerst, die grösste zuletzt:

$$3,89 \quad 9,88 \quad 20,45 \quad 23,19 \quad 26,77 \quad \dots \quad 2300,14$$

Man sieht daraus: 1/28 der Daten sind höchstens 3,89 Tage. Prognostiziert das Modell zukünftige Wartezeiten gut, so müsste die Wahrscheinlichkeit, dass man Wartezeiten von höchstens 3,89 Tage erhält, etwa 1/28 sein! Mit anderen Worten müsste das 1/28-Quantil des Prognosemodells ungefähr gleich der kleinsten Beobachtung sein. Analog müsste das 2/28-Quantil des Prognosemodells in etwa mit der zweitkleinsten und das 3/28-Quantil mit der drittallesten Beobachtung übereinstimmen. 100 % der beobachteten Wartezeiten sind höchstens 2300,14 Tage. Das 100 %-Quantil des Prognosemodells ist aber unendlich. Um unendlich nicht mit dem Wert 2300,14 Tagen vergleichen zu müssen, subtrahiert man von den Werten 1/28, 2/28, ... jeweils  $0,5 \cdot 1/28$ . Man vergleicht damit die 0,5/28-, 1,5/28-, 2,5/28-, ..., 27,5/28-Quantile des Prognosemodells mit der Rangliste der Beobachtungen.

Die Quantile des Prognosemodells lassen sich aus den 50 000 simulierten Werten für Zeiten zwischen aufeinanderfolgenden, starken Erdbeben bestimmen. Da  $0,5/28 = 0,0179$  ist, ist das 0,5/28-Quantil die  $0,0179 \cdot 50\,000 = 893$ . kleinste simulierte Zeit. Diese ist 8,65 Tage. Analog ist das 1,5/28-Quantil die  $1,5/28 \cdot 50\,000 = 2679$ . kleinste Wert der simulierten Zeiten. Dies beträgt 25,28 Tage. Man erhält:

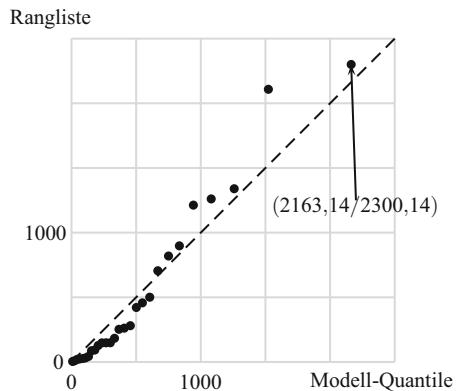
$$q_{0,5/28} = 8,65 \text{ Tage}, \quad q_{1,5/28} = 25,28 \text{ Tage}, \quad \dots, \quad q_{27,5/28} = 2163,14 \text{ Tage}$$

Vergleicht man die 28 Beobachtungen mit den Prognose-Quantilen, hat man

Quantil	8,65	25,28	44,15	62,68	...	2163,14
Rangliste	3,89	9,88	20,45	23,19	...	2300,14

Die Werte der vervollständigten Tabelle sind in Abb. 9.8 mit einem Streudiagramm dargestellt. Die Punkte müssen bei guter Modellwahl in etwa auf einer Geraden durch den Nullpunkt mit Steigung  $45^\circ$  liegen! Die Abbildung zeigt, dass sich dies so verhält. Damit

**Abb. 9.8** QQ-Plot der beobachteten Wartezeiten



wird deutlich, dass das Prognosemodell (und damit das Datenmodell) geeignet ist. Ersichtlich ist auch, dass kleine Modellquantile tendenziell grösser sind als die entsprechenden Werte aus den Beobachtungen. Im Modell werden daher kleine Zeiten zwischen Erdbeben überschätzt. Analog werden grosse Zeiten zwischen starken Erdbeben unterschätzt.  $\square$

Mit einer sorgfältigen statistischen Rechnung sollte beurteilt werden können, wie gut das Datenmodell ist. Dies kann, wie oben gezeigt, grafisch gemacht werden. Zusammengefasst hat man:

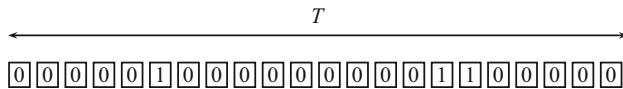
**Theorem 9.3 (Wie man Datenmodelle mit dem Quantil-Quantil-Plot analysiert)**

Das Streudiagramm bestehend aus der Rangliste der  $n$  Datenwerte und der  $(1 - 0,5)/n$ -,  $(2 - 0,5)/n$ -,  $(3 - 0,5)/n$ -, ...,  $(n - 0,5)/n$ -Quantile des Prognosemodells nennt man den Quantil-Quantil-Plot (engl. Quantile-Quantile-Plot (QQ-Plot)). Ist das Modell gut gewählt, sollten die Punktpaare des QQ-Plots etwa auf der Geraden durch den Nullpunkt mit Steigung  $45^\circ$  liegen.

## 9.2 Die Poissonverteilung

Diskrete positive Größen können im Gegensatz zu Wartezeiten nur Werte wie  $0, 1, 2, \dots$  annehmen. Dies ist der Fall, wenn Ereignisse über einen festen, endlichen Zeitraum oder einen beschränkten örtlichen Bereich gezählt werden. Die Anzahl Unfälle während eines Monats, jährliche Schadensfälle bei einer Versicherung oder die Anzahl Zerfälle von  $\alpha$ -Teilchen während einer Minute sind Beispiele für diskrete Größen. Die Plausibilität zu solchen Größen wird meist mit einem diskreten Wahrscheinlichkeitsmodell und einer Massenfunktion beschrieben. Wie lautet minimale Vorinformation zu solchen Größen? Um das zu bestimmen, kann man sich den Zeitraum (oder örtlichen Bereich)  $T$  in klei-

ne Intervalle  $\Delta t$  zerlegt denken, so dass nie mehr als ein solches Ereignis pro Intervall möglich ist:



In der Skizze sind  $M = T/\Delta t$  Intervalle dargestellt. Dabei ist  $\Delta t$  klein und damit die Anzahl Boxen  $M$  gross. Mit Nullen belegte Boxen bedeuten kein Ereignis. Mit Einern besetzte Boxen zeigen Ereignisse. In der Skizze sind drei Ereignisse dargestellt. Wie soll man die Wahrscheinlichkeit setzen, dass im gesamten Zeitintervall  $i$  Ereignisse stattfinden? Dies ist nicht schwierig, wenn man annimmt, dass die Einer und Nullen unabhängig voneinander sind. Minimale Vorinformation bedeutet dann, dass man aus einem Korb mit Kugeln, die mit Null oder Eins beschriftet sind, Kugeln blindlings zieht und in die Boxen legt.<sup>5</sup> Es gilt dann:<sup>6</sup>

**Theorem 9.4 (MaxEnt für die Poissonverteilung)**

Man zählt in einem endlichen Zeitfenster (oder endlichen Bereich) die Anzahl  $N$  Ereignisse zusammen. Als Information  $\mathcal{I}$  hat man: (1) Jedes Ereignis ist unabhängig und gleich wahrscheinlich und (2) Ereignisse treten nie gleichzeitig auf. Bei minimaler Vorinformation würde man  $N$  gemäss dem obigen Verfahren beschreiben. Das diskrete Wahrscheinlichkeitsmodell für  $N$  mit der grössten relativen Entropie, gegeben ein endlicher Erwartungswert  $\lambda$ , ist dann die Poissonverteilung:<sup>7</sup>

$$N \sim \text{Poisson}(\lambda) : \quad \mathbb{P}(N = k \mid \lambda, \mathcal{I}) = \frac{\lambda^k}{k!} \cdot \exp(-\lambda) \quad \text{für } k = 0, 1, 2, \dots$$

Der Parameter  $\lambda$  der Poissonverteilung ist eine Zahl grösser als null. Er kann kontinuierliche Werte annehmen, wie  $\lambda = 0,67$  oder  $\lambda = 13,584$ . Die Werte für die Zufallsgrösse

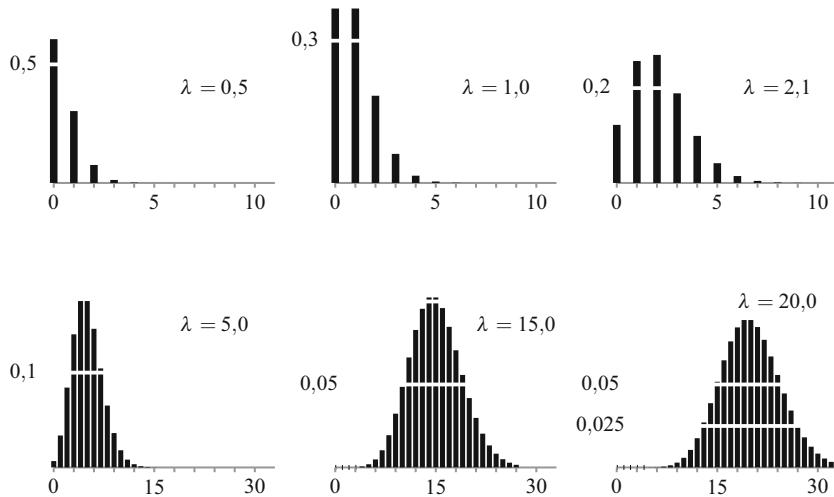
<sup>5</sup> Die Wahrscheinlichkeit  $m_i$ , dass im gesamten Zeitintervall  $i$  Ereignisse stattfinden, ist proportional zur Anzahl Möglichkeiten,  $i$  Einer und  $M - i$  Nullen in die  $M$  Boxen zu legen. Diese Anzahl rechnet man mit einem Binomialkoeffizienten. Es gilt

$$m_i \propto \binom{M}{i} = \frac{M!}{i! \cdot (M - i)!} \approx \frac{M^i}{i!}$$

Der approximative Ausdruck rechts ist nur für grosse  $M$  gültig und wird hier verwendet. Die Summe aller  $m_i$  muss 1 sein. Man erhält  $m_i \approx M^i / i! \cdot \exp(-M)$ .

<sup>6</sup> Für einen Beweis siehe [3]

<sup>7</sup>  $k!$  ist das Produkt der ersten  $k$  natürlichen Zahlen:  $1! = 1$ ,  $2! = 1 \cdot 2$ ,  $3! = 1 \cdot 2 \cdot 3$ . Für  $0!$  setzt man  $0! = 1$ .



**Abb. 9.9** Graphen der Massenfunktion der Poissonverteilung

$N$  sind diskret. Möglich sind nur Werte  $0, 1, 2, \dots$ . Visualisiert wird die Verteilung daher nicht mit einer Dichtefunktion, sondern der Massenfunktion. Abb. 9.9 zeigt Graphen der Massenfunktion der Poissonverteilung für verschiedene Werte des Parameters  $\lambda$ . Beachtenswert ist, dass die Graphen für  $\lambda > 5$  ein glockenförmiges Aussehen haben. Je grösser der Parameter  $\lambda$ , um so breiter wird die Verteilung. Die Breite wächst mit  $\sqrt{\lambda}$ . Ist  $\lambda > 5$ , dann ist die Wahrscheinlichkeit etwa 0,68, dass Werte zwischen  $\lambda - \sqrt{\lambda}$  und  $\lambda + \sqrt{\lambda}$  liegen. Mit einer Wahrscheinlichkeit von etwa 0,95 liegen Werte im Bereich  $\lambda \pm 1,96 \cdot \sqrt{\lambda}$ .

Wie die Poissonverteilung benutzt werden kann, um diskrete, nicht negative Größen zu beschreiben, zeigt das folgende Beispiel:

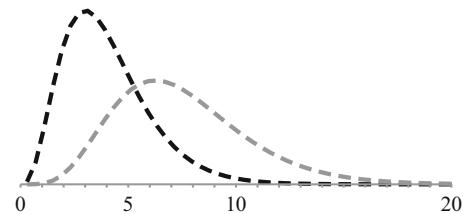
**Beispiel 9.4 (Anzahl grosser Schäden)** Eine Versicherungsfirma, die Schäden an Häusern infolge Feuer, Wassereinbrüchen oder Stürmen versichert, interessiert sich für die Anzahl grosser Schadensummen während eines Halbjahrs mit mehr als CHF 100 000,–. Tab. 9.1 zeigt dazu Daten in den Halbjahren 1998 bis 2002. Mit einem Streudiagramm der Beobachtungen sieht man, dass die Daten unter statistischer Kontrolle stehen, da kaum Trends und keine Zyklen ersichtlich sind. Zudem sind die Datenwerte unabhängig. Die Versicherungsfirma möchte die folgenden zwei Fragen beantworten:

- Wie lautet die zukünftige durchschnittliche Anzahl  $\lambda$  grosser Schadensummen pro Halbjahr?

**Tab. 9.1** Anzahl grosser Schadensummen (Daten aus einer Versicherungsfirma in der Schweiz)

Jahr/Halbjahr	1998/1	1998/2	1999/1	1999/2	2000/1	2000/2	2001/1	2001/2	2002/1	2002/2
Anzahl Schäden	2	4	7	7	3	5	8	4	8	5

**Abb. 9.10** Gammaverteilung als Strukturfunktion (Prior):  
Der Modus ist  $(a - 1)/b$  und die Breite ist proportional zu  $\sqrt{a}/b$  ( $a = 3, b = 1$  schwarz;  
 $a = 5, b = 0,8$  grau)



- (b) Wie gross ist die Anzahl  $N_{\text{zukunft}}$  grosser Schadensummen im nächsten Halbjahr?

Die Antwort zur Frage (a) erfolgt am besten, wenn man angibt, wo die nicht direkt messbare Grösse  $\lambda$  am plausibelsten liegt. Da  $\lambda$  kontinuierliche Werte annehmen kann, beschreibt man sie mit einem stetigen Wahrscheinlichkeitsmodell. Mit der Regel von Bayes kann die Vorinformation zu  $\lambda$  mit den Daten aktualisiert werden:

$$\underbrace{\text{pdf}(\lambda \mid \text{Daten, Vorinformation})}_{\text{A posteriori Verteilung}} \propto \underbrace{\mathbb{P}(\text{Daten} \mid \lambda)}_{\text{Likelihood}} \cdot \underbrace{\text{pdf}(\lambda \mid \text{Vorinformation})}_{\text{Prior}} \quad (9.2)$$

Um die Plausibilität von  $\lambda$  zu berechnen, braucht man seinen Prior. Im Versicherungswesen nennt man ihn nach H. Bühlmann (siehe [1]) die *Strukturfunktion*. Sie wird in der Regel mit einer Gammaverteilung dargestellt:

$$\text{Prior: } \lambda \sim \text{Gamma}(a, b) \quad (9.3)$$

Die Dichtefunktion lautet  $\text{pdf}(\lambda \mid \text{Vorinformation}) \propto \lambda^{a-1} \cdot \exp(-b \cdot \lambda)$ . Abb. 9.10 zeigt den Prior für verschiedene Werte der Konstanten  $a$  und  $b$ . Hängen die Schadensarten von sehr verschiedenen Faktoren (Infrastruktur, Risiko von Unwettern oder Überschwemmungen) ab, so ist  $\lambda$  sehr unsicher und man wählt eine breite A priori-Verteilung.<sup>8</sup> Die Strukturfunktion spiegelt daher das Risikoprofil der Schadensart ab. Ist  $a = b = 0$ , so entspricht dies Jeffreys' Prior

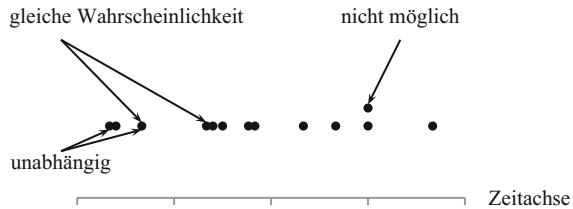
$$\text{pdf}(\lambda \mid \text{Information} = \text{min. Vorinformation}) \propto 1/\lambda$$

Dieser Fall – mit  $\lambda$  zwischen 0,5 und 100 – soll für die weitere Rechnung gelten, da keine weitere Information vorhanden ist.

Die Likelihood wird aus dem Datenmodell berechnet. Als Information dazu wird angenommen: (1) Jeder mögliche Schaden ist unabhängig von den anderen Schäden, und (2) Schäden treten nie gleichzeitig auf. Dies illustriert Abb. 9.11. Diese Information kann in ein Wahrscheinlichkeitsmodell umgesetzt werden, das eine maximale Entropie hat. Nach Theorem 9.4 folgt: Die Poissonverteilung beschreibt die Plausibilität zur diskreten Grösse

<sup>8</sup> Man nennt die Konstanten  $a$  und  $b$ , die im Prior sind, *Hyperparameter*.

**Abb. 9.11** Annahme zu den Schadensfällen: (1) Jeder mögliche Schaden unabhängig von den anderen Schäden und (2) Schäden treten nie gleichzeitig auf



„Anzahl Schäden pro Halbjahr“:

$$\text{Datenmodell: } \mathbb{P}(\text{Anzahl Schäden} = k | \lambda) = \frac{\lambda^k}{k!} \cdot \exp(-\lambda)$$

Die Likelihood kann man daraus berechnen. Die Wahrscheinlichkeit 2, 4, 7, ..., 5 Schadensfälle zu beobachten ist nach dem Multiplikationsgesetz gleich dem Produkt dieser Wahrscheinlichkeiten:

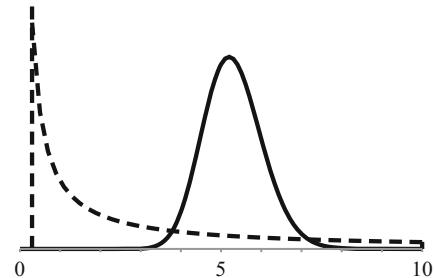
$$\begin{aligned} \mathbb{P}(\text{Daten} | \lambda) &= \mathbb{P}(2, 4, 7, \dots, 5 | \lambda) = \mathbb{P}(2 | \lambda) \cdot \mathbb{P}(4 | \lambda) \cdot \dots \cdot \mathbb{P}(5 | \lambda) \\ &= \frac{\lambda^2}{2!} \cdot \exp(-\lambda) \cdot \frac{\lambda^4}{4!} \cdot \exp(-\lambda) \cdot \dots \cdot \frac{\lambda^5}{5!} \cdot \exp(-\lambda) \end{aligned}$$

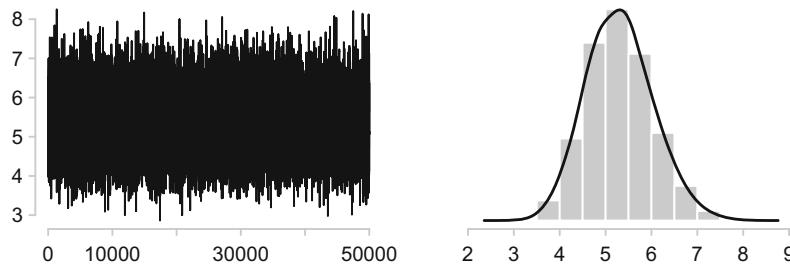
Implementiert man dies in Gleichung (9.2), erhält man die A posteriori-Verteilung von  $\lambda$ . Ihre Dichtefunktion lautet

$$\begin{aligned} \text{pdf}(\lambda | \text{Daten, min. Vor.}) &\propto \frac{\lambda^2}{2!} \cdot \exp(-\lambda) \cdot \frac{\lambda^4}{4!} \cdot \exp(-\lambda) \cdot \dots \cdot \frac{\lambda^5}{5!} \cdot \exp(-\lambda) \cdot \frac{1}{\lambda} \\ &\propto \lambda^{53} \cdot \exp(-10\lambda) \cdot \lambda^{-1} \end{aligned}$$

In Abb. 9.12 befindet sich ihr Graph. Der plausibelste Wert für  $\lambda$  ist der Modus  $\lambda_0$ . Er lässt sich mit analytischen Werkzeugen (der Ableitung) oder mit einer Wertetabelle schnell finden. Man erhält  $\lambda \approx \lambda_0 = 5.2$ . Wahrscheinlichkeitsintervalle kann man mit einer MCMC-Kette berechnen. Bei Statistikprogrammen genügt es dazu, die Daten einzugeben

**Abb. 9.12** Plausibilität zu den durchschnittlichen Schadensfällen pro Halbjahr: gestrichelt der Prior, ausgezogen die A posteriori-Verteilung





**Abb. 9.13** MCMC-Kette der A posteriori-Verteilung für die (zukünftige) durchschnittliche Anzahl Schäden  $\lambda$  pro Halbjahr

und das Datenmodell sowie den Prior zu nennen:

$$\begin{aligned} \text{Datenmodell: } & i\text{-te Beobachtung} \sim \text{Poisson}(\lambda) \\ \text{Prior: } & \ln \lambda \sim \text{Uniform}(\ln 0,5 ; \ln 100) \end{aligned}$$

Abb. 9.13 zeigt eine Kette von 50 000 Punkten und einer Akzeptanzrate von 39 %. Es sind 25 % der Punkte kleiner als 4,79 und 25 % sind grösser als 5,77. Es besteht daher eine Wahrscheinlichkeit von 0,5, dass die (zukünftige) durchschnittliche Schadenzahl pro Halbjahr zwischen 4,79 und 5,77 liegt. Mit einer Wahrscheinlichkeit von 0,95 liegt die durchschnittliche Schadenzahl zwischen 3,97 und 6,82. Damit ist die Frage (a) nach der durchschnittlichen zukünftigen Schadenzahl beantwortet.

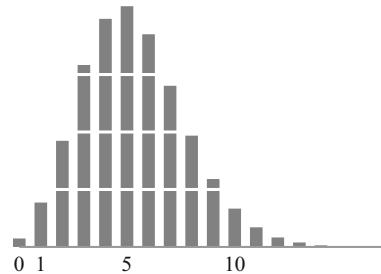
Mit der obigen Rechnung lässt sich nun die Frage (b) – Wie gross ist die Schadenzahl  $N_{\text{zukunft}}$  des nächsten Halbjahrs? – beantworten. Dazu arbeitet man mit dem Gesetz der Marginalisierung. Man benutzt (1) das Datenmodell für die Anzahl Schäden – die Poissonverteilung – und (2) den Posterior des Parameters  $\lambda$ . In das Datenmodell werden alle möglichen Werte von  $\lambda$  eingesetzt, gewichtet über den Posterior von  $\lambda$ . Dies ist das Monte-Carlo-Verfahren von Theorem 7.3. 50 000 Werte des Parameters  $\lambda$  sind aus der Kette der MCMC-Simulation bekannt. Mit diesen Werten lassen sich anschliessend Anzahl zukünftiger Schäden  $N_{\text{zukunft}}$  simulieren. Hier ein Pseudo-Code, der dies umsetzt:

```
for i = 1 to 50000 do
    lambda[i] = i-ter Punkt aus MCMC-Sim. für lambda
    Nzukunft[i] = erzeugt aus Poissonverteilung
                  mit Parameter lambda[i]
end
```

Abb. 9.14 visualisiert 50 000 simulierte Anzahl Schadensfälle pro Halbjahr. Es stellt die Massenfunktion des Prognosemodells dar. Die Wahrscheinlichkeit, dass sich weniger als fünf Schadensfälle pro Halbjahr ereignen werden, ist

$$\mathbb{P}(N_{\text{zukunft}} < 5 | \text{Daten}) = \frac{\text{Anzahl Simulationen} < 5}{50\,000} = \frac{19\,803}{50\,000} = 0,40$$

**Abb. 9.14** Prognosemodell für die Anzahl Schäden im nächsten Halbjahr



Die Wahrscheinlichkeit, dass im nächsten Halbjahr mehr als zehn grosse Schadensummen auftreten werden, beträgt 0,025. In Tab. 9.2 findet man weitere Werte des Prognosemodells. Die Wahrscheinlichkeit, dass im nächsten Halbjahr mehr als 10 Schadensfälle vorkommen werden, ist 0,0252.

Auch in diesem Fall ist es sinnvoll, das Datenmodell der Poissonverteilung zu kontrollieren. Dies kann man, wie schon beim Datenmodell zur Exponentialverteilung gezeigt, mit einem QQ-Plot durchführen. Um den QQ-Plot zu zeichnen, braucht man die Quantile aus dem Prognosemodell für zukünftige Anzahl Schadensfälle. Nötig sind dabei die  $0,5/10$ -,  $1,5/10$ -,  $2,5/10$ -, ...,  $9,5/10$ -Quantile, da  $n = 10$  Datenwerte vorliegen. Da  $0,5/10 \cdot 50\,000 = 2500$  ist, ist das  $0,5/10$ -Quantil der 2500. kleinste Wert der simulierten 50 000 Anzahl Schadensfälle. Er beträgt zwei. Analog ist das  $1,5/10$ -Quantil der 7500. kleinste Wert der simulierten Anzahl Schadensfälle. Er ist drei. Man erhält:

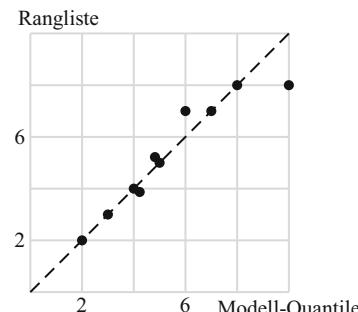
Quantil	2	3	4	4	5	5	6	7	8	10
Rangliste	2	3	4	4	5	5	7	7	8	8

Abb. 9.15 zeigt den daraus gebildeten QQ-Plot. Sie visualisiert, dass das Prognosemodell (und damit das Datenmodell der Poissonverteilung) plausibel wirkt. Zu bedenken ist,

**Tab. 9.2** Massenfunktion des Prognosemodells für die Anzahl Schäden  $N$  im nächsten Halbjahr

$N$	0	1	2	3	4	5	6	7	8	9	10
$\mathbb{P}$	0,0059	0,0309	0,0739	0,1266	0,1588	0,1675	0,1480	0,1122	0,0775	0,0469	0,0267

**Abb. 9.15** QQ-Plot zu den Anzahl Schäden: Da mehrere Punkte aufeinanderfallen, werden sie *gejittert*, d.h. leicht versetzt



dass sehr wenige Beobachtungen vorliegen. In der Tat sind QQ-Plots mit weniger als 10–15 Datenwerten nicht sehr aussagekräftig.  $\square$

Aus dem Beispiel zu den Schadensfällen folgt allgemein:

**Theorem 9.5 (Das Poissonmodell: Parameter berechnen und Messwerte prognostizieren)**

Gegeben seien  $x_1, x_2, \dots, x_n$  unabhängig modellierbare nicht negative, diskrete Werte, die unter statistischer Kontrolle sind. Sie streuen poissonverteilt um den Parameter  $\lambda$ :

Datenmodell:  $i$ -ter Messwert  $\sim \text{Poisson}(\lambda)$

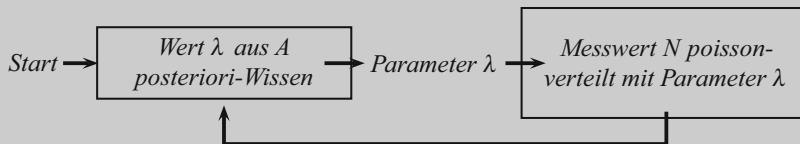
(A) Aus Vorinformation  $\mathcal{I}$  lässt sich mit den Daten die Plausibilität zu  $\lambda > 0$  aktualisieren:

$$\underbrace{\text{pdf}(\lambda | \text{Daten}, \mathcal{I})}_{\text{A posteriori-Plausibilität}} \propto \underbrace{\lambda^{x_1+x_2+\dots+x_n} \cdot \exp(-n \cdot \lambda)}_{\text{Likelihood}} \cdot \underbrace{\text{pdf}(\lambda | \mathcal{I})}_{\text{Prior}}$$

(B) Prognosen für Messwerte  $N$  werden mit dem Gesetz der Marginalisierung berechnet:

$$\mathbb{P}(N = k | \text{Daten}) = \int_0^{\infty} \underbrace{\frac{\lambda^k}{k!} \cdot \exp(-\lambda)}_{\text{Datenmodell}} \cdot \underbrace{\text{pdf}(\lambda | \text{Daten}, \mathcal{I})}_{\text{A posteriori-Plausibilität zu } \lambda} d\lambda$$

oder man arbeitet mit einer Simulation:



Hat man keine Vorinformation zum Parameter  $\lambda$ , so wählt man  $\text{pdf}(\lambda | \mathcal{I}) \propto 1/\lambda$ . In diesem Fall kann man die A posteriori-Verteilung von  $\lambda$  und das Prognosemodell explizit berechnen. Die Formeln finden sich in Anhang A in Abschn. A.2.

Die Likelihood-Funktion im Ausdruck für den Posterior von  $\lambda$  ist dominant, wenn man viele Datenwerte hat. Der Einfluss des Priors auf den Posterior ist dann klein und Wahrscheinlichkeitsintervalle für den Parameter  $\lambda$  lassen sich approximativ mit einfachen

Formeln angeben. Man hat mit einer Wahrscheinlichkeit von *etwa* 0,95:

$$\lambda = \bar{x} \pm 1,96 \cdot \text{SE}(\text{pois}) = \bar{x} \pm 1,96 \cdot \frac{\sqrt{\bar{x}}}{\sqrt{n}} \quad (9.4)$$

Dabei ist  $\bar{x}$  das arithmetische Mittel der  $n$  unabhängigen Datenwerte.

### 9.3 Zusammenfassung

Die vorigen Abschnitte zeigen beispielhaft, wie Datenmodelle gesetzt werden können. Insbesondere ist eine spezifische Methode angewendet worden: das Prinzip der maximalen Entropie. Dabei geht es darum, genau die Information in das Datenmodell zu stecken, welche vorhanden ist. Es gibt auch andere Verfahren, um ein Datenmodell oder eine A priori-Verteilung zu wählen. So findet man oft *physikalische Überlegungen* aus denen ein Datenmodell gewählt wird. Argumente, wie *empirischer Bayes* oder *konjugierte Verteilungen*, sind beliebt, um die A priori-Verteilung zu fixieren. Dazu ist das folgende Vorgehen hilfreich:

Prüfen Sie, ob Größen, die Sie untersuchen, in Forschungsberichten und in wissenschaftlichen Büchern schon mit einem Modell beschrieben sind. Sind Argumente angegeben, warum das Datenmodell verwendet wird? Denken Sie auch daran, das gewählte Datenmodell mit einem QQ-Plot qualitativ zu beurteilen. Dies kann helfen, ein besseres Modell zu finden. Beurteilen Sie auch, wie empfindlich die Resultate auf die Wahl des Priors sind.

### Reflexion

**9.1** Für die Passagierflugzeuge vom Typ Boeing 720 soll die durchschnittlich erwartbare Funktionsdauer  $\mu$  der Klimaanlage berechnet werden. Zudem soll prognostiziert werden, wie lange eine Klimaanlage ohne Unterbruch funktioniert. Dazu werden die Funktionsdauern – eine kontinuierliche Größe – von zwölf Klimaanlagen untersucht. Hier die Resultate (in Stunden) aus [2]:

7 230 100 98 5 3 43 85 130 18 487 91

Machen Sie eine sinnvolle Annahme für ein stetiges Datenmodell  $\mathbb{P}(T | \mu)$  der Funktionsdauer  $T$ , um den gesuchten Parameter  $\mu$  zu berechnen. Weiter ist es sinnvoll anzunehmen, dass  $\mu$  zwischen einer und 1000 Stunden liegt.

- (a) Kontrollieren Sie, ob die Daten unter statistischer Kontrolle sind.
- (b) Zeichnen Sie den Graphen der A posteriori-Dichtefunktion der mittleren erwartbaren Funktionsdauer  $\mu$ . Was ist der plausibelste Wert von  $\mu$ ? Beantworten Sie die folgenden Fragen: Wie lautet die Fünf-Zahlen-Zusammenfassung zu  $\mu$ , wie ein 50 % und ein 95 % Wahrscheinlichkeitsintervall?
- (c) Wie gross ist die Wahrscheinlichkeit, dass zukünftig gemessene Funktionsdauern (1) kleiner als 50 Stunden, (2) zwischen 100 und 150 Stunden und (3) grösser als 500 Stunden betragen werden?
- (d) Beurteilen Sie, ob Ihr gewähltes Datenmodell sinnvoll ist. Zeichnen Sie dazu einen QQ-Plot mit Hilfe des Prognosemodells.

**9.2** Eine Firma interessiert sich, wie lange es geht, bis ihre produzierten, wiederaufladbaren Batterien bei voller Belastung entladen sind. Die Verantwortliche der Produktion möchte dazu die durchschnittliche Entladungszeit  $\bar{T}_{\text{Prod}}$  aller produzierten Batterien bestimmen. Physikalisch bedingt muss sie zwischen 10 und 1000 Minuten liegen. Weiter will sie auch prognostizieren, in welchem Bereich gemessene Entladungszeiten  $T$  liegen werden. Das Wissen zu Entladungszeiten soll dabei mit einem stetigen Wahrscheinlichkeitsmodell beschrieben werden. Machen Sie dazu eine sinnvolle Annahme für das Datenmodell  $\mathbb{P}(T \mid \bar{T}_{\text{Prod}})$  der Entladungszeit  $T$ .

- (a) Berechnen Sie den gesuchten Parameter  $\bar{T}_{\text{Prod}}$  aus drei gemessenen Entladungszeiten: 301,2, 378,7 und 315,0 Minuten. Zeichnen Sie dazu den Graphen der A posteriori-Dichtefunktion von  $\bar{T}_{\text{Prod}}$ . Was ist der plausibelste Wert für die durchschnittliche Entladungszeit?
- (b) Wie lautet die Fünf-Zahlen-Zusammenfassung, wie ein 50 % und ein 95 % Wahrscheinlichkeitsintervall für  $\bar{T}_{\text{Prod}}$ ?
- (c) Wie gross ist die Wahrscheinlichkeit, dass zukünftig gemessene Entladungszeiten (1) kleiner als 300 Minuten, (2) zwischen 60 und 120 Minuten und (3) grösser als 360 Minuten betragen werden?
- (d) Führen Sie die Aufgaben (a)–(c) nocheinmal durch, wenn nur eine Messung von 310,5 Minuten vorliegt.
- (e) Rechnen Sie die Aufgabe (d) noch einmal durch: als Information haben Sie die drei Messwerte von (a) und den Messwert von (d).

**9.3** Beim Beispiel 3.20 von W. W. Scherkenbach aus der Motorenfabrik Ford werden bei Angestellten der Firma die Anzahl Fehler während eines Zeitfensters – wie Rechen- oder Zeichenfehler, Fehler beim Zusammenstellen von Produkten, etc. – gezählt. Gesucht wird dabei die – in Zukunft – erwartbare durchschnittliche Anzahl Fehler  $\lambda$  pro Zeitfenster und pro Person, sowie eine Prognose für zukünftige Beobachtungen diesbezüglich. Erfahrungen der Firma zeigen, dass  $\lambda$  zwischen eins und hundert liegen muss. Machen Sie eine sinnvolle Annahme für das diskrete Datenmodell  $\mathbb{P}(N \mid \lambda)$  der Anzahl Fehler  $N$ .

- (a) Berechnen Sie den Posterior von  $\lambda$ . Was ist der plausibelste Wert von  $\lambda$ ?
- (b) Wie lautet die Fünf-Zahlen-Zusammenfassung zu  $\lambda$ , wie Wahrscheinlichkeitsintervalle zum Niveau 0,5, 0,68 und 0,95?
- (c) Berechnen Sie ein Wahrscheinlichkeitsintervall für  $\lambda$  zum Niveau von etwa 0,95 mit der im Kapitel vorgestellten Standardfehler-Formel.
- (d) Wie gross ist die Wahrscheinlichkeit, dass zukünftige Beobachtungen der Anzahl Fehler (1) grösser als 10, (2) grösser als 20 sind?
- (e) Beurteilen Sie, ob das Datenmodell – die Poissonverteilung – sinnvoll ist. Zeichnen Sie dazu den QQ-Plot mit Hilfe des Prognosemodells aus der Aufgabe (d).
- (f) Ist es gerecht, Charlie einen Bonus zu verweigern? Berechnen Sie dazu die Wahrscheinlichkeit aus dem Prognosemodell, dass  $N_{\text{zukunft}}$  grösser 22 ist.

**9.4** Die Schweizerischen Bundesbahnen (SBB) erfassen im Rahmen ihres Qualitätsmanagements die Signalvorfälle. Hier die beobachteten Signalvorfälle in den Jahren 2008–2012 (Sonntagszeitung vom 13.1.2013):

Jahr	2008	2009	2010	2011	2012
Vorfälle	96	129	86	91	111

Gesucht ist (1) die zukünftige durchschnittliche Anzahl Signalvorfälle  $\lambda$  pro Jahr und (2) eine Prognose für die Anzahl Signalvorfälle  $N$  im Jahr 2013. Sie nehmen an, dass die Anzahl Signalvorfälle  $N$  mit einer Poisson-Verteilung modelliert werden. Sie haben weiter keine zusätzlichen Informationen zu  $\lambda$ . Sie haben weiter als zusätzliche Information nur, dass  $10 \leq \lambda \leq 1000$  ist.

- (a) Beurteilen Sie, ob die Beobachtungen unter statistischer Kontrolle sind.
- (b) Berechnen Sie den gesuchten Parameter  $\lambda$ . Was ist der plausibelste Wert von  $\lambda$ ?
- (c) Bestimmen Sie die Fünf-Zahlen-Zusammenfassung zu  $\lambda$ , sowie ein 50 % und ein 95 % Wahrscheinlichkeitsintervall.
- (d) Wie lauten Wahrscheinlichkeitsintervalle für  $\lambda$  zum Niveau von etwa 0,5 und etwa 0,95, berechnet mit dem Standardfehler? Vergleichen Sie die Resultate mit denen von Aufgabe (c).
- (e) Wie gross sind die Wahrscheinlichkeiten, dass die Anzahl Signalvorfälle  $N$  im Jahr 2013 (1) zwischen 70 und 130, (2) kleiner als 110 und (3) grösser als 90 sind?
- (f) Ist es sinnvoll das Datenmodell, mit einem QQ-Plot zu „überprüfen“?

**9.5** Eine Firma produziert Filtermembranen. Wegen varierender Produktionsbedingungen haben die Membranen verschiedene Poren-dichten. Ein Chiemelabor möchte aus einer Stichprobe die durchschnittliche Poren-dichte  $\bar{P}_{\text{Prod}}$  – die zwischen 100 und 2000 Poren pro  $\text{mm}^2$  liegt – der Filtermembranen der Gesamtproduktion präziser bestimmen und wissen, in welchem Bereich gezählte Poren-dichten liegen werden. Eine Person untersucht

eine Membran und zählt 1020 Poren pro  $\text{mm}^2$ . Machen Sie eine sinnvolle Annahme für das Datenmodell  $\mathbb{P}(P \mid \bar{P}_{\text{Prod}})$  der Porendichte  $P$ . Das Datenmodell soll diskret sein.

- Berechnen Sie die A posteriori-Verteilung des gesuchten Parameters  $\bar{P}_{\text{Prod}}$ . Was ist der plausibelste Wert von  $\bar{P}_{\text{Prod}}$ ?
- Wie lautet die Fünf-Zahlen-Zusammenfassung für  $\bar{P}_{\text{Prod}}$ , wie Wahrscheinlichkeitsintervalle zum Niveau 0,5 und 0,95?
- Wie gross sind die Wahrscheinlichkeiten, dass zukünftige Messungen der Porendichte (1) grösser als 1030 pro  $\text{mm}^2$ , (2) zwischen 1000 und 1030 pro  $\text{mm}^2$  sind?
- Führen Sie die Aufgaben (a) und (b) noch einmal durch, diesmal unter der Voraussetzung, dass vier gemessene Porendichten 1020, 1044, 1012 und 1025 pro  $\text{mm}^2$  vorliegen. Benutzen Sie dabei als Prior für den Parameter  $\bar{P}_{\text{Prod}}$  den Posterior aus Aufgabe (a).

**9.6** Rechnen Sie das Beispiel 9.4 zu den Anzahl grosser Schäden noch einmal durch, diesmal mit dem Prior

$$\lambda \sim \text{Gamma}(5; 0,8), \quad \text{also } \text{pdf}(\lambda \mid \text{Vorinformation}) \propto \lambda^4 \cdot \exp(-0,8 \cdot \lambda)$$

- Zeichnen Sie den Graphen des Priors.
- Wie lauten Wahrscheinlichkeitsintervalle zum Niveau 0,5 und 0,95 für die zukünftige durchschnittliche Anzahl  $\lambda$  grosser Schadensummen pro Halbjahr? Nennen Sie den plausibelsten Wert von  $\lambda$ .
- Wie gross ist die Wahrscheinlichkeit, dass die Anzahl  $N_{\text{zukunft}}$  grosser Schadensummen im nächsten Halbjahr grösser als 5 ist?

**9.7** Gletscherstürze und Gletscherhochwasser stellen Gefahren für Siedlungen, Verkehrs- und Wanderwege dar. Bei 49 Gletschern in der Schweiz besteht laut einer Studie der ETH Zürich die Gefahr, dass in naher Zukunft solche Katastrophen zu Todesopfern führen können. Tab. 9.3 zeigt Schadenereignisse mit Todesopfern aus dem Zeitabschnitt von 1595 bis 1992. Mit den Daten sollen die zukünftige durchschnittliche Anzahl Todesopfer bei Gletscherabbrüchen und Gletscherhochwassern berechnet werden.

- Stellen Sie die Daten mit einem Streudiagramm dar. Beurteilen Sie: Sind die Daten unter statistischer Kontrolle? Können sie als unabhängig betrachtet werden?
- Erklären Sie, warum es sinnvoll ist, die zukünftige Anzahl Todesopfer mit einer Poissonverteilung zu modellieren. Berechnen Sie den Parameter  $\lambda$  der Poissonverteilung. Was ist der plausibelste Wert von  $\lambda$ ? Geben Sie Wahrscheinlichkeitsintervalle zum Niveau 0,5 und zum Niveau 0,95 für  $\lambda$  an.
- Berechnen Sie ein Wahrscheinlichkeitsintervall für  $\lambda$  zum Niveau von etwa 0,95 mit der im Kapitel vorgestellten Standardfehler-Formel.

**Tab. 9.3** Todesopfer wegen Gletscherabbrüchen und Gletscherhochwassern aus dem Zeitabschnitt von 1595 bis 1992

Jahr	Gletscher	Was ist passiert?	Opfer
1595	Giétro	Hochwasser	140
1597	Hohmattu	Gletschersturz	81
1636	Bis	Eis-Schnee-Lawine	37
1720	Bis	Eis-Schnee-Lawine	12
1792	Altels	Gletschersturz	4
1818	Giétro	Hochwasser	44
1819	Bis	Eis-Schnee-Lawine	2
1829	Gruben	Hochwasser	1
1895	Altels	Gletschersturz	6
1901	Rossboden	Gletschersturz	2
1918	Rhone	Eissturz	1
1922	Unterer Grindelwald	Eissturz	1
1934	Rhone	Hochwasser	2
1941	Rhone	Eissturz	6
1948	Rosenlaui	Eissturz	3
1965	Allalin	Gletschersturz	88
1974	Breitlouwesen	Eissturz	2
1976	Hochfirn	Eissturz	3
1982	Giesen	Eissturz	1
1982	Oberer Grindelwald	Eissturz	1
1992	Challifirn	Eissturz	3

- (d) Wie gross ist die Wahrscheinlichkeit, dass bei einem zukünftigen Gletscherabbruch oder Gletscherhochwasser mehr als 50 Personen sterben? Was liefert das Modell für die Wahrscheinlichkeit, dass mehr als 150 Personen sterben könnten?
- (e) Beurteilen Sie, ob das Datenmodell in (b) sinnvoll ist. Zeichnen Sie dazu den QQ-Plot mit Hilfe des Prognosemodells.

---

## Literatur

1. H. Bühlmann, A. Gisler, *A Course in Credibility Theory and its Applications* (Springer-Verlag, Berlin, Heidelberg, New York, 2005)
2. F. Proschan, Theoretical explanation of observed decreasing failure rate. *Technometrics* **5**, 373–383 (1963)
3. D. S. Sivia, J. Skilling, *Data Analysis, a Bayesian Tutorial* (Oxford University Press, 2006)

*Hinab, hinab, hinab. Wollte das denn nie ein Ende nehmen? „Wie viele Meilen ich wohl schon gefallen bin?“ sagte sie laut. „Weit kann es nicht mehr sein bis zum Erdmittelpunkt. Das wären dann, ja: sechstausend Kilometer wären das, ungefähr wenigstens –“ (denn wohlgemerkt, Alice hatte mancherlei Dinge dieser Art in der Schule lernen müssen, und wenn dies auch keine sehr gute Gelegenheit war, ihr Wissen anzubringen, weil ihr nämlich keiner zuhörte, so war es doch eine gute Übung) „– ja, das dürfte wohl die richtige Entfernung sein – aber dann möchte ich doch gerne wissen, welchen Längengrad ich wohl inzwischen habe und welchen Breitengrad?“*

*Lewis Carroll, Alice im Wunderland (Insel Taschenbuch, 1973, S. 13.)*

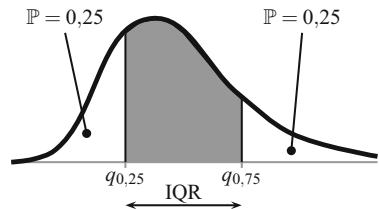
## Zusammenfassung

In der Produktionstechnik und in technischen Wissenschaften wird oft angenommen, dass mit einer Normalverteilung beschrieben werden kann, wie Messwerte streuen. Die Normalverteilung wird in diesem Kapitel vorgestellt. Zudem wird illustriert, warum es sinnvoll ist, diese Verteilung zu benutzen: Sie hat, gegeben der Erwartungswert und die Standardabweichung, eine maximale Entropie. Die Normalverteilung hat zwei Parameter: den Modus und die Standardabweichung. In vielen Anwendungen interessiert allerdings nur der Modus. Die Standardabweichung ist dann ein *Störparameter* (engl. *nuisance parameter*), der mit dem Gesetz der Marginalisierung eliminiert werden kann.

## 10.1 Die Streuung als Information

Im Qualitätsmanagement und in der Messtechnik ist der Begriff der Streuung dominierend. Streuung besagt beispielsweise, wie stark Messwerte von Messung zu Messung variieren. Bei einem Wahrscheinlichkeitsmodell misst Streuung, wie „breit“ die Verteilung ist. Ein solches Mass ist in Abb. 10.1 dargestellt: die Differenz zwischen dem 0,75-

**Abb. 10.1** Quartilsdifferenz  
IQR



und 0,25-Quantil, die *Quartilsdifferenz* (engl. *Interquartile range* [IQR]). Je grösser die Quartilsdifferenz ist, umso unsicherer kann man mit dem Modell eine Grösse lokalisieren oder prognostizieren.

Die Standardabweichung ist ein weiteres Mass, um Streuung zu quantifizieren. Sie misst, wie durchschnittlich Werte im Wahrscheinlichkeitsmodell vom Erwartungswert  $\mu$  abweichen. Es folgt die genaue Definition:

Die mit einem diskreten Wahrscheinlichkeitsmodell beschriebene Grösse  $X$  kann die Werte  $x_1, x_2, \dots$  mit Wahrscheinlichkeiten  $\mathbb{P}(X = x_1), \mathbb{P}(X = x_2), \dots$  annehmen. Die Differenz  $x_i - \mu$  misst, wie die einzelnen Werte vom Erwartungswert  $\mu$  abweichen. Damit sich negative und positive Abweichungen nicht aufheben, summiert man die Quadrate dieser Abweichungen auf und gewichtet mit ihren Auftretenswahrscheinlichkeiten:

$$\sigma^2 = (x_1 - \mu)^2 \cdot \mathbb{P}(X = x_1) + (x_2 - \mu)^2 \cdot \mathbb{P}(X = x_2) + \dots$$

Man nennt dies die *Varianz* (engl. *variance*) des Modells. Es ist die *im Schnitt erwartbare quadratische Abweichung vom Erwartungswert*. Die Wurzel der Varianz heisst die *Standardabweichung* (engl. *standard deviation*). Sie hat die gleiche Einheit wie die Grösse  $X$ . In der Messtechnik nennt man die Standardabweichung  $\sigma$  auch vereinfachend die *Unsicherheit* (engl. *uncertainty*) des Modells. Sie wird mit  $u$  notiert.<sup>1</sup> In der Wirtschaftswissenschaft spricht man auch von der *Volatilität* (engl. *volatility*).

Für stetige Wahrscheinlichkeitsmodelle wandelt man die obige Summe für die Varianz in ein Integral um:

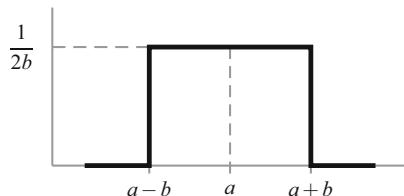
$$\text{Varianz} = \sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot \text{pdf}(x) \, dx$$

Dabei ist  $\text{pdf}(x)$  die Dichtefunktion und  $\mu$  der Erwartungswert des Modells. Die Varianz eines Wahrscheinlichkeitsmodells kann auch unendlich sein. Dies ist der Fall, wenn grosse Werte mit hoher Wahrscheinlichkeit vorkommen.

**Beispiel 10.1 (Messgerät)** Die Frequenz eines Stromkreises mit einer Wechselspannung von 250 V und einer Frequenz von 4 Hz werde mit dem Gerät, das in Beispiel 1.2 in Kap. 1

<sup>1</sup> Der Begriff ‚Unsicherheit‘ darf hier nicht mit dem umgangssprachlichen Wort ‚Unsicherheit‘ verwechselt werden. Er bezeichnet hier die Standardabweichung.

**Abb. 10.2** Dichtefunktion der Gleichverteilung



vorgestellt ist, gemessen. Aus der Gebrauchsanleitung erfährt man: Die Typ B Unsicherheit der Messung wird durch eine Gleichverteilung mit Werten zwischen 3,996 Hz und 4,004 Hz modelliert. Sie wird in der Messtechnik mit der Standardabweichung quantifiziert:<sup>2</sup>

$$\text{Messunsicherheit des Geräts} = \pm u(\text{zukünftige Messung})$$

Der Erwartungswert der Gleichverteilung auf dem Intervall  $[a - b, a + b]$  ist der Mittelpunkt  $a$  des Intervalls (siehe Abb. 10.2). Die Varianz des Modells ist

$$\text{Varianz} = \int_{a-b}^{a+b} (x - a)^2 \cdot \text{pdf}(x) dx = \int_{a-b}^{a+b} (x - a)^2 \cdot \frac{1}{2b} dx = \frac{b^2}{3}$$

Die Standardabweichung des Modells ist damit  $b/\sqrt{3}$ . Beim erwähnten Messgerät beträgt sie – es ist  $b = (4,004 \text{ Hz} - 3,996 \text{ Hz})/2 = 0,004 \text{ Hz}$  –:

$$\text{Unsicherheit Typ B} = \sigma = 0,004 \text{ Hz}/\sqrt{3} = 0,0023 \text{ Hz}$$

In der Messtechnik wird mit der *mittleren quadratischen Abweichung* (engl. *mean square error* MSE) charakterisiert, wie Messwerte streuen und wie gross ihre systematischen Fehler sind. Diese ist

$$\text{MSE einer Messung} = (\text{systematischer Fehler Messung})^2 + \sigma^2$$

Dabei ist  $\sigma$  die Typ B Unsicherheit. In technischen Berichten findet man, wenn nach den Richtlinien des GUM gearbeitet wird, Angaben zu Messfehlern, die in Vielfachen von  $\sqrt{\text{MSE}}$  angegeben sind. □

Oft nimmt man in technischen Untersuchungen an, dass kontinuierliche Messwerte mit einem Datenmodell mit endlichem Erwartungswert und endlicher Standardabweichung beschrieben werden können. Jemand möchte sein Wissen zu solchen Grössen mit grösstmöglicher Entropie beschreiben. Wie soll er dies tun? Die Antwort ist:

<sup>2</sup> GUM: *Guide to the Expression of Uncertainty in Measurement*, 1993 Geneva, Switzerland. Herausgegeben von ISO.

**Theorem 10.1 (MaxEnt für die Normalverteilung)**

Die Plausibilität zu einer Grösse  $X$  soll mit einem stetigen Wahrscheinlichkeitsmodell beschrieben werden. Als Information 1 hat man: (1)  $X$  kann beliebige Werte annehmen und (2) das Modell hat einen endlichen Erwartungswert  $\mu$  und eine endliche Standardabweichung  $\sigma$ . Dann gilt: Das Modell mit (1) und (2), das die grösste Shannon-Entropie hat, ist die Normalverteilung mit Modus  $\mu$  und Standardabweichung  $\sigma$ .

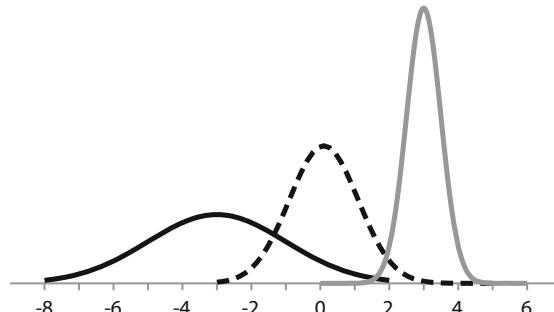
## 10.2 Die Normalverteilung

Die Normalverteilung – auch *Gauss-Verteilung* genannt – ist ein stetiges Wahrscheinlichkeitsmodell. Sie hat die Dichtefunktion

$$\text{pdf}_{\text{Gauss}}(x | \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp \left\{ -0,5 \cdot \left( \frac{x - \mu}{\sigma} \right)^2 \right\}$$

Beschreibt man eine Grösse  $X$  mit der Normalverteilung, so schreibt man angelehnt an die Notation in Statistikprogrammen:<sup>3</sup>  $X \sim \text{Normal}(\mu, \sigma)$ . In der Formel befinden sich die Parameter  $\mu$  und  $\sigma > 0$ . Der Parameter  $\mu$  ist der Modus und  $\sigma$  ist die Standardabweichung. Abb. 10.3 zeigt Graphen der Dichtefunktion der Normalverteilung. Die Normalverteilung ist glockenförmig, symmetrisch und die Spitze liegt beim Modus  $\mu$ . Der Modus  $\mu$  ist auch der Median und der Erwartungswert. Die Standardabweichung  $\sigma$  kann man, wie in Abb. 10.4 gezeigt, aus der glockenförmigen Kurve ablesen. Auf halber Höhe der Spitze hat die Kurve eine Breite von  $2,35 \cdot \sigma$ . Je grösser  $\sigma$  ist, umso breiter ist der Graph der Dichtefunktion.<sup>4</sup>

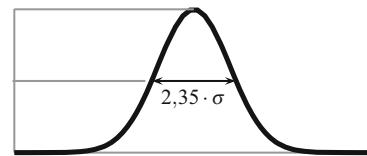
**Abb. 10.3** Dichtefunktion der Normalverteilung mit Modus  $\mu$  und Standardabweichung  $\sigma$ :  
ausgezogene Linie  $\mu = -3$  und  $\sigma = 2$ ; gestrichelt  $\mu = 0,1$  und  $\sigma = 1$ ; graue Linie  $\mu = 3$  und  $\sigma = 0,5$



<sup>3</sup> Eine oft verwendete mathematische Notation ist  $X \sim \mathcal{N}(\mu, \sigma^2)$ .

<sup>4</sup> Die Shannon-Entropie der Normalverteilung lautet  $0,5 + 0,5 \cdot \ln(2\pi) + \ln \sigma$ . Je grösser die Streuung  $\sigma$  der Normalverteilung, um so grösser ist auch die Shannon-Entropie.

**Abb. 10.4** Halbwertsbreite (engl. *full width at half maximum (FWHM)*): bei der Normalverteilung  $\text{FWHM} = 2 \cdot \sqrt{2 \ln 2} \cdot \sigma = 2,355 \cdot \sigma$



Wahrscheinlichkeitsintervalle für eine normalverteilte Grösse lassen sich in der Form  $\mu \pm k \cdot \sigma$  angeben. Die Zahl  $k$  nennt man den *Erweiterungsfaktor*. In der Messtechnik werden Wahrscheinlichkeitsintervalle oft mit diesem Faktor charakterisiert. Mit einer Wahrscheinlichkeit von 0,6828 (oder zum Faktor  $k = 1$ ) liegen Werte zwischen  $\mu - \sigma$  und  $\mu + \sigma$ . Analog hat man für eine Grösse  $M$ , die normalverteilt modelliert wird (für einen Beweis, siehe [3]):

$$\begin{aligned}\mathbb{P}(\mu - 2\sigma < M \leq \mu + 2\sigma) &= 0,955 & [k = 2] \\ \mathbb{P}(\mu - 3\sigma < M \leq \mu + 3\sigma) &= 0,997 & [k = 3]\end{aligned}$$

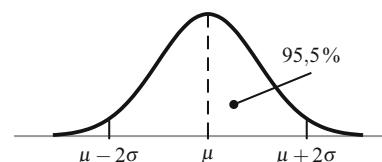
Illustriert ist dies in Abb. 10.5. Eine wichtige Tatsache zu diesem Modell ist: Werte, die mehrere Standardabweichungen weg vom Erwartungswert liegen, sind sehr unwahrscheinlich. Ihre Wahrscheinlichkeiten nehmen *exponentiell* in Funktion der Standardabweichung ab. So hat man

$$\mathbb{P}(M > \mu + 4 \cdot \sigma) = 3,2 \cdot 10^{-5}, \quad \mathbb{P}(M > \mu + 20 \cdot \sigma) = 2,8 \cdot 10^{-89}$$

Es besteht damit die Gefahr, die *Plausibilität extremer Werte stark zu unterschätzen*, wenn man Messwerte mit der Normalverteilung modelliert! Bei Produktionen, die unter statistischer Kontrolle liegen, sind solche extremen Werte hingegen kaum erwartbar. Bei Grössen aus der Ökonomie, wie Schadenssummen, Wartezeiten, möglichen Gewinnen, Preisen von komplexen Finanzprodukten oder bei Daten aus der Meteorologie, wie Geschwindigkeiten von Windböen oder Regenintensitäten, ist es nicht immer sinnvoll anzunehmen, dass die Messungen oder Beobachtungen mit einem Modell mit endlicher Varianz oder endlichem Erwartungswert beschrieben werden können. Hier führen andere Annahmen mit dem Prinzip der maximalen Entropie zu Potenz-, Exponential- oder Extremwertverteilungen.

Die Normalverteilung wird in Produktionsprozessen oft eingesetzt. So will man sicherstellen, dass die verlangte Qualität eines Produkts erfüllt wird. Dabei gilt:

**Abb. 10.5** Im Bereich  $\mu \pm 2\sigma$ : Werte mit einer Wahrscheinlichkeit von 0,955



**Tab. 10.1** Spezifikationen aus dem Jahr 2008 für in der Schweiz produzierte Uhren

	Gangabweichung pro Tag		Gangabweichung pro Woche	
	LSL	USL	LSL	USL
Uhr mit Handaufzug	-5 s	+19 s	-15 s	+80 s
Uhr mit Automatik Uhrwerk	-6 s	+20 s	-2 s	+90 s
Uhr mit Gangzertifizierung	-3 s	+10 s	-10 s	+60 s
Uhr mit Quarzwerk	-0,1 s	+0,15 s	-0,7 s	+1,0 s

Qualität bedeutet, dass produzierte Bauteile innerhalb einer geforderten Toleranz liegen. Diese werden mit einer oberen und einer unteren *Spezifikationsgrenze* USL und LSL (engl. *upper specification limit* und *lower specification limit*) definiert. Die beiden Grenzen nennt man auch die *Toleranzgrenzen* UTG und OTG.

**Beispiel 10.2 (Uhrwerke)** Uhrwerke werden einreguliert, entmagnetisiert und kontrolliert. Wegen Produktionsschwankungen ist es aber nicht möglich, dass Uhren völlig exakt laufen. Gangabweichungen sind normal und in der Produktion Normen unterworfen. Tab. 10.1 zeigt Spezifikationen aus dem Jahr 2008. Uhren mit Gangzertifizierung, sogenannte Chronometer, erfüllen die Spezifikationsgrenzen unter genau vorgegebenen Versuchsbedingungen, die im DIN 8319 oder im ISO 3159 Reglement fixiert sind. □

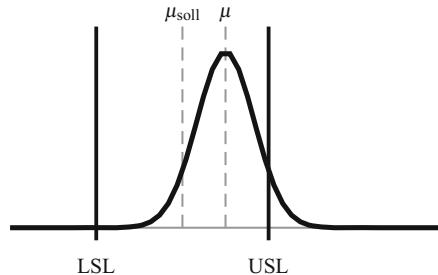
**Beispiel 10.3 (WC-Papierrollen)** Eine schweizerische Ladenkette verkaufte im Jahr 2009 WC-Papierrollen der Marke „Soft Comfort“. Auf den Verkaufspackungen waren die in Tab. 10.2 dargestellten Spezifikationsgrenzen für die Rollenlänge und die Qualität des Papiers angegeben. □

Ein unter statistischer Kontrolle liegender Produktionsprozess kann nur dann wirtschaftlich erfolgreich sein, wenn die meisten produzierten Elemente innerhalb der Spezifikationsgrenzen liegen. Ausserhalb der Spezifikationsgrenzen anfallende Bauteile führen nämlich zu unnötigen Kosten: Das Material der Ausschussware, ihre Kontrolle durch Personen und die für die Ausschussware benötigte Arbeitszeit sind nicht gratis. Was ein *produktionsfähiger Prozess* ist, wird mit einem Wahrscheinlichkeitsmodell festgesetzt, das prognostiziert, wo zukünftige Messwerte der Produktionsgrösse liegen werden. Im Bereich des TQM (engl. *Total quality management*) setzt man voraus, dass das Datenmodell

**Tab. 10.2** Spezifikation zur WC-Papierrollen (Migros 2009)

	LSL	USL
Rollenlänge	26,2 m	27,8 m
Dichte Papier	48,5 g/m <sup>2</sup>	53,5 g/m <sup>2</sup>

**Abb. 10.6** Produktionsszenario mit tendenziell zu grossen Werten



für die Messwerte einen endlichen Erwartungswert und eine endliche Standardabweichung hat. Nach MaxEnt wird dieses Wissen mit der Normalverteilung plausibilisiert. Abb. 10.6 zeigt ein mögliches Szenario. Beim dargestellten Szenario fallen die Messwerte tendenziell zu gross aus: die Produktion ist nicht bei der Sollgrösse  $\mu_{\text{soll}}$  zentriert, sondern liegt bei  $\mu$  und ist leicht höher.

Das „Sechs Sigma“ Programm betrachtet beim Normalverteilungsmodell zwei bis drei Parameter, um zu quantifizieren, wie gross der Anteil der Produktion innerhalb der Spezifikationsgrenzen liegen wird. Es ist ein Instrument aus der Prozesskontrolle und arbeitet mit Prozessfähigkeiten. Die Firma Motorola benutzt das Instrument und erhielt dafür 1988 den Malcolm Baldrige National Quality Award (siehe [2]). Seither ist das „Sechs Sigma“ Programm sehr beliebt. Die zitierte *Prozessfähigkeit* – oder das *Prozesspotential* –  $C_p$  (engl. *process capability*) misst die Standardabweichung  $\sigma$  des Prozesses in Relation zur Spezifikationsbreite:

$$C_p = \frac{\text{Spezifikationsbreite der Produktionsgrösse}}{6 \cdot \sigma} = \frac{\text{USL} - \text{LSL}}{6 \cdot \sigma}$$

Je kleiner die Prozessfähigkeit  $C_p$  ist, desto weniger effizient ist ein Produktionsprozess. Der Parameter  $C_p$  ist nämlich klein, wenn die verlangte Spezifikation der Produktionsgrösse klein ist und/oder ihre Streuung gross ist. Oft gilt eine unter statistischer Kontrolle liegende Produktionsgrösse als prozessfähig, wenn die Streuung  $\sigma$  *mindestens sechsmal* innerhalb der Spezifikationsgrenzen Platz hat. Dies bedeutet, dass  $C_p > 1$  ist! Bei der Firma Motorola, der Begründerin des  $6\sigma$ -Qualitätsmanagements, wird verlangt, dass ein Produktionsprozess eine Prozessfähigkeit von mindestens zwei aufweist. Die Breite des Toleranzbereichs beträgt  $\pm 6 \cdot \text{Standardabweichung}$ , daher wird dies als Sechs Sigma bezeichnet (siehe [1]). Die unteren und oberen *kritischen Prozessfähigkeiten* messen zusätzlich, wie weit der Produktionsschwerpunkt von den beiden Spezifikationsgrenzen entfernt ist:

$$C_{pkl} = \frac{\mu - \text{LSL}}{3 \cdot \sigma}, \quad C_{pku} = \frac{\text{USL} - \mu}{3 \cdot \sigma}$$

Aus den beiden Werten kann man berechnen, wie gross die (durchschnittlich erwartbare) Anzahl ppm pro eine Million produzierter Ware ist, welche nicht innerhalb der Spezifika-

tionsgrenzen liegt. Ist  $\text{pdf}_{\text{normiert}}(x)$  die Dichtefunktion der Normalverteilung mit  $\mu = 0$  und  $\sigma = 1$ , so ist

$$\text{ppm} = \left[ 1 - \int_{-3C_{pkl}}^{3C_{pku}} \text{pdf}_{\text{normiert}}(x) dx \right] \cdot 10^6$$

Ist  $C_{pkl} = C_{pku} = 1,5$ , so erhält man  $\text{ppm} = 6,8$ . Auf eine Million produzierter Ware erfüllen damit etwa sieben Elemente die Spezifikation nicht.

Der kleinere der beiden Werte  $C_{pkl}$  und  $C_{pku}$  wird als die kritische Prozessfähigkeit  $C_{pk}$  bezeichnet. Die kritische Prozessfähigkeit ist am kleinsten, wenn der Produktionsschwerpunkt  $\mu$  in der Mitte des Spezifikationsbandes liegt. Bei der Firma Motorola wird verlangt, dass der  $C_{pk}$ -Index mindestens 1,5 ist. Dies bedeutet: Der Schwerpunkt der Produktionsgrösse ist mindestens  $4,5 \cdot \sigma$  von den Spezifikationsgrenzen entfernt.

Bei  $C_p \geq 2$  und  $C_{pk} \geq 1,5$  sind höchstens 3,4 Defekte pro eine Million (ppm) produzierter Waren zu erwarten. Ein solcher Prozess wird bei der Firma Motorola als  $6\sigma$ -fähiger Prozess definiert. Eine derart tiefe Defektquote scheint übertrieben. Man stelle sich aber vor, dass ein Produkt aus 100 Komponenten bestehe und das Produkt nur funktioniere, wenn alle Komponenten nicht defekt sind. Ist die Produktion so eingestellt, dass „nur“ 93,32 % der Komponenten in Ordnung sind, dann beträgt die Wahrscheinlichkeit, dass ein Produkt aus der Produktion funktioniert gemäss der Multiplikationsregel nur noch

$$0,9332 \cdot 0,9332 \cdot \dots \cdot 0,9332 = (0,9332)^{100} = 0,00099$$

Die Produktion wird unbrauchbar. Bei einem  $6\sigma$ -fähigen Prozesses hat man:

$$(0,99999660)^{100} = 0,99966$$

Ein anderer Ansatz des Qualitätsmanagements besteht darin, einen Produktionsprozess über seine Kosten zu steuern. Wird ein Produkt ausserhalb der Spezifikationen geliefert, fallen Reparaturkosten an. Die Reparaturkosten der Gesamtproduktion sind umso höher, je weiter der Schwerpunkt  $\mu$  der Produktion vom Sollschwerpunkt  $\mu_{\text{soll}}$  entfernt ist und je höher die Standardabweichung der Produktion ist. Beides erhöht nämlich, die Wahrscheinlichkeit defekte Ware zu produzieren. Der japanische Ingenieur G. Taguchi modelliert die Reparaturkosten  $K$  mit

$$K \propto (\mu - \mu_{\text{soll}})^2 + \sigma^2$$

Man nennt diesen Ausdruck die (quadratische) *Risiko-Funktion* (engl. *risc-function*) der Produktion. Das Ziel des Qualitätsmanagements ist es dann, die Kosten K möglichst klein zu halten, also den Schwerpunkt der Produktion nahe beim Sollwert und gleichzeitig die Standardabweichung  $\sigma$  klein zu halten.

## 10.3 Normalverteilung als Datenmodell

Wie im obigen Abschnitt beschrieben, nimmt man oft an, dass Messwerte aus der Technik oder der Produktionstechnik mit einer Normalverteilung beschrieben werden können.<sup>5</sup> Wie man daraus den Modus des Datenmodells berechnet oder weitere Messwerte prognostiziert, zeigt das folgende Beispiel:

**Beispiel 10.4 (Chloridgehalt)** Um den Chloridgehalt einer Kalium-Chlorid-Lösung zu bestimmen, kann man die Chloridionen mit Silberionen ausfällen. Dabei werden die Messwerte streuen, da Unsicherheiten in den Messapparaten bestehen und Kovariablen<sup>6</sup> vorhanden sind. Hier sieben Messwerte (in mol/m<sup>3</sup>) aus dem Labor Chemie der Berner Fachhochschule:

102,8    103,3    102,3    103,6    104,3    101,5    102,1

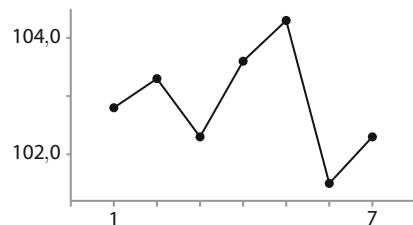
Das Streudiagramm in Abb. 10.7 zeigt, dass die Messwerte unter statistischer Kontrolle sind. Es gibt kaum Trends nach höheren oder tieferen Messwerten, aussergewöhnlich hohe oder tiefe Messwerte sind nicht vorhanden. Die Daten scheinen unabhängig zu sein.

Ein Chemiker möchte aus diesen Daten die folgenden zwei Fragen beantworten:

- (a) Wie lautet der (nicht direkt messbare) Chloridgehalt Cl der Lösung?
- (b) Wo liegen weitere, mit derselben Methode ermittelte Messwerte mit hoher Wahrscheinlichkeit?

Um die Fragen zu beantworten, braucht man ein Datenmodell. Dieses besagt, wie die (kontinuierlichen) Messwerte um den gesuchten Chloridgehalt streuen. Wir nehmen an, dass das Datenmodell eine endliche Standardabweichung  $\sigma$  haben soll. Mit MaxEnt heisst dies: Am sinnvollsten ist es anzunehmen, dass die Messwerte normalverteilt um den Chloridgehalt Cl variieren. Abb. 10.8 zeigt das Modell für die Messwerte. Die Wahrscheinlichkeit, einen Messwert  $x$  zu erhalten, kann deshalb mit der Dichtefunktion der

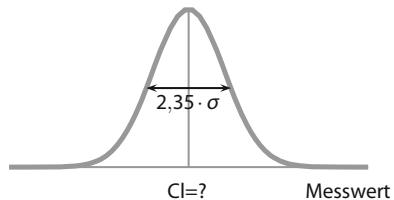
**Abb. 10.7** Streudiagramm der sieben Messwerte



<sup>5</sup> Zumindest, wenn der Prozess unter statistischer Kontrolle ist.

<sup>6</sup> Zum Beispiel die Verteilung der Chloridionen oder Temperaturschwankungen in der Lösung

**Abb. 10.8** Datenmodell: So streuen die Messwerte, wenn der Chloridgehalt bekannt wäre; gesucht ist aber Cl



Normalverteilung beschrieben werden:

$$\text{pdf}(\text{Messwert} = x \mid \text{Cl und } \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp \left\{ -0,5 \cdot \left( \frac{x - \text{Cl}}{\sigma} \right)^2 \right\} \quad (10.1)$$

Mit der Regel von Bayes kann man nun die Parameter Cl und  $\sigma$  berechnen. Der hier wichtige Parameter ist der gesuchte, nicht direkt messbare Chloridgehalt Cl. Der Parameter  $\sigma$  ist für den Chemiker weniger interessant. Man sagt auch, dass  $\sigma$  ein *Störparameter* (engl. *nuisance parameter*) ist. Da die beiden Parameter kontinuierliche Werte annehmen können, werden sie mit einem stetigen Wahrscheinlichkeitsmodell beschrieben. Ihre gemeinsame Dichtefunktion lautet mit der Regel von Bayes

$$\underbrace{\text{pdf}(\text{Cl}, \sigma \mid \text{Daten, } \mathcal{K})}_{\text{A posteriori Wissen zu Parametern}} \propto \underbrace{\mathbb{P}(\text{Daten} \mid \text{Cl}, \sigma)}_{\text{Likelihood}} \cdot \underbrace{\text{pdf}(\text{Cl}, \sigma \mid \mathcal{K})}_{\text{Prior}}$$

Dabei ist  $\mathcal{K}$  die vorhandene Vorinformation oder das Vorwissen. Sie sei minimal zum Chloridgehalt Cl (der sicherlich zwischen 50 und 150 mol/m<sup>3</sup> ist) und zur Streuung  $\sigma$  (die zwischen 0,1 und 20 mol/m<sup>3</sup> ist). Der Parameter Cl ist ein Lageparameter. Die Streuung  $\sigma$  beschreibt, wie breit die Glockenkurve ist. Sie ist ein Skalierungsparameter. Theorem 8.2 führt dazu, die A priori-Plausibilität zum Chloridgehalt gleichverteilt und die Standardabweichung  $\sigma$  mit der Dichtefunktion  $1/\sigma$  zu beschreiben. Der Prior lautet deshalb

$$\text{pdf}(\text{Cl}, \sigma \mid \mathcal{K}) = \text{pdf}(\text{Cl}, \sigma \mid \text{min. Vorinformation}) \propto 1 \cdot \frac{1}{\sigma}$$

Die Likelihood-Funktion bestimmt man mit dem Datenmodell für die Messwerte. Bei unabhängigen sieben Messwerten Cl<sub>1</sub>, Cl<sub>2</sub>, ..., Cl<sub>7</sub> hat man mit dem Multiplikationsgesetz

$$\mathbb{P}(\text{Daten} \mid \text{Cl}, \sigma) = \mathbb{P}(\text{Cl}_1 \mid \text{Cl}, \sigma) \cdot \mathbb{P}(\text{Cl}_2 \mid \text{Cl}, \sigma) \cdot \dots \cdot \mathbb{P}(\text{Cl}_7 \mid \text{Cl}, \sigma)$$

Die einzelnen Wahrscheinlichkeiten auf der rechten Seite der Gleichung sind proportional zu Dichtefunktion des Datenmodells. Die Messwerte eingesetzt, erhält man mit Gleichung (10.1) für die Likelihood:

$$\begin{aligned} \mathbb{P}(\text{Daten} \mid \text{Cl}, \sigma) &\propto \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right)^7 \exp \left\{ -\frac{1}{2} \left( \frac{102,8 - \text{Cl}}{\sigma} \right)^2 \right\} \dots \\ &\dots \cdot \exp \left\{ -\frac{1}{2} \left( \frac{102,1 - \text{Cl}}{\sigma} \right)^2 \right\} \end{aligned}$$

Die sieben Faktoren mit der Exponentialfunktion lassen sich zusammenfassen:

$$\mathbb{P}(\text{Daten} \mid \text{Cl}, \sigma) \propto \frac{1}{\sigma^7} \cdot \exp \left\{ -\frac{1}{2} \frac{(102,8 - \text{Cl})^2 + \dots + (102,1 - \text{Cl})^2}{\sigma^2} \right\}$$

Im Argument der Exponentialfunktion befindet sich die Summe der Quadratabweichungen von den Messwerten zum gesuchten Chloridgehalt Cl, dividiert durch die Varianz  $\sigma^2$  des Datenmodells. Diese Summe taucht in vielen Rechnungen auf. Sie wird deshalb meist abgekürzt mit  $\chi^2$  geschrieben:

$$\chi^2 = \frac{(102,8 - \text{Cl})^2 + (103,3 - \text{Cl})^2 + \dots + (102,1 - \text{Cl})^2}{\sigma^2}$$

Damit sind alle Bestandteile der Regel von Bayes zusammengestellt. Die Plausibilität zum Chloridgehalt und zur Streuung  $\sigma$  ist durch die gemeinsame A posteriori-Verteilung bestimmt:

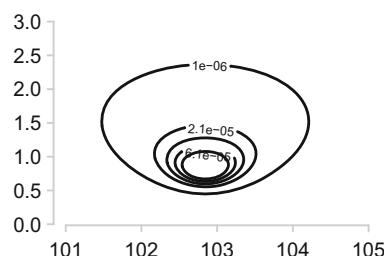
$$\text{pdf}(\text{Cl}, \sigma \mid \text{Daten, min. Vor.}) \propto \frac{1}{\sigma^7} \cdot \exp \{-0,5 \cdot \chi^2\} \cdot \frac{1}{\sigma} \quad \text{für } 50 \leq \text{Cl} \leq 150$$

Visualisieren kann man diese Verteilung mit ihren Höhenlinien (siehe Abb. 10.9). Die horizontale Achse steht für den Chloridgehalt, die vertikale Achse für die Standardabweichung  $\sigma$ . Der Graph der Dichtefunktion hat eine Spitze beim Modus. Er befindet sich für Cl dort, wo das Argument in der Exponentialfunktion maximal ist. Wegen des Minuszeichens bedeutet dies, dass  $\chi^2$  minimal sein muss. Dies heißt: Die Summe der Quadrat-Abweichungen der Messwerte zu Cl dividiert durch die Varianz  $\sigma^2$  muss minimal werden. Man sagt, dass man den plausibelsten Wert hier mit der *Methode der kleinsten Quadrate* (engl. *least square method*) bestimmt. Das Minimum ist schnell berechnet. Ableiten des Arguments nach Cl und Null Setzen ergibt:

$$\frac{2 \cdot (102,8 - \text{Cl}) + 2 \cdot (103,3 - \text{Cl}) + \dots + 2 \cdot (102,1 - \text{Cl})}{\sigma^2} = 0$$

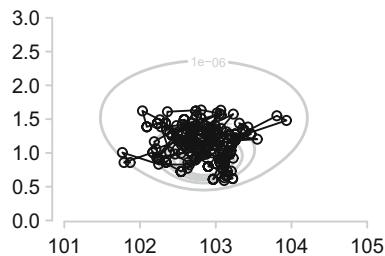
Der plausibelste Wert für den Chloridgehalt Cl ist damit gleich dem arithmetischen Mittel 102,84 mol/m<sup>3</sup> der sieben Messwerte.

**Abb. 10.9** Höhenlinien der gemeinsamen A posteriori-Dichtefunktion von Cl und  $\sigma$



**Abb. 10.10** MCMC-

Simulation: die ersten 200  
Punkte einer Kette von 50 000  
Punkten



Um die Plausibilität zum Chloridgehalt Cl zu bestimmen, muss man den Parameter  $\sigma$  aus der obigen A posteriori-Verteilung eliminieren. Man kann dazu mit Theorem 6.1 arbeiten und das folgende Integral auswerten:

$$\text{pdf(Cl} | \text{Daten, min. Vorinformation}) \propto \int_0^{\infty} \frac{1}{\sigma^7} \cdot \exp \{-0.5 \cdot \chi^2\} \cdot \frac{1}{\sigma} d\sigma$$

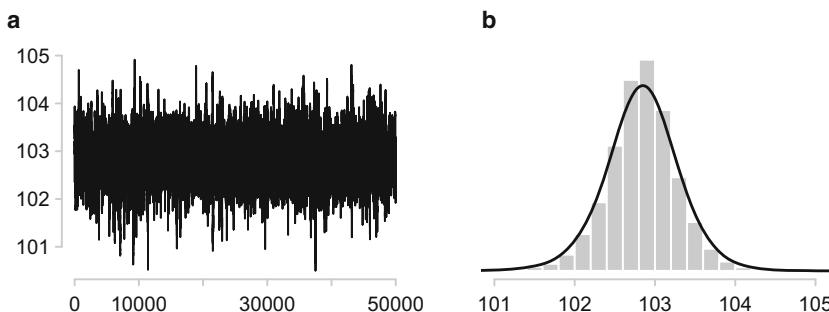
Am einfachsten aber macht man dies nach dem Verfahren von Theorem 6.2. Man konstruiert mit einer MCMC-Simulation Punkte der gemeinsamen A posteriori-Verteilung der beiden Parameter Cl und  $\sigma$ . Die simulierten Cl-Koordinaten liefern die Plausibilität zu Cl. Mit einem Statistikprogramm kann dies durchgeführt werden. Man gibt die Messwerte ein und nennt das Datenmodell sowie die A priori-Verteilungen der Parameter Cl und  $\sigma$ :

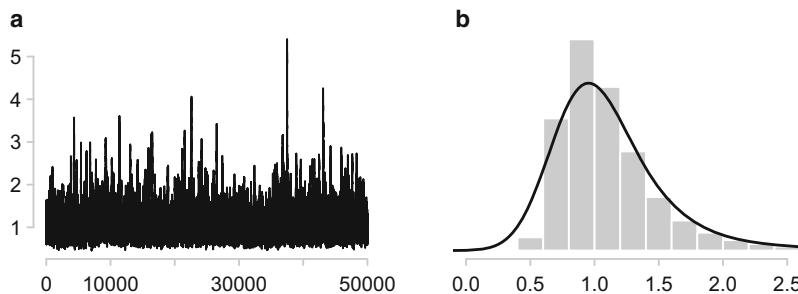
Datenmodell:  $i$ -ter Messwert  $\sim \text{Normal(Cl}, \sigma)$

Prior:  $\text{Cl} \sim \text{Uniform}(50 ; 150)$

Prior:  $\ln \sigma \sim \text{Uniform}(\ln(0,1) ; \ln(20))$

Daraus bildet das Programm die A posteriori-Verteilung der Parameter Cl und  $\sigma$  mit der Formel Likelihood  $\times$  Prior. Abb. 10.10 zeigt die ersten 200 Punkte einer Kette von 50 000 Punkten mit einer Akzeptanzrate von 63 %.

**Abb. 10.11** MCMC-Kette der A posteriori-Verteilung: **a** die Cl-Koordinaten, **b** die berechnete Randverteilung von Cl



**Abb. 10.12** MCMC-Kette der A posteriori-Verteilung: **a** die  $\sigma$ -Koordinaten, **b** die berechnete Randverteilung von  $\sigma$

Abb. 10.11 zeigt die gesamte Kette. Links sind die Cl-Koordinaten der vollständigen Kette dargestellt. Rechts ist die mit der Kette berechnete Randverteilung von Cl visualisiert. Sie ist unimodal symmetrisch.<sup>7</sup> Es besteht eine Wahrscheinlichkeit von 0,5, dass der Chloridgehalt zwischen 102,6 mol/m<sup>3</sup> und 103,1 mol/m<sup>3</sup> liegt. Mit einer Wahrscheinlichkeit von 0,95 befindet er sich zwischen 101,9 mol/m<sup>3</sup> und 103,7 mol/m<sup>3</sup>.

Auch die  $\sigma$ -Koordinaten der Kette lassen sich visualisieren (siehe Abb. 10.12). Damit wird plausibel, in welchem Bereich die (unwichtige) Standardabweichung liegt. Die Verteilung ist unimodal schief. Der Modus – der „wahrscheinlichste“ Wert – ist  $\sigma_0 = 0,89$  mol/m<sup>3</sup>. Aus den 25 % kleinsten und grössten Werten der simulierten  $\sigma$ -Werte erhält man ein Wahrscheinlichkeitsintervall von 0,5. Es besteht aus den Zahlen 0,84 mol/m<sup>3</sup> und 1,26 mol/m<sup>3</sup>. Ein Wahrscheinlichkeitsintervall für  $\sigma$  zum Niveau 0,95 setzt sich aus dem 2,5 % kleinsten Wert (0,62 mol/m<sup>3</sup>) und 2,5 % grössten Wert (2,07 mol/m<sup>3</sup>) der simulierten  $\sigma$ -Werte zusammen. Damit ist die Frage (a) nach dem Chloridgehalt (und der Streuung) beantwortet.

Die Frage (b), in welchem Bereich weitere mit der gleichen Methode ermittelte Messwerte liegen werden, kann man nun beantworten. Dies lässt sich nach Theorem 7.3 durchführen. Man mittelt das Datenmodell über den Posterior der Parameter Cl und  $\sigma$ . Die dazu benutzte Monte-Carlo-Simulation ist in Abb. 10.13 dargestellt. 50 000 Werte der Parameter Cl und  $\sigma$  hat man aus der MCMC-Simulation für die gemeinsame A posteriori-Verteilung der beiden Parameter. Hier ein Pseudo-Code, der die Simulation illustriert:

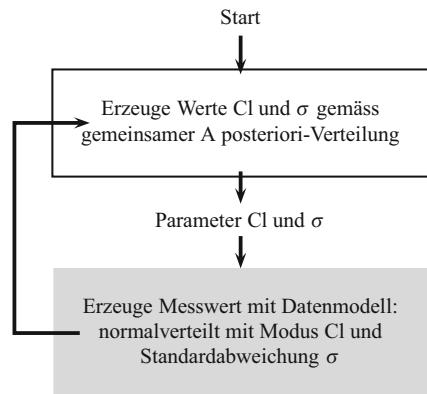
```

for i = 1 to 50000 do
    (Cl[i], Sigma[i]) = i-ter Punkt aus MCMC-Simulation
    für Cl und sigma
    Messwert[i] = erzeugt aus Normalverteilung, mit
    Modus Cl[i] und Streuung Sigma[i]
end

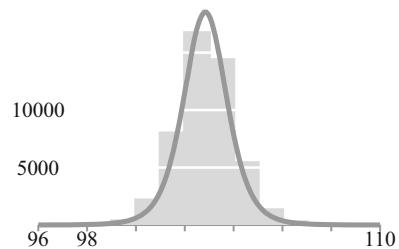
```

<sup>7</sup> Im Gegensatz zur Normalverteilung strebt der Posterior für Cl nicht exponentiell schnell gegen Null. Man hat bei den sieben Messungen  $\text{pdf}(\text{Cl} \mid \text{Daten}) \propto \text{Cl}^{-7}$ , wenn Cl gross ist. Damit sind auch entfernte Werte vom Modus viel plausibler als beim Modell der Normalverteilung. Siehe dazu den Anhang A.

**Abb. 10.13** Monte-Carlo-Simulation für das Prognosemodell



**Abb. 10.14** Prognosemodell für den nächsten Messwert



In Abb. 10.14 ist die Verteilung der 50 000 simulierten Messwerte dargestellt. Es stellt das Prognosemodell für weitere, mit derselben Methode ermittelte Chloridwerte dar. Ein Beispiel dazu: 31 370 der simulierten Messwerte liegen zwischen 102 und 104 mol/m<sup>3</sup>. Die Wahrscheinlichkeit, dass der nächste Messwert in diesem Bereich liegen wird, ist daher

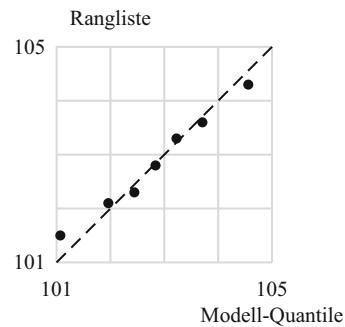
$$\mathbb{P}(102 \text{ mol/m}^3 \leq \text{Messwert} \leq 104 \text{ mol/m}^3 | \text{Daten}) \approx \frac{31\,370}{50\,000} = 0,63$$

Damit ist auch die Frage (b) beantwortet.

Die durchgeführten Rechnungen hängen vom gewählten Datenmodell, der Normalverteilung, ab. Es ist daher – wie schon bei den Datenmodellen zur Exponential- und Poissonverteilung von Kap. 9 – empfehlenswert, das Modell kritisch zu begutachten. Man kann dies mit dem QQ-Plot (Theorem 9.3) tun. Man hat sieben Messwerte. Der QQ-Plot besteht darin, die Rangliste der Messwerte gegen die 0,5/7-, 1,5/7-, ..., 6,5/7-Quantile des Prognosemodells zu zeichnen. Bei 50 000 simulierten Werten des Prognosemodells ist das 0,5/7-Quantil der  $0,5/7 \cdot 50\,000 = 3571$ . kleinste simulierte Wert. Vergleicht man diese Quantile mit der Rangliste der sieben Messwerte, erhält man

Prognose-Quantil	101,1	102,0	102,4	102,8	103,2	103,7	104,6
Rangliste	101,5	102,1	102,3	102,8	103,3	103,6	104,3

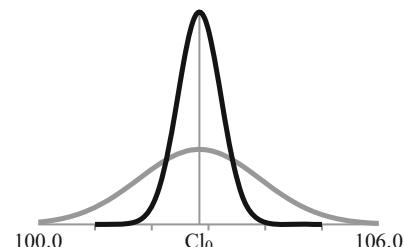
**Abb. 10.15** QQ-Plot der sieben Messwerte mit dem Datenmodell der Normalverteilung



Die Werte sind in Abb. 10.15 dargestellt. Das Modell scheint gut. Weil wenig Messungen vorliegen, ist der QQ-Plot leider wenig aussagekräftig.  $\square$

**Eine Klarstellung** Im obigen Beispiel arbeitet man mit verschiedenen Wahrscheinlichkeitsmodellen. Das erste Modell ist das *Datenmodell*. Es ist eine Normalverteilung. Es beschreibt, wie Messungen um den Chloridgehalt Cl streuen würden, wenn Cl und  $\sigma$  bekannt wären. Aus den Daten und dem Datenmodell wird die *A posteriori-Verteilung* für den Chloridgehalt Cl und die Standardabweichung  $\sigma$  (zwei nicht direkt messbare Größen) berechnet. Mit einer Simulation oder dem Gesetz der Marginalisierung erhält man aus dem Datenmodell und der A posteriori-Verteilung das *Prognosemodell*. Es besagt, wo weitere Messwerte (zukünftige Werte einer unsicheren Größe) liegen werden. Die A posteriori-Verteilung und das Prognosemodell sind keine Normalverteilungen. Wichtig ist, die A posteriori-Verteilung für den Lageparameter Cl nicht mit dem Prognosemodell für weitere Messungen zu verwechseln. So weiß man, dass der Chloridgehalt zwischen 101 mol/m<sup>3</sup> und 105 mol/m<sup>3</sup> liegt. Das Prognosemodell ist viel „breiter“: Weitere Messwerte können kleiner als 101 mol/m<sup>3</sup> sein. Sie liegen mit hoher Wahrscheinlichkeit zwischen 98 und 108 mol/m<sup>3</sup>. Um dies zu verdeutlichen, sind die beiden Wahrscheinlichkeitsmodelle in Abb. 10.16 in einem gemeinsamen Bild dargestellt. Mit einer grauen Linie ist das Prognosemodell für weitere Messwerte und in schwarzer Farbe die Plausibilität zum Chloridgehalt gezeichnet. Beachten Sie: Je mehr unabhängige Messwerte vorhanden sind, umso präziser kann der Chloridgehalt berechnet werden. Ist  $n$  die Anzahl Messungen, so

**Abb. 10.16** Plausibilität zum Chloridgehalt Cl in schwarzer Farbe, Prognosemodell für weitere Messungen in grauer Farbe



nimmt die Breite der schwarzen Kurve proportional zu  $1/\sqrt{n}$  ab. Die graue Kurve bleibt breit. Mit zunehmendem  $n$  wird sie nie schmäler als die Streuung  $\sigma$  des Datenmodells.

Das obige Beispiel zeigt allgemein:

**Theorem 10.2 (Die Normalverteilung: Parameter berechnen und Messwerte prognostizieren)**

Gegeben seien  $x_1, x_2, \dots, x_n$  unabhängig modellierbare kontinuierliche Datenwerte, die unter statistischer Kontrolle sind. Man nimmt an, dass sie um einen Wert  $\mu$  mit endlicher Streuung  $\sigma$  variieren. Mit MaxEnt beschreibt man die Plausibilität dazu mit der Normalverteilung:

$$\text{Datenmodell: } i\text{-ter Messwert} \sim \text{Normal}(\mu, \sigma)$$

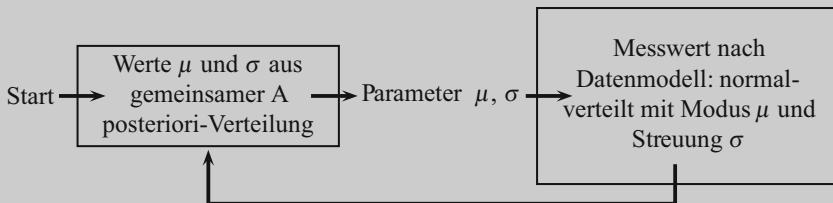
(A) Aus Vorinformation  $\mathcal{I}$  lässt sich mit den Datenwerten die Plausibilität zu  $\mu$  und zu  $\sigma$  aktualisieren:

$$\underbrace{\text{pdf}(\mu, \sigma \mid \text{Daten}, \mathcal{I})}_{\text{Posterior}} \propto \underbrace{\frac{1}{\sigma^n} \cdot \exp(-0,5 \cdot \chi^2)}_{\text{Likelihood}} \cdot \underbrace{\text{pdf}(\mu, \sigma \mid \mathcal{I})}_{\text{Prior}}$$

Dabei ist

$$\chi^2 = \frac{(x_1 - \mu)^2 + \dots + (x_n - \mu)^2}{\sigma^2}$$

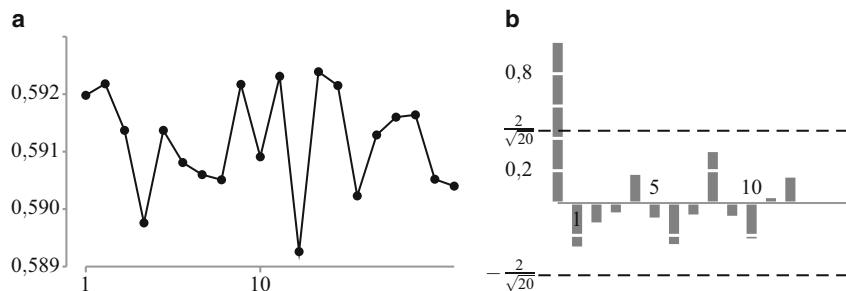
(B) Weitere Messwerte können mit einer Simulation prognostiziert werden:



Der Ausdruck  $\chi^2$  lässt sich mit dem arithmetischen Mittel  $\bar{x}$  und der empirischen Standardabweichung  $s$  der Messwerte kürzer schreiben:

$$\chi^2 = \frac{(n-1) \cdot s^2 + n \cdot (\mu - \bar{x})^2}{\sigma^2}$$

Die Likelihood hängt nur von  $\bar{x}$  und  $s$  ab. Diese zwei Zahlen genügen also, um sie vollständig zu beschreiben. Man sagt, dass  $\bar{x}$  und  $s$  suffizient für die Normalverteilung sind.



**Abb. 10.17** Streudiagramm (a) und Graph der Autokorrelationsfunktion (b) der Druckmesswerte aus der Vakuumkammer

Bei manchen Untersuchungen hat man außer den Daten nur minimale Vorinformation zu den Parametern  $\mu$  und  $\sigma$  der Normalverteilung. Die Verteilungen der Parameter und die Prognoseverteilung können dann explizit berechnet werden. Die erhaltenen Formeln finden sich in Anhang A in Abschn. A.3. Insbesondere ist der plausibelste Wert von  $\mu$  gleich dem arithmetischen Mittel der Datenwerte.

Hat man viele Datenwerte, so wird die Likelihood im Ausdruck für den Posterior von  $\mu$  und  $\sigma$  dominant. Der Prior beeinflusst dann die Plausibilität zu den beiden Parameter kaum. In diesem Fall lassen sich Wahrscheinlichkeitsintervalle zu  $\mu$  und  $\sigma$  mit einfachen Formeln angeben. So ist mit einer Wahrscheinlichkeit von etwa 0,95:<sup>8</sup>

$$\mu = \bar{x} \pm 1,96 \cdot \text{SE}(\text{Normal}, \mu) = \bar{x} \pm 1,96 \cdot \frac{s}{\sqrt{n}} \quad (10.2)$$

Dabei ist  $\bar{x}$  das arithmetische Mittel und  $s$  die empirische Standardabweichung der  $n$  unabhängigen Datenwerte. Man nennt dies die Gleichung von *de Moivre* und den Term  $s/\sqrt{n}$  den *geschätzten Standardfehler* (engl. *estimated standard error*) des arithmetischen Mittels. Weiter ist mit einer Wahrscheinlichkeit von etwa 0,95:

$$\sigma = s \pm 1,96 \cdot \text{SE}(\text{Normal}, \sigma) = s \pm 1,96 \cdot \frac{s}{\sqrt{2 \cdot n}} \quad (10.3)$$

**Beispiel 10.5 (Druck in einer Vakuumkammer)** Abb. 10.17a zeigt das Streudiagramm von zwanzig Messungen, um den Unterdruck in einer Vakuumkammer zu bestimmen. In der Abb. 10.17b findet man den Graphen der Autokorrelationsfunktion. Die Messwerte dazu finden sich in Tab. 1.7. Die Abb. 10.17 zeigt, dass die Messwerte unter statistischer Kontrolle sind, da keine Trends nach höheren oder tieferen Werten vorhanden sind und keine Werte ausserhalb der Kontrollgrenzen von 0,588 bar und 0,594 bar liegen. Es ist

<sup>8</sup> Man benutzt das Verfahren von Laplace, dass in Kap. 14 vorgestellt wird.

plausibel, die Messwerte unabhängig zu modellieren. Der Graph der Autokorrelationsfunktion weist darauf hin, die Daten trendfrei und unabhängig sind.

Um den Unterdruck  $p$  in der Vakuumkammer aus den Messwerten zu berechnen, nimmt eine Ingenieurin an, dass die (kontinuierlichen) Messwerte um  $p$  streuen und dass sie eine endliche Standardabweichung  $\sigma$  haben. Mit MaxEnt beschreibt sie daher ihr Wissen durch das Datenmodell

$$i\text{-ter Messwert Unterdruck} \sim \text{Normal}(p, \sigma)$$

Das arithmetische Mittel  $\bar{p}$  der zwanzig Messwerte beträgt 0,59117 bar. Die empirische Standardabweichung  $s$  der Messwerte ist 0,0009 bar. Die Ingenieurin besitzt keine weitere Information zu  $p$  und  $\sigma$ . Somit ist die A posteriori-Verteilung für den Unterdruck  $p$  und die Streuung  $\sigma$

$$\text{pdf}(p, \sigma \mid \text{Daten, min. Vorinformation}) \propto \frac{1}{\sigma^{20}} \cdot \exp(-0,5 \cdot \chi^2) \cdot \frac{1}{\sigma}$$

Dabei ist

$$\chi^2 = \frac{(20 - 1) \cdot s^2 + 20 \cdot (\mu - \bar{p})^2}{\sigma^2} = \frac{1,542 \cdot 10^{-5} + 20 \cdot (\mu - 0,59117)^2}{\sigma^2}$$

Mit einer MCMC-Simulation erhält man: Ein Wahrscheinlichkeitsintervall zum Niveau 0,95 für  $p$  ist

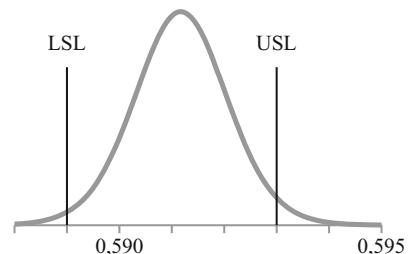
$$p = (0,5912 \pm 0,004) \text{ bar}$$

Ein Wahrscheinlichkeitsintervall zum Niveau von 0,95 für  $\sigma$  ist

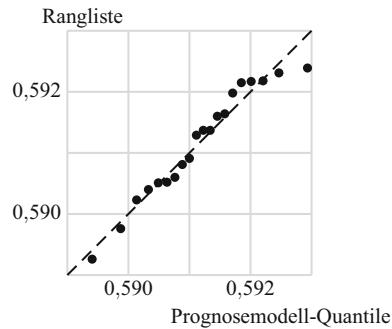
$$\sigma = (0,0009 \pm 0,0003) \text{ bar}$$

Die Hersteller der Unterdruckkammer wollen, dass mit der gleichen Methode *gemessene* Unterdrücke mit grosser Wahrscheinlichkeit innerhalb der Spezifikationsgrenzen von LSL = 0,589 bar und USL = 0,593 bar liegen. Wie gross ist diese Wahrscheinlichkeit? Um dies zu bestimmen, braucht man die Verteilung für weitere Messwerte. Dies ist das Prognosemodell. Man erhält es mit dem Verfahren von Theorem 10.2. In Abb. 10.18 ist

**Abb. 10.18** Prognosemodell für zukünftige Messwerte in der Vakuumkammer



**Abb. 10.19** QQ-Plot der Druckmesswerte beim Datenmodell der Normalverteilung



der Graph dieses Prognosemodells gezeichnet. Es besteht eine Wahrscheinlichkeit von 0,95, dass zukünftige Messwerte  $p_{\text{zuk}}$  im folgenden Bereich sind:

$$p_{\text{zuk}} = (0,591 \pm 0,002) \text{ bar}$$

Die erhaltenen Resultate basieren auf der Annahme, dass das Wissen zu möglichen Messwerten mit der Normalverteilung modellierbar ist. Es ist daher empfehlenswert, das Modell der Normalverteilung mit dem QQ-Plot zu beurteilen. Wie üblich zeichnet man dazu die Rangliste der Messwerte gegen die entsprechenden Quantile des Prognosemodells auf. Der QQ-Plot in Abb. 10.19 zeigt, dass das Modell der Normalverteilung gut ist.  $\square$

**Beispiel 10.6 (Kontrolle „Produktion von Widerständen“)** Eine Firma produziert Widerstände für Stromkreise. Gemäß ihren Angaben betragen die Toleranzgrenzen bei einem bestimmten Typ von Widerständen:  $6,79 \pm 0,02 \text{ k}\Omega$ . Angenommen wird, dass produzierte Widerstände mit einem Modell beschrieben werden können, das einen endlichen Erwartungswert  $\mu$  und eine endliche Streuung  $\sigma$  hat. Mit MaxEnt setzt man damit:

$$\text{Datenmodell: } i\text{-ter Messwert Widerstand} \sim \text{Normal}(\mu, \sigma)$$

Die Firma will, dass die Prozessfähigkeit  $C_p$  mindestens eins ist. Dies bedeutet:

$$C_p = \frac{\text{USL} - \text{LSL}}{6 \cdot \sigma} = \frac{0,04 \text{ k}\Omega}{6 \cdot \sigma} > 1, \text{ also } \sigma < \frac{0,04 \text{ k}\Omega}{6} = 0,0067 \text{ k}\Omega$$

Bevor dies mit einer Stichprobe überprüft wird, habe man minimale Vorinformation zu den beiden Parametern:

$$\text{Prior: } \mu \sim \text{Uniform}(5 ; 10), \quad \ln \sigma \sim \text{Uniform}(\ln(10^{-5}) ; \ln(10))$$

Mit einer Zufallsauswahl wurden an der Berner Fachhochschule fünf Widerstände aus einer unter statistischer Kontrolle stehenden Produktion gemessen:

$$6,789 \text{ k}\Omega, \quad 6,793 \text{ k}\Omega, \quad 6,787 \text{ k}\Omega, \quad 6,796 \text{ k}\Omega, \quad 6,793 \text{ k}\Omega$$

Das arithmetische Mittel  $\bar{R}$  der Messwerte ist gleich  $6,7916 \text{ k}\Omega$  und die empirische Standardabweichung  $s$  lautet  $0,00358 \text{ k}\Omega$ . Der Posterior für  $\mu$  und  $\sigma$  lautet damit:

$$\text{pdf}(\mu, \sigma \mid \text{Daten, min. Vorinformation}) \propto \frac{1}{\sigma^5} \cdot \exp(-0,5 \cdot \chi^2) \cdot \frac{1}{\sigma}$$

Dabei ist

$$\chi^2 = \frac{(5-1) \cdot s^2 + 5 \cdot (\mu - \bar{R})^2}{\sigma^2} = \frac{5,12 \cdot 10^{-5} + 5 \cdot (\mu - 6,7916)^2}{\sigma^2}$$

Mit einer MCMC-Simulation kann man die A posteriori-Verteilung von  $\sigma$  berechnen. Der plausibelste Wert für  $\sigma$  ist  $0,0032 \text{ k}\Omega$ . Die Wahrscheinlichkeit, dass  $\sigma$  kleiner als  $0,0067 \text{ k}\Omega$  ist, beträgt  $0,89$ . Mit einer Wahrscheinlichkeit von  $0,89$  ist also die Prozessfähigkeit  $C_p$  grösser als eins.

Wie viel Prozent der produzierten Widerstände werden innerhalb der Toleranzgrenzen liegen? Dazu benutzt man das Prognosemodell. Mit der Monte-Carlo-Simulation aus Theorem 10.2 berechnet man

$$\mathbb{P}(\text{LSL} \leq \text{produzierter Widerstand} \leq \text{USL} \mid \text{Daten, min. Vor.}) = 0,9929$$

Über 99 % der Produktion wird die Toleranz erfüllen.  $\square$

In den vorigen Beispielen hat man neben den Daten nur minimale Vorinformation. Das folgende Beispiel zeigt, dass präzisere Vorinformation sehr sinnvoll sein kann, um Parameter möglichst präzis zu bestimmen.

**Beispiel 10.7 (Chloridgehalt)** Die in Beispiel 10.4 vorgestellte Art, mit sieben Messungen

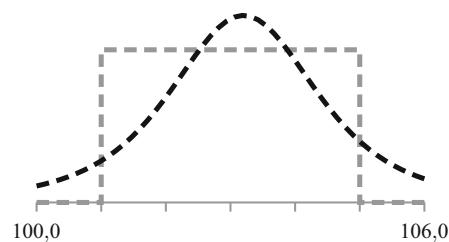
$$102,8 \quad 103,3 \quad 102,3 \quad 103,6 \quad 104,3 \quad 101,5 \quad 102,1$$

den Chloridgehalt einer Kalium-Chlorid-Lösung zu bestimmen, ist aufwändig. Es wird daher ineffizient, den Chloridgehalt mit mehr als sieben Messungen präziser zu berechnen. Eine Chemikerin kann den Chloridgehalt präziser bestimmen, indem sie zuerst den Chloridgehalt mit einer billigen und einfachen Methode ungefähr lokalisiert. Mit dieser Methode wird die Lösung viermal gemessen:

$$105,8 \quad 104,1 \quad 100,2 \quad 102,7$$

Das arithmetische Mittel der Messwerte ist  $103,2 \text{ mol/m}^3$ . Die empirische Standardabweichung beträgt  $2,4 \text{ mol/m}^3$  und ist 2,5-mal grösser als mit der aufwändigeren Methode. Um den Chloridgehalt damit vorweg zu bestimmen, nimmt die Chemikerin als Datenmodell die Normalverteilung. Sie hat weiter keine Information zu Cl und Streuung  $\sigma_1$  der

**Abb. 10.20** A priori-Plausibilität zum Chloridgehalt Cl: *grau* bei minimaler Vorinformation, *schwarz* mit Zusatzinformation aus den vier Messwerten



Methode. Weiter sind die Messwerte unabhängig. Daraus erhält sie als Plausibilität zum Chloridgehalt Cl und  $\sigma_1$ :

$$\text{pdf}(\mu, \sigma_1 \mid \text{Information}) \propto \frac{1}{\sigma_1^4} \cdot \exp(-0,5 \cdot \chi_{\text{vor}}^2) \cdot \frac{1}{\sigma_1}$$

Dabei ist

$$\chi_{\text{vor}}^2 = \frac{(4-1) \cdot 2,4^2 + 4 \cdot (\mu - 103,2)^2}{\sigma_1^2}$$

Abb. 10.20 zeigt den Graphen der mit einer MCMC-Simulation ermittelten Randverteilung von Cl. Die Verteilung dient der Chemikerin nun als Prior, um den Chloridgehalt aus den Messungen mit der aufwändigen Methode mit der Regel von Bayes zu aktualisieren. Mit Theorem 10.2 folgt:

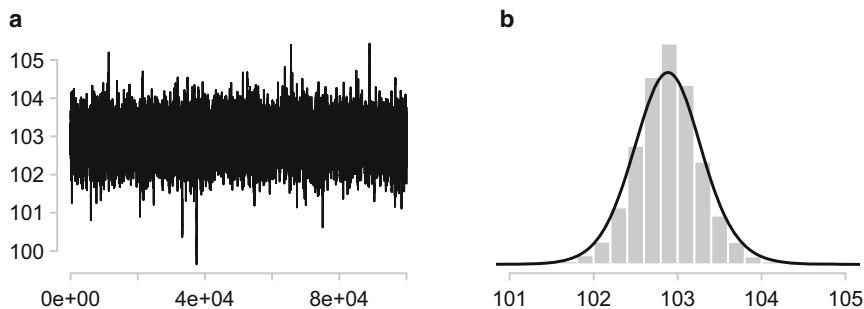
$$\text{pdf}(\text{Cl}, \sigma, \sigma_1 \mid \text{Daten, Infor.}) \propto \underbrace{\frac{1}{\sigma^7} \cdot \exp\{-0,5 \cdot \chi^2\}}_{\text{Likelihood}} \cdot \underbrace{\frac{1}{\sigma_1^4} \cdot \exp(-0,5 \cdot \chi_{\text{vor}}^2) \cdot \frac{1}{\sigma_1} \cdot \frac{1}{\sigma}}_{\text{Prior}}$$

Dabei ist  $\sigma$  die Streuung der aufwändigen Methode und  $\chi^2$  wird mit den neuen Messwerten gebildet:

$$\chi^2 = \frac{(102,8 - \text{Cl})^2 + (103,3 - \text{Cl})^2 + \dots + (102,1 - \text{Cl})^2}{\sigma^2}$$

Mit einer MCMC-Simulation lassen sich Wahrscheinlichkeitsintervalle für Cl bestimmen. In einem Statistikprogramm gibt man dazu die zwei Messreihen ein und nennt alle Datenmodelle sowie A priori-Verteilungen zu den Parametern:

- |                               |   |
|-------------------------------|---|
| Datenmodell erste Messreihe:  | $i$ -ter Messwert Reihe 1 $\sim \text{Normal}(\text{Cl}, \sigma_1)$ |
| Datenmodell zweite Messreihe: | $i$ -ter Messwert Reihe 2 $\sim \text{Normal}(\text{Cl}, \sigma_2)$ |
| Prior:                        | $\text{Cl} \sim \text{Uniform}(101, 105)$                           |
| Prior:                        | $\ln \sigma_1 \sim \text{Uniform}(\ln(10^{-4}); \ln(10^2))$         |
| Prior:                        | $\ln \sigma_2 \sim \text{Uniform}(\ln(10^{-4}); \ln(10^2))$         |



**Abb. 10.21** MCMC-Kette der A posteriori-Verteilung: **a** die Cl-Koordinaten, **b** die berechnete Randverteilung von Cl

Damit ist die gesamte Information vorhanden, um die MCMC-Simulation zu starten. Abb. 10.21 zeigt eine Kette von 100 000 Punkten. Der plausibelste Wert  $\text{Cl}_0$  für den Chloridgehalt ist

$$\text{Cl}_0 = 102,9 \text{ mol/m}^3$$

Mit einer Wahrscheinlichkeit von 0,95 befindet sich Cl zwischen 102,1 mol/m<sup>3</sup> und 103,7 mol/m<sup>3</sup>.  $\square$

## Reflexion

**10.1** Die Plausibilität zu einer stetigen Grösse  $K$  beschreibt jemand mit einer Normalverteilung. Der Modus der Verteilung ist  $\mu = 10,0$  und die Standardabweichung  $\sigma$  beträgt 3,4.

- (a) Zeichnen Sie den Graphen der Dichtefunktion mit einem Statistikprogramm.
- (b) Berechnen Sie mit einem Taschenrechner oder einem Statistikprogramm die folgenden Wahrscheinlichkeiten:

$$\mathbb{P}(7,1 \leq K \leq 11,2 | \mu, \sigma), \quad \mathbb{P}(K < 14,3 | \mu, \sigma), \quad \mathbb{P}(K \geq 9,3 | \mu, \sigma)$$

- (c) Bestimmen Sie mit einem Statistikprogramm das 0,7- und das 0,96-Quantil des Wahrscheinlichkeitsmodells.
- (d) Erzeugen Sie mit einem Statistikprogramm 4000 unabhängige Werte, die normalverteilt mit Modus 10,0 und Standardabweichung 3,4 sind. Visualisieren Sie die Werte mit einem Histogramm.
- (e) Bestimmen Sie von Hand ein 0,68-, ein 0,955- und ein 0,997-Wahrscheinlichkeitsintervall für die Grösse  $K$ .

**10.2** Für einen Produktionsprozess gelte  $C_{pk} = 0,7$ . Wie weit ist der Schwerpunkt der Produktionsgrösse mindestens von den Spezifikationsgrenzen entfernt?

**10.3** Für einen Produktionsprozess, der mit einer Normalverteilung modelliert ist, gelte  $C_{pkl} = 0,8$  und  $C_{pku} = 0,5$ .

- (a) Visualisieren Sie die Situation mit einer Grafik.
- (b) Wie gross ist die (durchschnittlich erwartbare) Anzahl ppm pro eine Million produzierter Ware, welche nicht innerhalb der Spezifikationsgrenzen liegt?

**10.4** Eine Person möchte ihr Gewicht  $G$  bestimmen. Sie steht dazu viermal auf eine Personenwaage und liest folgende Zahlen ab:

$$75,5 \text{ kg} \quad 74,8 \text{ kg} \quad 75,2 \text{ kg} \quad 75,7 \text{ kg}$$

Die Person weiss, dass ihr Gewicht zwischen 60 und 90 Kilo liegt. Weiter geht sie davon aus, dass die Messwerte, gegeben  $G$ , normalverteilt um  $G$  streuen. Weitere Informationen hat die Person nicht.

- (a) Überprüfen Sie, ob die Messwerte unter statistischer Kontrolle und unabhängig modellierbar sind.
- (b) Wie lautet der Ausdruck  $\chi^2$ ?
- (c) Die Plausibilität zu  $G$  aus den Messungen kann mit dem Posterior von  $G$  beschrieben werden. Bestimmen Sie diese Verteilung und zeichnen Sie ihren Graphen. Berechnen Sie Wahrscheinlichkeitsintervalle zum Niveau 0,5 und 0,95 für das Gewicht  $G$ . Was ist der plausibelste Wert für das Gewicht der Person? (Arbeiten Sie dazu mit einer MCMC-Simulation oder mit Hilfe der Formeln zur  $t$ -Verteilung in Anhang A.)
- (d) Bestimmen Sie ein Wahrscheinlichkeitsintervall zum Niveau von etwa 0,95 für das Gewicht  $G$  mit der Gleichung von de Moivre.
- (e) Die Person will weitere Messungen mit der Waage machen. In welchem Bereich werden diese Messwerte (1) mit einer Wahrscheinlichkeit von 0,5 und (2) mit einer Wahrscheinlichkeit von 0,95 liegen? (Arbeiten Sie dazu mit einer Monte-Carlo-Simulation oder mit Hilfe der Formeln zur  $t$ -Verteilung in Anhang A.)

**10.5** Eine Person möchte den Umfang  $d$  eines Balls berechnen. Sie misst dazu zweimal den Umfang und erhält 80,3 cm und 77,8 cm. Sie geht davon aus, dass mit diesem Band abgelesene Messwerte normalverteilt um den Umfang  $d$  streuen. Vor den Messungen hat die Person minimale Vorinformation zu  $d$ :  $70 \text{ cm} \leq d \leq 90 \text{ cm}$ .

- (a) Wie lautet der Ausdruck  $\chi^2$ ?
- (b) Die Plausibilität zu  $d$  aus den Messungen kann mit ihrer A posteriori-Verteilung beschrieben werden. Zeichnen Sie den Graphen dieser Verteilung (entweder aus einer MCMC-Simulation oder mit Hilfe der  $t$ -Verteilung aus Anhang A).

- (c) Was ist der plausibelste Wert  $d_0$  für den Umfang  $d$ ? Bestimmen Sie Wahrscheinlichkeitsintervalle zum Niveau 0,5 und 0,95 für  $d$ .
- (d) Jemand will weitere Messungen des Umfangs durchführen. In welchem Bereich werden diese Messwerte mit einer Wahrscheinlichkeit von 0,5, bzw. einer Wahrscheinlichkeit von 0,95 liegen?

**10.6** Die Firma Tillamook-Cheese in Oregon (USA) produziert Käselaibe mit einer mittleren Masse von etwa 19 kg. Zur Qualitätskontrolle müssen neben der Zusammensetzung der Laibe auch die mittlere Masse und die Streuung der Tagesproduktion bestimmt werden. Als Annahme gilt: die Messwerte streuen normalverteilt um die mittlere Masse der Produktion. Tab. 1.13 in Kap. 1 zeigt 20 Massen. Weitere Information ist nicht vorhanden.

- (a) Überprüfen Sie, ob die Stichprobenwerte unter statistischer Kontrolle sind. Zeichnen Sie dazu ein Streudiagramm mit geschätzten Kontrollgrenzen. Plotten Sie auch den Graphen der Autokorrelationsfunktion um zu testen, ob die Messwerte als trendfrei und unabhängig betrachtet werden können.
- (b) Bestimmen Sie die Plausibilität zur mittleren Masse  $\bar{m}_{\text{Tag}}$  der Tagesproduktion. Zeichnen Sie dazu den Graphen der A posteriori-Dichtefunktion. Nennen Sie den Modus und Wahrscheinlichkeitsintervalle zum Niveau 0,5 und 0,95 für  $\bar{m}_{\text{Tag}}$ .
- (c) Wie lautet die A posteriori-Dichtefunktion für die Streuung  $\sigma$  der Normalverteilung? Zeichnen Sie den Graphen dieser Verteilung. Was ist der „wahrscheinlichste“ Wert für  $\sigma$ ? Berechnen Sie Wahrscheinlichkeitsintervalle zum Niveau 0,5 und 0,95 für  $\sigma$ .
- (d) Berechnen Sie approximative Wahrscheinlichkeitsintervalle für  $\bar{m}_{\text{Tag}}$  und  $\sigma$  zum Niveau 0,95 mit den Gleichungen (10.2) und (10.3).
- (e) In welchem Bereich werden weitere gemessene Massen mit einer Wahrscheinlichkeit von 0,5 bzw. von 0,95 liegen?
- (f) Überprüfen Sie das Datenmodell. Benutzen Sie dazu das Prognosemodell für zukünftige Messwerte und einen QQ-Plot. Ist das Modell sinnvoll?
- (g) Die Spezifikation für die produzierten Massen lautet: UTG = 41,6 und OTG = 42,8. Wie gross ist die Wahrscheinlichkeit, dass  $C_p > 1$  ist? (Tipp: Was bedeutet  $C_p > 1$  für die Streuung  $\sigma$ ? Benutzen Sie dann (c).)

**10.7** Um die Durchschnittsmasse  $\bar{m}_{\text{Becken}}$  von 150 Fischen in einem Aufzuchtbecken zu berechnen, wurden zehn Fische durch Ziehen ohne Zurücklegen gewogen. Die Massen in Gramm sind:

898 1198 1294 1516 1196 2050 644 1450 1452 1200

Angenommen wird, dass die Messwerte um die Durchschnittsmasse streuen und mit einem Modell mit endlicher Standardabweichung beschrieben werden können. Zusätzlich weiss man, dass  $\bar{m}_{\text{Becken}}$  zwischen 300 und 3000 gr ist.<sup>9</sup>

---

<sup>9</sup> KTI-Nr. 9029.1 PFIWI-IW, Berner Fachhochschule, 2007.

- (a) Kontrollieren Sie, ob die Messwerte unter statistischer Kontrolle sind.
- (b) Bestimmen Sie die A posteriori-Verteilung der Durchschnittsmasse der 150 Fische. Zeichnen Sie den Graphen der Dichtefunktion von  $\bar{m}_{\text{Becken}}$ . Wie lautet der Modus  $m_0$  dieser Verteilung? Bestimmen Sie Wahrscheinlichkeitsintervalle zum Niveau 0,5 und 0,95 für die Durchschnittsmasse der 150 Fische.
- (c) Zeichnen Sie den Graphen der A posteriori-Dichtefunktion für die Standardabweichung  $\sigma$ . Was ist der „wahrscheinlichste“ Wert für  $\sigma$ ? Berechnen Sie Wahrscheinlichkeitsintervalle zum Niveau 0,5 und 0,95 für  $\sigma$ .
- (d) Eine Person zieht zufällig einen weiteren Fisch aus dem Becken. In welchem Bereich wird die dabei gemessene Fischmasse mit einer Wahrscheinlichkeit von 0,9 liegen? Wie gross ist die Wahrscheinlichkeit, dass die gemessene Fischmasse zwischen 700 und 1200 g liegen wird?
- (e) Überprüfen Sie mit einem QQ-Plot und dem Prognose-Modell von (d): Ist die Modellannahme sinnvoll, dass gemessene Fischmassen normalverteilt um die mittlere Fischmasse streuen?

**10.8** Ultrakurze Laserimpulse mit einer Dauer von einigen Femtosekunden können in Halbleitern Terahertz-Pulse im Picosekundenbereich erzeugen und damit Löcher bohren. Dies erlaubt es, Eigenschaften des Materials über den mittleren Durchmesser  $\mu_D$  der Bohrlöcher zu bestimmen. Der Durchmesser wurde in [4] mittels einer Versuchsreihe im Laserlabor der Berner Fachhochschule an einem mit ps-Pulsen mit einer Wellenlänge von 532 nm bestimmt. Als Modellannahme gilt: Einzelne gebrannte Durchmesser streuen normalverteilt um den mittleren Durchmesser. Hier die Resultate von sieben gebrannten Löchern (in  $\mu\text{m}$ ):

102,70 111,18 108,36 109,78 97,02 106,94 106,00

Als Vorinformation hat man, dass der mittlere Durchmesser  $\mu_D$  zwischen 20 und 200  $\mu\text{m}$  beträgt. Weitere Information zu den Durchmessern liegt nicht vor.

- (a) Überprüfen Sie, ob die Messwerte unter statistischer Kontrolle sind.
- (b) Bestimmen Sie die Plausibilität zu  $\mu_D$ . Zeichnen Sie dazu den Graphen der A posteriori-Dichtefunktion, nennen Sie den Modus und bestimmen Sie Wahrscheinlichkeitsintervalle zum Niveau 0,5 und 0,95.
- (c) Wie lautet die A posteriori-Dichtefunktion für  $\sigma$ ? Zeichnen Sie dazu den Graphen dieser Verteilung. Was ist der „wahrscheinlichste“ Wert für  $\sigma$ ?
- (d) In welchem Bereich liegen mit einer Wahrscheinlichkeit von 0,5 und 0,95 gemessene Durchmesser der Bohrlöcher?

**10.9** Bei einem Produktionsverfahren wird Wasser benutzt. Das nach der Produktion ausgeschüttete Wasser enthält Silberionen. Dabei interessiert der durchschnittliche Gehalt  $\mu_S$  an Silberionen im Abwasser. Als Datenmodell habe man die Normalverteilung. Hier die

Resultate von fünf Messungen des Gehalts an Silberionen (in  $\mu\text{g}/\text{L}$ ):

9,8   10,2   10,7   9,5   10,5

Aus Erfahrung weiss man, dass  $\mu_S \leq 50 \mu\text{g}/\text{L}$  ist.

- (a) Überprüfen Sie, ob die Messwerte unter statistischer Kontrolle sind.
- (b) Bestimmen Sie die Plausibilität  $\mathbb{P}(\mu_S | \text{Daten})$ . Zeichnen Sie dazu den Graphen der A posteriori-Dichtefunktion, nennen Sie den Modus und bestimmen Sie Wahrscheinlichkeitsintervalle zum Niveau 0,5 und 0,95.
- (c) Wie lautet die A posteriori-Dichtefunktion für die Standardabweichung  $\sigma$ ? Zeichnen Sie den Graphen dieser Verteilung. Was ist der „wahrscheinlichste“ Wert für  $\sigma$ ?
- (d) Berechnen Sie Wahrscheinlichkeitsintervalle zum Niveau von 0,5, 0,9 und 0,95 für die Standardabweichung  $\sigma$  mit den in Anhang A erwähnten Formeln.
- (e) In welchem Bereich würden mit einer Wahrscheinlichkeit von 0,5 und 0,95, weitere Messungen des Gehalts an Silberionen liegen?

---

## Literatur

1. M. J. Harry, *The Nature of Six Sigma Quality* (Motorola Inc., Governement Electronics Group, 1988)
2. T. Pyzdek, *The Complete Guide to the CQE* (Quality Publishing Inc., Tucson, 1996)
3. J. A. Rice, *Mathematical Statistics and Data Analysis* (Duxbury Press, 1995)
4. R. Schor, Bohren von Lithium Niobat mit ps-Pulsen, Diplomarbeit Maschinentechnik, Berner Fachhochschule, Burgdorf (2006)

[...] „und was für einen Zweck haben schliesslich Bücher“,  
sagte sich Alice, „in denen überhaupt keine Bilder  
und Unterhaltungen vorkommen?“  
Lewis Carroll, Alice im Wunderland (Insel Taschenbuch, 1973, S. 11)

## Zusammenfassung

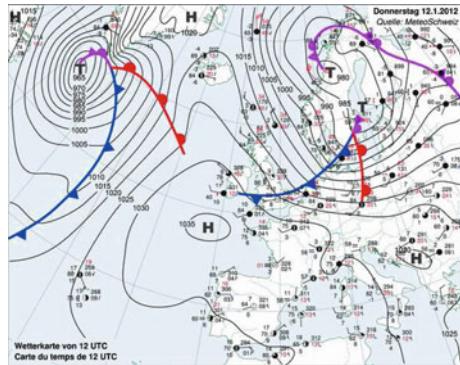
In den vorigen Kapiteln wird erklärt, wie Resultate zu nicht direkt messbaren Größen dargestellt werden können. Man kann den Graphen der A posteriori-Dichtefunktion darstellen, den plausibelsten Wert angeben oder Wahrscheinlichkeitsintervalle nennen. Die Resultate hängen von den Daten und von Vorinformation ab. Neben den Rechnungen ist es sinnvoll, die verwendeten Daten zu visualisieren. In diesem Kapitel werden einfache und prägnante grafischen Darstellungen von univariaten Datenwerten vorgestellt. Dies kann auch nützlich sein, um Fragen wie: „War das Experiment unter statistischer Kontrolle?“, „Sind extreme Werte vorhanden?“, oder „Ist das gewählte Modell gut?“ zu beantworten. Die in den Daten steckende Information kann auch helfen, ein gutes Datenmodell zu wählen.

## 11.1 Erste Beispiele zu grafischen Darstellungen

Wetterkarten, wie in Abb. 11.1 dargestellt, entstehen aus zahlreichen Messungen zu Luftdruck, Temperatur, Wind, Bewölkung und Niederschlägen. Benutzerinnen und Benutzer mit Kenntnissen in Wetterphänomenen und Physik (z. B. bewegen sich Winde parallel zu den Isobarenkurven von Hoch- zu Tiefdruckgebieten, drehen Tiefdruckwirbel in der nördlichen Halbkugel im Gegenuhrzeigersinn) können daraus prognostizieren, wie sich das Wetter in den nächsten ein bis zwei Tagen entwickeln wird. Dieses Beispiel zeigt, wie Daten – hier Wetterdaten – effizient und übersichtlich kommuniziert werden können. Kaum ein anderes statistisches Werkzeug ist wirkungsvoller. Es erlaubt nicht nur die Daten klar, präzise und effizient darzustellen, sondern auch neues Wissen zu erzeugen.

**Tab. 11.1** Eine Kreuztabelle

	Gruppe 1	Gruppe 2	...	Gruppe $n$	Total
Resultat 1					
Resultat 2					
...					
Resultat $m$					
Total					

**Abb. 11.1** Wetterkarte mit Isobaren und Fronten (Quelle MeteoSchweiz)

Die einfachste Art, Daten darzustellen, ist, diese in *Kreuztabellen* oder *Kontingenztafeln* (engl. *two-way tables*) zu platzieren. Eine Kreuztabelle findet sich in Tab. 11.1.

**Beispiel 11.1 (Bildungsstand)** Tab. 11.2 zeigt, wie der Bildungsstand eines Teils der Schweizer Bevölkerung verteilt ist. Die dargestellten Zahlen sind, da sie aus einem Zensus stammen, exakte Werte. Die Tabelle zeigt, dass in der Westschweiz mehr Personen ihren abschliessenden Bildungsstand in staatlichen Institutionen (obligatorische Schule, Sekundarstufe II, Allgemeinbildung und Universität) als der Deutschschweiz absolvieren. □

**Beispiel 11.2 (Tarife Telefonbetreiber)** Abb. 11.2 zeigt Preise von drei Telefonbetreibern in der Schweiz, wie sie in verschiedenen Werbekampagnen im Jahr 2001 visualisiert wurden. Die Abbildung stellt sechs Zahlen und drei Telefonbetreiber mit unnötig viel

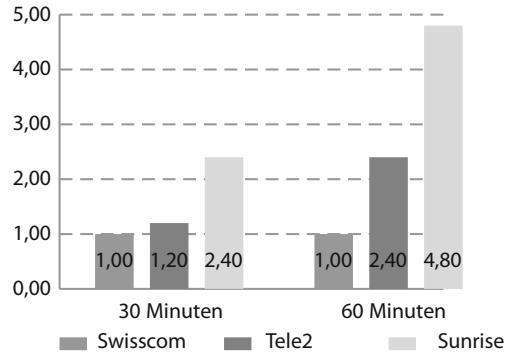
**Tab. 11.2** Bildungsstand der 25- bis 64-Jährigen in der Schweiz im Jahr 1998 (Statistisches Jahrbuch der Schweiz, 2000)

	Deutschschweiz	Westschweiz
Obligatorische Schule	18 %	23 %
Sekundarstufe II (Berufsausbildung)	52 %	44 %
Sekundarstufe II (Allgemeinbildung)	7 %	9 %
Tertiärstufe (ausseruniversitär)	14 %	11 %
Tertiärstufe (universitär)	9 %	14 %

**Tab. 11.3** Gesprächskosten in CHF, dargestellt mit einer Kreuztabelle

	30 Minuten	60 Minuten
Swisscom	1,00	1,00
Tele 2	1,20	2,40
Sunrise	2,40	4,80

**Abb. 11.2** Gesprächskosten Fernbereich Schweiz (19.00–20.00 Uhr) im Jahr 2001 in CHF



Druckerschwärze dar. Nach E. R. Tufte besitzt die Grafik eine *Datendichte* von

$$\frac{\text{Anzahl Daten}}{\text{Fläche der Grafik}} = \frac{6 \text{ Preise} + 2 \text{ Gesprächsdauern} + 3 \text{ Firmen}}{6,7 \text{ cm} \times 4,8 \text{ cm}} = 0,34 \text{ Daten/cm}^2$$

Dies ist eine kleine Zahl. Einfacher ist es, die Preise wie in Tab. 11.3 darzustellen. Die Datendichte ist höher. Zudem entfällt der zeitliche Aufwand, die wenigen Zahlen mit Stabdiagrammen, Achsen und Graustufen umzusetzen. □

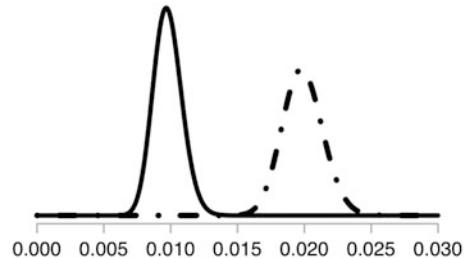
**Beispiel 11.3 (Aspirin)** Die Daten der Untersuchung aus [3], ob tiefdosiertes Aspirin einen Einfluss auf nicht-tödliche Herzinfarkte bei Patienten mit kardiovaskulären Risiken hat, sind in Tab. 11.4 gut lesbar. Aus der Stichprobe, bestehend aus 17 187 Personen, lässt sich berechnen, wie Aspirin auf Patienten mit kardiovaskulären Risiken wirkt. Diese ist durch den folgenden Quotienten gegeben

$$V_{\text{Herz}} = \frac{A_{\text{Asp}}}{A_{\text{oAsp}}} = \frac{\text{Anteil Herzinfarkte bei Aspirineinnahme}}{\text{Anteil Herzinfarkte ohne Aspirineinnahme}}$$

**Tab. 11.4** Studie zu Aspirin aus Second International Study of Infarct Survival (ISIS-2, siehe [3])

	Aspirin	Placebo
nicht-tödlicher Herzinfarkt	83	170
nicht-tödlicher Schlaganfall	27	51
Tod infolge kardiovaskularem Ereignis	804	1016
Anzahl Personen	8587	8600

**Abb. 11.3** Plausibilität zu den Anteilen  $A_{\text{Asp}}$  (*ausgezogene Linie*) und  $A_{\text{oAsp}}$



Gemäss Theorem 5.3 sind die A posteriori-Verteilungen der Anteile  $A_{\text{Asp}}$  und  $A_{\text{oAsp}}$  Beta-Verteilungen:

$$\text{pdf}(A_{\text{Asp}} | \text{Daten, min. Vorinformation}) \propto A_{\text{Asp}}^{83} \cdot (1 - A_{\text{Asp}})^{8587-83} \cdot 1$$

$$\text{pdf}(A_{\text{oAsp}} | \text{Daten, min. Vorinformation}) \propto A_{\text{oAsp}}^{170} \cdot (1 - A_{\text{oAsp}})^{8600-170} \cdot 1$$

Die Graphen der beiden Verteilungen sind in Abb. 11.3 dargestellt. Ersichtlich ist, dass die Anteile mit hoher Plausibilität in verschiedenen Regionen liegen. Daher dürfte  $V_{\text{Herz}}$  ungleich eins sein. Die Plausibilität für diesen Parameter lässt sich mit einer Monte-Carlo-Simulation berechnen. Man erzeugt mit einem Statistikprogramm 100 000 Werte der Beta-Verteilungen von  $A_{\text{Asp}}$  und  $A_{\text{oAsp}}$ . Daraus rechnet man 100 000 Werte von  $V_{\text{Herz}} = A_{\text{Asp}}/A_{\text{oAsp}}$  aus. Aus dem 2500 kleinsten und 97 500 grössten Wert der simulierten Quotienten, erhält man ein Wahrscheinlichkeitsintervall für  $V_{\text{Herz}}$  zum Niveau 0,95:

$$0,38 < V_{\text{Hertz}} < 0,63$$

Es ist plausibel, dass  $V_{\text{Herz}}$  kleiner als Eins ist. Daher kann dank Aspirin die Herzinfarktrate bei Patienten mit kardiovaskulären Risiken reduziert werden. □

**Tab. 11.5** Angebot von ProSpeciaRara-Gemüse in der Schweiz (Coop, 2009)

**Tab. 11.6** Angebot von ProSpeciaRara-Gemüse in der Schweiz (Coop, 2009)

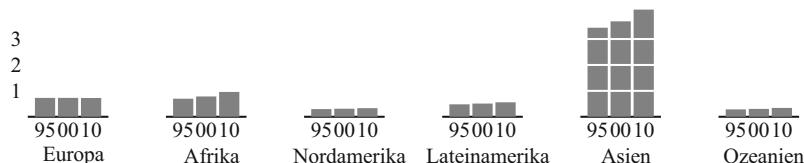
	März	Apr	Mai	Juni	Juli	Aug	Sept	Okt	Nov	Dez	Jan	Feb	März
Radieschen Eiszapfen													
Spinat rote Gartenmelde													
Tomate Coeur de Boeuf													
Hornpeperoni grün													
Tomate Orange													
Hornpeperoni farbig													
Mangold rot													
Karotten Küttiger													
Kartoffel blaue Schweden													
Pastinake													
Randen Chioggia													
Suppengemüse													
Federkohl													
Kartoffel Parli													

**Beispiel 11.4 (Gemüseernte)** Tab. 11.5 zeigt, wann besondere und seltene Gemüsesorten – sogenannte ProSpeciaRara-Gemüse – erhältlich sind. In der Tabelle sind die Monate über ein Jahr dargestellt und vertikal befinden sich in alphabetischer Reihenfolge die Gemüsesorten. Dies macht die Tabelle unübersichtlich. Besser kann man die Tabelle für die Konsumenten darstellen, wenn man die Gemüsesorten nach ihrer Saison gliedert und zudem den Monat März doppelt darstellt (siehe Tab. 11.6). Nun ist gut ersichtlich, dass die Gemüsesorten im Frühling nur kurz erhältlich sind, während das Wintergemüse wie Karotten oder Federkohl fast ein halbes Jahr zu kaufen ist. □

**Beispiel 11.5 (Bevölkerungsstand)** Tab. 11.7 zeigt 252 Werte zum Bevölkerungsstand und zur Altersstruktur auf prägnante Art. Vergleicht man die Darstellung der Daten in einer Tabelle mit jener in Stabdiagrammen (vgl. Abb. 11.4), ist die tabellarische Darstellung vorzuziehen. □

Um einen ersten Überblick über Datenwerte zu erhalten und um extrem grosse oder kleine Werte schnell aufzufinden, kann man diese nach ihrer Grösse ordnen:

**Beispiel 11.6 (Unwetterschäden)** Die Daten zu den Unwetterschäden von Beispiel 1.6 zeigt Abb. 11.5. Sichtbar ist, dass die Beobachtungen kaum Trends besitzen. Auffallend

**Abb. 11.4** Bevölkerungsstand in Millionen in allen Weltteilen, dargestellt mit Balkendiagrammen

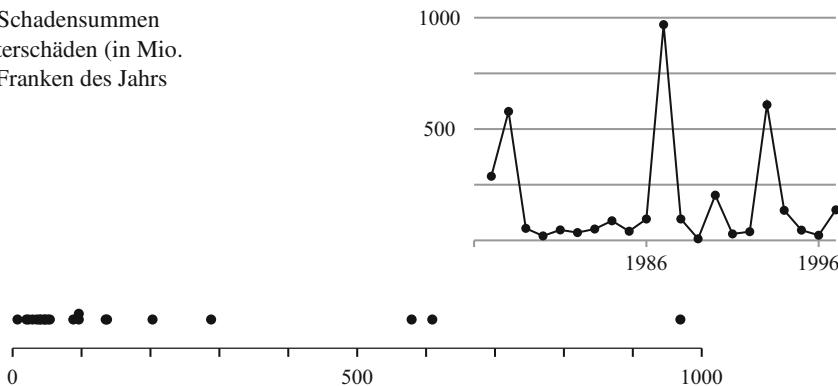
**Tab. 11.7** Bevölkerungsstand in Millionen in allen Weltteilen (The 1998 Revision, United Nations und Bundesamt für Statistik Schweiz)

Weltteil	Bevölkerungsstand in Millionen			Anteil im Alter von ... in %					
				unter 15 Jahren			65 und mehr Jahren		
	1995	2000	2010	1995	2000	2010	1995	2000	2010
Welt total	5666	6055	6795	31,2	29,7	26,5	6,6	6,9	7,6
Europa	728	729	724	19,2	17,5	15,2	13,9	14,7	16,2
Nordeuropa	94	94	95	19,4	18,3	16,6	15,5	15,6	16,9
Südeuropa	143	144	143	17,0	15,7	14,2	15,0	16,5	18,4
Westeuropa	181	183	185	17,7	17,0	15,1	15,1	15,9	18,0
Osteuropa	310	307	301	21,0	18,2	15,4	12,3	13,0	13,7
Afrika	697	784	973	43,6	42,5	39,9	3,1	3,2	3,2
Nordafrika	157	173	206	38,4	35,6	31,2	3,8	4,0	4,4
Südafrika	43	47	50	37,0	35,9	33,7	3,4	3,6	3,7
Westafrika	196	222	281	45,4	43,9	41,5	2,9	3,0	3,2
Ostafrika	217	247	310	45,9	45,5	43,3	2,8	2,7	2,6
Mittelafrika	84	96	126	46,4	46,9	44,8	3,1	3,1	3,0
Nordamerika	297	310	332	22,0	21,2	18,9	12,5	12,5	13,3
Lateinamerika	480	519	558	33,7	31,6	27,8	5,1	5,4	6,4
Karibik	36	38	40	31,1	29,5	25,8	6,6	6,9	8,1
Zentralamerika	123	135	158	37,1	34,8	30,4	4,2	4,5	5,5
Südamerika	321	346	359	33,8	30,5	26,9	5,2	5,6	6,5
Asien	3436	3683	4136	31,8	29,9	25,8	5,4	5,9	6,8
Westasien	168	188	231	36,7	35,0	32,6	4,4	4,8	5,2
Südasien	1365	1491	1729	37,0	34,8	29,8	4,3	4,6	5,2
Südostasien	480	519	588	34,0	31,4	26,9	4,3	4,7	5,6
Ostasien	1422	1485	1588	25,4	23,9	20,0	6,8	7,7	9,2
Ozeanien	28,5	30,4	34,2	25,9	25,2	23,3	9,9	9,9	10,7
Australien und Neuseeland	21,6	22,7	24,8	21,8	21,0	19,0	12,0	12,0	13,3
Melanesien	5,8	6,5	7,9	39,1	37,8	35,0	3,1	3,3	3,8
Mikronesien	0,5	0,5	0,7	39,6	38,4	35,3	3,2	3,5	4,0
Polynesien	0,6	0,6	0,7	36,7	35,3	30,8	3,9	4,4	5,3

sind drei grosse Schadensummen. Um diese anders zu visualisieren, kann man die Beobachtungen der Grösse nach arrangieren und wie in Abb. 11.6 darstellen. Sichtbar sind die vielen Schadensummen kleiner als 100 Mio. CHF und die drei grossen Schadensummen über 500 Mio. CHF. □

**Beispiel 11.7 (Energieverbrauch)** Der Grösse nach sortierte Werte, integriert in grafische Elemente, eignen sich auch, um zwei Messreihen zu vergleichen. Dies verdeutlicht Abb. 11.7. Die Verteilung in Prozent des Energieverbrauchs in der Schweiz nach Energie-

**Abb. 11.5** Schadensummen von Unwetterschäden (in Mio. Schweizer Franken des Jahres 1980)

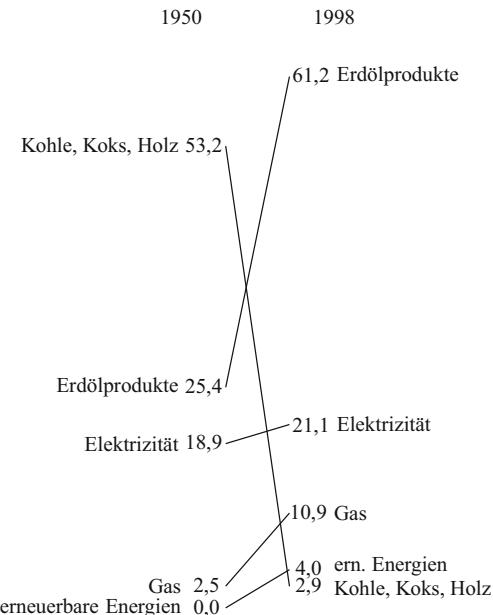


**Abb. 11.6** Schadensummen von Unwettern in Mio. Franken des Jahres 1980

trägern in den Jahren 1950 und 1998 zeigt die starke Zunahme des Energieträgers Erdöl. Kohle, Koks und Holz spielen kaum mehr eine Rolle. □

Die Darstellung des Energieverbrauchs ist nach Tufte ([9], Seite 159) multifunktional. Sie vereinfacht, Merkmale der Datenmenge zu erkennen: Einerseits befindet sich die gesamte Information in einer Grafik und andererseits werden die Daten durch die Linien getrennt. Das menschliche Auge folgt so den wichtigsten Merkmalen der Daten: Zu-

**Abb. 11.7** Verteilung des Energieverbrauchs in der Schweiz nach Energieträgern (in %) in den Jahren 1950 und 1998



oder Abnahme sowie Größenordnung der einzelnen Energieträger. Tufte spricht von der *sehenden Architektur* (engl. *viewing architecture*) einer Grafik, um komplexe Information organisiert darzustellen. Geschickte grafische Darstellungen erlauben es daher, Daten klar, präzise und effizient zu präsentieren. Dabei sollen sie den Leser oder die Leserin auf Trends und qualitativ hervorragende Merkmale der Daten hinweisen und Verzerrungen vermeiden. Man spricht von *explorativer Datenanalyse*.

Hervorragende grafische Darstellungen besitzen ein ansprechendes Format, d. h. eine Größe, die das Ablesen von Daten unterstützt. Sie haben oft eine größere horizontale als vertikale Ausdehnung (der goldene Schnitt Länge zu Breite =  $1,618 : 1$  wird als angenehm empfunden). Sie bestehen aus Zahlen, Wörtern und Malelementen, zeichnen sich durch eine sorgfältige und professionelle Ausführung aus und besitzen kaum inhaltslose Dekorationen („Chartchunk“).

Daten grafisch darzustellen ist nicht einfach. Dazu müssen sie auf verschiedene Art und Weise dargestellt werden, bis die für die Untersuchung definierten Merkmale zum Vorschein treten. Natürlich sollte dabei bedacht werden, dass Bilder nur ein erster Schritt sind, um gesuchte Größen zu berechnen. Entscheide aus Bildern der explorativen Datenanalyse werden daher erst gezogen, wenn die gesuchten Größen mit ihrer Präzision und dazugehöriger Plausibilität bestimmt sind. Der Statistiker J. W. Tukey benutzt dazu in [10] einen Vergleich mit der Kriminaljustiz: Unterschieden wird zwischen der Suche nach Beweisen, die von der Polizei und der Staatsanwalt unternommen wird, und der Evaluation der Beweise, die von Gerichten und Richtern durchgeführt wird. Explorative Datenanalyse entspreche der Suche nach Beweisen. Die schliessende Statistik evaluiere anschliessend die dargestellten Merkmale mit Präzisionen und Plausibilitäten. Es ist daher empfehlenswert, die Daten-Grafiken mit der nötigen Vorsicht zu interpretieren. Es ist sinnvoll, neben visualisierten Daten auch den Posterior oder Wahrscheinlichkeitsintervalle der berechneten Parameter anzugeben.

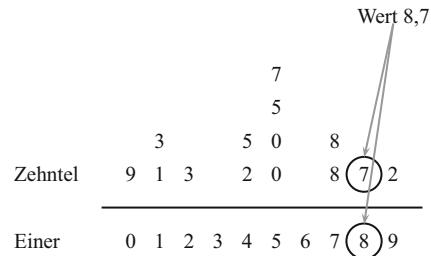
## 11.2 Darstellen, wie Daten verteilt sind

Stamm-Blatt-Diagramme, Strichlisten, Stabdiagramme und Histogramme sind beliebt, um zu visualisieren, wie univariante Datenwerte verteilt sind.

Das *Stamm-Blatt-Diagramm* (engl. *stem and leaf diagram*) erlaubt es, Daten effizient, klar und ohne Verlust darzustellen. Es lässt sich auch von Hand mit Hilfe eines karierten Blattes zeichnen. Man teilt dazu die Datenwerte in einen Stamm und entsprechende Blätter auf. Hat man die Zahlen

1,3 4,5 5,0 5,5 0,9 2,3 8,7 1,1 5,0 5,7 7,8 4,2 7,8 9,2

können die Einer als Stamm und die Zehntel als Blatt benutzt werden. Das zugehörige Stamm-Blatt-Diagramm findet sich in Abb. 11.8. Man überblickt rasch, wie die Daten verteilt sind.

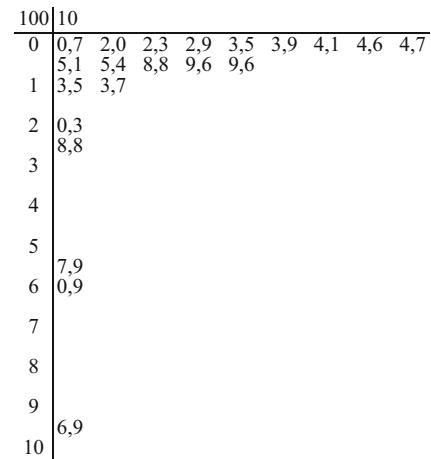
**Abb. 11.8** Ein Stamm-Blatt-Diagramm

**Beispiel 11.8 (Unwetterschäden)** Die Schadensummen der 21 Unwetter, vorgestellt in Beispiel 1.6, variieren zwischen 7 und 969 Mio. CHF. Als Stamm bieten sich 100 Mio. CHF an. Als Blätter ist es sinnvoll, 10 Mio. CHF zu wählen. Der Wert  $x_2 = 579$  Mio. CHF wird so beim Stamm 5 und beim Blatt 7,9 eingetragen. Das so entstandene Stamm-Blatt-Diagramm in Abb. 11.9 zeigt, dass Schadensummen über 200 Mio. CHF selten sind. Zudem ist ersichtlich, dass Schadensummen durch ein unimodals schiefes Wahrscheinlichkeitsmodell beschrieben werden sollten. Man kann daher versuchen, Schadensummen – eine kontinuierliche Grösse – mit einer Exponentialverteilung zu modellieren. Ist  $\mu$  der zukünftige Durchschnittsschaden, so heisst dies

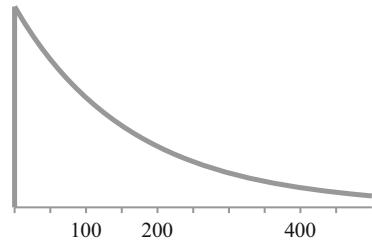
$$\text{Datenmodell: } i\text{-te beobachtete Schadensumme} \sim \text{Exponential}(1/\mu)$$

Theorem 9.2 zur Exponentialverteilung sagt, wie man den Posterior für den Skalierungsparameter  $\mu$  berechnen kann. Ist vor der Datensammlung nur minimale Information zum Parameter  $\mu$  vorhanden, so hat man für die A posteriori-Verteilung von  $\mu$

$$\text{pdf}(\mu \mid \text{Daten, min. Vorinformation}) \propto \frac{1}{\mu^{21}} \cdot \exp(-3592 \text{ Mio. CHF}/\mu) \cdot \frac{1}{\mu}$$

**Abb. 11.9** Stamm-Blatt-Diagramm der Schadensummen der Unwetterschäden aus Beispiel 1.6

**Abb. 11.10** Prognosemodell für zukünftige Schadensummen von Unwettern



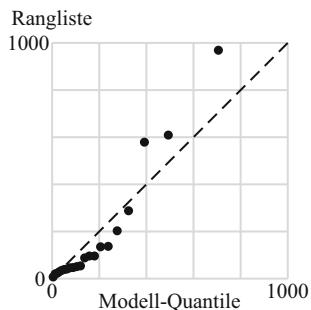
Dabei ist 3592 Mio. CHF die Gesamtsumme aller beobachteten Schäden. Der plausibelste Wert für den zukünftigen Durchschnittsschaden ist  $\mu_0 = 162,9$  Mio. CHF. Man hat 21 Beobachtungen und einen flachen Prior. Daher ist die Likelihood-Funktion im Posterior von  $\mu$  dominant. Ein Wahrscheinlichkeitsintervall für  $\mu$  zum Niveau von etwa 0,95 ist somit gemäss Gleichung (9.1)

$$\mu = \bar{x} \pm 1,96 \cdot \frac{\bar{x}}{\sqrt{21}} = (171 \pm 73) \text{ Mio. CHF}$$

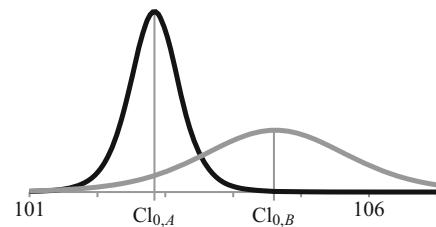
Die Dichtefunktion um zukünftige, einzelne Schadensummen zu prognostizieren, kann man mit Theorem 9.2 berechnen. Ihr Graph ist in Abb. 11.10 visualisiert. Aus dem Prognosemodell lassen sich die 0,5/21-, 1,5/21-, ... und 20,5/21-Quantile berechnen und der QQ-Plot erstellen (Abb. 11.11). Es wird ersichtlich, dass das gewählte Modell nicht ideal ist. Kleine Schadensummen werden mit einer zu hohen Wahrscheinlichkeit besetzt, grosse Schadensummen werden unterschätzt. Die berechneten Resultate sind daher eher fragwürdig. Hier hilft ein Blick in die Literatur oder zusätzliches Wissen zu Schadensummen und ihrer Streuung. Dies führt meist dazu, Gamma-Verteilungen oder Extremwertverteilungen als Datenmodell oder einen geschickteren Prior für  $\mu$  zu wählen.  $\square$

**Beispiel 11.9 (Chloridgehalt)** In Beispiel 10.4 ist gezeigt, wie der Chloridgehalt Cl in einer Kalium-Chlorid-Lösung aus Messungen bestimmt werden kann. Benutzt wurde dabei ein chemisch-mechanisches Verfahren (Methode A). Hier noch einmal die Messwerte

**Abb. 11.11** QQ-Plot der Schadensummen beim Datenmodell zur Exponentialverteilung



**Abb. 11.12** Plausibilität zum Chloridgehalt mit den Daten aus der Methode A (schwarze Linie) und mit den Daten aus der Methode B (graue Linie)



(in mol/m<sup>3</sup>):

$$102,8 \quad 103,3 \quad 102,3 \quad 103,6 \quad 104,3 \quad 101,5 \quad 102,1$$

Eine zweite Methode (Methode B) verwendet ein chemisch-elektrisches Verfahren. Mit ihr wurde die Lösung zehnmal gemessen:

$$108,7 \quad 110,9 \quad 101,1 \quad 102,5 \quad 100,1 \quad 101,9 \quad 105,8 \quad 106,0 \quad 104,1 \quad 105,1$$

Um den Chloridgehalt zu bestimmen, kann man als Datenmodell annehmen, dass Datenwerte des Chloridgehalts normalverteilt um den Chloridgehalt streuen:

$$\text{Datenmodell: } i\text{-ter Messwert} \sim \text{Normal}(\text{Cl}, \sigma)$$

Hat man keine weitere Information zu den beiden Parametern Cl und  $\sigma$  des Modells und nimmt man an, dass die Messwerte unabhängig sind, so erhält man nach Theorem 10.2 Wahrscheinlichkeitsintervalle für Cl zum Niveau 0,95.<sup>1</sup>

$$\text{Cl}_{\text{Methode } A} = (102,8 \pm 0,9) \text{ mol/m}^3 \quad \text{Cl}_{\text{Methode } B} = (104,6 \pm 2,4) \text{ mol/m}^3$$

Die beiden Graphen der A posteriori-Verteilungen zu Cl findet man in Abb. 11.12. Auffallend ist, dass die Methode B – trotz gröserer Anzahl Messungen – den Chloridgehalt weniger präzis bestimmt als die Methode A. Dies liegt daran, dass die Werte bei der Methode B stark streuen. Ein Stamm-Blatt-Diagramm der Messwerte visualisiert dies. Wie dabei Stamm und Blätter gewählt werden sollen, ist nicht sofort ersichtlich. Die Daten haben als kleinste Einheit 0,1 mol/m<sup>3</sup>. Die letzte Ziffer der Daten scheint zufällig verteilt zu sein: es besteht die Möglichkeit, die Blätter 10 mol/m<sup>3</sup> zu wählen. Der Stamm besteht dann aus den Klassen 100 mol/m<sup>3</sup> und 110 mol/m<sup>3</sup>. So wird der Wert 103,6 mol/m<sup>3</sup> beim

<sup>1</sup> Man kann auch mit der Formel (A.1) aus Abschn. A.3 rechnen

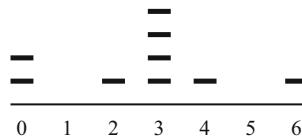
$$\text{Cl}_{\text{Methode } A} = \overline{\text{Cl}_A} \pm t_0(0,95) \cdot s_A / \sqrt{7} = (102,8 \pm 0,9) \text{ mol/m}^3 \quad [k = 1,96]$$

Dabei ist  $\overline{\text{Cl}_A}$  das arithmetische Mittel und  $s_A$  die empirische Standardabweichung der Datenwerte.

**Abb. 11.13** Stamm-Blatt-Diagramm der Chloridwerte

		<b>Methode A</b>		<b>Methode B</b>	
		1 mol/m <sup>3</sup>	mol/m <sup>3</sup>	1 mol/m <sup>3</sup>	
			100	0,1	
			1,5	1,1	1,9
2,8	2,3	2,1		2,5	
		3,6	3,3		
			4,3	4,1	
				5,1	5,8
				6,0	
					8,7
			110	0,9	

**Abb. 11.14** Eine Strichliste



Stamm 100 und beim Blatt 3,6 eingetragen. Um eventuelle Unterschiede der Methoden *A* und *B* zu veranschaulichen, können die Stamm-Blatt-Diagramme „*Rücken an Rücken*“ – wie in Abb. 11.13 – gezeichnet werden. Der Vorteil ist augenfällig: Es wird deutlich, dass eine gering gestreute Datenmenge einer breit gestreuten gegenüber steht. □

Stamm-Blatt-Diagramme erfüllen wichtige Ziele der explorativen Datenanalyse: Die Datenwerte und ihre Verteilungen werden effizient, übersichtlich und ohne Verzerrungen gezeigt. Dem Leser oder der Leserin fallen – bei guter Wahl der Stammbreite – wichtige Trends der Datenreihe auf.

Die Zahlen in Stamm-Blatt-Diagrammen können nicht in beliebig kleine Einheiten unterteilt werden. Daher eignen sich Stamm-Blatt-Diagramme weniger, um grosse Datenmengen zu visualisieren. Hier kann ein Histogramm helfen (siehe weiter unten).

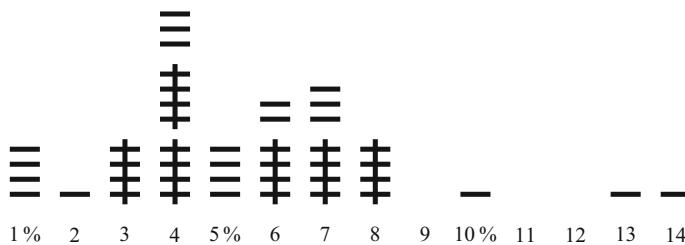
Für *diskrete* Werte können Strichlisten zeigen, wie Daten verteilt sind. Hat man

4    3    3    2    0    0    3    6    3

so zeigt die Strichliste in Abb. 11.14, wie die Datenwerte verteilt sind.

**Beispiel 11.10 (Nicht keimende Blumenzwiebeln)** Für das Beispiel 1.7 der Gärtnerei, wo der Anteil nicht keimender Blumenzwiebeln in Kisten gezählt wird, ist die Strichliste in Abb. 11.15 erstellt worden. Sie zeigt eine Häufung zwischen drei und acht Prozent. Auffallend sind zwei Kisten mit dreizehn und vierzehn Prozent nicht keimender Blumenzwiebeln.

Den Anteil *A* – eine Zahl zwischen Null und Eins – nicht keimender Blumenzwiebeln der *Gesamtproduktion* kann man aus den gemessenen Werten berechnen. Untersucht wurden 50 Kisten mit jeweils 100 Blumenzwiebeln. 268 der 5000 Blumenzwiebeln waren



**Abb. 11.15** Strichliste der 50 beobachteten Anteile an nicht keimenden Blumenzwiebeln

nicht keimend. Hat man keine weitere Information zu  $A$  und setzt weiter voraus, dass die Datenwerte unabhängig sind, ist nach Theorem 5.3:

$$\text{pdf}(A \mid \text{Daten, min. Vorinformation}) \propto A^{268} \cdot (1 - A)^{5000 - 268} \cdot 1$$

Dies ist eine Beta-Verteilung mit Kennzahlen 269 und 4733. Der wahrscheinlichste Wert für  $A$  ist der Modus  $A_0$  der Verteilung. Er lautet  $268/5000 = 0,0536$ . Ein Wahrscheinlichkeitsintervall für  $A$  zum Niveau 0,5 ist  $0,0516 \leq A \leq 0,0559$ . Es besteht eine Wahrscheinlichkeit von 0,95, dass  $A$  zwischen 4,8 % und 6,0 % ist. Wegen der vielen Messwerte ist die bestimmte Präzision zu  $A$  skeptisch zu hinterfragen. Es ist plausibel, dass Abhängigkeiten bei der Keimfähigkeit der Blumenzwiebeln vorhanden sind.  $\square$

Arbeitet man von Hand schnell mit Strichlisten, so können wegen Unaufmerksamkeit Fehler entstehen. Mögliche Fehler bei Notierungen sind:

||||| oder ||||| oder +++

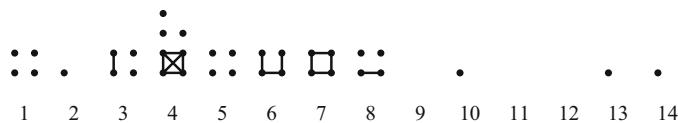
Bei schnellem Arbeiten empfiehlt daher J. W. Tukey (siehe [10]) die Striche durch die Zeichen in Abb. 11.16 zu ersetzen. Beim Beispiel zu den Blumenzwiebeln erhält man so die Strichliste in Abb. 11.17.

Um zu veranschaulichen, wie verschieden zwei Klassen von Datenwerten verteilt sind, können Strichlisten ähnlich wie Stamm-Blatt-Diagramme Rücken an Rücken gezeichnet werden.

*Stabdiagramme* sind ebenfalls beliebt, um zu visualisieren, wie diskrete Datenwerte verteilt sind. Sie entstehen, indem man die Strichliste durch Stäbe ersetzt. Die Stäbe sollten dabei voneinander getrennt sein.

1:	•	3:	••	5:	••	7:	□□	9:	□□
2:	•	4:	••	6:	••	8:	□□	10:	□□

**Abb. 11.16** Strichlistensymbole bei schnellem Arbeit nach Tukey

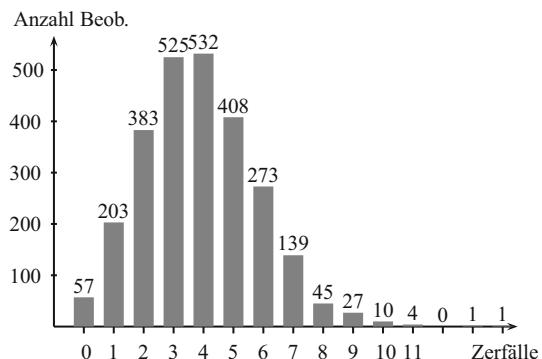


**Abb. 11.17** Strichliste der 50 beobachteten Anteile an nicht keimenden Blumenzwiebeln, mit Symbolen nach Tukey

**Beispiel 11.11 (Zerfall von Polonium)** Die Physiker E. Rutherford und M. Geiger mas- sen die Anzahl Funken, die durch den radioaktiven Zerfall von Polonium entstanden. Zählte man während je 72 Sekunden den  $\alpha$ -Teilchenzerfall des Poloniums, ergaben sich gemäss [6] die 2608 Werte in Tab. 11.8. Abb. 11.18 zeigt die Verteilung der Messwerte mit einem Stabdiagramm. Die Verteilung ist unimodal schief und rechtsschwänzig. Das Stabdiagramm ist aber mit unnötigen Angaben überladen. So sind die Höhen der Stäbe durch zwei Informationen gegeben: die Höhe der Stäbe selbst und die Zahl oberhalb der Stäbe. Auf solche redundanten Elemente sollte man verzichten. Die Aufmerksamkeit der Betrachterinnen und Betrachter sollte bei der *Verteilung* der Daten liegen. Das Stabdiagramm in Abb. 11.19 ist dem ersten Stabdiagramm vorzuziehen, da es die Verteilung in den Vordergrund stellt. So sind die Koordinatenachsen auf das Wesentliche reduziert und der „Tintenverbrauch“, um das Stabdiagramm zu drucken, ist verkleinert. □

**Beispiel 11.12 (Auflagenzahlen)** Im Jahr 1998 illustrierte eine Zeitung ihre Auflagenzahlen mit einem Stabdiagramm, wie es in Abb. 11.20 dargestellt ist. Die Figur besitzt grosse Mängel. So werden (1) die Auflagenzahlen verzerrt dargestellt und (2) die Zeitabschnitte sind ungleich gewählt, um die Zunahme der Auflage visuell zu verstärken. Weiter ist die Grafik höher als breit. Der Wachstumseffekt wird durch diesen grafischen

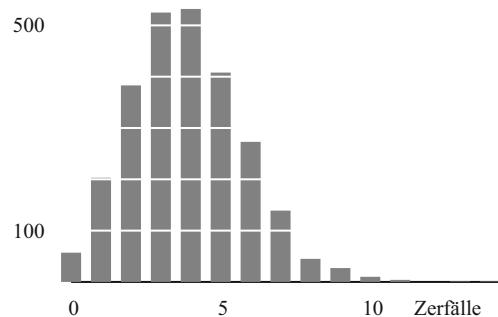
**Abb. 11.18** Stabdiagramm visualisiert, wie die 2608 Messwerte verteilt sind



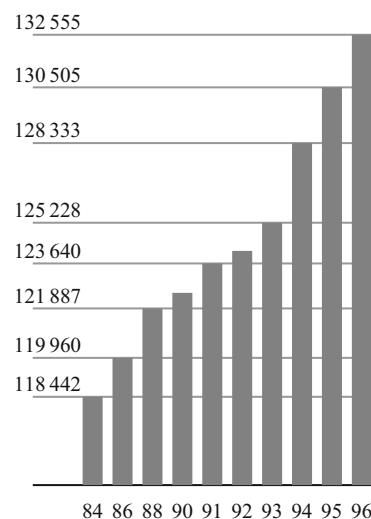
**Tab. 11.8** Anzahl Zerfälle von  $\alpha$ -Teilchen des Poloniums (aus [6])

Zerfälle	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Häufigkeit	57	203	383	525	532	408	273	139	45	27	10	4	0	1	1

**Abb. 11.19** Stabdiagramm der 2608 Messwerte ohne redundante Angaben



**Abb. 11.20** Verzerrte Visualisierung der Auflage einer Zeitung in den Jahren 1984 bis 1996



„Wolkenkratzereffekt“ dramatisch erhöht. Der Effekt der Verzerrung lässt sich mit dem *Lügenfaktors* nach Tufte (siehe [9]) messen:

$$\text{Lügenfaktor} = \frac{\text{Mass des Effekts gezeigt in der Grafik}}{\text{Mass des Effekts der Datenwerte}}$$

Das Wachstum der Auflage 1984 bis 1996 beträgt

$$\frac{132\,555 - 118\,442}{118\,442} \times 100 \% = 11,92 \%$$

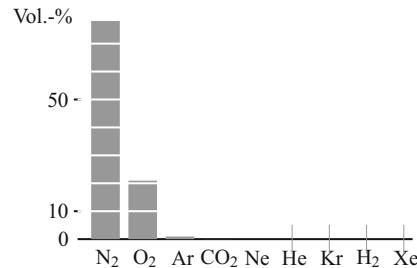
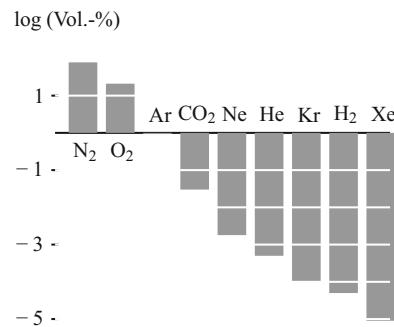
Im Stabdiagramm ist das dargestellte Wachstum durch die Länge der Stäbe gegeben:

$$\frac{6,1 \text{ cm} - 1,2 \text{ cm}}{1,2 \text{ cm}} \times 100 \% = 408 \%$$

Der Lügenfaktor beträgt  $408 \% / 11,92 \% = 34,2$ . □

**Tab. 11.9** Bestandteile von trockener Luft (Knaurs Lexikon 1996)

Bestandteil	N <sub>2</sub>	O <sub>2</sub>	Ar	CO <sub>2</sub>	Ne	He	Kr	H <sub>2</sub>	Xe
Volumen-%	78,10	20,93	0,93	0,03	0,0018	0,0005	0,0001	0,00005	0,000009

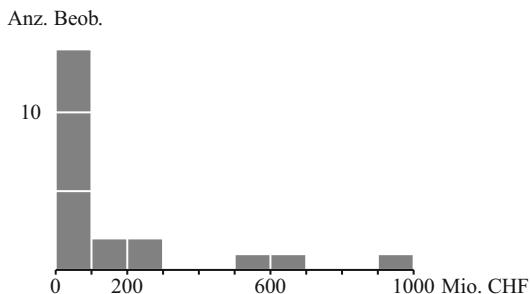
**Abb. 11.21** Luftbestandteile von trockener Luft (Knaurs Lexikon 1996)**Abb. 11.22** Logarithmierte Werte der Luftbestandteile

**Beispiel 11.13 (Luftbestandteile)** Dieses Beispiel illustriert, wie durch eine Transformation der Daten ein „unbalanciertes“ Stabdiagramm lesbarer wird. Tab. 11.9 zeigt, aus welchen Stoffen trockene Luft besteht. Ein Stabdiagramm (Abb. 11.21) kann nicht sichtbar machen, wie Luftbestandteile verteilt sind. Logarithmiert man die Volumen-Prozente, so werden kleine Wert sichtbarer. Dies zeigt Abb. 11.22. □

*Histogramme* (engl. *Chart bar*) sind wohl das bekannteste Instrument um viele univariat-stetige Datenwerte zu visualisieren. Sie entstehen, indem man aus dem Stamm *Klassen* macht und aus den Blättern Stäbe. Im Gegensatz zu Stabdiagrammen werden die Stäbe von Histogrammen miteinander verbunden.

**Beispiel 11.14 (Unwetterschäden)** Abb. 11.23 zeigt mit einem Histogramm die Verteilung der Schadensummen in Mio. CHF des Jahrs 1980. Die Klassen im Histogramm haben eine Länge von 100 Mio. CHF und die Klassenmitten sind die (einfachen) Zahlen 50 Mio. CHF, 150 Mio. CHF bis 950 Mio. CHF. Dabei wird eine Schadensumme von

**Abb. 11.23** Histogramm mit der Verteilung der Schadensummen (in Mio. CHF des Jahres 1980)



100 Mio. CHF in die erste Klasse (0–100 Mio. CHF) gelegt und eine solche von 200 Mio. CHF in die zweite Klasse getan.  $\square$

Erstellt man Histogramme, so gelten die folgenden Richtlinien:

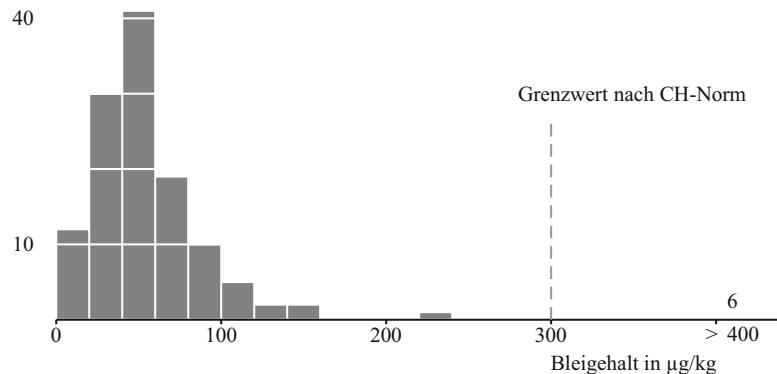
- (1) Die Klassenintervalle wählt man möglichst gleich gross. In der Regel gehört der obere Wert des Intervalls zur Klasse, der untere Wert nicht.
- (2) Die Klassenmitten (oder die Klassengrenzen) wählt man einfach.
- (3) Die Datenwerte sollen nicht auf wenige, sondern auf zahlreiche Stäbe verteilt werden.<sup>2</sup>

Eine bedeutende Regel, die eingehalten werden sollte, wenn man Histogramme erstellt, ist: *Die Flächen (und nicht etwa die Höhen) der Balken sind proportional zu der Anzahl der Datenwerte*. Dies spielt eine Rolle, wenn die Klassenintervalle ungleich gewählt werden.

**Beispiel 11.15 (Bleigehalte in Weinen)** Beim Beispiel 2.2 hat man 128 Messwerte. Ein Histogramm (oder ein Stamm-Blatt-Diagramm) visualisiert, wie die Daten verteilt sind. Um ein Histogramm zu zeichnen, sind ungefähr  $\sqrt{128} \approx 11$  Klassen nötig. Bis auf fünf Weinproben sind die Bleigehalte zwischen 0 und 240 µg/kg. Eine Klassenbreite von  $(240 - 0)/11 = 21,81 \approx 20$  µg/kg ist damit sinnvoll. Man erhält das Histogramm in Abb. 11.24. Die Verteilung hat einen langen Schwanz rechts. Sie ist *rechtsschwänzig*. Die sechs „extrem“ grossen Bleiwerte von über 400 µg/kg wurden zusammengefasst, damit das Bild nicht zu breit wird! Es wäre nicht richtig, über der Zahl 400 einen Balken der Höhe sechs zu zeichnen. Dieses würde einen falschen Eindruck dazu geben, wie die Daten verteilt sind. Wie in Abb. 11.24 gezeigt, schreibt man an die entsprechende Stelle die entsprechende Zahl.

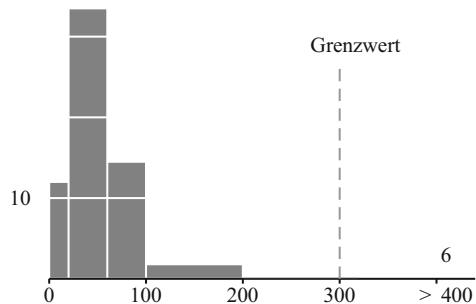
Das Histogramm in Abb. 11.25 zeigt die Verteilung der Bleigehalte mit den unterschiedlich grossen Klassen 0–20, 20–60, 60–100, 100–200 und 200–300. In der

<sup>2</sup> Eine meist funktionierende Faustregel, um dies zu verhindern, ist: Anzahl Klassen mindestens  $\approx \sqrt{n}$ , dabei ist  $n$  die Anzahl Datenwerte (falls  $n > 1000$  ist  $10 \log_{10} n$  eine geeignete Wahl).



**Abb. 11.24** Das Histogramm zeigt die Verteilung der 128 Bleigehalte aus dem Beispiel 2.2

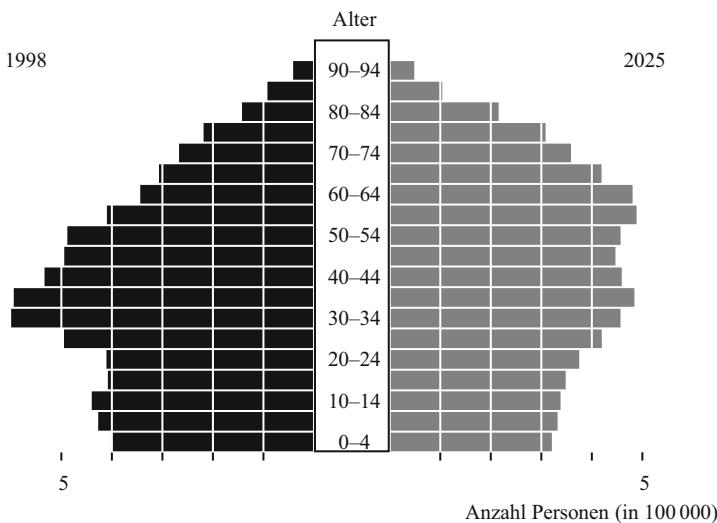
**Abb. 11.25** Ein Histogramm mit unterschiedlich grossen Klassen



Klasse 100–200 befinden sich neun Weine. Diese Klasse ist fünfmal so gross wie die kleinste Klasse 0–20. Daher beträgt die Höhe des zugehörigen Balkens nicht 9, sondern  $9/5 = 1,8$ .  $\square$

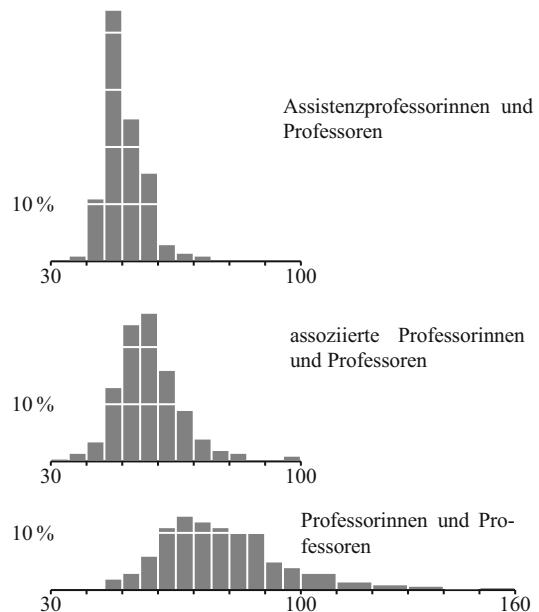
**Beispiel 11.16 (Altersstruktur)** Um die Altersstruktur von Bevölkerungsgruppen zu zeigen, wie etwa diejenige in der Schweiz im Jahre 1998 und eine Prognose für das Jahr 2025, können Histogramme Rücken an Rücken gelegt werden (Abb. 11.26). Die Abbildung zeigt, wie gross der prozentuale Anteil der über 60-jährigen Personen der Wohnbevölkerung im Jahre 2025 sein könnte.  $\square$

**Beispiel 11.17 (Einkommensverteilung)** Abb. 11.27 zeigt die Verteilung von Jahresgehältern im Jahr 2001 in 56 Mathematikdepartementen der USA – bestehend aus 206 Assistenz-, 417 assoziierten und 955 Professorinnen und Professoren – mit Histogrammen. Weil die Stichprobenumfänge verschieden gross sind, sind die Höhen der Stäbe relativ in Prozent angegeben. Ordentliche Professorinnen und Professoren aus der Stichprobe haben ein grösseres durchschnittliches Einkommen, aber eine markant grössere Streuung als Assistenzprofessorinnen und -professoren.  $\square$



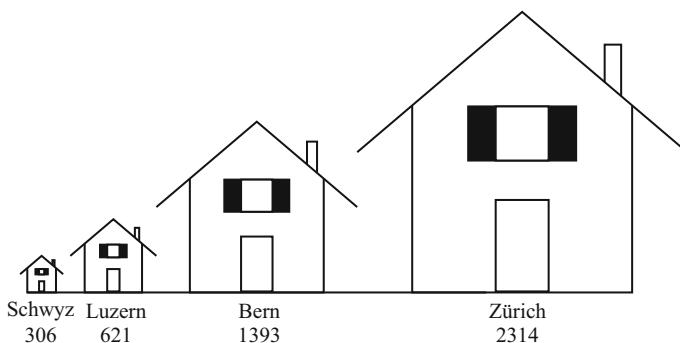
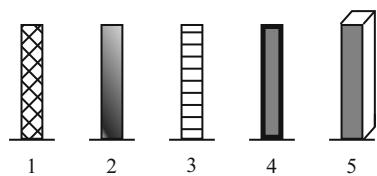
**Abb. 11.26** Altersstruktur der Schweiz in den Jahren 1989 und 2025 (Daten, Statistisches Jahrbuch der Schweiz, 2000)

**Abb. 11.27** Verteilung der Jahresgehalts (in tausend Dollar) für verschiedene Gruppen (Notices of the AMS, 48 (2), 2001)



Es ist empfehlenswert die Balken von Histogrammen schmucklos zu halten. Scheinbar „kunstvolle“, die Betrachterin oder den Betrachter ablenkende Ausschmückungen, drei-dimensionalen Bilder wie in Abb. 11.28, sowie Verzerrungen bringen keinen Mehrwert. Die drei ersten Balken in der Abbildung sind durch unnötigen grafischen Plunder

**Abb. 11.28** Balkenstäbe mit unerwünschten inhaltslosen Dekorationen



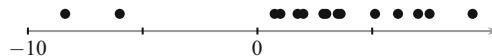
**Abb. 11.29** Gebaute neue Einfamilienhäuser im Jahr 1998 in vier verschiedenen Kantonen (Lügenfaktor 7,56)

von Computerprogrammen gekennzeichnet. Der Balken ganz rechts beinhaltet die dritte Dimension, die keine Information trägt. Auch beim scheinbar schlichten vierten Balken wird die Höhe auf vier (!) Arten wiederholt dargestellt: durch die Höhen der linken und rechten Striche, durch die Position des horizontalen Strichs und schliesslich durch die Schattierung des Balkens. Technisch wird dabei Tinte verschwendet.

**Beispiel 11.18 (Einfamilienhäuser)** Im Jahre 1998 wurden im Kanton Schwyz 306, im Kanton Luzern 612, im Kanton Bern 1393 und im Kanton Zürich 2314 neue Einfamilienhäuser erstellt. Stellt man die Daten mit einem dekorierten Stabdiagramm wie in Abb. 11.29 dar, erhält man einen Lügenfaktor von 7,56. Die Höhen der verzierten Stäbe (= Häuser) stimmen mit den Grössen der Daten überein, durch die grafische Gestaltung der „Balken“ als Häuser und den damit verbundenen zweidimensionalen Effekt sind die dargestellten Flächen nicht mehr proportional zu den Daten. □

### 11.3 Ausreisser und Extremwerte

Aussergewöhnliche Datenwerte, die weit von der Hauptmasse der Daten entfernt sind, nennt man *Ausreisser* oder *Extremwerte*. Es kann interessant sein, diese genauer zu untersuchen. Bei Produktionsgrössen können sie bedeuten, dass die Produktion nicht mehr unter statistischer Kontrolle liegt. Solche Datenwerte können aber auch Resultate zu nicht direkt messbaren Grössen stark beeinflussen.



**Abb. 11.30** 15 Messwerte (in Inch) von Wachstumsdifferenzen der Pflanze *Zea mays*

**Beispiel 11.19 (Wachstum von Mais)** Charles Darwin untersuchte im Jahr 1878, ob Pflanzen, die von anderen Pflanzen befruchtet werden, mehr wachsen als Pflanzen, die sich selber bestäuben. Dazu muss die *durchschnittliche Differenz*  $\Delta$  zwischen dem Wachstum von Pflanzen bei Fremdbestäubung und demjenigen bei Selbstbestäubung bestimmt werden. Darwin untersuchte 15 Maispflanzen (*Zea mays*). Er erhielt die folgenden Wachstumsdifferenzen (in Inch, aus [8]):

$$\begin{array}{cccccccccc} 6,125 & -8,375 & 1,000 & 2,000 & 0,750 & 2,875 & 3,500 & 5,125 \\ 1,750 & 3,625 & 7,000 & 3,000 & 9,375 & 7,500 & -6,000 \end{array}$$

Auffallend sind die Datenwerte  $-8,375$  und  $-6,000$ , die sich deutlich von den anderen Werten unterscheiden. Dies zeigt auch Abb. 11.30 der markierten Werte auf einer Zahlengeraden.

Um den gesuchten Parameter  $\Delta$  zu berechnen, könnte man annehmen, dass die Messwerte normalverteilt um  $\Delta$  streuen. Hat man minimale Vorinformation zu  $\Delta$  und der Streuung der Normalverteilung, so ist die A posteriori-Verteilung mit Theorem 10.2 berechenbar. Der plausibelste Wert  $\Delta_0$  ist das arithmetische Mittel der 15 Datenwerte:  $\Delta_0 = 2,62$ . Ein Wahrscheinlichkeitsintervall zum Niveau 0,95 ist

$$\Delta = 2,62 \pm 2,61$$

Was passiert, wenn man ohne die beiden aussergewöhnlichen Werte  $-8,375$  und  $-6,000$  rechnet? Der plausibelste Wert ist nun 4,13 und man hat

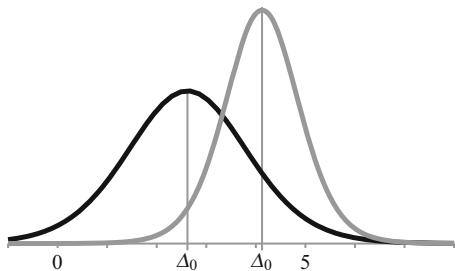
$$\Delta = 4,13 \pm 1,63$$

Beide Abschätzungen und Abb. 11.31 mit den A posteriori-Verteilungen zeigen deutlich, wie stark die aussergewöhnlichen Werte auf das Resultat von  $\Delta$  wirken. Einerseits wird  $\Delta$  deutlich kleiner und andererseits bewirken sie eine kleinere Präzision für  $\Delta$ . Dies ist problematisch.  $\square$

Aussergewöhnliche Werte sollten daher analysiert werden:

- (1) War das Messgerät nicht geeicht oder kalibriert? Sind die aussergewöhnlichen Werte entstanden, weil das Messgerät falsch abgelesen wurde? Hat man Komastellen falsch aufgeschrieben? Man spricht hier von *groben Fehlern* (engl. *gross errors*). Solche Datenwerte müssen entfernt werden.

**Abb. 11.31** Plausibilität zu  $\Delta$ :  
in schwarzer Farbe mit allen  
Messwerten, in grauer Farbe  
ohne die beiden aussergewöhnlichen  
Messwerte



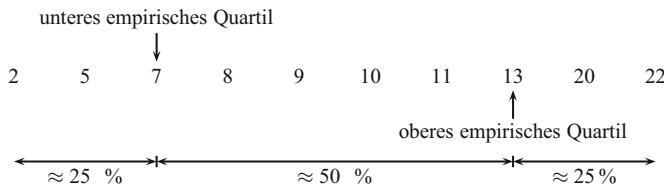
- (2) Hat man die Faktoren, die im Ursache-Wirkungs-Diagramm aufstellt wurden, unter Kontrolle? Sind Faktoren oder Kovariablen, die einen wesentlichen Einfluss auf die Messgröße haben, vergessen worden? Kann man daraus etwas lernen, um die Messgröße besser zu verstehen?
- (3) Denken Sie auch an das scheinbar Unmögliche: Sind vielleicht die aussergewöhnlichen Datenwerte die „normalen“ Werte und die „normalen“ Werte durch nicht geeichte oder falsch kalibrierte Geräte entstanden?

**Beispiel 11.20 (Wachstum von Mais)** Darwin stellte fest, dass bei den im ersten Topf gesetzten Pflanzen eine krank war, eine nach dem Versuch schnell starb und eine Pflanze nie ihre normale Größe erreichte. Es ist daher plausibel, diese Messungen auszuschliessen. Dann verbleiben die Werte

2,000	0,750	2,875	3,500	5,125	1,750
3,625	7,000	3,000	9,375	7,500	– 6,000

mit nur einem aussergewöhnlichen Wert. Um die ursprüngliche Anzahl von Datenwerten beizubehalten, müssen zusätzliche drei Messungen durchgeführt werden. Hat man Glück, entstehen keine weiteren Extremwerte, hat man Pech (Krankheit der Pflanze, unerklärbares Absterben) ist man in der gleichen Situation wie vorher. Das Problem aber bleibt, Daten ausgedünnt zu haben (die Reduktion von 15 auf 12 Messungen). Dies führt zum Verlust von Information und zu geringerer Präzision für die gesuchte Größe  $\Delta$ .  $\square$

Aussergewöhnliche Werte können ein wichtiges Merkmal einer Messgröße sein. Krankheiten bei Pflanzen, die nie ihre erwartete Länge erreichen, sind üblich und sollten daher in der Rechnung bleiben. Extreme Werte bei Produktionsprozessen weisen darauf hin, dass der Prozess schlecht unter Kontrolle ist oder mehr streut als erwartet. Untersucht man Unwetterschäden, so sind extreme Werte die interessantesten. Aussergewöhnliche



**Abb. 11.32** Bestimmung von empirischen Quartilen

Werte können zudem auf einen neuen, unbekannten Effekt hinweisen. Die Relevanz von solchen Werten betont der Statistiker J. A. Rice in [5]:

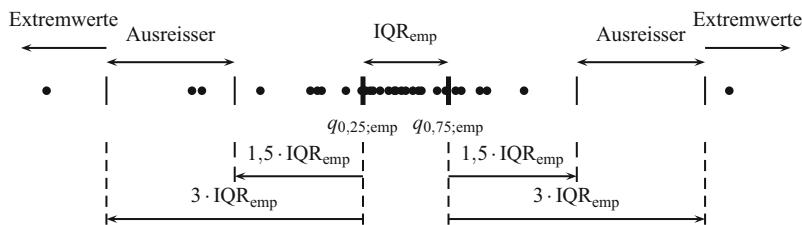
Although outliers are often unexplainable aberrations, an examination of them and their causes can sometimes deepen an investigator's understanding of the phenomena under study.

Es ist daher nicht zu empfehlen, die extremen Werte einfach zu entfernen. Überprüft man Daten auf aussergewöhnliche Werte, dient dies dazu, sich zu überlegen, (a) warum sie entstanden sind, (b) welchen *Effekt* sie auf statistische Rechnungen haben und (c) wie man diesen Effekt durch geschickt gewählte Modelle minimieren kann.

Wann sind Werte wirklich aussergewöhnlich? Dies wird in der Regel mit der beobachteten Streuung der Datenwerte definiert. So werden bei kleiner Streuung der Daten schon Werte nahe der Hauptmasse als extrem bezeichnet. Als Streumass dient hier die *empirische Quartilsdifferenz* (engl. *empirical interquartile range* [IQR<sub>emp</sub>]). Sie ist die Differenz zwischen dem oberen empirischen Quartil  $q_{0.75;\text{emp}}$  und dem unteren empirischen Quartil  $q_{0.25;\text{emp}}$ . Das obere empirische Quartil ist eine Zahl derart, dass 25 % der Datenwerte grösser gleich und 75 % kleiner oder gleich diese Zahl sind. Zwischen dem oberen und dem unteren empirischen Quartil befinden sich 50 % der Datenwerte. Besteht die Rangliste aus den zehn Zahlen 2, 5, 7, 8, 9, 10, 11, 13, 20 und 22, so hat man  $25\% \cdot 10 = 2,5$ . Als unteres empirisches Quartil kann man daher den drittkleinsten Wert der Daten nehmen. Analog ist das obere empirische Quartil der drittgrösste Wert der Daten. Beim Beispiel beträgt die empirische Quartilsdifferenz deshalb  $13 - 7 = 6$ . Abb. 11.32 illustriert dies.

Datenwerte werden als *Ausreisser* (engl. *outlier*) und *Extremwerte* (engl. *outer fence*) gemäss Abb. 11.33 bezeichnet. Ist beispielsweise ein Wert grösser als das obere empirische Quartil plus dreimal die empirische Quartilsdifferenz IQR<sub>emp</sub>, so nennt man ihn einen Extremwert. In der Abbildung hat man zwei Ausreisser, sowie einen tiefen und einen hohen Extremwert.<sup>3</sup>

<sup>3</sup> Für Grössen, die mit einer Normalverteilung modelliert werden, bedeuten die Regeln: (1) Ausreisser: Werte, die mindestens  $2,9 \cdot \sigma$  aber nicht mehr als  $4,7 \cdot \sigma$  vom Modus entfernt sind, (2) Extremwert: Datenwerte, die mehr als  $4,7 \cdot \sigma$  vom Modus entfernt sind.



**Abb. 11.33** Definition von Ausreisern und Extremwerten in Funktion der empirischen Quartilsdifferenz  $IQR_{\text{emp}}$

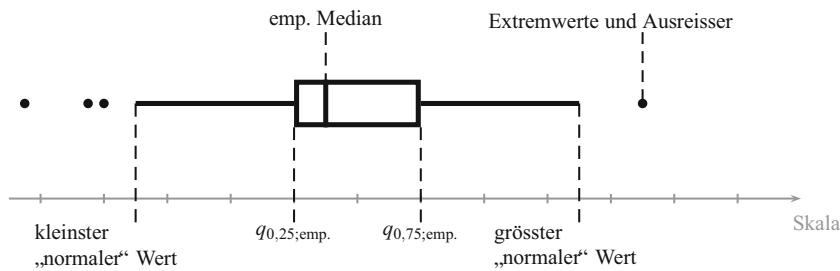
**Beispiel 11.21 (Unwetterschäden)** Das Stamm-Blatt-Diagramm in Abb. 11.9 zeigt drei grosse isolierte Schadensummen 579 Mio. CHF, 609 Mio. CHF und 969 Mio. CHF. Es sind 21 Beobachtungen vorhanden. 25 % von 21 sind 5,25. Damit sind 25 % der Beobachtungen kleiner als der sechst kleinste Wert der Schadensummen. Dieser ist das untere empirische Quartil: 39 Mio. CHF. Analog ist die sechstgrösste Schadensumme das obere empirische Quartil. Sie beträgt 137 Mio. CHF. Die Quartilsdifferenz ist deshalb  $137 \text{ Mio. CHF} - 39 \text{ Mio. CHF} = 98 \text{ Mio. CHF}$ . Eine beobachtete grosse Schadensumme ist ein Extremwert, wenn sie grösser als  $137 + 3 \cdot 98 = 431 \text{ Mio. CHF}$  ist. Dies sind die Schadensummen 579 Mio. CHF, 969 Mio. CHF und 609 Mio. CHF. Eine grosse Schadensumme ist ein Ausreisser, wenn sie zwischen  $137 + 1,5 \cdot 98 = 284 \text{ Mio. CHF}$  und  $137 + 3 \cdot 98 = 431 \text{ Mio. CHF}$  liegt. Es gibt einen solchen Wert: 288 Mio. CHF. □

**Beispiel 11.22 (Wachstum von Mais)** Hier hat man 15 Datenwerte. 25 % von 15 ist 3,75. Der viertkleinste und der viertgrösste Messwert bilden daher die empirischen unteren und oberen Quartile. Die Werte sind 1 Inch und 6,125 Inch. Die empirische Quartilsdifferenz ist 5,125 Inch. Werte kleiner als  $1 - 3 \cdot 5,125 = -14,375 \text{ Inch}$  sind Extremwerte. Es hat also keine Extremwerte. Werte zwischen  $-14,375 \text{ Inch}$  und  $1 - 1,5 \cdot 5,125 = -6,875 \text{ Inch}$  sind Ausreisser. Es hat einen solchen Wert:  $-8,375 \text{ Inch}$ . Der negative Messwert  $-6,000 \text{ Inch}$  ist ein gewöhnlicher Datenwert. □

Da Ausreisser oder Extremwerte Resultate zu Größen stark beeinflussen können, ist es sinnvoll, solche Werte in Berichten zu benennen. Oft werden dazu *Kistendiagramme*, auch *Box & Whisker Plots*<sup>4</sup> genannt, benutzt. Sie visualisieren die Extremwerte und Ausreisser, die beiden empirischen Quantile, sowie den empirischen Median (der Wert in der Mitte der Rangliste der Datenwerte).

Bei Statistikprogrammen sehen Kistendiagramme meist wie in Abb. 11.34 aus.

<sup>4</sup> Box & Whisker Plots wurden von J. W. Tukey erfunden, siehe [10].



**Abb. 11.34** Box & Whisker Plot, wie ihn viele Statistikprogramme zeichnen

**Beispiel 11.23 (Nicht keimende Blumenzwiebeln)** Beim Beispiel 1.7 der Gärtnerei werden Anteile nicht keimender Blumenzwiebeln gezählt. Der empirische Median ist der Wert in der Mitte der Rangliste der 50 Messungen. Da 50 eine gerade Zahl ist, besteht diese Mitte aus zwei Zahlen, dem 25. und 26. Wert der Rangliste. Als empirischen Median wählt man daher den Mittelwert der beiden Werte:

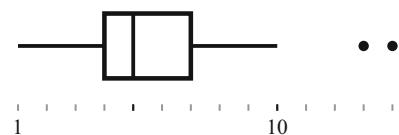
$$\text{emp. Median} = \frac{5 + 5}{2} = 5$$

Das obere empirische Quartil ist sieben, das untere Quartil ist vier und die Quartilsdifferenz beträgt drei. Ausreisser und Extremwerte sind Anteile, die grösser als  $7 + 1,5 \cdot 3 = 11,5$  sind. Es sind die zwei grössten beobachteten Anteile 13 % und 14 %. Der grösste „normale“ Anteil ist 10 %. Damit erhält man den Box & Whisker Plot in Abb. 11.35. □

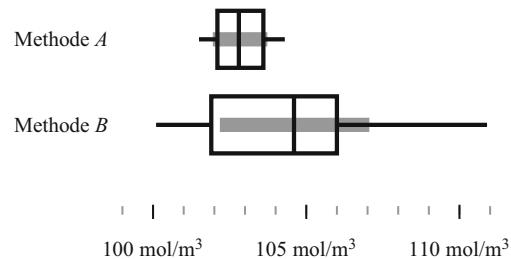
**Beispiel 11.24 (Chloridgehalt)** Die Stamm-Blatt-Diagramme in Abb. 11.13 zeigen, wie die gemessenen Chloridgehalte der Methode A und B verteilt sind. Man kann die Daten auch mit Box & Whisker Plots visualisieren. Um die Datenwerte vergleichen zu können, sind diese in Abb. 11.36 untereinander gezeichnet. Die Abbildung zeigt, wie unterschiedlich die Messwerte der Methoden A und B streuen. Mit den grauen Balken sind die Resultate zum gesuchten Chloridgehalt Cl eingebaut. Es sind Wahrscheinlichkeitsintervalle zum Niveau 0,95 für Cl.

□

**Abb. 11.35** Box & Whisker Plot der nicht keimenden Blumenzwiebeln



**Abb. 11.36** Box & Whisker Plots der beiden Messreihen zum Chloridgehalt



Von E. R. Tufte wurden in [9] einfache Versionen des Box & Whisker Plots vorgeschlagen, welche die gleiche Information weniger aufwändig darstellen:

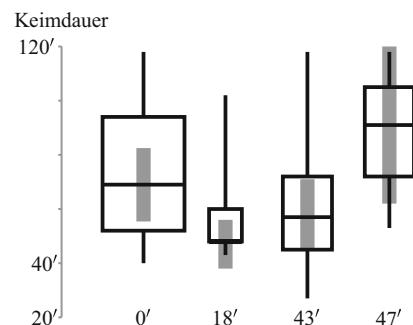
oder



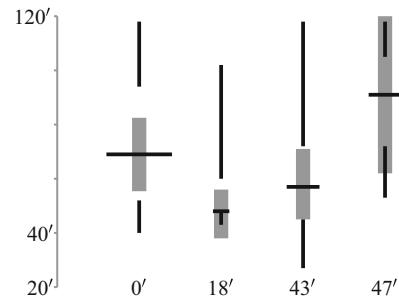
Im ersten Bild stellt die Lage des Punktes den empirischen Median dar und die beiden Striche beginnen beim kleinsten (bzw. grössten) „normalen“ Wert und enden beim unteren (bzw. oberen) empirischen Quartil.

**Beispiel 11.25 (Keimdauer)** Abb. 11.37 und Abb. 11.38 zeigen verschiedene Box & Whisker Plots aus der Untersuchung [1] zur Keimdauer einer Pflanze in Funktion der Zeit (0, 18, 43 oder 47 Minuten), die die Fruchtkerne im Magen einer Eidechse lagen. Abb. 11.37 ist in traditioneller Form nach Tukey realisiert. Die Breite der Boxen ist dabei proportional zur Wurzel der Anzahl Messungen (hier 24, 5, 9 und 7). Abb. 11.38 ist in moderner Form nach E. R. Tufte realisiert. In beiden Grafiken sind mit grauen Balken Wahrscheinlichkeitsintervalle zum Niveau 0,68 für die *durchschnittliche* Keimdauer eingezeichnet.

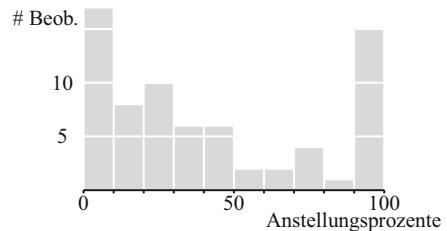
**Abb. 11.37** Keimdauern einer Pflanze in Funktion der Zeit, die die Fruchtkerne im Magen einer Eidechse lagen



**Abb. 11.38** Keimdauren dargestellt mit Box & Whisker Plots nach Tufte



**Abb. 11.39** Verteilung der Stellen nach Anstellungsprozenten im Lehrerseminar Liestal (2001)



**Abb. 11.40** Darstellung der Verteilung der Stellen nach Anstellungsprozenten im Lehrerseminar Liestal



Box & Whisker Plots eignen sich in der Regel nicht, um multimodale Verteilungen darzustellen. Abb. 11.39 zeigt die Verteilung der Anstellungsprozente der 71 Dozierenden am Lehrerseminar Liestal im Wintersemester 2000/01. Die Verteilung ist „U-förmig“. Der Box & Whisker Plot der Anstellungsprozente in Abb. 11.40 gibt ein falsches Bild der Verteilung. Multimodal verteilte Datenwerte lassen sich kaum durch Box & Whisker Plots darstellen.

## 11.4 Robustere Datenmodelle

Ausreisser und Extremwerte können Resultate zu gesuchten Parametern stark beeinflussen. So verändert bei dem in diesem Kapitel besprochenen Beispiel zum Wachstum von Mais ein Ausreisser den plausibelsten Wert für die durchschnittliche Wachstumsdifferenz um fast 50 %. Es kann daher sinnvoll sein, Datenmodelle so zu wählen, dass Ausreisser und Extremwerte das Resultat nicht stark ändern. Solche Modelle nennt man *robuste* Modelle. Es werden drei solche Modelle vorgestellt, die in [7] mathematisch hergeleitet werden:

- (1) Das Datenmodell der Normalverteilung bewirkt, dass Datenwerte, die weit vom Modus  $\mu$  entfernt sind, nur mit kleinsten Wahrscheinlichkeiten auftreten können. Dramatisch wird die Situation, wenn die Streuung  $\sigma$  klein angenommen wird. Dies hat man bei minimaler Vorinformation:  $\text{pdf}(\sigma) \propto 1/\sigma$ . Statt Plausibilität zur Streuung  $\sigma$  zu formulieren, kann man deshalb versuchen, den *minimalen* Wert  $\sigma_{\min}$  für die Streuung  $\sigma$  zu beschreiben. Dies führt in dieser Situation zum Datenmodell

$$\text{pdf}(\text{Messwert} = x | \mu, \sigma_{\min}) = \frac{1}{\sqrt{2\pi \cdot \sigma_{\min}^2}} \cdot \frac{1 - \exp\{-0,5(x - \mu)^2 / \sigma_{\min}^2\}}{(x - \mu)^2 / \sigma_{\min}^2}$$

In Abb. 11.41 findet sich ein Graph dieses Modells. Im Gegensatz zur Normalverteilung sind nun auch Werte weit weg vom Modus  $\mu$  plausibel.

- (2) Man kann versuchen, statt mit der Normalverteilung mit einer symmetrischen Verteilung mit Modus  $\mu$  zu arbeiten, die nicht exponentiell schnell abnimmt (und damit aussergewöhnliche Messwerte eher zulässt). Ein solches Modell ist die  $t$ -Verteilung mit Dichtefunktion

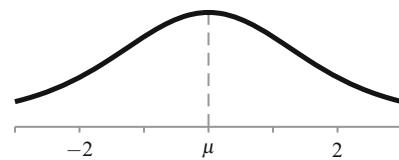
$$\text{pdf}(\text{Messwert} = x | \mu, \sigma, \beta) \propto \left(1 + \frac{(x - \mu)^2}{\beta \cdot \sigma^2}\right)^{-(\beta+1)/2}$$

Kurz und prägnant kann man dies wie folgt schreiben:

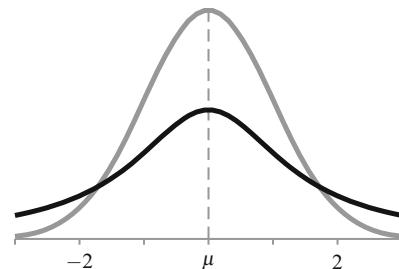
$$\text{Datenmodell: } i\text{-ter Messwert} \sim \text{StudentT}(\beta, \mu, \sigma)$$

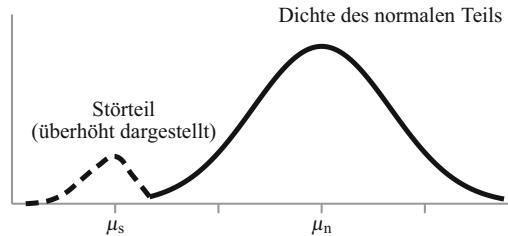
Dabei nennt man  $\beta \geq 1$  den Freiheitsgrad der Verteilung. Ist  $\beta = 1$ , so nennt man dies die *Cauchyverteilung*. Ist  $\beta = \infty$ , so erhält man wieder die Normalverteilung. Abb. 11.42 zeigt diese beiden Datenmodelle als Extremfälle der  $t$ -Verteilung.

**Abb. 11.41** Korrigiertes Modell mit  $\mu = 0$  und  $\sigma_{\min} = 1$



**Abb. 11.42** Normalverteilung (graue Kurve) und Cauchyverteilung (schwarze Kurve) mit  $\mu = 0$  und  $\sigma = 1$  als extreme Fälle der  $t$ -Verteilung



**Abb. 11.43** Mischverteilung

- (3) Ein drittes Modell teilt die Datenwerte in zwei Kategorien auf. Die „normalen“ Werte treten dabei mit hoher Wahrscheinlichkeit  $1 - \epsilon$  und die aussergewöhnlichen mit Wahrscheinlichkeit  $\epsilon < 0,5$  auf (siehe Abb. 11.43). Die Dichtefunktion dieses Datenmodells ist damit

$$\begin{aligned} \text{pdf}(\text{Messwert} = x \mid \epsilon, \mu_n, \mu_s, \sigma_n, \sigma_s) \\ = (1 - \epsilon) \cdot \underbrace{\text{pdf}_n(x \mid \mu_n, \sigma_n)}_{\text{Normalverteilung}} + \epsilon \cdot \underbrace{\text{pdf}_s(x \mid \mu_s, \sigma_s)}_{\text{Normalverteilung}} \end{aligned}$$

Man nennt dieses dritte Modell auch eine *Mischverteilung* (engl. *mixture model*), die mathematisch oft kurz

$$i\text{-ter Messwert} \sim (1 - \epsilon) \cdot \text{Normal}(\mu_n, \sigma_n) + \epsilon \cdot \text{Normal}(\mu_s, \sigma_s)$$

notiert wird. Sowohl die normalen, wie die aussergewöhnlichen Werte werden oft mit Normalverteilungen mit Moden  $\mu_n$ ,  $\mu_s$  und Standardabweichungen  $\sigma_n$  und  $\sigma_s$  modelliert. Der Nachteil ist, dass fünf Parameter bestimmt werden müssen. Bei wenigen Messungen und minimaler Vorinformation zu den Parametern ist es kaum möglich, die Parameter präzise zu lokalisieren. Verwendet wird diese Methode daher, wenn viele Datenwerte vorliegen oder Vorwissen zu den Parametern besteht.

**Beispiel 11.26 (Wachstum von Mais)** Mit der obigen ersten Methode kann man versuchen, die durchschnittliche Grössendifferenz  $\Delta$  der Pflanzen bei Fremdbestäubung und bei Selbstbestäubung zu berechnen. Die Regel von Bayes sagt

$$\underbrace{\text{pdf}(\Delta, \sigma_{\min} \mid \text{Daten, Vorwissen})}_{\text{A posteriori-Verteilung}} \propto \underbrace{\mathbb{P}(\text{Daten} \mid \Delta, \sigma_{\min})}_{\text{Likelihood}} \cdot \underbrace{\text{pdf}(\Delta, \sigma_{\min} \mid \text{Vorwissen})}_{\text{Prior}}$$

Ausser den Messungen liege keine weitere Information vor. Daher hat man:

$$\text{pdf}(\Delta, \sigma_{\min} \mid \text{Vorwissen}) \propto 1/\sigma_{\min}$$

Mit dem Modell von Punkt (1) ist:

$$\mathbb{P}(\text{Messwert} = x \mid \Delta, \sigma_{\min}) \propto \frac{1}{\sigma_{\min}} \cdot \frac{1 - \exp\{-0,5 \cdot (x - \Delta)^2 / \sigma_{\min}^2\}}{(x - \Delta)^2 / \sigma_{\min}^2}$$

Nimmt man an, dass die Datenwerte unabhängig modelliert werden können, so ist die Likelihood ein Produkt von 15 solcher Faktoren. Der erste Faktor mit  $x = 6,125$  (dem 1. Messwert), der zweite Faktor mit  $x = -8,375$  (dem 2. Messwert) und so fort. Damit ist die gemeinsame A posteriori-Verteilung für die Parameter  $\Delta$  und  $\sigma$  mit der Regel von Bayes bestimmt. Sie lautet:

$$\text{pdf}(\Delta, \sigma_{\min} | \text{Dat., min. Vor.}) \propto \left( \frac{1}{\sigma_{\min}} \right)^{16} \cdot \prod_{i=1}^{15} \frac{1 - \exp\{-0,5 \cdot (x_i - \Delta)^2 / \sigma_{\min}^2\}}{(x_i - \Delta)^2 / \sigma_{\min}^2}$$

Das Zeichen  $\prod$  bedeutet das Produkt der Faktoren mit  $x_1 = 6,125, x_2 = -8,375, \dots, x_{15} = -6,000$ . Mit einer MCMC-Simulation (und einer Kette von 100 000 Punkten) wird diese Verteilung handhabbar. Daraus lässt sich die Randverteilung für  $\Delta$  bestimmen. Mit einer Wahrscheinlichkeit von 0,95 ist

$$\Delta = 3,36 \pm 1,74$$

Der plausibelste Wert für  $\Delta$  ist 3,36. Er ist nicht mehr das arithmetische Mittel der fünfzehn Messwerte.

Was passiert, wenn man mit diesem Modell ohne die zwei aussergewöhnlichen Werte von  $-8,375$  und  $-6,000$  arbeitet? Man erhält ein Wahrscheinlichkeitsintervall von

$$\Delta = 3,54 \pm 1,50$$

Im Gegensatz zur Normalverteilung beeinflussen die beiden aussergewöhnlichen Werte das Resultat zu  $\Delta$  nun kaum mehr! □

Die obige Diskussion des Maisbeispiels zeigt, dass eine vertiefte Analyse der Daten notwendig ist, wenn Parameter mit arithmetischen Mitteln geschätzt werden und extreme Werte auftauchen. Der Statistiker F. Hampel betont in [2], dass aussergewöhnlich grosse oder kleine Datenwerte nicht mit groben Fehlern zu verwechseln sind. So sind zwar die meisten groben Fehler auch Ausreisser oder Extremwerte, aber sie können auch in den gewöhnlichen Messwerten versteckt sein. Umgekehrt sind Ausreisser oder Extremwerte oft grobe Fehler, aber sie können auch „normale“ Messwerte sein. F. Hampel weist darauf hin, dass aussergewöhnlich grosse oder kleine Datenwerte, die keine groben Fehler sind, wertvoll sind: Sie zeigen neue, unerwartete Phänomene. Als Beispiel nennt er die Entdeckung des Ozonlochs oberhalb der Antarktis, wo die gemessenen Ozonwerte im Vergleich zu alten Messwerten Ausreisser waren.

## 11.5 Weiterführende Literatur

Ausführliche und weiterführende Informationen, wie man Daten darstellen und Resultate von statistischen Rechnungen präsentieren kann, findet man in:

1. J. Bertin, *La Sémiologie Graphique* (Paris, Mouton, 1969)
2. E. R. Tufte, *The Visual Display of Quantitative Information*, Seventeenth printing (Graphics Press, Cheshire, Conn., 1983)
3. E. R. Tufte, *Envisioning Information*, Seventh printing (Graphics Press, Cheshire, Conn., 1990)
4. E. R. Tufte, *Visual Explanation, Images and Quantities, Evidence and Narrative*, Fourth printing (Graphics Press, Cheshire, Conn., 1997)
5. J. W. Tukey, *Exploratory Data Analysis* (Addison & Wesley, Reading, Mass., 1977)
6. H. Wainer, *Graphic Discovery, A Trout in the Milk and Other Visual Adventures* (Princeton University Press, Princeton and Oxford, 2005)
7. H. Wainer, *Picturing the Uncertain World, How to Understand, Communicate, and Control Uncertainty through Graphical Display* (Princeton University Press, Princeton and Oxford, 2009)

## Reflexion

**11.1** Ein Fischzüchter will wissen, wie gross die durchschnittliche Masse  $\bar{m}_{\text{Fische}}$  der Fische in einem Wasserbecken ist. Dazu wurden 15 Fische gewogen. Hier die Resultate (in g) – die Urliste wird entlang der Zeilen gelesen –:<sup>5</sup>

898	2050	1198	644	1294	1450	1516	1452
1196	1200	650	1206	836	1290	1380	

Weiter weiss man, dass die Fischmassen zwischen 100 und 5000 g liegen müssen.

- (a) Speichern Sie die Datenwerte in einem Statistikprogramm.
- (b) Kontrollieren Sie mit einem Statistikprogramm: Sind die Daten unter statistischer Kontrolle? Ist es plausibel anzunehmen, dass die Messwerte unabhängig modelliert werden können?
- (c) Ordnen Sie die Daten mit Hilfe eines Stamm-Blatt-Diagramms und stellen Sie sie in einem Histogramm dar. Was fällt auf? Beschreiben Sie Ihre Beobachtungen. Wie ist die Stichprobe verteilt?
- (d) Zeichnen Sie mit einem Statistikprogramm den Box & Whisker Plot der Stichprobe. Sind Extremwerte oder Ausreisser vorhanden?

---

<sup>5</sup> Experiment aus einem Versuch zur Bestimmung des Wachstums von Stören bei der Firma Tropenhaus Frutigen AG, 2007.

**Tab. 11.10** 40 Größen von 40 Zellkulturen (in mm<sup>2</sup>)

60,9	37,4	74,5	115,1	15,0	91,5	35,4	78,8	44,3	49,3
51,8	10,5	106,7	7,0	56,6	61,8	121,3	62,1	94,5	45,5
47,6	22,6	36,5	81,1	93,5	30,5	31,4	39,3	71,4	55,4
27,0	37,7	56,2	165,8	27,8	19,3	29,0	63,3	49,5	52,8

- (e) Berechnen Sie mit einem geeigneten Modell die durchschnittliche Masse  $\bar{m}_{\text{Fische}}$ . Geben Sie an: plausibelster Wert, Wahrscheinlichkeitsintervalle zum Niveau 0,5 und 0,95. Überprüfen Sie auch Ihr gewähltes Modell mit einem QQ-Plot.

**11.2** Tab. 11.10 gibt die Größenordnung in mm<sup>2</sup> von 40 Zellkulturen an.

- (a) Ordnen Sie die Werte mit Hilfe eines Stamm–Blatt-Diagramms. Zeichnen Sie auch ein Histogramm. Was fällt auf? Beschreiben Sie Ihre Beobachtungen.  
 (b) Wie lauten der empirische Median, die empirischen Quartile und die empirische Quartilsdifferenz?  
 (c) Wie sind die Datenwerte verteilt? Wie beurteilen Sie den Wert 165,8? Ist er ein Ausreisser oder ein Extremwert? Zeichnen Sie den Box & Whisker Plot der Messwerte.

**11.3** Sie haben 1500 Datenwerte zwischen 10,8 cm und 11,9 cm. Wie wählen Sie die Klasseneinteilung für ein Histogramm der Daten? Finden Sie heraus, wie Sie bei Ihrem Statistikprogramm die Klasseneinteilung bei Histogrammen selber wählen können. Probieren Sie dies an Beispielen aus!

**11.4** Eine Maschine stellt elektrische Widerstände her. Die Spezifikation lautet: (100,0 ± 3,0) Ω. Zur Qualitätskontrolle wurden 20 Widerstände gemessen:

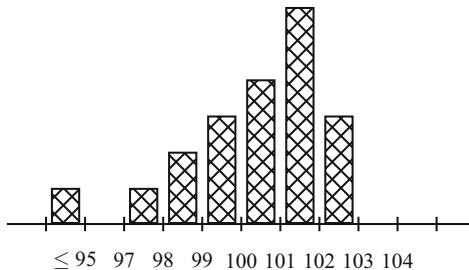
101,8 100,0 102,0 101,5 98,2 102,6 100,0 102,2 101,7 97,8  
 100,0 100,1 101,5 100,9 100,1 102,4 91,1 101,3 98,4 100,7

Um die Messwerte grafisch darzustellen, wurde Abb. 11.44 gezeichnet.

- (a) Finden Sie sechs unprofessionelle Umsetzungsvariablen in der Grafik, die eine erfahrene Benutzerin von statistischen Werkzeugen nicht wählen würde.  
 (b) Zeichnen Sie einen Box & Whisker Plot, um die Daten zu visualisieren. Sind Ausreisser oder Extremwerte vorhanden?

**11.5** Tab. 11.11 zeigt dreissig aufeinanderfolgende Werte von Regenmengen im Monat März für Minneapolis/St. Paul aus [4]. Daraus möchte man die durchschnittliche Regenmenge pro Tag für zukünftige Monate März bestimmen.

**Abb. 11.44** Ein Histogramm, das nicht professionell erstellt ist



- (a) Untersuchen Sie die Beobachtungen mit grafischen Werkzeugen der EDA. Arbeiten Sie dabei mit mehreren grafischen Darstellungen. Beschreiben Sie Ihre Beobachtungen.
- (b) Hat es Trends in den Datenwerten? Sind die Beobachtungen unter statistischer Kontrolle? Können die Datenwerte als unabhängig betrachtet werden?
- (c) Ihrer Chefin müssen Sie Ihre Untersuchungen und Beobachtungen vorstellen. Sie dürfen dabei eine grafische Darstellung benutzen. Welche Art wählen Sie?
- (d) Berechnen Sie mit einem geeigneten Modell die durchschnittliche Regenmenge pro Tag für den Monat März. Geben Sie an: plausibelster Wert, Wahrscheinlichkeitsintervalle zum Niveau 0,5 und 0,95.

**11.6** Im Rahmen der Untersuchung der Biodiversität in der Schweiz interessiert man sich für die durchschnittliche Anzahl  $\mu_A$  und  $\mu_B$  von Tagfalterarten pro Region im Ackerland und in den Bergen. Dazu wurden im Jahr 2003 in je 23 Regionen die Tagfalterarten gezählt. Die Resultate sind in Tab. 11.12 dargestellt.

- (a) Zeichnen Sie Stamm-Blatt-Diagramme der Daten.
- (b) Zeichnen Sie die Box & Whisker Plots für diese Daten in einem Bild. Sind Ausreisser und Extremwerte vorhanden?
- (c) Berechnen Sie mit einem geeigneten Modell die Parameter  $\mu_A$  und  $\mu_B$ .
- (d) Wie gross ist die Wahrscheinlichkeit, dass  $\mu_A$  grösser als  $\mu_B$  ist. Benutzen Sie dazu das Resultat von (c) und eine Monte-Carlo-Simulation.
- (e) Ihrem Auftraggeber müssen Sie Ihre Untersuchungen und Beobachtungen aus der EDA vorstellen. Sie dürfen dabei eine grafische Darstellung zeigen. Welche Art wählen Sie?

**Tab. 11.11** Dreissig aufeinanderfolgende Werte von Regenmengen (in Inches) im Monat März für Minneapolis/St. Paul (Die Tabelle wird entlang der Spalten gelesen.)

0,77	4,75	0,32	3,00	1,20	1,89	2,81	2,10	0,47	1,43
3,37	1,74	2,48	0,59	3,09	1,95	0,90	1,87	0,57	1,62
1,31	2,20	0,81	0,96	0,81	1,51	1,20	2,05	1,18	1,35

**Tab. 11.12** Tagfalterartenzahlen in Flächen der Schweiz (Biodiversitätsmonitoring Schweiz BDM, Hintermann & Weber AG, Reinach (Basel), Herbst 2007)

Ackerland

44	51	21	34	60	60	35	26	55	43	34	59	26	57	49	57	36	72	51	57	54	51	53
----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----

Bergregion

47	25	24	22	20	57	32	37	20	64	24	25	23	23	23	24	58	57	42	47	38	40	42
----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----

**11.7** Teure mechanische Uhren werden nach ihrer Fertigstellung auf ihre Genauigkeit kontrolliert. Bei der Uhrenfabrik Zenith International S. A. wird der tägliche Gang der Uhren während 15 Tagen beobachtet. Gemessen wird die Abweichung in Sekunden von der exakten Zeit. Der Gangschein Nummer 4924839 vom 16.7.1995 der Uhr mit Werknummer 6157 zeigt ihren täglichen Gang:<sup>6</sup>

$$\begin{array}{cccccccccc} -1,0 & -1,9 & +3,0 & +0,1 & +1,0 & +1,0 & +4,1 & +3,8 \\ +4,1 & +1,0 & +9,0 & +5,0 & -2,0 & +1,1 & +3,0 \end{array}$$

- (a) Zeichnen Sie ein Stamm-Blatt-Diagramm der Datenwerte. Gibts es Werte, die auffallen?
- (b) Ist es sinnvoll, für die Daten einen Box & Whisker Plot zu zeichnen? Ist der Zahlenwert +9,0 ein Ausreisser oder Extremwert?
- (c) Sind die Messwerte unter statistischer Kontrolle? Ist es sinnvoll anzunehmen, dass sie unabhängig modelliert werden können?
- (d) Berechnen Sie mit dem Normalverteilungsmodell die durchschnittliche Gangabweichung  $\mu$  pro Tag der Uhr. Gesucht ist der plausibelste Wert von  $\mu$ , sowie Wahrscheinlichkeitsintervalle zum Niveau 0,5 und 0,95.
- (e) Was passiert mit dem Resultat von (d), wenn Sie den Wert +9,0 weglassen?
- (f) Führen Sie die Aufgabe von (d) mit einem „robusten“ Modell durch. Was erhalten Sie nun? Was, wenn Sie mit diesem Modell den Wert +9,0 weglassen?

## Literatur

1. A. M. Catilla, Does passage time through the lizard *Podarcis liofordis* gets affect germination performance in the plant *withanic fructescens*? *Acta Oecologica* **21**(2), 119–124 (2000)
2. F. Hampel, Robust inference. Research Report **93**, ETH Zürich (2000)
3. C. H. Hennekens, Aspirin in Chronic Cardiovascular Disease and Acute Myocardial Infarction. *Clin. Cardiol.* **13**, V-62–66 (1990)
4. D. Hinkley, On quick choice of power transformation. *Applied Statistics* **26**, 67–69 (1977)
5. J. A. Rice, *Mathematical Statistics and Data Analysis* (Duxbury Press, 1995)
6. E. Rutherford, M. Geiger, The Probability Variations in the Distribution of Alpha Particles. *Philosophical Magazine, Series 6*, **20**, 698–704 (1910)

<sup>6</sup> Angaben aus einem Kontrollblatt aus dem Uhrenmuseum in La Chaux-de-Fonds.

7. D. S. Sivia, J. Skilling, *Data Analysis, a Bayesian Tutorial* (Oxford University Press, 2006)
8. R. G. Staudte, S. J. Sheather, *Robust Estimation and Testing* (John Wiley & Sons, Inc., 1990)
9. E. R. Tufte, *The Visual Display of Quantitative Information* (Graphics Press, Cheshire, Connecticut, 1983)
10. J. W. Tukey, *Exploratory Data Analysis* (Addison-Wesley Publishing Company, 1977)

„Wie viele Stunden Unterricht hattet ihr denn am Tag?“ fragte Alice, um schnell das Thema zu wechseln.  
„Zehn Stunden am ersten Tag“, sagte die Falsche Suppenschildkröte; „neun am nächsten und so fort.“  
„Einen schönen Stundenplan müsst ihr da gehabt haben!“ rief Alice;  
„der wurde ja von Tag zu Tag leerer!“  
„Es waren ja auch lauter Lehrer im Haus“, bemerkte der Greif, „da war das ganz unvermeidlich.“  
Dieser Gedanke war Alice neu, und sie dachte eine Weile darüber nach, bis sie schliesslich sagte: „Der elfte Tag war dann also schulfrei?“  
„Das ist doch klar“, sagte die Falsche Suppenschildkröte.  
„Und was passierte dann am zwölften?“ fragte Alice eifrig weiter.  
*Lewis Carroll, Alice im Wunderland (Insel Taschenbuch, 1973, S. 100)*

## Zusammenfassung

Der Wert einer Grösse kann von vielen verschiedenen Faktoren und Kovariablen abhängen. Oft geht es darum, den Wert der Grösse aus wenigen dieser Faktoren zu prognostizieren. So möchte jemand den Preis eines Gebrauchtwagens aus dem Kilometerstand des Wagens berechnen. Oder ein Arzt will das Lungenvolumen aus dem Alter und der Körpergrösse eines Patienten bestimmen. Man spricht in solchen Fällen auch von *statistischem Lernen* (engl. *statistical learning*). Häufig will man nur den durchschnittlich erwartbaren Wert der Zielgrösse in Funktion der wenigen Faktoren, den sogenannten abhängigen Grössen, berechnen. Dies tun Regressionsmodelle. Solche Modelle werden in diesem Kapitel vorgestellt.

## 12.1 Streudiagramme

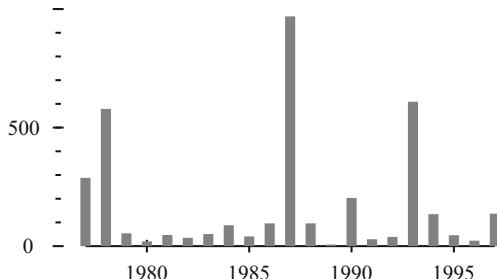
Trend- oder Streudiagramme sind in den vorangehenden Kapiteln benutzt worden, um zu visualisieren, ob ein Experiment unter statistischer Kontrolle liegt. Ein Spezialfall eines Streudiagramms ist eine *Kontrollkarte* (engl. *control chart*). Ein solches findet sich in Abb. 1.5. Streudiagramme werden auch verwendet, um zu analysieren, wie Größen von Faktoren und Kovariablen abhängen. In vielen Untersuchungen hängen Beobachtungsmerkmale von der Kovariablen ‚Zeit‘ ab. Unter der Zeit versteht man eine messbare Zeitperiode, wie Jahr, Monat, Stunden, Minuten, .... Man spricht von *Zeitreihen*. Viele Zeitreihen finden sich in Zeitschriften, in Zeitungen und in statistischen Jahrbüchern. Üblicherweise stellt man sie grafisch dar, indem man die beobachteten Werte gegen die Zeit aufträgt, sei es mit einem Streu- oder mit einem Balkendiagramm. Man erhält ein Merkmal-Zeit-Diagramm.<sup>1</sup> Hier folgen Beispiele dazu:

**Beispiel 12.1 (Unwetterschäden)** Das Beispiel 1.6 zu den Schadensummen von Unwettern ist eine Zeitreihe über eine Periode von mehr als 20 Jahren. Um das Merkmal-Zeit-Diagramm zu zeichnen, sollten die Schadensummen zuerst inflationsbereinigt werden. Dies wurde in Kap. 1 durchgeführt. Das Balkendiagramm in Abb. 12.1 verdeutlicht die grossen Schadensummen in den Jahren 1978, 1987 und 1993. □

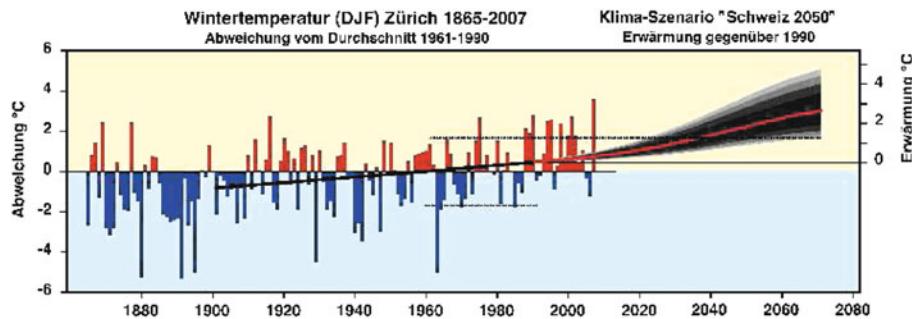
Eine weitere Zeitreihe ist in Abb. 1.4 dargestellt. Die Abbildung zeigt die Preisentwicklung von 1860 bis 2013 des Preises für ein Barrel Rohöl.

**Beispiel 12.2 (Temperaturprognose)** Meteorologen, Biologen und andere Wissenschafter versuchen Lufttemperaturen in der Schweiz zu prognostizieren. Dazu sind hochwertige Messdaten nötig. Das Messnetz der MeteoSchweiz liefert solche seit 1864 für alle grossen Klimaregionen der Schweiz. Das Merkmal-Zeit-Diagramm in Abb. 12.2 zeigt, dass der Temperaturverlauf während der Wintermonate für Zürich während der beobachteten 150 Jahre starken jährlichen Schwankungen unterworfen war. Die blauen und

**Abb. 12.1** Schadensummen von Unwettern in der Schweiz (in realen Mio. CHF des Jahres 1980)



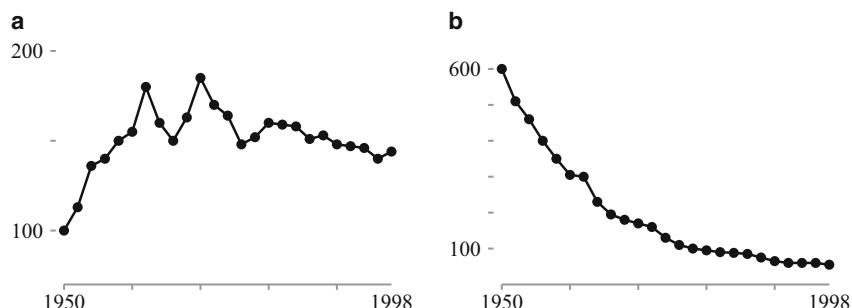
<sup>1</sup> Merkmal-Zeit-Diagramme wurden erst, dies mag erstaunen, gegen Ende des 18. Jahrhunderts von William Playfair (1759–1823) erfunden. William Playfair gilt als Vater von Datengrafiken. Siehe dazu [11].



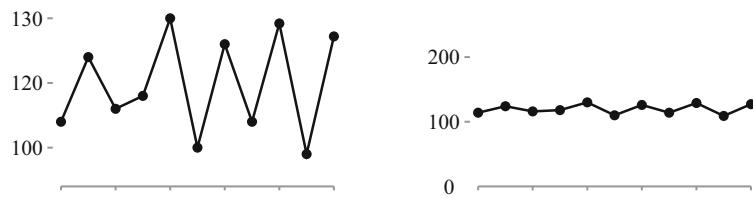
**Abb. 12.2** Temperaturverlauf während der Wintermonate für Zürich von 1860 bis 2010 und eine gerechnete Prognose (aus [7], MeteoSchweiz)

roten Stäbe zeigen gemessene Temperaturabweichungen von der Durchschnittstemperatur aus den Jahren 1961–1990. Die Temperaturen sind ab 1980 tendenziell gestiegen. Es wurden fast ausschliesslich positive Abweichungen von der Durchschnittstemperatur in der Zeitspanne 1961–1990 registriert. Die grauen Keile stellen dar, wie die durchschnittliche Temperatur während der Wintermonate bis zum Jahr 2070 prognostiziert wird, gegeben ein Wahrscheinlichkeitsmodell und die Daten. □

**Beispiel 12.3 (Strassenverkehrsunfälle)** Abb. 12.3 zeigt zweimal die gleiche Zeitreihe. Dargestellt wird, wie sich die Anzahl Strassenverkehrsunfälle mit Personenunfällen in der Schweiz von 1950 bis 1998 entwickelt hat. Abb. 12.3a illustriert, dass die Anzahl Verkehrsunfälle zuerst stark zugenommen, und sich dann, ab 1980 stabilisiert hat. Im Jahr 1980 ist die Anzahl Strassenverkehrsunfälle höher als im Jahr 1950. Abb. 12.3b visualisiert die Anzahl Verkehrsunfälle pro 10 000 Fahrzeuge. Dies zeigt, dass der Anteil jener Fahrzeuge, die in Verkehrsunfällen beteiligt sind, seit 1950 stark abgenommen hat. □



**Abb. 12.3** a) indexierte (1950 = 100) und b) pro 10 000 Motorfahrzeuge, beobachtete Strassenverkehrsunfälle in der Schweiz (Daten aus Statistisches Jahrbuch Schweiz, 2000)



**Abb. 12.4** Zweimal die gleiche Zeitreihe, dargestellt mit verschiedenen skalierten vertikalen Achse

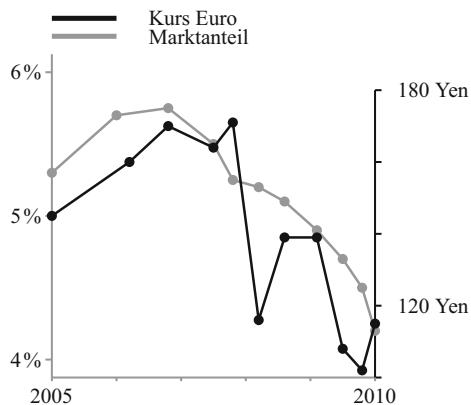
In vielen Zeitreihen werden Geldwerte geplottet. Die Darstellung der *inflationskorrigierten* und *standardisierten* Geldwerte ist dabei der nominalen Darstellung der Geldwerte vorzuziehen. So ist es zum Beispiel wenig sinnvoll, die Staatsausgaben der Schweiz in den Jahren 1950–2000 in einer Zeitreihe mit Einheit Franken darzustellen. Standardisierte Einheiten für diese Zeitreihe könnten reale Franken des Jahrs 2002, reale Franken des Jahrs 2002 pro Einwohner oder prozentuale Anteile am Bruttoinlandsprodukt sein.

Stellt man Zeitreihen grafisch dar, stellt sich die Frage, wie die Streuung der Datenwerte dargestellt werden soll. Dies illustriert Abb. 12.4. Wie die vertikale Achse skaliert und wie ihr Nullpunkt gesetzt werden soll, ist nicht einfach festzulegen. Die Skalierung sollte sorgfältig gewählt werden. Eine gute Regel des Statistikers B. Cleveland in [5] besagt, dass die Datenreihe die Grafik möglichst ausfüllen sollte. Dies bedeutet, dass der Nullpunkt der vertikalen Achse nicht auf der Höhe der horizontalen Achse sein muss. In Abb. 12.4 wird dies an der linken Grafik ersichtlich. Sie erfüllt die oben genannte Regel.

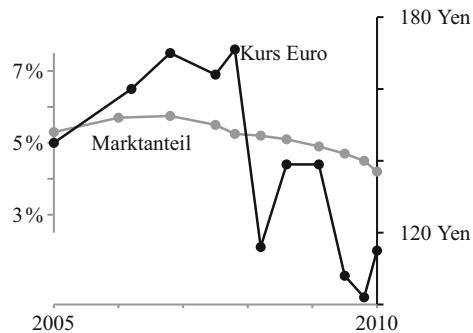
Besonders schwierig kann es sein, die vertikalen Achsen zu skalieren, wenn Zeitreihen zu unterschiedlichen Größen in einem Bild kombiniert werden:

**Beispiel 12.4 (Kursentwicklung und Marktanteil)** In Abb. 12.5 sind zwei Zeitreihen dargestellt, einmal der Kurs des Euro in Yen und einmal der Marktanteil des japanischen Autoproduzenten Toyota in Europa. Sie suggerieren, dass die beiden Größen stark zu-

**Abb. 12.5** Marktanteil von Toyota-Autos und Euro-Kurs in Yen (Nach J. Bertin gelingt die Rezeption von Grafiken einfacher, wenn die Kurven einzeln beschriftet sind (vgl. Abb. 12.6) als wenn Legenden erstellt werden.)



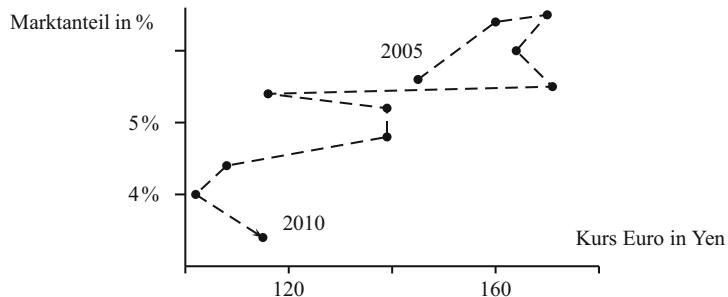
**Abb. 12.6** Marktanteil von Toyota-Autos und Euro-Kurs in Yen



sammenhängen. Skaliert man die vertikale Achse des Marktanteils anders, ergibt sich ein andersartiger Eindruck, wie dies Abb. 12.6 zeigt. Die wohl objektivste Art zu visualisieren, wie zwei Größen zusammenhängen, ist ein Streudiagramm (vgl. Abb. 12.7). Mit dem empirischen Korrelationskoeffizienten  $\rho_{\text{emp}}$  nach Pearson – siehe Abschn. 6.3 – lässt sich messen, wie stark die beiden Variablen verbunden sind. Es ist  $\rho_{\text{emp}} = 0,82$ . Der Wert ist grösser als  $2/\sqrt{11} = 0,60$ . Der Marktanteil dürfte damit vom Kurs des Euro in Yen abhängen. □

Größen können nicht nur von der Zeit, sondern auch von anderen Variablen abhängen. So dehnt sich die Länge eines Kabels in Funktion der Temperatur, oder der Preis eines Gebrauchtwagens hängt vom Kilometerstand des Fahrzeugs ab. Hier Beispiele dazu:

**Beispiel 12.5 (Akkommodationsbreite)** In medizinischen Schriften findet man Untersuchungen zur Akkommodation des Sehens. Die Akkommodation ist die Anpassung der Augenlinsen an die Entfernung der zu betrachtenden Gegenstände. Wie maximal sich die Augenlinse anpassen kann, wird als Akkommodationsbreite bezeichnet. Die Akkommodationsbreite nimmt mit zunehmendem Alter ab. Bei Kleinkindern beträgt sie durchschnittlich etwa 14 Dioptrien. Personen im hohen Alter haben eine Akkommodationsbrei-

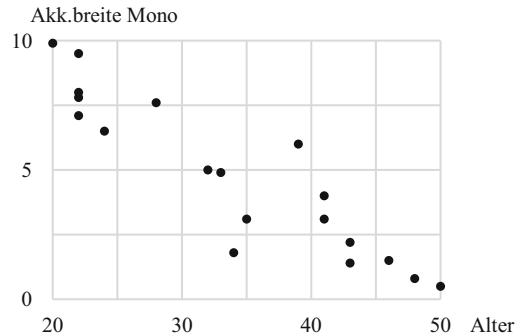


**Abb. 12.7** Marktanteil von Toyota-Autos und Euro-Kurs in Yen, dargestellt mit einem Streudiagramm

**Tab. 12.1** Akkommodationsbreiten (in Dioptrien) von 19 Personen beim Mono- und Stereosehen (aus [9])

Alter	20	22	22	22	22	24	28	32	33	34	35	39	41	41	43	43	46	48	50
Mono	9,9	9,5	7,1	7,8	8,0	6,5	7,6	5,0	4,9	1,8	3,1	6,0	4,0	3,1	1,4	2,2	1,5	0,8	0,5
Stereo	10,1	9,9	7,3	8,2	8,7	7,8	7,7	5,1	5,3	4,7	4,0	6,0	4,3	3,8	3,0	3,5	2,4	1,0	0,9

**Abb. 12.8** Akkommodationsbreite des Mono-Sehens bei 19 Personen zwischen 20 und 50 Jahren (aus [9])



te von weniger als 0,5 Dioptrien. Die Akkommodationsbreite hängt aber auch von anderen Kovariablen ab, wie der Gesundheit, der genetischen Veranlagung, den Umwelteinflüssen und den Sehgewohnheiten – um nur einige zu nennen. Eine Fachperson könnte versuchen, die Akkommodationsbreite in Funktion aller dieser Kovariablen zu rechnen. Dies ist jedoch kaum umsetzbar. Einerseits dürfte das Modell ausserordentlich komplex ausfallen. Andererseits fehlen Daten, um ein solches Modell präzis zu entwerfen. Man begnügt sich daher, die Auswirkung der wichtigsten Faktoren auf die Akkommodationsbreite zu berechnen. Man kann versuchen, die folgenden Fragen zu beantworten:

- (a) Wie lauten die wichtigsten Faktoren, die auf die Akkommodationsbreite wirken?
- (b) Lässt sich, ausgehend von den wichtigsten Faktoren, zumindest eine *nicht direkt messbare Grösse*, wie die durchschnittliche Akkommodationsbreite, in Funktion dieser Faktoren *berechnen*? Wie genau und wie sicher ist eine solche Angabe?
- (c) Lässt sich die Akkommodationsbreite bei einer nichtuntersuchten Person in Funktion der wichtigen Kovariablen *prognostizieren*?

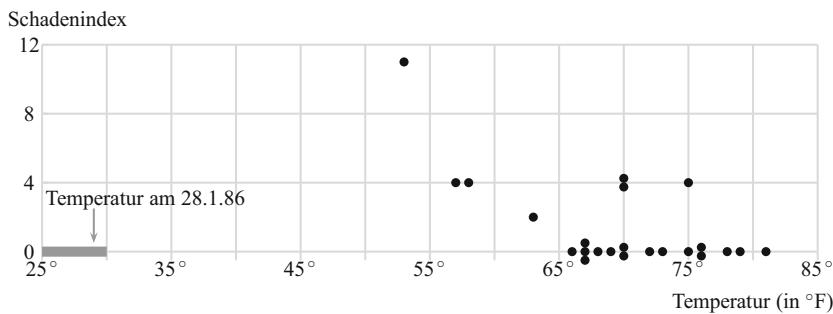
Die Antwort auf die Frage (a) ist schwierig. In Produktionsprozessen, wo mit kontrollierten Faktoren gearbeitet werden kann, helfen faktorielle Pläne und Pareto-Diagramme – siehe Kap. 2 – eine Antwort zu finden. Sind unkontrollierte Faktoren vorhanden, wie hier bei der Akkommodationsbreite, kann die Frage (a) im Rahmen eines einführenden Statistikkurses nicht beantwortet werden. Statistische Rechnungen haben aber gezeigt, dass das Alter die grösste Wirkung auf die Akkommodationsbreite hat. Man kann mit diesem Wissen versuchen, die Fragen (b) und (c) zu beantworten. Dazu braucht man Daten, wie sie Tab. 12.1 zeigt. Gegeben sind Akkommodationsbreiten von 19 Personen im Alter von 20 bis 50 Jahren. Das Streudiagramm in Abb. 12.8 visualisiert die Daten beim Monose-

hen. Die Akkommodationsbreite nimmt bei den Probanden mit zunehmenden Alter ab. Ist dies auch so für die *Grundgesamtheit* aller Menschen, die zwischen 20 und 50 Jahre alt sind? Man kann dazu versuchen, die *durchschnittliche* Akkommodationsbreite  $\mu_{\text{Akk}}(A)$  aller Personen in Funktion ihres Alters  $A$  zu beschreiben. Dies nennt man ein *Regressionsmodell* (engl. *regression model*). Die Arbeit besteht darin  $\mu_{\text{Akk}}(\text{Alter } A)$  zu rechnen: Was ist sein plausibelster Wert? Wie lauten Wahrscheinlichkeitsintervalle zum Niveau 0,5 und 0,95?  $\square$

**Beispiel 12.6 (Space Shuttle Challenger)** Am 28. Januar 1986 explodierte nach einer Minute Flugdauer der Space Shuttle Challenger. Sieben Astronauten starben. Die Explosion erfolgte wegen zweier undichter Gummiringdichtungen an den am Shuttle angebrachten Festkörperraketen. Die Ringe verloren ihre Dichtigkeit, weil der Space Shuttle Challenger an einem sehr kalten Tag gestartet war und die Ringe Temperaturen von weniger als  $30^{\circ}\text{F}$  ausgesetzt waren.

Die Ingenieure der NASA waren sich vor dem Start des Space Shuttle Challenger bewusst, dass die Dichtungseigenschaften der Gummiringe von der Temperatur abhängen. Um zu analysieren, ob die Temperatur einen Einfluss auf die Häufigkeit der Defekte der Gummiringe hat, wurden die Temperaturen und die Schäden der Gummiringe aller Starts des Space Shuttles vor dem 28.1.1986 analysiert. Gemäss der Kommission, die den Unfall untersuchte<sup>2</sup>, wurde der mögliche Schadenindex jedoch nie in Funktion der Temperatur mit einem Streudiagramm visualisiert. Daher liess sich kaum überblicken, wie sich die Ringe bei einer Starttemperatur unter  $30^{\circ}\text{F}$  verhalten.

Tab. 12.2 zeigt die Temperaturen und den Schadenindex von allen Starts der Space Shuttles vor dem 28.1.1986. Abb. 12.9 ermöglicht die Daten mit einem Streudiagramm zu



**Abb. 12.9** Lufttemperatur und Schadenindex bei 24 Starts von Space Shuttles (Es fallen verschiedene Punkte aufeinander, wie die Datenpaare (76/0) und (76/0). Um solche Daten ersichtlich zu machen, wurden die Punkte leicht versetzt (engl. gejittert).)

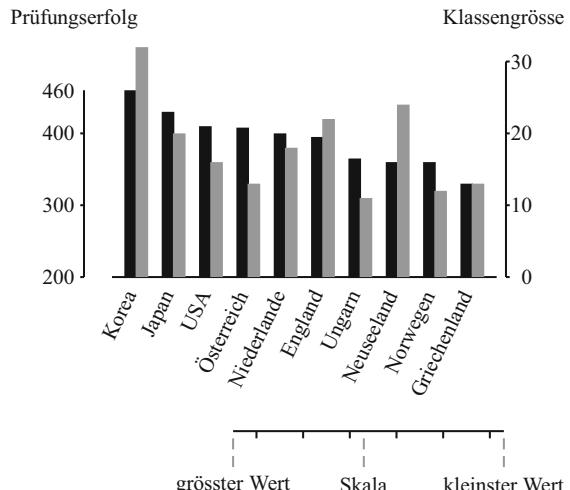
<sup>2</sup> Report of the Presidential Commission on the Space Shuttle Challenger Accident. Washington D.C. (1986)

**Tab. 12.2** Lufttemperatur (in °F) und Schadenindex bei 24 Starts von Space Shuttles (aus [10])

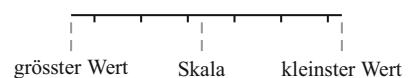
Flug	Datum	Temperatur	Schadenindex
51-C	24.01.85	53°	11
41-B	02.03.84	57°	4
61-C	12.01.86	58°	4
41-C	06.04.84	63°	2
1	12.04.86	66°	0
6	04.04.83	67°	0
51-A	08.11.84	67°	0
51-D	12.04.85	67°	0
5	11.11.82	68°	0
3	22.03.82	69°	0
2	12.11.81	70°	4
9	28.11.83	70°	0
41-D	30.08.84	70°	4
51-G	17.06.85	70°	0
7	18.06.83	72°	0
8	30.08.83	73°	0
51-B	29.04.85	75°	0
61-A	30.10.85	75°	4
51-I	27.08.85	76°	0
61-B	26.11.85	76°	0
41-G	05.10.84	78°	0
51-J	03.10.85	79°	0
4	27.06.82	80°	?
51-F	29.07.85	81°	0

überblicken. Das Streudiagramm deckt die Gefahr auf, den Space Shuttle bei tiefer Temperatur zu starten, da schon bei 52° ein Schadenindex von 12 beobachtet wurde. Man könnte versuchen, den durchschnittlich erwartbaren Schadenindex  $\mu_{\text{Schaden}}$  mit einem Modell in Funktion der Temperatur zu beschreiben. Ein solches Regressionsmodell kann aber wegen der vorhandenen Daten nur für Temperaturen zwischen 50°F und 85°F sinnvoll eingesetzt werden. Eine *Extrapolation* auf eine Temperatur von 30°F ist äußerst fragwürdig, wie dies E. R. Tufte in [10] sagt, da die Starttemperatur von 29°F fast sechs empirische Standardabweichungen vom Mittelwert der früheren Starttemperaturen entfernt ist. Der Wirkungsraum der 24 Starts war also zu klein, um den durchschnittlichen Schadenindex bei einer Starttemperatur von 29°F zu bestimmen. Das obige einfache Streudiagramm hätte die Entscheidungsträger überzeugt, den Shuttle nicht zu starten. Der Entscheid, den Shuttle zu starten, basierte auf unzweckmäßigen Grafiken. Eine ausführliche und spannende Diskussion von E. R. Tufte zum Space Shuttle Start vom 28. Januar 1986 findet man in [10].

**Abb. 12.10** Klassengrösse und Lernerfolg bei zehn Ländern, OECD Untersuchung *Education at Glance* aus dem Jahr 2005



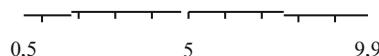
**Abb. 12.11** Skala nach Tufte



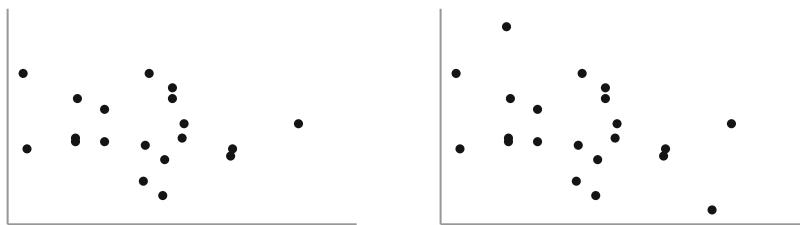
**Beispiel 12.7 (Prüfungserfolg)** In der in Beispiel 1.5 vorgestellten Studie, wird untersucht, ob der Lernerfolg bei Schülerinnen und Schülern der Mittelstufe von der Klassengrösse abhängt. In verschiedenen Zeitungen wurden die Daten zu Klassengrösse und Prüfungserfolg wie in Abb. 12.10 mit verschachtelten Balkendiagrammen dargestellt. Die Balken zeigen in schwarzer Farbe den Prüfungserfolg und in grauer Farbe die Klassengrösse. Das in Abb. 1.3 in Kap. 1 dargestellte Streudiagramm visualisiert die Resultate der Untersuchung prägnanter. Aussergewöhnlich ist Korea. Da die Punkte im Streudiagramm mit den Ländern gekennzeichnet sind, spricht von einem *kodierten* Streudiagramm. □

Streudiagramme können ästhetisch optimiert werden. Traditionelle Koordinatenachsen mit Pfeilen und Querstrichen lenken meist von den Daten ab und beinhalten keine Informationen. Daher empfiehlt E. R. Tufte, Koordinatenachsen wie in Abb. 12.11 zu realisieren. Auch Kennzahlen der einzelnen Merkmale, wie etwa der empirische Median oder die empirischen Quartile, können in den Koordinatenachsen eingetragen werden. Beim Beispiel 12.5 beträgt beim Mono-Sehen der empirische Median 4,9, das obere empirische Quartil ist 7,6 und das untere empirische Quartil ist 1,8. Eine Skala mit dieser Information könnte wie in Abb. 12.12 aussehen.

Einzelne univariante Datenwerte verändern die grafischen Darstellungen, wie Box & Whisker Plots oder Histogramme, kaum. Streudiagramme sind aber empfindlicher auf die Auswirkung einzelner Beobachtungen. Abb. 12.13 illustriert eine solche Situation.



**Abb. 12.12** Mögliche Skala beim Beispiel 12.5 zur Akkommodationsbreite



**Abb. 12.13** Zwei Streudiagramme: rechtes Diagramm mit zwei zusätzlichen Messpunkten

Im rechten Streudiagramm wurden zwei Messwerte addiert. Im rechten Streudiagramm, aber nicht im linken, scheint die Zielgrösse (vertikale Achse) mit zunehmendem Wert der erklärenden Variable tendenziell abzunehmen.

## 12.2 Beispiele von Regressionsmodellen

Ein Regressionsmodell bestimmt den *durchschnittlichen* Wert einer unsicheren Zielgrösse  $Z$  in Funktion einer oder mehrerer erklärender Variablen  $X, Y, \dots$ . Dies bedeutet insbesondere, dass die Zielgrösse  $Z$  bei gegebenen Werten der Kovariablen  $X = x, Y = y, \dots$  einen endlichen Erwartungswert  $\mu_Z(x, y, \dots)$  haben soll. Um diesen durchschnittlichen Wert ausrechnen zu können, braucht man ein Datenmodell. Es besagt, wie Werte der Zielgrösse  $Z$  um  $\mu_Z(x, y, \dots)$  streuen. Dazu wollen wir annehmen, dass die Streuung  $\sigma(x, y, \dots)$  des Datenmodells endlich ist. Weiter sei die Zielgrösse  $Z$  kontinuierlich. Mit dem Theorem 10.1 der maximalen relativen Entropie heisst dies: Diese Information wird mit der grössten Unordnung mit der Normalverteilung charakterisiert:

Datenmodell:  $i$ -ter Messwert  $Z \sim \text{Normal}(\mu_i, \sigma)$  mit  $\mu_i = \mu_Z(x_i, y_i, \dots)$

Besteht minimale Vorinformation zu den beiden Parametern  $\mu_Z$  und  $\sigma$ , so ist nach der Diskussion zum Modell der Normalverteilung von Kap. 10: Das arithmetische Mittel der  $Z$ -Messwerte zu gegebenen Werten der erklärenden Variablen ist der plausibelste Wert von  $\mu_Z(x, y, \dots)$ :

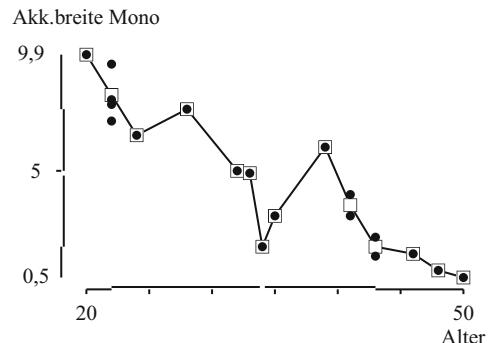
$$\mu_Z(x, y, \dots) \approx \bar{Z}_{\text{wenn } X=x, Y=y, \dots}$$

Das Verfahren wird am Beispiel der Akkommodationsbreite illustriert:

**Beispiel 12.8 (Akkommodationsbreite)** Beim Beispiel 12.5 will man den plausibelsten Wert für die durchschnittliche Akkommodationsbreite  $\mu_{\text{Akk}}$  beim Mono-Sehen in Funktion des Alters berechnen. Im Alter von 22 Jahren hat man vier Messwerte für Akkommodationsbreiten:

$$9,5 \quad 7,1 \quad 7,8 \quad 8,0$$

**Abb. 12.14** Schätzung der durchschnittlichen Akkommodationsbreite  $\mu_{\text{Akk}}$  des Monosehens mit arithmetischen Mitteln der Messwerte



Der plausibelste Wert für die durchschnittliche Akkommodationsbreite im Alter von 22 Jahren kann daher mit

$$\mu_{\text{Akk}}(22 \text{ Jahre}) \approx \frac{9,5 + 7,1 + 7,8 + 8,0}{4} = 8,1$$

geschätzt werden. Im Alter von 20 Jahren hat man nur eine Beobachtung, die Zahl 9,9. Daher ist

$$\mu_{\text{Akk}}(20 \text{ Jahre}) \approx \frac{9,9}{1} = 9,9$$

Analog kann für andere Altersangaben gerechnet werden. Die so bestimmten Werte für das Regressionsmodell können ins Streudiagramm der Daten eingezeichnet und durch Geradenstücke verbunden werden. Dies zeigt Abb. 12.14. Das Regressionsmodell hängt sehr stark von den einzelnen Messwerten ab. Insbesondere ist dies dort ausgeprägt, wo bei gegebenem Alter nur ein Messwert vorhanden ist. So beispielsweise im Alter von 38 Jahren, wo der Messwert 6,0 Dioptrien beträgt. Es ist voraussehbar, dass mit diesem Regressionsmodell weitere gemessene Akkommodationsbreiten von Personen, die 38 Jahre alt sind, nicht gut prognostiziert werden können. □

Das obige Regressionsmodell besteht aus 13 Geradenstücken. Jedes Geradenstück ist charakterisiert durch zwei Parameter: Steigung und Achsenabschnitt. Das Modell besitzt also 26 Parameter. Dies führt dazu, dass es stark auf einzelne Messwerte reagiert. Man sagt, das man eine *Überanpassung* (engl. *overfitting*) hat. Überanpassung findet statt, wenn ein Regressionsmodell zu viele Parameter hat. Überanpassung entsteht auch, wenn das Regressionsmodell zu viele erklärende Variablen enthält. Beim Beispiel der Akkommodationsbreite könnte man ein Regressionsmodell entwickeln, dass viele Kovariablen enthält: Alter, Gesundheit der Person, Sehgewohnheiten wie Fernsehen oder Lesen, oder Essgewohnheiten. Diese Kovariablen erlauben es, ein Regressionsmodell anzufertigen, dass alle Messpunkte gut trifft. Die Gefahr ist dann aber gross, dass das Modell zu spezifisch auf den Daten aufbaut. Mit einem solchen Modell prognostizierte Akkommodationsbreiten können leider schlecht ausfallen.<sup>3</sup>

<sup>3</sup> Siehe dazu auch das Beispiel 15.2, das in Kap. 15 diskutiert wird.

Um ein Regressionsmodell zu erhalten, das weniger auf einzelne Messwerte reagiert, kann man mit dem arithmetischen Mittel über Altersbereiche von  $\pm h$  Jahren arbeiten:

$$\mu_{\text{Akk}}(\text{Alter} = x) \approx \text{Mittelwert der Messungen im Fenster } x - h \text{ und } x + h$$

Sind im Fenster zwischen  $x - h$  und  $x + h$  keine Messwerte, so wird dort  $\mu_{\text{Akk}}$  nicht geschätzt. Man nennt das Verfahren auch *Kernel-Regression* mit  $h$  als *Bandbreite*.

**Beispiel 12.9 (Akkommodationsbreite)** Wendet man hier die Kernel-Regression mit einer Bandbreite von  $h = 5$  Jahren an, erhält man die folgenden plausibelsten Werte für die durchschnittliche Akkommodationsbreite:

$$\begin{aligned}\mu_{\text{Akk}}(20 \text{ Jahre}) &\approx \text{Mittelwert Akk.breite zwischen 15 und 25 Jahren} \\ &= \frac{9,9 + 9,5 + 7,1 + 7,8 + 8,0 + 6,5}{6} = 8,13\end{aligned}$$

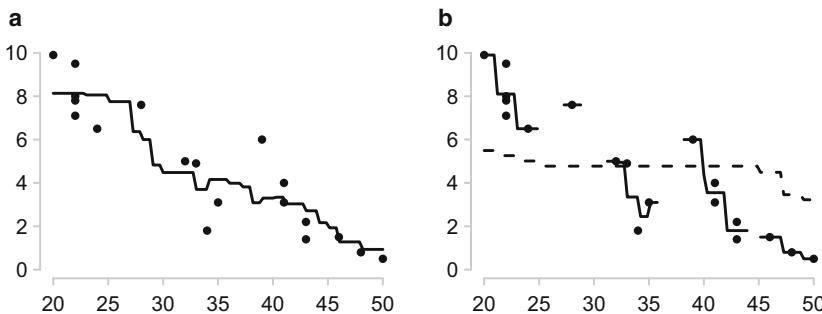
Analog ist

$$\begin{aligned}\mu_{\text{Akk}}(28 \text{ Jahre}) &\approx \text{Mittelwert Akk.breite zwischen 23 und 32 Jahren} \\ &= \frac{6,5 + 7,6 + 5,0}{3} = 6,37\end{aligned}$$

Analog kann man für weitere Altersangaben rechnen. Die berechneten Schätzungen können ins Streudiagramm der Datenwerte eingezeichnet werden. Dies zeigt das linke Streudiagramm in Abb. 12.15. Das rechte Streudiagramm in dieser Abbildung zeigt zwei Kernel-Regressionsmodelle mit verschiedenen Bandbreiten: mit zwei Jahren, ausgezogen dargestellt, und mit 25 Jahren gestrichelt. Bei der Bandbreite von zwei Jahren erfolgt in Bereichen, in denen keine Messwerte vorhanden sind, keine Schätzung für das Regressionsmodell. Das Streudiagramm zeigt, wie stark das Modell von der gewählten Bandbreite abhängt. Bei zu kleiner Bandbreite fliessen die einzelnen Messwerte stark ins Modell ein. Das Modell ist überangepasst. Bei zu grosser Bandbreite fällt das Regressionsmodell unterangepasst aus: das Alter hat kaum Einfluss auf die Akkommodationsbreite. □

In der Praxis sind andere Kernel-Regressionsmodelle verbreitet, die nicht mit einfachen arithmetischen Mitteln arbeiten. Vielmehr werden gewichtete Mittelwerte über einem Fenster, zentriert in  $x$ , und mit Bandweite  $h$ , gerechnet. Je weiter ein Messpunkt im Fenster vom Zentrum  $x$  entfernt ist, umso weniger fliest er in den Mittelwert ein. Man spricht auch von der *Nadaraya-Watson-Kernel-Regression*.

Gerade bei Grössen, bei denen keine physikalischen Gesetze vorhanden sind, um analytische Regressionsmodelle zu konstruieren, werden solche Kernel-Modelle angewendet. Dies geschieht meist erfolgreich, wenn sehr viele Messwerte oder Beobachtungen vorliegen. In vielen Bereichen der Technik und der Wirtschaftswissenschaften ist es aber üblich, den Zusammenhang zwischen Grössen mit einfachen, analytischen Regressionsmodellen



**Abb. 12.15** Regressionsmodelle für die durchschnittliche Akkommodationsbreite des Mono-Sehens: **a** Kernelregression mit Bandbreite 5 Jahren, **b** mit Bandbreite 2 Jahren (*ausgezogen*) und 25 Jahren (*gestrichelt*)

zu beschreiben. So kann jemand versuchen, den durchschnittlichen Wert  $\mu_Z(x, y)$  einer Zielgröße  $Z$  aus erklärenden Variablen  $X$  und  $Y$  mit Modellen wie

$$\mu_Z(X = x, Y = y) = a + b \cdot x + c \cdot x \cdot y + d \cdot y^2$$

oder

$$\mu_Z(X = x, Y = y) = a \cdot x + \sin(b \cdot x \cdot y + c)$$

zu beschreiben. In diesen Modellen sind  $a$ ,  $b$ ,  $c$  und  $d$  unbekannte, aus den Daten zu berechnende Parameter. In der statistischen Literatur unterscheidet man zwischen *linearen* und *nicht-linearen Regressionsmodellen*. Für eine univariante Zielgröße  $Z$  heisst das Regressionsmodell mit Parametern  $a, b, c, \dots$  linear, wenn es in der Form

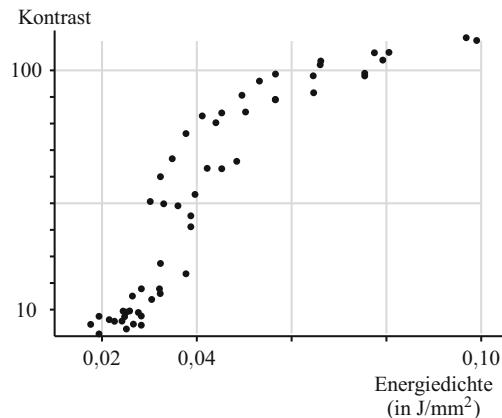
$$\mu_Z(X = x, Y = y) = a \cdot \{\dots\} + b \cdot \{\dots\} + c \cdot \{\dots\} + \dots$$

geschrieben werden kann. Dabei dürfen sich in den geschwungenen Klammern keine Parameter befinden. Der Begriff linear bezieht sich also auf die Parameter. Das obige erste Modell ist linear, das zweite nicht.

**Beispiel 12.10 (Lasermarkierung)** Konsumgüterpackungen werden mit Strichcodes markiert. Diese können produziert werden, indem mit Lasern aufgetragene Farbschichten abgetragen werden. Dieses Verfahren ist in Beispiel 1.4 vorgestellt. Eine hohe Laserstärke (oder Energiedichte) trägt eine Farbschicht mehr ab und erhöht den Kontrast der Markierung. Das Streudiagramm in Abb. 12.16 zeigt, wie dieser Effekt bei 54 Messungen aussieht. Wegen Kovariablen, wie Unregelmässigkeiten in der Farbschicht, im Papier und im Brennvorgang, wie sie auch im Ursache-Wirkungsdiagramm in Abb. 1.2 aufgeführt sind, misst man bei gegebener Energiedichte nicht immer den gleichen Kontrast.

Das Ziel der Untersuchung war es, ein Regressionsmodell zu konstruieren, dass von der Energiedichte  $E$  des Lasers auf den durchschnittlichen Farbkontrast  $\mu_{\text{Kontrast}}(E)$  rechnet.

**Abb. 12.16** Kontrast der Markierung in Funktion der Energiedichte des Lasers bei 54 Messungen, aus [1]



Physikalische Überlegungen führen zur Fermi- oder Sigmoidfunktion<sup>4</sup>

$$\mu_{\text{Kontrast}}(E) = \frac{a}{1 + \exp\{-(E - \rho)/\kappa\}}$$

Das Regressionsmodell besitzt die drei Parameter  $a$ ,  $\rho$  und  $\kappa$ . Es ist nicht linear. Aus den Daten möchte man die Parameter berechnen und sagen, wie genau und plausibel die Resultate sind.  $\square$

**Beispiel 12.11 (Ausspülen von Milchrohren)** In der Schweiz befinden sich im Jahr 2008 ungefähr 260 Milchsammelwagen bei verschiedenen Firmen im Einsatz. Mit den Sammelwagen wird die Milch bei den Milchsammelstellen oder direkt beim Landwirt abgeholt. Um die Qualität der Milch zu überprüfen, wird mit einem automatischen Probeentnahmesystem eine Probe pro Sammelstelle oder Landwirt entnommen. Im Labor kann die Milch dann analysiert werden. Untersucht wird die Probe auf den Fettgehalt und auf Verschmutzungen wie Antibiotika. Bei den Proben ist wichtig, dass keine Durchmischung mit vorheriger Milch stattfindet. Das Probeentnahmesystem wird ein- bis zweimal jährlich auf seine Funktionsfähigkeit und Genauigkeit geprüft. Dieser Prozess heisst Homologierung. Durch das Bundesamt für Veterinärwesen ist der Ablauf der Homologierung in der „Technische[n] Weisung für die Durchführung der Qualitätskontrolle der Verkehrsmilch“ operationell definiert.

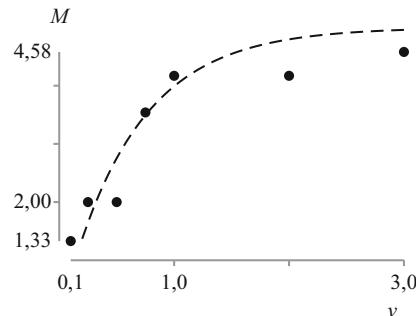
Rohre im Milchwagen müssen nach jeder Leerung mit Wasser ausgespült werden. Wenn die Strömungsgeschwindigkeit des Spülwassers tief ist, wird nicht das ganze Rohr

<sup>4</sup> Die Funktionsgleichung der Sigmoidfunktion  $\text{sigm}(x)$  lautet

$$\text{sigm}(x) = \frac{a}{1 + \exp\{b \cdot (x - c)\}}$$

Dabei sind  $a$ ,  $b$  und  $c$  feste Zahlen.

**Abb. 12.17** Milchkonzentration  $M$  (in  $10^{-4}$  ml/ml) in Funktion der Spülgeschwindigkeit  $v$  (in m/s) bei sieben Messungen (aus [2])



mit Wasser gefüllt: Ein gewisser Teil der Milch bleibt im System hängen und durchmischt sich mit der nachfolgend eingepumpten Milch. Mit sieben Messungen wurde in [2] versucht, die ausgespülte Milchkonzentration  $M$  in Funktion der Spülgeschwindigkeit  $v$  zu modellieren. Die Messwerte sind:

$v$	0,10	0,25	0,50	0,75	1,00	2,00	3,00
$M$	1,33	2,00	2,00	3,54	4,17	4,17	4,58

Eine hohe Milchkonzentration  $M$  bedeutet, dass besser ausgespült wird. Physikalische Gesetze weisen darauf hin, die durchschnittliche Milchkonzentration  $\mu_M(v)$  mit einer Sättigungsfunktion, die von der Spülgeschwindigkeit  $v$  abhängt, zu modellieren:

$$\mu_M(v) = A \cdot (1 - \exp\{-c \cdot v\})$$

Dabei sind  $A$  und  $c$  unbekannte Parameter. Das Streudiagramm in Abb. 12.17 zeigt den Graphen des nicht linearen Regressionsmodells mit  $A = 5,0$  und  $c = 1,6$  s/m. Was sind die plausibelsten Werte für die Parameter  $A$  und  $c$ ? Wie lauten Wahrscheinlichkeitsintervalle für diese Parameter? □

**Beispiel 12.12 (Akkommodationsbreite)** Um eine Überanpassung wie beim Beispiel 12.8 zu vermeiden, kann man versuchen, ein Regressionsmodell zu konstruieren, das wenige Parameter hat. Im Gegensatz zu den obigen zwei Beispielen ist das Modell nicht aus physikalischen Überlegungen ableitbar. Ein Blick auf das Streudiagramm in Abb. 12.8 kann helfen, ein einfaches Modell zu gestalten. Die Abbildung zeigt, dass die durchschnittliche Akkommodationsbreite  $\mu_{\text{Akk}}(\text{Alter})$  beim Mono-Sehen linear mit dem Alter abnehmen könnte:

$$\mu_{\text{Akk}}(\text{Alter}) = a + b \cdot \text{Alter}$$

Dies ist ein Regressionsmodell mit zwei Parametern. Die Parameter  $a$  und  $b$  sind unbekannt. Sie müssen aus den Daten berechnet werden. Das Regressionsmodell hängt weiter

linear von den beiden Parametern ab. Es ist also ein lineares Regressionsmodell. Der Parameter  $b$  ist besonders interessant. Er codiert die Steigung der Geraden. Er sagt damit, um wie viele Dioptrien pro Jahr die mittlere Akkommodationsbreite abnimmt.  $\square$

**Beispiel 12.13 (Kanalwärmetauscher)** In Beispiel 2.13 wird untersucht, wie Wärmetauschelemente in Abwasserkanälen verformt werden. Die Verformung hängt bei den Wärmetauschern hauptsächlich von drei Faktoren ab: Der Dicke  $S$  des Siggenblechs, der Dicke  $D$  des Deckblechs und der Breite  $R$  des Rinnendeckblechs. Die Faktoren  $S$ ,  $D$  und  $R$  nehmen die Werte  $-1$  und  $+1$  an. Das Paretodiagramm in Abb. 12.18 visualisiert die gemessenen durchschnittlichen Auswirkungen der drei Faktoren sowie ihrer Interaktionen. Man möchte berechnen, wie gross die durchschnittliche Verformung  $\mu_{\text{Verformung}}$  des Wärmetauschelements in Funktion der Faktoren und ihrer Interaktionen ist. In der Produktionstechnik sind dazu einfache, lineare Regressionsmodelle beliebt. In derartige Modelle fliessen die aus dem Paretodiagramm herausgelesenen wichtigsten Faktoren und Interaktionen ein.<sup>5</sup> Hier eine Auswahl:

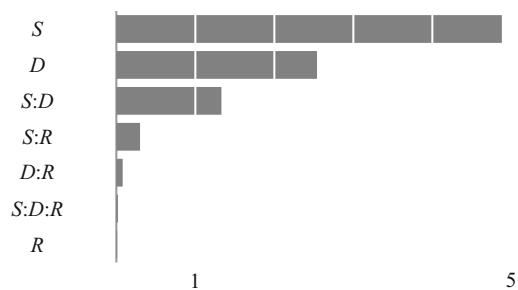
- (1) Ein äusserst einfaches Modell ist ein Modell, das die durchschnittliche Verformung linear mit den Hauptfaktoren verbindet:

$$\mu_{\text{Verformung}}(S, D) = a_0 + a_1 \cdot S + a_2 \cdot D$$

Dabei sind  $a_0$ ,  $a_1$  und  $a_2$  unbekannte, aus den Messungen zu bestimmende Parameter. Ist  $S = 1$  und  $D = -1$ , so erhält man eine durchschnittliche Verformung von  $a_0 + a_1 - a_2$ .

- (2) Ein schon komplexeres Modell ist ein Modell 2. Ordnung: Dieses Modell berücksichtigt, neben linearen Auswirkungen der einzelnen Faktoren, auch die aus dem

**Abb. 12.18** Paretodiagramm:  
die Verformung wird vor allem  
von den Faktoren  $S$  und  $D$   
und ihrer Interaktion  $S : D$   
beeinflusst



<sup>5</sup> In vielen Arbeiten werden Modelle nicht in Funktion der physikalischen Relevanz der Faktoren, sondern in Funktion der „statistischen Signifikanz“ ausgewählt. Statistische Signifikanz bedeutet, dass ein Faktor wesentlich ist, wenn er mit hoher Plausibilität eine Wirkung auf die Zielgröße hat. Dabei ist es unbedeutend, wie stark die Wirkung ist. Es gibt Statistiker, die mit diesem Verfahren nicht einverstanden sind. Siehe dazu auch [8] auf Seite 22.

Paretodiagramm herausgelesenen, wesentlichen Interaktionen zwischen zwei Faktoren

$$\mu_{\text{Verformung}}(S, D, R) = a_0 + a_1 \cdot S + a_2 \cdot D + a_3 \cdot S \cdot D + a_4 \cdot S \cdot R$$

Auch hier sind  $a_0, a_1, \dots, a_4$  zu bestimmende Parameter. Das Regressionsmodell ist linear in den Parametern.  $\square$

Obwohl die vorgestellten Modelle kaum Zielgrößenwerte in präziser Weise in Funktion der Kovariablen berechnen, werden sie in der Praxis oft benutzt. Sie sind nämlich sehr nützlich, um Zielgrößen in Funktion ihrer Faktoren approximativ zu beschreiben. Auch komplexere Modelle, die vielleicht schwierig zu begründen sind, werden kaum mehr Einsicht in die Beziehung zwischen der Zielgröße und den Faktoren bringen. Zudem besteht die Gefahr, dass sie überangepasst sind. Dazu meint der Statistiker G. Box in [3]:

All models are wrong but some are useful.

Es kann sehr vorteilhaft sein, mit analytischen Regressionsmodellen zu arbeiten. Der im Schnitt erwartbare Wert der Zielgröße ist einfach zu berechnen. Nichtsdestoweniger bestehen auch Risiken. Das Modell kann schlecht, unterangepasst oder überangepasst sein. Dies führt dazu, dass berechnete Prognosewerte systematisch falsch sind. Regressionsmodelle sollten daher kritisch beurteilt werden: Gibt es physikalische oder technische Argumente, die gegen die getroffene Wahl des Modells sprechen?

Zudem ist bei einfachen Regressionsmodellen die Versuchung gross, Werte ausserhalb der Messreihen zu prognostizieren. Bei der Akkommodationsbreite könnte man Alter von 60 oder gar 80 Jahren in die lineare Gleichung einsetzen. Vor solchen Extrapolationen wird gewarnt. Siehe dazu auch das nächste Kapitel.<sup>6</sup>

---

## Reflexion

**12.1** Wie lautet der plausibelste Wert für die durchschnittliche Akkommodationsbreite  $\mu_{\text{Akk}}$  des Mono-Sehens beim vorgestellten (und überangepassten) Modell im Alter von 35, 37 und 38 Jahren?

---

<sup>6</sup> H. Wainer diskutiert vertieft ein Regressionsmodell, das die Entwicklung der Siegerlaufzeit von Marathonläufern modelliert. Dabei wird ein Kernel-Modell gegen ein einfaches Modell mit zwei Parametern gestellt. Siehe dazu [11].

**12.2** Bestimmen Sie ein Regressionsmodell für die durchschnittliche Akkommodationsbreite des Stereosehens mit dem in diesem Kapitel vorgestellten Verfahren benutzt. Die Daten finden sich in Tab. 12.1. Benutzen Sie dabei (a) Mittelwerte und (b) Mittelwerte über ein Fenster mit verschiedenen Bandbreiten. Wann ist das Modell überangepasst, wann unterangepasst?

**12.3** Bei einem Terrassenhaus in Morges hängt der Gasverbrauch für die Heizung hauptsächlich von der durchschnittlichen Monatsaussentemperatur  $T$  der Wintermonate ab. Die Daten in Tab. 12.3 während dreier Winter illustrieren dies. Man möchte daher von  $T$  auf den durchschnittlichen Verbrauch  $\mu_{\text{Verbrauch}}(T)$  rechnen.

- (a) Stellen Sie die Daten mit einem Streudiagramm dar.
- (b) Bestimmen Sie verschiedene Regressionsmodelle, indem Sie über verschiedene Fenster mitteln. Arbeiten Sie dazu mit einem Statistikprogramm und auch mit gewichteten Mittelwerten und verschiedenen Bandbreiten, sodass eine „glattes“ Regressionsmodell für  $\mu_{\text{Verbrauch}}(T)$  entsteht.

**Tab. 12.3** Gasverbrauch bei einem Terrassenhaus und durchschnittlichen Monatsaussentemperatur der Wintermonate (H. Hausmann, Berner Fachhochschule, Burgdorf)

Monat	Temperatur $T$ (in °C)	Verbrauch (in m <sup>3</sup> /Tag)
September 2006	18,0	6,7
Oktober 2006	13,6	11,9
November 2006	7,7	19,1
Dezember 2006	3,4	24,8
Januar 2007	4,3	23,5
Februar 2007	5,6	22,8
März 2007	6,7	20,4
April 2007	14,3	11,6
September 2007	14,7	11,9
Oktober 2007	10,8	15,5
November 2007	4,2	22,7
Dezember 2007	2,3	25,8
Januar 2008	3,8	24,5
Februar 2008	3,9	23,7
März 2008	6,0	21,1
April 2008	9,2	18,3
September 2008	14,1	13,7
Oktober 2008	11,2	14,3
November 2008	6,3	20,8
Dezember 2008	1,6	27,2
Januar 2009	-0,2	29,8
Februar 2009	1,8	26,4
März 2009	6,3	20,3
April 2009	12,4	13,0

**Tab. 12.4** Übertragungsverhalten  $U$  einer pneumatischen Druckmessleitung beim Startvorgang einer Maschine, der 20 Sekunden dauert (aus [4])

$t$	0,0	0,5	1,0	1,5	2,0	2,5	3,0	3,5	4,0	4,5
$U$	0,8776	0,9315	0,9255	0,9527	0,9654	0,9830	1,0343	1,0112	1,0141	1,0514
$t$	5,0	5,5	6,0	6,5	7,0	7,5	8,0	8,5	9,0	9,5
$U$	1,0018	1,0254	0,9863	0,9853	0,9697	1,0131	0,9982	0,9614	0,9859	0,9615
$t$	10,0	10,5	11,0	11,5	12,0	12,5	13,0	13,5	14,0	14,5
$U$	0,9490	0,989	0,9677	0,9793	1,0160	1,0070	1,0273	1,0447	1,0078	1,0610
$t$	15,0	15,5	16,0	16,5	17,0	17,5	18,0	18,5	19,0	19,5
$U$	1,0562	1,0555	1,0323	1,0568	1,0560	1,0564	1,0393	1,0431	1,0481	1,0467

- (c) Sie wollen ein einfaches Regressionsmodell für  $\mu_{\text{Verbrauch}}(T)$  mit wenigen Parametern aufstellen. Was für ein Modell schlagen Sie vor?

**12.4** Um das Übertragungsverhalten  $U$  von pneumatischen Druckmessleitungen beim Startvorgang einer Maschine, der 20 Sekunden dauert, zu modellieren, wurde die Daten in Tab. 12.4 gemessen. Das Ziel ist es, ein Regressionsmodell zu bestimmen, das das durchschnittliche Übertragungsverhalten  $\mu_U(t)$  in Funktion der Zeit  $t$  des Startvorgangs angibt.

- (a) Zeichnen Sie die Daten in einem Streudiagramm.  
 (b) Bestimmen Sie verschiedene Regressionsmodelle, indem Sie über verschiedene Fenster mitteln. Arbeiten Sie dazu mit einem Statistikprogramm und auch mit gewichteten Mittelwerten und verschiedenen Bandbreiten, sodass weder ein über- noch ein unterangepasstes Regressionsmodell für  $\mu_U(t)$  entsteht.

**12.5** Eine Biologin möchte den Herzschlag eines Vogels aus dem Körpergewicht prognostizieren. Um dies zu tun, benutzt sie die Messungen in Tab. 12.5. Visualisieren Sie die Daten mit einem kodierten Streudiagramm.

**Tab. 12.5** Herzschlag (in Schlägen/min) und Körpergewicht (in g) von verschiedenen Vogelarten (aus [6])

	Körpergewicht	Herzschlag
Blaumeise	9,9	963
Distelfink	12,6	754
Fink	12,9	831
Gimpel	23,1	729
Grünling	22,0	697
Kohlmeise	7,7	1037
Lasurmeise	15,8	798

## Literatur

1. D. Bättig, M. Buri, P. A. C. Gane, B. Neuenschwander, H. Scheidiger, D. C. Spielmann, Mechanism of Post-Print Laser Marking on Coated Substrates: factors controlling ink ablation in the application of calcium carbonate, *Journal of Graphic Technology* **1**, 37–48 (2002)
2. P. Bernhard, Prüfprozedur der Probenahme auf Milchsammelwagen, Bachelorarbeit Maschinenbau, Berner Fachhochschule, Burgdorf (2008)
3. G. P. E. Box, Robustness in the Strategy of Scientific Model Building, In: R. L. Launer, G. N. Wilkinson (eds.), *Robustness in Statistics* (Academic Press, New York, 1979) 201–236
4. St. Brechbühl, Modellierung des Ansprechverhaltens pneumatischer Messungen, Bachelorarbeit, Berner Fachhochschule Burgdorf (2008)
5. W. S. Cleveland, *The elements of graphing data* (Hobart Press, Summit, NJ, 1994)
6. G. Fels, M. und H. Grah, F. Liesenfeld, *Der Organismus* (Ernst Klett Verlag, 1974)
7. C. Frei, M. Croci-Maspoli, C. Appenzeller, *Die Klimaentwicklung der Schweiz* (OcCC, 2008)
8. A. Gelman, J. Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models* (Cambridge University Press, 2007)
9. Y. Otake, M. Miyao, S. Ishihara, M. Kashiwamata, T. Kondo, H. Sakakibara, S. Yamada, An experimental study on the objective measurement of accommodative amplitude under binocular and natural viewing conditions, *Tohoku J. Exp. Med.* **170**, 93–102 (1993)
10. E. Tufte, *Visual and Statistical Thinking: Displays of Evidence for Making Decisions* (Graphics Press, P. O. Box 430, Cheshire, CT 06410, 1997)
11. H. Wainer, *Graphic Discovery, A Trout in the Milk and Other Visual Adventures* (Princeton University Press, 2008)

„Wie lautet euer Urteil?“ fragte der König der Schöffen.  
„Halt, noch nicht!“ rief das Weisse Kaninchen dazwischen, „vor dem Urteil kommt noch allerlei anderes!“  
Lewis Carroll, Alice im Wunderland (Insel Taschenbuch, 1973, S. 112)

## Zusammenfassung

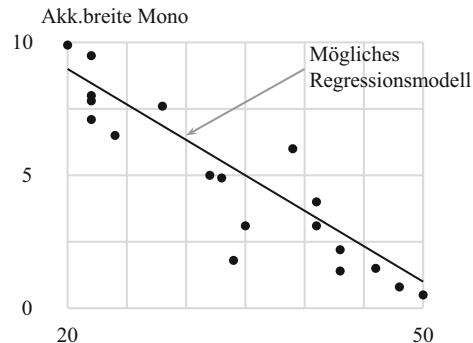
Im vorigen Kapitel ist erklärt, was ein Regressionsmodell ist. In diesem Kapitel wird gezeigt, wie die Parameter eines Regressionsmodells berechnet werden können. Um anzugeben, wie präzis und wie plausibel Resultate sind, braucht man ein Datenmodell für die Messwerte oder Beobachtungen. Auch zukünftige Messwerte oder Beobachtungen lassen sich prognostizieren. Dies geschieht mit dem Gesetz der Marginalisierung.

## 13.1 Beispiele mit der Normalverteilung

Ein Regressionsmodell, um von Faktoren oder Kovariablen auf eine Zielgröße zu rechnen, gibt den im Schnitt erwartbaren Wert der Zielgröße bei gegebenen Werten der erklärenden Variablen an. Fließen zu viele Faktoren, Kovariablen oder Parameter in ein Regressionsmodell zu, besteht die Gefahr der Überanpassung. Fehlen Faktoren oder Kovariablen, die einen wesentlichen Einfluss auf die Zielgröße haben, hat man eine Unteranpassung. In beiden Fällen ist das Regressionsmodell nicht brauchbar. Beim Beispiel 12.5 der Akkommotionsbreite des vorigen Kapitels ist  $\mu_{\text{Akk}}(\text{Alter}) = a + b \cdot \text{Alter}$  ein sinnvolles Modell für die durchschnittliche Akkommotionsbreite. Das Modell enthält die zwei Parameter  $a$  und  $b$ , sowie die wichtigste Kvariable, das Alter. Beim Beispiel 12.11 der Ausspülung von Milchrohren wird die durchschnittlich ausgespülte Milchkonzentration mit dem Regressionsmodell  $\mu_M(v) = A \cdot (1 - \exp\{-c \cdot v\})$  beschrieben. Das Modell hat die Parameter  $A$  und  $c$ , sowie die Spülgeschwindigkeit  $v$  als Faktor. Wie man die Parameter eines Regressionsmodells berechnen kann, wird im Folgenden illustriert:

**Tab. 13.1** Akkommodationsbreite des Mono-Sehens bei 19 Personen zwischen 20 und 50 Jahren

Alter	20	22	22	22	22	24	28	32	33	34	35	39	41	41	43	43	46	48	50
Akk.breite	9,9	9,5	7,1	7,8	8,0	6,5	7,6	5,0	4,9	1,8	3,1	6,0	4,0	3,1	1,4	2,2	1,5	0,8	0,5

**Abb. 13.1** Mögliche Regressionsmodell für die durchschnittliche Akkommodationsbreite beim Mono-Sehen für Personen zwischen 20 und 50 Jahren

**Beispiel 13.1 (Akkommodationsbreite)** Tab. 13.1 zeigt die 19 Messwerte mit denen man die Akkommodationsbreite in Funktion des Alters modellieren will. Dabei interessieren zwei Fragen:

- (a) Wie lautet die durchschnittliche Akkommodationsbreite  $\mu_{\text{Akko}}$  in Funktion des Alters? Wie genau und wie sicher ist eine solche Angabe?
- (b) Lässt sich die Akkommodationsbreite einer Person bei gegebenem Alter prognostizieren?

Um die Frage (a) zu beantworten, braucht man ein Regressionsmodell. Gegeben die Information aus dem Streudiagramm in Abb. 13.1 scheint eine lineare Funktion sinnvoll:

$$\mu_{\text{Akko}}(\text{Alter}) = a + b \cdot \text{Alter}$$

Interessant ist der Parameter  $b$ . Er gibt an, um wie viel sich die durchschnittliche Akkommodationsbreite pro Jahr ändert.

Mit einem Datenmodell werden die Parameter  $a$  und  $b$  berechnet. Das Datenmodell besagt, wie Akkommodationsbreiten streuen. Die untersuchende Ärztegruppe nimmt an, dass Akkommodationsbreiten einen endlichen Wert  $\mu_{\text{Akko}}(\text{Alter})$  haben. Weiter sei die Streuung der Akkommodationsbreiten endlich. Da Akkommodationsbreiten kontinuierliche Werte annehmen können, bedeutet diese Information nach dem Prinzip der maximalen Entropie und Theorem 10.1: Plausiblerweise streuen Akkommodationsbreiten normalverteilt um den durchschnittlichen Wert. Abb. 13.1 zeigt, dass die Streuung  $\sigma$  der gemessenen Akkommodationsbreiten kaum von der Kovariablen – dem Alter – abhängt. Damit ist  $\sigma$  eine feste, unbekannte Zahl. Ein sinnvolles Datenmodell für die  $i$ -te gemessene Akkommodationsbreite in Funktion des gegebenen  $i$ -ten Alters ist damit:

$$i\text{-te Akkommodationsbreite} \sim \text{Normal}(\mu_i, \sigma) \quad \text{mit } \mu_i = a + b \cdot (i\text{-tes Alter})$$

Die Dichtefunktion des Datenmodells lautet:

$$\text{pdf}(\text{Akko} = z \mid a, b, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp \left\{ -0,5 \cdot \left( \frac{z - \mu_{\text{Akko}}(\text{Alter})}{\sigma} \right)^2 \right\} \quad (13.1)$$

Das Datenmodell besitzt die Parameter  $a, b$  und die Streuung  $\sigma$ . Der Parameter  $\sigma$  ist wenig interessant. Er ist ein *Störparameter*. Mit der Regel von Bayes lassen sich die Plausibilitäten zu den drei Parametern mit den Daten aus der Vorinformation aktualisieren:

$$\underbrace{\text{pdf}(a, b, \sigma \mid \text{Daten, Vorinfo.})}_{\text{Posterior}} \propto \underbrace{\mathbb{P}(\text{Daten} \mid a, b, \sigma)}_{\text{Likelihood}} \cdot \underbrace{\text{pdf}(a, b, \sigma \mid \text{Vorinfo.})}_{\text{Prior}}$$

Vor dem Experiment hatten die Mediziner folgende Informationen: die durchschnittliche Akkommodationsbreite kann nicht grösser als 20 Dioptrien sein, sie kann mit dem Alter nicht zunehmen und wird nicht mit mehr als 10 Dioptrien pro Jahr abnehmen. Also müssen  $-10 \leq b \leq 0$  und  $0 \leq a \leq 20$  sein. Weiter sind die Parameter  $a$  und  $b$  Lageparameter. Die Streuung  $\sigma$  ist ein Skalierungsparameter. Deshalb setzt man den Faktor ganz rechts

$$\text{pdf}(a, b, \sigma \mid \text{Vorinformation}) \propto \begin{cases} 1 \cdot 1 \cdot 1/\sigma = 1/\sigma & \text{für } 0 \leq a \leq 20, -10 \leq b < 0 \\ 1 \cdot 0 \cdot 1/\sigma = 0 & \text{sonst} \end{cases}$$

Den andere Faktor, die Likelihood, berechnet man aus dem Datenmodell. Bei unabhängigen Messwerten folgt aus dem dem Multiplikationsgesetz

$$\mathbb{P}(\text{Daten} \mid a, b, \sigma) = \mathbb{P}(\text{Akko}_1 \mid a, b, \sigma) \cdot \dots \cdot \mathbb{P}(\text{Akko}_{19} \mid a, b, \sigma)$$

Die einzelnen Faktoren erhält man aus Gleichung (13.1). Für den ersten Messwert (Akkommodationsbreite von 9,9 bei Alter 20) setzt man  $z = 9,9$  und das Alter 20 ein. Man erhält für die Likelihood:

$$\begin{aligned} \mathbb{P}(\text{Daten} \mid a, b, \sigma) &\propto \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp \left\{ -0,5 \cdot \left( \frac{9,9 - a - b \cdot 20}{\sigma} \right)^2 \right\} \cdot \dots \\ &\quad \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp \left\{ -0,5 \cdot \left( \frac{0,5 - a - b \cdot 50}{\sigma} \right)^2 \right\} \end{aligned}$$

Die 19 Faktoren lassen sich zusammenfassen:

$$\mathbb{P}(\text{Daten} \mid a, b, \sigma) \propto \frac{1}{\sigma^{19}} \cdot \exp \left\{ -0,5 \cdot \chi^2 \right\}$$

Dabei ist, wie in Kap. 10 gesagt,  $\chi^2$  die Summe der Quadratabweichungen von den Messwerten zur gesuchten durchschnittlichen Akkommodationsbreite  $\mu_{\text{Akko}}$ , dividiert durch

die Streuung des Datenmodells:

$$\chi^2 = \left( \frac{9,9 - a - b \cdot 20}{\sigma} \right)^2 + \dots + \left( \frac{0,5 - a - b \cdot 50}{\sigma} \right)^2$$

Damit sind alle Bestandteile der Regel von Bayes vorhanden. Die Plausibilität zu den Parametern  $a$ ,  $b$  und  $\sigma$  gibt die gemeinsame A posteriori-Verteilung. Ihre Dichtefunktion lautet:

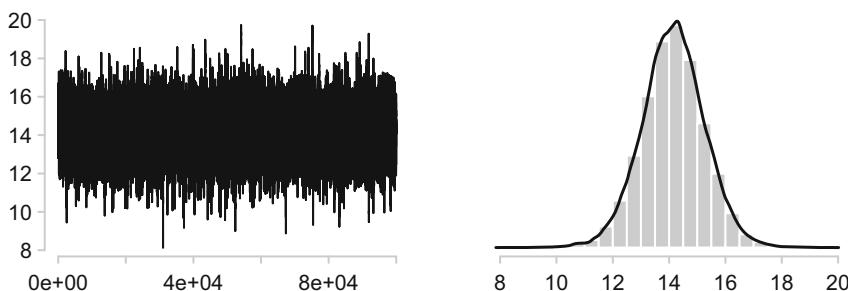
$$\text{pdf}(a, b, \sigma \mid \text{Daten, min. Vorinformation}) \propto \frac{1}{\sigma^{19}} \cdot \exp\{-0,5 \cdot \chi^2\} \cdot \frac{1}{\sigma} \quad \text{für } b \leq 0$$

Da diese Verteilung von drei Parametern abhängt, lässt sie sich kaum grafisch darstellen. Bestimmen und visuell darstellen lassen sich ihre Randverteilungen. Sie geben an, in welchem Bereich die Parameter  $a$ ,  $b$  und (vielleicht weniger wichtig)  $\sigma$  plausibel sind. Um sie zu berechnen, konstruiert man mit einer MCMC-Simulation Punkte der gemeinsamen A posteriori-Verteilung. Bei einem Statistikprogramm werden dazu die Messwerte eingegeben und das Daten- und das Regressionsmodell sowie die A priori-Verteilungen der Parameter genannt:

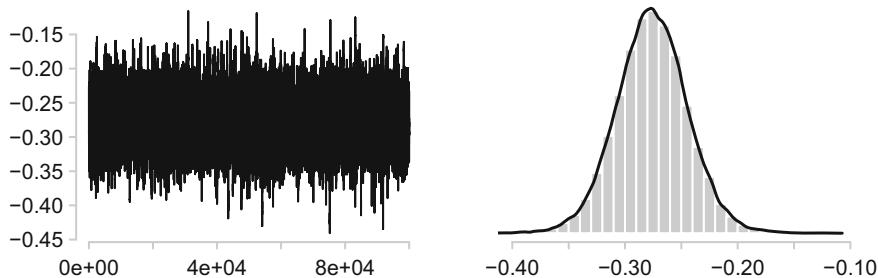
- |                    |  |
|--------------------|--|
| Datenmodell:       | $i$ -te Akkommodationsbreite $\sim \text{Normal}(\mu_i, \sigma)$ |
| Regressionsmodell: | $\mu_i = a + b \cdot (i\text{-tes Alter})$                       |
| Prior:             | $a \sim \text{Uniform}(0; 20)$                                   |
| Prior:             | $b \sim \text{Uniform}(-10; 0)$                                  |
| Prior:             | $\ln \sigma \sim \text{Uniform}(\ln(10^{-3}); \ln(10^3))$        |

Daraus bildet ein Statistikprogramm mit dem Produkt Likelihood  $\times$  Prior die gemeinsame A posteriori-Verteilung der Parameter  $a$ ,  $b$  und  $\sigma$ . Abb. 13.2 zeigt die  $a$ -Koordinaten einer MCMC-Kette mit 100 000 Punkten. Rechts in der Abbildung ist die mit der Kette berechnete Randverteilung von  $a$  visualisiert. Sie ist unimodal symmetrisch. Der Modus  $a_0$  liefert den plausibelsten Wert für  $a$ :

$$a \approx a_0 = 14,15 \text{ Dioptrien}$$



**Abb. 13.2** MCMC-Kette mit Akzeptanzrate 0,49 für den Posterior des Parameters  $a$  des Regressionsmodells



**Abb. 13.3** MCMC-Kette für den Posterior des Parameters  $b$  des Regressionsmodells

Es besteht eine Wahrscheinlichkeit von 0,5, dass  $a$  zwischen 13,43 Dioptrien und 14,87 Dioptrien liegt. Mit einer Wahrscheinlichkeit von 0,95 befindet sich  $a$  zwischen 11,92 Dioptrien und 16,34 Dioptrien.

Abb. 13.3 zeigt die  $b$ -Koordinaten der MCMC-Kette und die A posteriori-Verteilung des Parameters  $b$ . Der plausibelste Wert für  $b$  ist der Modus  $b_0$  der Verteilung:

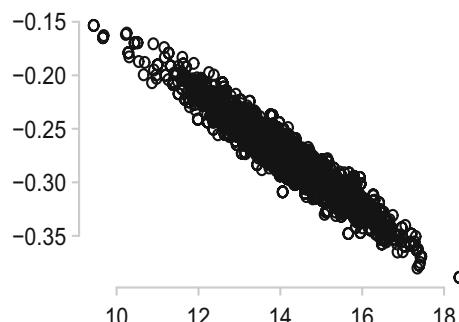
$$b \approx b_0 = -0,28 \text{ Dioptrien/Jahr}$$

Es besteht eine Wahrscheinlichkeit von 0,5, dass  $b$  zwischen  $-0,30$  Dioptrien/Jahr und  $-0,26$  Dioptrien/Jahr liegt. Mit einer Wahrscheinlichkeit von 0,95 befindet er sich zwischen  $-0,34$  Dioptrien/Jahr und  $-0,21$  Dioptrien/Jahr. Man zeichnet meistens das Regressionsmodell mit den plausibelsten Werten von  $a$  und  $b$  im Streudiagramm ein (siehe Abb. 13.7).

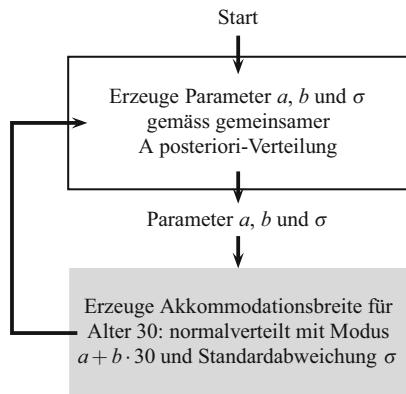
Das Plausibilität zu den beiden Parametern  $a$  und  $b$  ist nicht unabhängig. Dies zeigt Abb. 13.4. Dargestellt sind 5000 ( $a/b$ )-Koordinaten der MCMC-Kette. Die beiden Parameter sind negativ korreliert.

Die Frage (b) kann man nun beantworten: Lässt sich die Akkommodationsbreite einer Person bei gegebenem Alter prognostizieren? Beispielsweise: Wo liegt mit hoher Wahrscheinlichkeit die Akkommodationsbreite einer dreissigjährigen Person? Dazu braucht

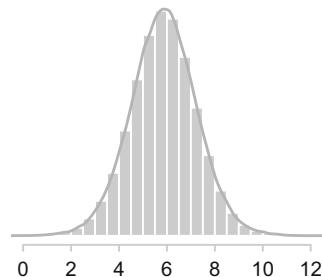
**Abb. 13.4** Gemeinsamer Posterior der Parameter  $a$  und  $b$ , dargestellt durch 5000 Punkte der MCMC-Kette



**Abb. 13.5** Monte-Carlo-Simulation, um das Prognosemodell für Akkommodationsbreiten im Alter von 30 zu bestimmen



**Abb. 13.6** Prognosemodell für die Akkommodationsbreite im Alter von 30 Jahren



man das Gesetz der Marginalisierung. Es besagt, dass man das Datenmodell für Akkommodationsbreiten über den gemeinsamen Posterior der Parameter  $a$ ,  $b$  und  $\sigma$  mittelt. Am einfachsten macht man dies mit dem Monte-Carlo-Verfahren von Theorem 7.3. Die dazugehörige Simulation ist in Abb. 13.5 für ein Alter von 30 Jahren dargestellt. 100 000 Werte der Parameter  $a$ ,  $b$  und  $\sigma$  lassen sich aus der MCMC-Simulation für die gemeinsame A posteriori-Verteilung der drei Parameter ablesen. Hier ein Pseudo-Code, der die Simulation für Alter 30 Jahre illustriert:

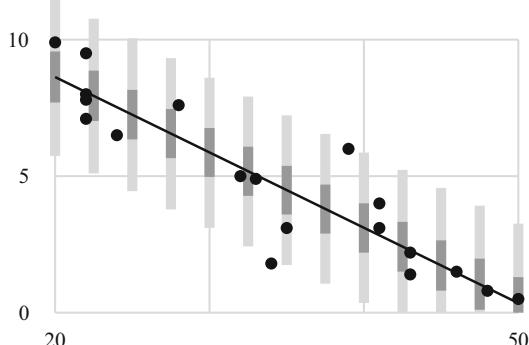
```

for i = 1 to 100000 do
  (a[i], b[i], Sigma[i]) = i-ter Punkt aus
    MCMC-Simulation für a, b und sigma
  Akkommodation30[i] = Wert normalverteilt, mit
    Modus a[i]+b[i]*30 und Streuung Sigma[i]
end
  
```

In Abb. 13.6 sind 100 000 simulierte Akkommodationsbreiten dargestellt. Der 2,5 % kleinste und 2,5 % grösste Wert sind 3,13 Dioptrien und 8,63 Dioptrien. Damit besteht eine Wahrscheinlichkeit von 0,95, dass die Akkommodationsbreite einer dreissigjährigen Person zwischen 3,13 Dioptrien und 8,63 Dioptrien liegt. Es besteht eine Wahrscheinlichkeit von 0,5, dass die Akkommodationsbreite zwischen 4,97 Dioptrien und 6,76 Dioptrien ist. Man kann dieses Verfahren für die Alter 20, 22,5, 25, ..., 47,5 und 50 wiederho-

**Abb. 13.7** Streudiagramm der Messwerte mit eingezeichnetem plausibelsten Wert des Regressionsmodells (= der durchschnittlich erwartbaren Akkommodationsbreite in Funktion des Alters) und Prognosebänder für zukünftige Messwerte, zum Niveau 0,5 und 0,95

Akkommodationsbreite Mono



len. Die entsprechenden Wahrscheinlichkeitsintervalle lassen sich in das Streudiagramm der Daten einbauen. Dies zeigt Abb. 13.7. In heller Farbe sind die Prognoseintervalle zum Niveau 0,95, in dunkelgrauer Farbe diejenigen zum Niveau 0,5 gezeichnet. Man spricht auch von *Prognosebändern*. Die Prognosebänder sind breit, etwa  $\pm 3$  Dioptrien. Dies bedeutet, dass das einfache Regressionsmodell Akkommodationsbreiten nicht präzis prognostizieren kann. Der Grund ist klar: neben dem Alter wirken andere Kovariablen auf die Akkommodationsbreite. Auffallend ist, dass ein (tiefer) Messwert nicht im Prognoseband zum Niveau 0,95 liegt. Dies ist ein aussergewöhnlicher Messwert, eine Person mit unerwartet tiefer Akkommodationsbreite. Mit den dargestellten Prognosebändern ist auch die Frage (b) beantwortet.

Das Prognosemodell kann benutzt werden, um zu beurteilen, ob das Regressions- und das Datenmodell sinnvoll sind. Da kaum Messwerte ausserhalb des Prognosebands zum Niveau 0,95 liegen, werden die vorhandenen Messwerte rückwirkend gut prognostiziert. Das Daten- und das Regressionsmodell sind daher brauchbar.<sup>1</sup> □

Das obige Beispiel zeigt allgemein:

**Theorem 13.1 (Regressionsmodell mit Normalverteilung: Parameter berechnen und Prognosebänder bestimmen)**

Um von einer Grösse  $X$  auf eine stetige Zielgrösse  $Z$  zu rechnen, hat man  $(x_1/z_1), (x_2, z_2), \dots, (x_n/z_n)$  unabhängig modellierbare bivariate Datenwerte, die unter statistischer Kontrolle sind. Weiter hat man ein Regressionsmodell  $\mu_Z(X = x)$  mit Parametern  $\theta_1, \theta_2, \dots$ . Man nimmt an, dass die  $z$ -Werte normalverteilt um das

<sup>1</sup> Weitere Ausführungen dazu findet man im Abschn. 13.3.

*Regressionsmodell streuen:*

$$\text{Datenmodell: } i\text{-ter Messwert} \sim \text{Normal}(\mu_Z(X = x_i), \sigma)$$

*Die Streuung  $\sigma$  des Regressionsmodells hänge nicht von der erklärenden Variable  $X$  ab. Dann gilt:*

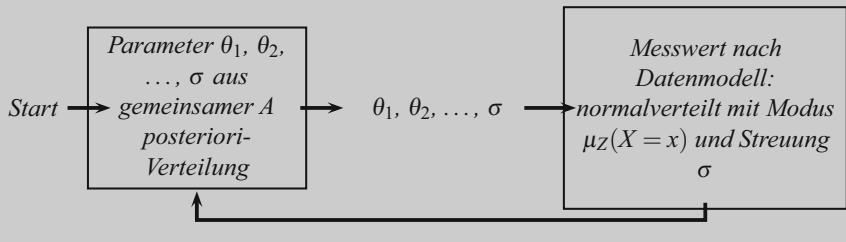
- (A) *Mit den Daten lässt sich aus der Vorinformation  $\mathcal{I}$  die gemeinsame Plausibilität zu den Parametern des Regressionsmodells und zu  $\sigma$  aktualisieren. Man hat*

$$\underbrace{\text{pdf}(\theta_1, \theta_2, \dots, \sigma | \text{Daten}, \mathcal{I})}_{\text{Posterior}} \propto \underbrace{\frac{1}{\sigma^n} \cdot \exp\{-0.5 \cdot \chi^2\}}_{\text{Likelihood}} \cdot \underbrace{\text{pdf}(\theta_1, \theta_2, \dots, \sigma | \mathcal{I})}_{\text{Prior}}$$

*Dabei ist*

$$\chi^2 = \frac{(z_1 - \mu_Z(x_1))^2 + \dots + (z_n - \mu_Z(x_n))^2}{\sigma^2}$$

- (B) *Weitere Messwerte  $Z$  bei gegebenem  $X = x$  können mit der folgenden Simulation prognostiziert werden:*



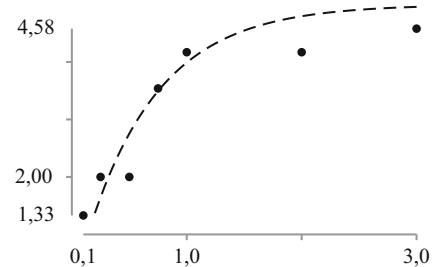
Hier ein zweites Beispiel dazu:

**Beispiel 13.2 (Ausspülen von Milchrohren)** In Beispiel 12.11 will man die durchschnittliche ausgespülte Milchkonzentration  $\mu_M$  in Funktion der Spülgeschwindigkeit  $v$  beschreiben. Dazu hat man eine Sättigungsfunktion als Regressionsmodell:

$$\mu_M(v) = A \cdot (1 - \exp\{-c \cdot v\})$$

Die Messwerte, mit denen die Parameter dieses Modells berechnet werden sollen, finden sich in Beispiel 12.11. Abb. 13.8 zeigt die sieben Messwerte und den Graphen des Regressionsmodells mit Werten  $A = 5,0$  und  $c = 1,6 \text{ s/m}$ . Um die plausibelsten Werte für  $A$  und  $c$  zu berechnen, wird angenommen, dass die kontinuierlichen Werte der Konzentration normalverteilt um  $\mu_M(v)$  streuen. Zudem hänge die Streuung  $\sigma$  nicht von der Spülgeschwindigkeit ab. Weiter hat man vor den Messungen nur minimale Information zu

**Abb. 13.8** Die sieben Messwerte mit dem Graphen des Regressionsmodells mit Werten  $A = 5,0$  und  $c = 1,6 \text{ s/m}$



den Parametern  $A$ ,  $c$  und  $\sigma$ . Nach Theorem 13.1 ist der gemeinsame Posterior von  $A$ ,  $c$  und  $\sigma$ :

$$\text{pdf}(A, c, \sigma \mid \text{Daten, min. Vorinformation}) \propto \underbrace{\frac{1}{\sigma^7} \cdot \exp\{-0,5 \cdot \chi^2\}}_{\text{Likelihood}} \cdot \underbrace{\frac{1}{\sigma}}_{\text{Prior}}$$

Dabei ist

$$\chi^2 = \left[ \frac{1,33 - A \cdot (1 - \exp\{-c \cdot 0,10\})}{\sigma} \right]^2 + \dots + \left[ \frac{4,58 - A \cdot (1 - \exp\{-c \cdot 3,00\})}{\sigma} \right]^2$$

Man kann 200 000 Punkte der gemeinsamen  $A$  posteriori-Verteilung von  $A$ ,  $c$  und  $\sigma$  mit einer MCMC-Kette abtasten. Die simulierten Koordinaten von  $A$  liefern die Plausibilität zu  $A$ . Bei einem Statistikprogramm gibt man dazu die Daten ein und nennt das Daten- und das Regressionmodell sowie die  $A$  priori-Verteilungen der Parameter:

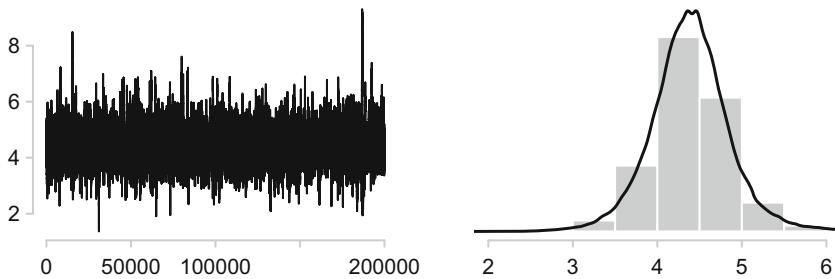
Datenmodell:	$i$ -ter Messwert $\sim \text{Normal}(\mu_i, \sigma)$
Regressionsmodell:	$\mu_i = A \cdot (1 - \exp\{-c \cdot v_i\})$
Prior:	$A \sim \text{Uniform}(0 ; 10)$
Prior:	$c \sim \text{Uniform}(0 ; 10)$
Prior:	$\ln \sigma \sim \text{Uniform}(\ln(10^{-3}) ; \ln(10^3))$

Das Regressionsmodell ist nicht linear in den Parametern  $A$  und  $c$ . Meist werden daher sehr lange Ketten gebraucht, um den Posterior der Parameter abzutasten (siehe [6], Seite 262). Hängen die Parameter sehr stark voneinander ab, so kann der Posterior meist nur unvollständig abgetastet werden. Es ist dann empfehlenswert, die Parameter des Regressionsmodell anders zu wählen. Für den Parameter  $A$  erhält man die  $A$  posteriori-Dichtefunktion in Abb. 13.9. Der plausibelste Wert für  $A$  ist der Modus  $A_0$  der Verteilung:

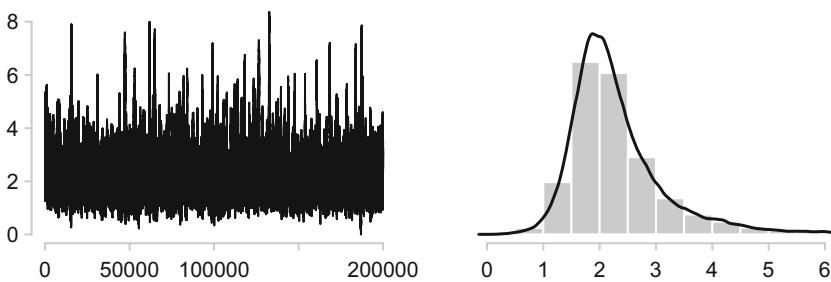
$$A \approx A_0 = 4,48$$

Es besteht eine Wahrscheinlichkeit von 0,5, dass  $A$  zwischen 4,0 und 4,6 liegt. Ein Wahrscheinlichkeitsintervall für  $A$  zum Niveau 0,95 ist:  $3,1 \leq A \leq 5,3$ . Für den zweiten Parameter  $c$  im Regressionsmodell erhält man aus der MCMC-Simulation die  $A$  posteriori-Verteilung in Abb. 13.10. Der plausibelste Wert für  $c$  ist der Modus  $c_0$  der Verteilung:

$$c \approx c_0 = 1,94 \text{ s/m}$$



**Abb. 13.9** MCMC-Kette der A posteriori-Verteilung für den Parameter  $A$  des Regressionsmodells (bei einer Akzeptanzrate von 0,37)



**Abb. 13.10** MCMC-Kette für den Posterior des Parameters  $c$

Es besteht eine Wahrscheinlichkeit von 0,5, dass  $c$  zwischen 1,8 s/m und 2,1 s/m liegt. Ein Wahrscheinlichkeitsintervall für  $c$  zum Niveau 0,95 besteht aus den Werten 1,15 s/m und 4,7 s/m.

In Abb. 13.11 sind Werte ( $A/c$ ) aus der MCMC-Kette dargestellt. Dies zeigt die gemeinsame A posteriori-Verteilung der beiden Parameter  $A$  und  $c$ . Damit könnte man Bereiche auswählen, in denen  $A$  und  $c$  mit hoher Wahrscheinlichkeit auftreten. Auffallend ist: die beiden Parameter sind, gegeben die Daten, korreliert. Der Korrelationskoeffizient beträgt  $-0,63$ . Dies ist keine starke Korrelation.

Auch Prognosebänder kann man jetzt berechnen. Wie in Theorem 13.1 erklärt, geht man wie folgt vor: Mit den 200 000 Werten der Parameter  $A$ ,  $c$  und  $\sigma$  aus der MCMC-Simulation, kann man jeweils für  $v = 0,1 \text{ m/s}, 0,3 \text{ m/s}, 0,5 \text{ m/s}, \dots, 2,9 \text{ m/s}$  und  $3,1 \text{ m/s}$  mit dem Regressionsmodell und dem Datenmodell Konzentrationen simulieren. Bei  $v = 0,3 \text{ m/s}$  sieht dies so aus:

```
for i = 1 to 200000 do
  (A[i], c[i], Sigma[i]) = i-ter Punkt aus
    MCMC-Simulation für A, c und sigma
  Konzentration0.3[i] = erzeugt aus Normalverteilung,
    mit Modus A[i] * (1-exp(-c[i]*0.3))
```

```
    und Streuung Sigma[i]
end
```

Die entsprechenden Prognosebänder zum Niveau 0,5 und 0,95 lassen sich in das Streudiagramm der Daten einbauen. Dies visualisiert Abb. 13.12. In heller Farbe sind Prognoseintervalle zum Niveau 0,95, in dunkelgrauer Farbe Prognoseintervalle zum Niveau 0,5 gezeichnet. Das Prognosemodell zeigt, dass das Daten- und das Regressionsmodell sinnvoll scheinen. Alle Messwerte sind innerhalb der Prognosebänder zum Niveau 0,95. Die Messwerte werden – rückwirkend – gut prognostiziert.  $\square$

## 13.2 Die Methode der kleinsten Quadrate

Die plausibelsten Werte für die Parameter eines Regressionsmodells werden in Spezialfällen mit der Methode der kleinsten Quadrate berechnet. Dies wird in diesem Abschnitt gezeigt.

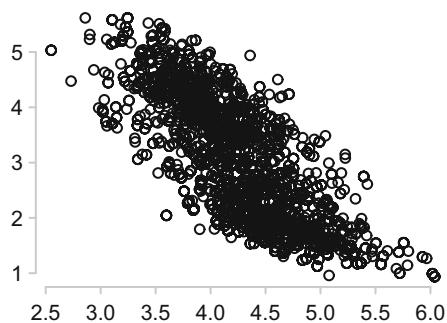
Ein Regressionsmodell, um von einer erklärenden Grösse  $X$  auf den durchschnittlichen Wert  $\mu_Z$  einer abhängigen Zielgrösse  $Z$  zu rechnen, sei gegeben:

$$\mu_Z(X = x) = \mu_Z(x | \theta) = \mu_Z(x | \theta_1, \theta_2, \dots)$$

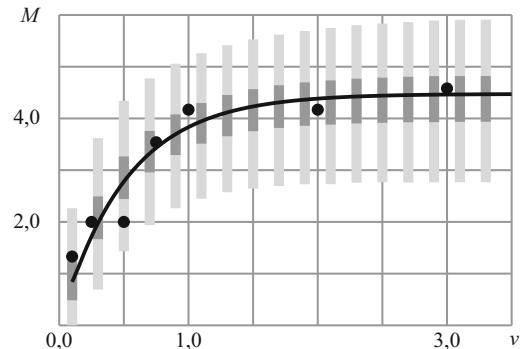
Das Regressionsmodell ist eine Funktion, die unbekannte Parameter  $\theta_1, \theta_2, \dots$  besitzt. Das Datenmodell, das plausibilisiert, wie Messwerte bei gegebenem Wert von  $X = x$  streuen, sei die Normalverteilung. Dabei hänge die Streuung  $\sigma$  nicht von der erklärenden Variable ab. Bei  $n$  gemessenen Paaren  $(x_1/z_1), (x_2/z_2), \dots, (x_n/z_n)$ , die unabhängig seien, lautet der Posterior für die Parameter des Regressionsmodells:

$$\text{pdf}(\theta, \sigma | \text{Daten, Vorinform.}) \propto \underbrace{\frac{1}{\sigma^n} \cdot \exp \left\{ -\frac{1}{2} \cdot \chi^2 \right\}}_{\text{Likelihood}} \cdot \underbrace{\text{pdf}(\theta, \sigma | \text{Vorinform.})}_{\text{Prior}}$$

**Abb. 13.11** Gemeinsame A posteriori-Verteilung der Parameter  $A$  und  $c$  des Regressionsmodells



**Abb. 13.12** Streudiagramm der Messwerte mit den plausibelsten Werten von  $A$  und  $c$  des Regressionsmodells (= die durchschnittlich erwartbare Konzentration in Funktion der Spülgeschwindigkeit) und Stäben, welche die Prognosebänder zum Niveau 0,5 und 0,95 für weitere Messwerte der Konzentration bilden



Dabei ist

$$\chi^2 = \left[ \frac{z_1 - \mu(x_1 | \theta)}{\sigma} \right]^2 + \left[ \frac{z_2 - \mu(x_2 | \theta)}{\sigma} \right]^2 + \cdots + \left[ \frac{z_n - \mu(x_n | \theta)}{\sigma} \right]^2$$

Die vorkommende Differenz  $z_i - \mu(x_i | \theta)$  ist die Differenz zwischen dem Messwert  $z_i$  und dem Modellwert des Regressionsmodells, wenn  $X = x_i$  ist. Man nennt dies das  $i$ -te *Residuum*. Bezeichnet wird es mit  $r_i$ . Die Residuen sind im Streudiagramm der Messwerte, wie in Abb. 13.13 dargestellt, einfach visualisierbar. Es ist damit:

$$\chi^2 = \frac{1}{\sigma^2} \cdot (r_1^2 + r_2^2 + \cdots + r_n^2)$$

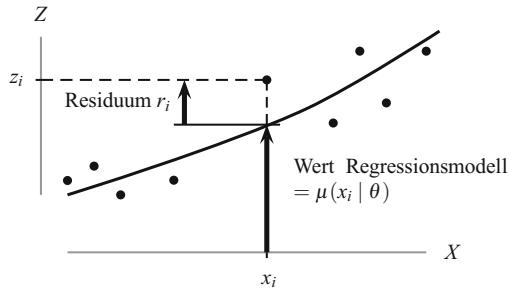
Um die plausibelsten Werte der Regressionsparameter zu finden, braucht man das Maximum der A posteriori-Verteilung. Hängt der Prior nicht von den Regressionsparametern ab (z. B. bei minimaler Vorinformation) oder ist der Likelihood dominant (z. B. bei vielen Datenwerten), so wird diese maximal, wenn der Ausdruck  $\chi^2$  minimal wird. Mit anderen Worten: *Die Summe der Residuen im Quadrat muss minimal werden!* Man spricht – wie im Kap. 10 erläutert – von der Methode der kleinsten Quadrate.<sup>2</sup> Diese Methode ist in vielen Statistikprogrammen implementiert. Bei Regressionsmodellen, die linear von den Parametern abhängen, existieren sogar explizite Formeln für (a) die plausibelsten Werte und (b) für Wahrscheinlichkeitsintervalle der Parameter des Regressionsmodells.<sup>3</sup> Die

<sup>2</sup> Die Feststellung, dass beim Datenmodell der Normalverteilung mit konstanter Streuung, die plausibelsten Werte der Parameter durch die Methode der kleinsten Quadrate bestimmt werden, stammt von den Mathematikern C. F. Gauss (1809) und A. M. Legendre (1806).

<sup>3</sup> Für das einfachste Regressionsmodell einer Geraden  $\mu_Z(x) = a + b \cdot x$  hat man für die plausibelsten Werte  $a_0$  und  $b_0$  der Parameter  $a$  und  $b$  die Matrix-Formel

$$\begin{pmatrix} a_0 \\ b_0 \end{pmatrix} = (X^T \cdot X)^{-1} \cdot X^T \cdot Z \quad \text{mit} \quad X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad Z = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix}$$

**Abb. 13.13** Definition eines Residuums



Methode der kleinsten Quadrate lässt sich auch bei Regressionsmodellen, die nicht linear von ihren Parametern abhängen, umsetzen. Zwar kann man hier die plausibelsten Werte der Parameter nicht mehr explizit berechnen. Algorithmen sind in der Regel fähig, diese zu bestimmen, wenn man weiß, wo die plausibelsten Werte etwa liegen.

**Beispiel 13.3 (Akkommodationsbreite)** Um die durchschnittliche Akkommodationsbreite  $\mu_{\text{Akko}}$  zu berechnen, wird das lineare Regressionsmodell

$$\mu_{\text{Akko}}(\text{Alter}) = a + b \cdot \text{Alter}$$

benutzt. Gegeben die 19 Messwerte, das Datenmodell der Normalverteilung mit konstanter Streuung und minimale Vorinformation zu den Parametern der Modelle, lauten die plausibelsten Werte  $a_0$  und  $b_0$  von  $a$  und  $b$ :

$$a_0 = 14,15 \text{ Dioptrien}, \quad b_0 = -0,28 \text{ Dioptrien/Jahr}$$

Diese Werte können mit der Methode der kleinsten Quadrate berechnet werden. Dazu muss die Summe

$$[9,9 - (a + b \cdot 20)]^2 + [9,5 - (a + b \cdot 22)]^2 + \dots + [0,5 - (a + b \cdot 50)]^2$$

die von  $a$  und  $b$  abhängt, minimiert werden. Dies lässt sich mit einem Statistikprogramm schnell ausführen. Ebenso lassen sich mit einem Statistikprogramm Wahrscheinlichkeitsintervalle explizit bestimmen. So erhält man mit einer Wahrscheinlichkeit von je 0,95, dass

$$\begin{aligned} 11,9 \text{ Dioptrien} &\leq a \leq 16,3 \text{ Dioptrien} \\ -0,34 \text{ Dioptrien/Jahr} &\leq b \leq -0,21 \text{ Dioptrien/Jahr} \end{aligned}$$

---

In den Matrizen  $X$  und  $Z$  befinden sich die Datenwerte  $(x_1/z_1), (x_2/z_2), \dots, (x_n/z_n)$ . Dabei ist  $X^T$  die transponierte Matrix von  $X$ . Weiter lassen sich Wahrscheinlichkeitsintervalle für  $a$  und  $b$  mit  $t$ -Verteilungen angeben. Ähnliche Formeln mit Matrizen und  $t$ -Verteilungen hat man für alle linearen Regressionsmodelle.

**Tab. 13.2** Alle (approximativen) Residuen beim linearen Regressionsmodell für die Akkommodationsbreite beim Mono-Sehen

Alter	20	22	22	22	22	24	28	32	33	34
Residuum	1,3	1,4	-1,0	-0,3	-0,1	-1,0	1,2	-0,3	-0,1	-3,0
Alter	35	39	41	41	43	43	46	48	50	
Residuum	-1,4	2,6	1,2	0,3	-0,9	-0,1	0,1	-0,1	0,2	

Der erste Messwert ist eine Person mit 20 Jahren und einer Akkommodationsbreite von 9,9 Dioptrien. Das zugehörige Residuum  $r_1$  lautet damit

$$r_1 = 9,9 \text{ Dioptrien} - (a + b \cdot 20 \text{ Jahr})$$

Die Regressionsparameter  $a$  und  $b$  sind jedoch nicht exakt bestimmt. Man kann das Residuum  $r_1$  approximativ bestimmen, indem man für  $a$  und  $b$  die plausibelsten Werte einsetzt. Damit ist

$$r_1 \approx 9,9 \text{ Diop.} - (14,15 \text{ Diop.} - 0,28 \text{ Diop./Jahr} \cdot 20 \text{ Jahr}) = 1,3 \text{ Diop.}$$

Die restlichen achtzehn Residuen berechnen sich analog. Man erhält die Werte in Tab. 13.2. Das augenfälligste Residuum ist -3,0 Dioptrien bei einer Person im Alter von 34 Jahren. Der Messpunkt befindet sich ausserhalb des Prognosebands zum Niveau 0,95. Es ist daher üblich, den Messpunkt als aussergewöhnlich zu bezeichnen.  $\square$

**Beispiel 13.4 (Kanalwärmetauscher)** In Beispiel 2.13 wird untersucht, wie Wärmetauschelemente in Abwasserkanälen verformt werden. Die Verformung hängt von drei Faktoren ab: Der Dicke  $S$  des Siggenblechs, der Dicke  $D$  des Deckblechs und der Breite  $R$  des Rinnendeckblechs. Die Faktoren  $S$ ,  $D$  und  $R$  nehmen dabei die Werte -1 und +1 an. Man kann versuchen, mit dem linearen Regressionsmodell

$$\mu_{\text{Verformung}}(S, D) = a_0 + a_1 \cdot S + a_2 \cdot D + a_3 \cdot S \cdot D$$

die durchschnittliche Verformung  $\mu_{\text{Verformung}}$  zu berechnen. Es enthält die aus dem Pareto-Diagramm herausgelesenen Hauptfaktoren und ihre Interaktion (siehe Abb. 12.18). Die plausibelsten Werte der Parameter  $a_0$ ,  $a_1$ ,  $a_2$  und  $a_3$  können mit der Methode der kleinsten Quadrate ermittelt werden. Man erhält:

$$a_0 \approx 4,63 \quad a_1 \approx -2,44 \quad a_2 \approx -1,27 \quad a_3 \approx 0,67$$

Ebenso lassen sich, da das Regressionsmodell linear in den Parametern ist, Wahrscheinlichkeitsintervalle mit einem Statistikprogramm explizit berechnen. So hat man zum Niveau 0,95:

Parameter	untere Grenze	obere Grenze
$a_0$	4,55	4,72
$a_1$	-2,52	-2,35
$a_2$	-1,35	-1,18
$a_3$	0,59	0,75

Man kann vier Prognosebereiche für weitere Messwerte der Verformung rechnen, da man vier Möglichkeiten für  $D$  und  $S$  hat:  $(+1, +1)$ ,  $(+1, -1)$ ,  $(-1, +1)$  und  $(-1, -1)$ . Sie lassen sich mit Theorem 13.1 bestimmen. Abb. 13.14 visualisiert die Daten und die Prognosebänder zum Niveau 0,95. Alle Messpunkte liegen in den Prognosebändern. Die ausgezogene Gerade ist das Regressionsmodell mit  $D = +1$ :

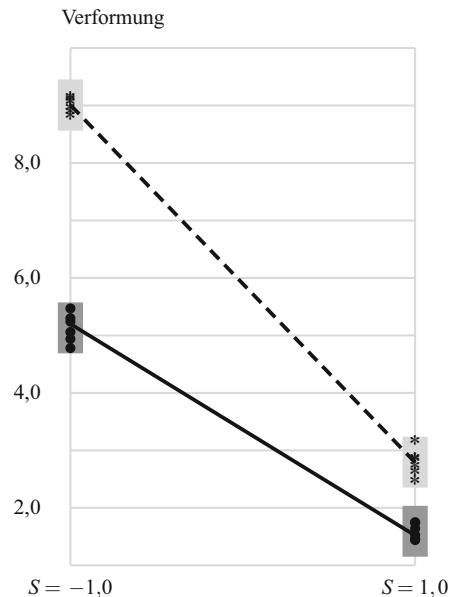
$$\mu_{\text{Verformung}}(S, D = +1) = 4,63 - 2,44 \cdot S - 1,27 \cdot 1 + 0,6 \cdot S \cdot 1 = 3,36 - 1,84 \cdot S$$

Die gestrichelte Gerade zeigt das Regressionsmodell bei  $D = -1$ . Es lautet:

$$\mu_{\text{Verformung}}(S, D = -1) = 5,90 - 3,11 \cdot S$$

Wegen der Interaktion zwischen den Faktoren haben die beiden Geraden verschiedene Steigungen.  $\square$

**Abb. 13.14** Messpunkte und lineares Regressionsmodell:  
Punkte sind die Messwerte bei  
 $D = +1$ , Sterne zeigen die  
Messwerte, wenn  $D = -1$   
ist; dunkle Prognosebänder für  
 $D = +1$ , helle für  $D = -1$ .



### 13.3 Kritische Überlegungen

Es ist sicher sinnvoll, die Rechnungen zu einem Regressionsmodell zu „überprüfen“. Es kann passieren, dass das Regressionsmodell schlecht, über- oder unterangepasst ist. Es gibt Verfahren, mit den man dies beurteilen kann: Kreuzvalidierung (siehe Abb. 9.7) und Plausibilitätskriterien sind zwei Beispiele. Eine Einführung zu Plausibilitätskriterien findet man in Kap. 14. Lohnenswert ist zuerst, das Folgende zu überlegen:

- (1) *Zum Regressionsmodell und zum Datenmodell:* Beschreibt das Regressionsmodell die gesuchte Zielgröße vernünftig? Ist das Modell mathematisch oder physikalisch begründbar? Sind die wesentlichen erklärenden Variablen im Modell vorhanden? In welchem Bereich der erklärenden Variablen soll das Modell angewendet werden? Prognostizieren das Daten- und das Regressionsmodell weitere Datenwerte in sinnvollen Bereichen?
- (2) *Zur Likelihood-Funktion:* Sind Abhängigkeiten zwischen den Datenwerten vorhanden? Wenn ja, sind diese in der Likelihood-Funktion eingebaut?

Wie man versuchen kann, solche Fragen zu beantworten, zeigen die weiteren Beispiele:

**Beispiel 13.5 (Akkommodationsbreite)** Zur Frage (1) *Sind das Regressions- und das Datenmodell sinnvoll gewählt?* zeigt Abb. 13.7 mit dem berechneten Regressionsmodell und den eingetragenen Prognosebändern, dass die Messwerte innerhalb der Prognosebänder liegen. Die Messwerte werden gut prognostiziert. Damit ist das Regressions- und Datenmodell brauchbar. Besser abgestützt wären die Modelle mit einer Kreuzvalidierung: Dazu benötigt man weitere Messwerte.

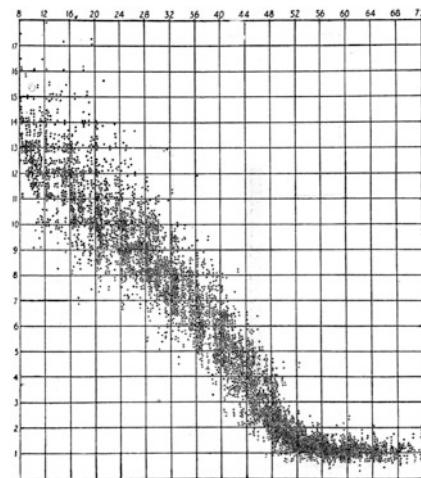
Die Prognosebänder ab etwa Alter 45 decken negative Bereiche der Akkommodationsbreite ab. Dies ist physikalisch nicht möglich. Das Regressionsmodell ist daher für Personen, die älter als 45 Jahre sind, nicht mehr benutzbar. Dies zeigt auch Abb. 13.15 aus der Untersuchung [7] von A. Duane aus dem Jahr 1922 von 4200 Personen. Es ist also nicht anzuraten, die durchschnittliche Akkommodationsbreite mit dem einfachen Regressionsmodell für über 45-jährige Personen zu extrapoliieren. Regressionsmodelle, die die durchschnittliche Akkommodationsbreite  $\mu_{\text{Akko}}$  für Personen in allen Altern rechnen, findet man in der Fachliteratur (siehe etwa [1]). Ein Beispiel ist eine Sigmoidfunktion:

$$\mu_{\text{Akko}}(\text{Alter}) = \frac{a}{1 + \exp\{b \cdot (\text{Alter} - c)\}}$$

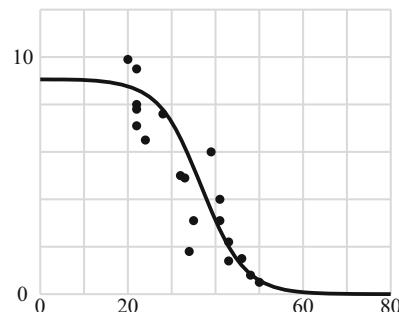
Der Graph dieses Regressionsmodells findet sich in Abb. 13.16.

Zur Frage (2): *Ist die Likelihood richtig gerechnet?* Es wurde angenommen, dass die Messwerte unabhängig sind. Ist das sinnvoll? Man kann dazu die gemessenen Akkommodationsbreiten in der Reihenfolge ihrer Messungen betrachten. Diese Messwerte haben

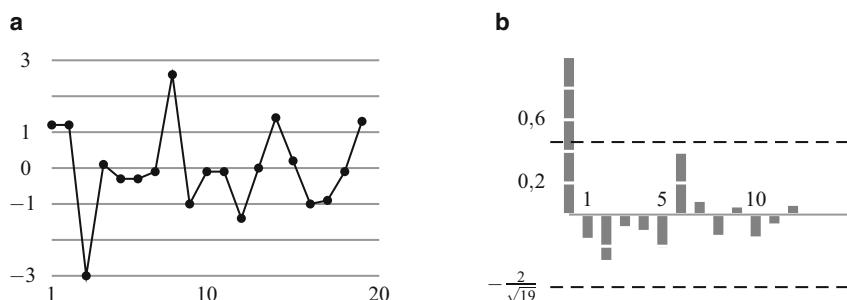
**Abb. 13.15** Akkommodationsbreite bei 4200 Personen  
(aus [7])



**Abb. 13.16** Sigmoidfunktion  
 $\mu_{\text{Akko}}(\text{Alter}) = a/(1 + \exp\{b \cdot (\text{Alter} - c)\})$  mit  $a = 9$ ,  
 $b = 36,8$  und  $c = 0,2$



einen Trend: je älter eine Person ist, umso kleiner ist tendenziell die Akkommodationsbreite. Um diesen Trend zu entfernen, betrachtet man die approximativen Residuen in der Reihenfolge ihrer Messung. Diese finden sich in Tab. 13.3. Das erste approximative



**Abb. 13.17** Streudiagramm der Residuen in der Reihenfolge der Messungen (a) und Graph der Autokorrelationsfunktion der Residuen (b)

**Tab. 13.3** Approximative Residuen aus dem linearen Regressionsmodell und Versuchsreihenfolge (Nr.)

Alter	20	22	22	22	22	24	28	32	33	34
Nr.	19	14	16	5	18	9	2	6	10	3
Residuum	1,3	1,4	-1,0	-0,3	-0,1	-1,0	1,2	-0,3	-0,1	-3,0
Alter	35	39	41	41	43	43	46	48	50	
Nr.	12	8	1	13	17	7	4	11	15	
Residuum	-1,4	2,6	1,2	0,3	-0,9	-0,1	0,1	-0,1	0,2	

Residuum beträgt 1,2, das zweite ist 1,2 und das dritte lautet -3,0. Mit der Autokorrelationsfunktion in Abb. 13.17 lässt sich überblicken, ob die Residuen trendfrei und voneinander unabhängig sind: In Abb. 13.17a finden sich die approximativen Residuen in der Reihenfolge ihrer Messung und in Abb. 13.17b ist der Graph der Autokorrelationsfunktion dargestellt. Für den Lag  $r = 0$  ist der Wert der Autokorrelationsfunktion eins. Kein anderer Wert ist ausserhalb des Bandes, das durch die Zahlen  $-2/\sqrt{19}$  und  $2/\sqrt{19}$  begrenzt wird. Daher scheint es sinnvoll anzunehmen, dass die Messwerte unabhängig sind.  $\square$

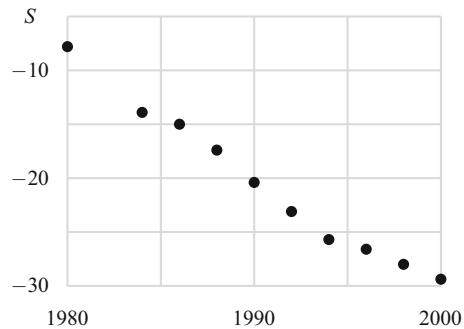
**Beispiel 13.6 (Vermessung Schüttdamm)** Der Schüttdamm Châtelard-Barberine der Schweizerischen Bundesbahnen SBB speichert Wasser zur Elektrizitätserzeugung. Der Damm besteht aus aufgeschüttetem Material, das sich vertikal setzt. Mit jährlichen Messungen will man die Setzung  $\mu_S(t)$  des Schüttdamms in Funktion der Zeit  $t$  berechnen. Gemäss physikalischen Untersuchungen in [9] lassen sich Setzungen von Schüttämmen mit einer Abklingfunktion modellieren. Ein plausibles Regressionsmodell für  $\mu_S(t)$  ist:

$$\mu_S(t) = a \cdot \exp\{-b \cdot (t - 1980)\} + c$$

Besonders interessant ist der Parameter  $c$ . Er gibt an, wie gross die Setzung langfristig sein wird. Um die Parameter des Regressionsmodells zu berechnen, hat man als Information die Daten in Tab. 13.4. Die Daten sind in Abb. 13.18 visualisiert. Ein Ingenieur nimmt an, dass gemessene (kontinuierliche) Werte der Setzung mit endlicher Streuung um das Regressionsmodell variieren. Nach MaxEnt wird damit das Wissen zu  $S$  mit der Normalverteilung beschrieben. Die Streuung  $\sigma$  scheint, wie Abb. 13.18 zeigt, nicht von der Zeit  $t$  abzuhängen. Der Ingenieur wählt sie daher konstant. Vor den Daten hat er keine weitere Information zu den Parametern. Weiter nimmt er vorerst an, dass die Messwerte unabhängig sind. Die plausibelsten Werte  $a_0$ ,  $b_0$  und  $c_0$  für die Parameter des Regressionsmodells berechnet man damit mit der Methode der kleinsten Quadrate. Die Werte müssen mit einem Computerprogramm berechnet werden, da das Regressionsmodell nicht linear in den Parametern ist. Man erhält:

$$a_0 = 46,82 \text{ mm}, \quad b_0 = 0,032 \text{ Jahr}^{-1}, \quad c_0 = -54,47 \text{ mm}$$

**Abb. 13.18** Streudiagramm der Messwerte aus Tab. 13.4



**Tab. 13.4** Setzung (in mm) des Schüttdamms während der Jahre 1980–2000 (aus [3])

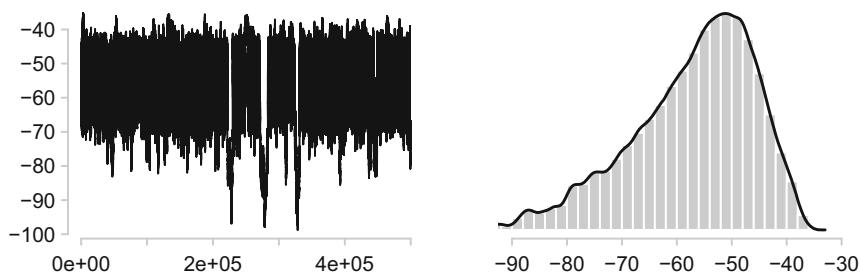
Jahr	1980	1984	1986	1988	1990	1992	1994	1996	1998	2000
Setzung $S$	-7,80	-13,90	-15,00	-17,40	-20,40	-23,10	-25,70	-26,60	-28,00	-29,38

Mit einer MCMC-Kette (nach Theorem 13.1) lassen sich die A posteriori-Verteilungen für die Parameter und damit Wahrscheinlichkeitsintervalle bestimmen. So zeigt Abb. 13.19 die Verteilung von  $c$ . Die Verteilung ist unimodal schief. Mit einer Wahrscheinlichkeit von 0,5 liegt  $c$  zwischen  $-62,9$  mm und  $-48,5$  mm. Es besteht eine Wahrscheinlichkeit von 0,95, dass  $c$  zwischen  $-81,2$  mm und  $-40,0$  mm ist.

Auch Prognosebänder für Messungen der Setzung zum Niveau 0,5 und 0,95 lassen sich mit dem erwähnten Verfahren von Theorem 13.1 bestimmen. Sie sind, zusammen mit dem Regressionsmodell, in Abb. 13.20 visualisiert.

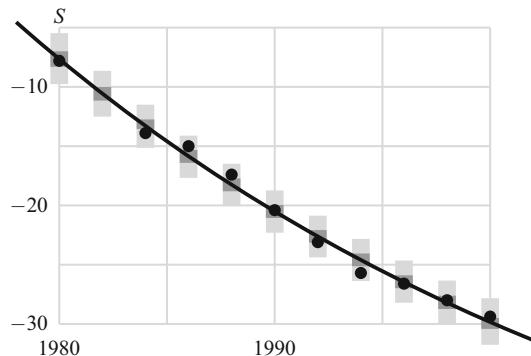
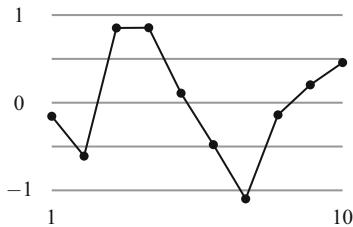
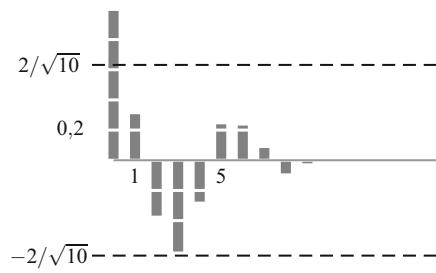
Mit den obigen Rechnungen lassen sich die Fragen (1) und (2), die am Anfang des Abschnitts gestellt wurden, beantworten:

- (1) Sind das Regressions- und das Datenmodell sinnvoll gewählt? Abb. 13.20 mit dem berechneten Regressionsmodell und den eingetragenen Prognosebändern zeigt, dass die Messwerte innerhalb der Prognosebänder liegen. Damit sind das Regressions- und



**Abb. 13.19** MCMC-Kette mit Akzeptanzrate 0,25 für den Posterior des Parameters  $c$

**Abb. 13.20** Regressionsmodell und Prognosebänder zum Niveau 0,5 und 0,95 zur Setzung in Funktion des Jahres

**a****b**

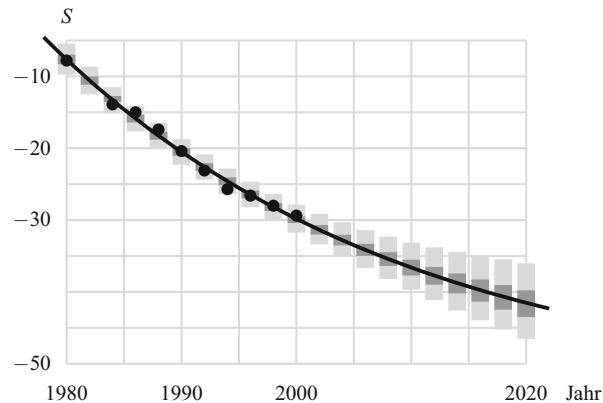
**Abb. 13.21** Streudiagramm der Residuen (a) und Graph der Autokorrelationsfunktion der Residuen (b)

Datenmodell gut brauchbar. Weiter ist das Regressionsmodell physikalisch erklärbar.  
Dies ist ein weiterer Pluspunkt.

- (2) *Ist die Likelihood richtig gerechnet?* Es wurde angenommen, dass die Messwerte unabhängig modelliert werden können. Dies kann mit dem Residuenplot und der Autokorrelationsfunktion der approximativen Residuen in Abb. 13.21 beurteilt werden. In Abb. 13.21a finden sich die approximativen Residuen in der Reihenfolge ihres Messzeitpunkts. Abb. 13.21b stellt den Graphen der Autokorrelationsfunktion dar. Es sind keine Werte ausserhalb des Bandes, das durch die Zahlen  $-2/\sqrt{10}$  und  $2/\sqrt{10}$  begrenzt wird. Daher ist es sinnvoll anzunehmen, dass die Messwerte unabhängig sind.

Da das Regressionsmodell physikalisch erklärt werden kann, lässt sich mit ihm prognostizieren, in welchem Bereich zukünftige Messungen der Setzung liegen werden. Abb. 13.22 zeigt Prognosebereiche zum Niveau 0,5 und 0,95 bis zum Jahr 2020. Sie sind mit dem üblichen Simulationsverfahren gerechnet. Im Jahr 2020 werden gemessene Setzungen mit einer Wahrscheinlichkeit von 0,95, zwischen  $-46,2 \text{ mm}$  und  $-36,2 \text{ mm}$  liegen.  $\square$

**Abb. 13.22** Prognosebereiche zum Niveau 0,5 und 0,95 für Messungen der Setzung beim Schüttdamm bis zum Jahr 2020



**Beispiel 13.7 (Lasermarkierung)** In Beispiel 12.10 will man den durchschnittlichen Kontrast  $\mu_{\text{Kontrast}}$  mit dem Regressionsmodell – einer Sigmoidfunktion –

$$\mu_{\text{Kontrast}}(E) = \frac{a}{1 + \exp\{-(E - \rho)/\kappa\}}$$

in Funktion der Energie  $E$  des Lasers beschreiben. Das Regressionsmodell ist eine Konsequenz von physikalischen Überlegungen. Es ist daher sinnvoll gewählt. Um die Parameter des Regressionsmodells zu berechnen, sind 54 Messungen vorhanden. Als Datenmodell wählten die Experimentierer die Normalverteilung mit konstanter Streuung  $\sigma$ . Vor dem Experiment hatten die Experimentierer keine weitere Information. Die plausibelsten Werte  $a_0$ ,  $\rho_0$  und  $\kappa_0$  für die Parameter  $a$ ,  $\rho$  und  $\kappa$  des Regressionsmodells berechnet man damit mit der Methode der kleinsten Quadrate. Man erhält:

$$a_0 = 101,8, \quad \rho_0 = 0,0380, \quad \kappa_0 = 0,00641$$

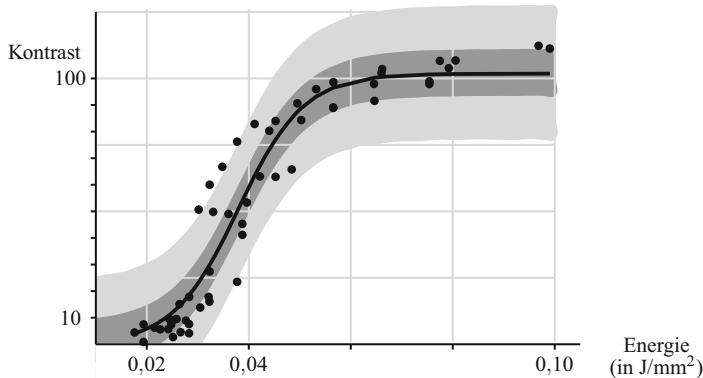
Mit Theorem 13.1 lassen sich die A posteriori-Verteilungen für diese Parameter und damit Wahrscheinlichkeitsintervalle bestimmen. So hat man mit einer Wahrscheinlichkeit von 0,95, dass

$$95,2 \leq a \leq 110,0$$

Mit einer Wahrscheinlichkeit von 0,95 liegt  $\rho$  zwischen 0,0361 und 0,0404. Es besteht eine Wahrscheinlichkeit von 0,95, dass  $0,005 \leq \kappa \leq 0,008$  ist.

Auch Prognosebänder zum Niveau 0,5 und 0,95 lassen sich mit dem Verfahren von Theorem 13.1 bestimmen. Sie sind, zusammen mit dem Regressionsmodell, in Abb. 13.23 dargestellt. Mit den obigen Rechnungen lassen sich wiederum die Fragen (1) und (2), die am Anfang des Abschnitts gestellt wurden, beantworten:

(1) Sind das Regressions- und das Datenmodell sinnvoll gewählt? Abb. 13.23 zeigt, dass viele Messwerte nicht mehr in den Prognosebändern zum Niveau 0,95 liegen. Dies

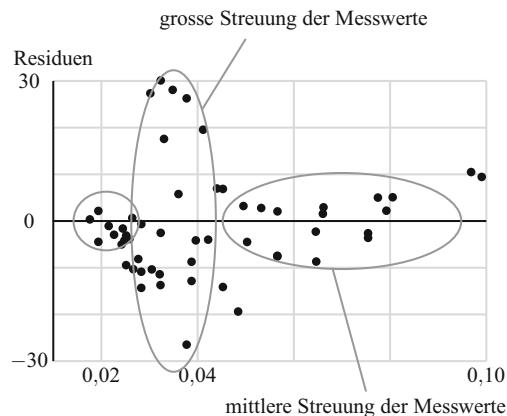
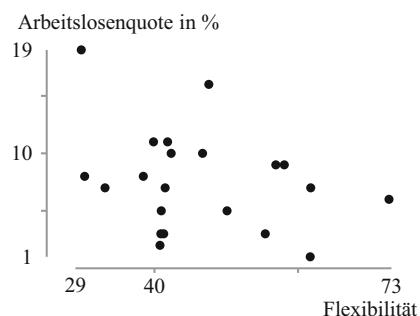


**Abb. 13.23** Regressionsmodell mit den plausibelsten Parametern und mit Prognosebänder (50 und 95 %) für weitere Messungen

im Bereich, indem der Kontrast über die Energiedichte gesteuert werden kann. Das Datenmodell ist damit nicht wirklich geeignet. Wie kann ein besseres Modell konstruiert werden? Dazu muss die Frage beantwortet werden, warum das Modell „versagt“ hat. Die Experimentierer wussten, dass bei kleinen Energien die Lasermarkierung kaum sichtbar ist. Bei mittleren Energien entstehen Streuungseffekte, aufgrund der uneinheitlichen Farbdicke auf den Papieren. Bei grossen Energien wird das Papier praktisch durchgebrannt, die Streuung des Kontrasts wird daher klein. Dieses Wissen lässt sich mit den Messungen visualisieren. Man kann dazu die approximativen Residuen gegen die Energie in einem Streudiagramm darstellen. Dies zeigt Abb. 13.24. Im Bereich zwischen  $0,025 \text{ J mm}^{-2}$  und  $0,04 \text{ J mm}^{-2}$  streuen die Residuen stark. Die Standardabweichung der Zielgrösse hängt somit vom gewählten Energiebereich ab. Dieses Wissen ist im Datenmodell der Normalverteilung mit konstanter Streuung  $\sigma$  nicht eingebaut. Daher ist das Modell schlecht. Ein besseres Modell arbeitet mit einem Datenmodell, dass die Standardabweichung in Funktion von  $E$  beschreibt. Eine solches Modell sprengt jedoch den Rahmen eines einführenden Kurses in Statistik. □

**Beispiel 13.8 (Arbeitslosigkeit und Flexibilität)** Beim Beispiel 6.3 möchte man mit einem Regressionsmodell von der Flexibilität des Arbeitsmarkts auf die mittlere Arbeitslosenquote rechnen. Dazu dienen die Beobachtungen von 21 Industrieländern in den Jahren 1984 bis 1990, die in Tab. 6.4 dargestellt sind. Die dargestellten Daten in Abb. 13.25 zeigen, dass ein lineares Modell der Form

$$\mu_A(\text{Flexibilität}) = a + b \cdot \text{Flexibilität}$$

**Abb. 13.24** Residuenplot des Regressionsmodells**Abb. 13.25** Streudiagramm der Daten zur Arbeitslosenquote und zur Flexibilität des Arbeitsmarkts

für die durchschnittliche Arbeitslosenquote  $\mu_A$  plausibel scheint. Der Parameter  $b$  zeigt dabei den Einfluss der Flexibilität auf die Arbeitslosenquote.<sup>4</sup>

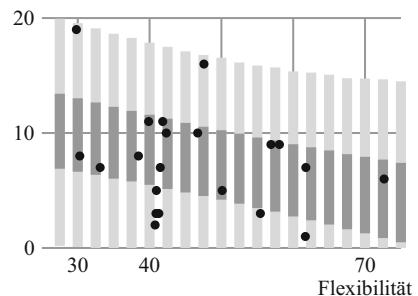
Mit Theorem 13.1 kann man die A posteriori-Verteilung für den wichtigen Parameter  $b$  berechnen und Prognosebänder bestimmen. Als Datenmodell wählt man die Normalverteilung mit konstanter Streuung  $\sigma$ . Vor der Datensammlung habe man keine weitere Information zu den Parametern. In Abb. 13.26 findet sich das berechnete Regressionsmodell und die Prognosebänder zum Niveau 0,5 und 0,95. Alle Daten werden durch das Regressions- und Datenmodell gut prognostiziert. Die Modelle sind daher sinnvoll. Abb. 13.27 zeigt den Posterior für den Parameter  $b$ . Der plausibelste Wert  $b_0$  ist  $-0,13$ .

<sup>4</sup> Eine Bemerkung: Die Daten sind negativ verbunden. Der empirische Korrelationskoeffizient ist  $-0,33$ . Der Wert  $-0,33$  ist nicht kleiner als  $-2/\sqrt{21} = -0,44$ . Man kann daher Flexibilität und Arbeitslosigkeit als unabhängig betrachten. Mit anderen Worten:

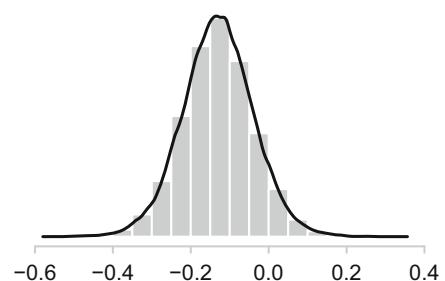
$$\mathbb{P}(\text{Arbeitslosigkeit} \mid \text{Flexibilität, Daten}) = \mathbb{P}(\text{Arbeitslosigkeit} \mid \text{Daten})$$

Es dürfte schwierig sein, mit den Daten Aussagen zur Arbeitslosigkeit in Industrieländern auf Grund der Flexibilität des Arbeitsmarkts zu formulieren.

**Abb. 13.26** Bänder zum Niveau 0,5 und 0,95, um die Arbeitslosenquote in Funktion der Flexibilität des Arbeitsmarkts zu prognostizieren



**Abb. 13.27** A posteriori-Verteilung des Parameters  $b$ , berechnet aus einer Kette von 100 000 Punkten



Daher ist:

$$b \approx b_0 = -0,13$$

Eine um 10 Punkte höhere Flexibilität vermindert die durchschnittliche Arbeitslosenquote um 1,3 %. Der Parameter  $b$  ist jedoch unpräzis bestimmt. Mit einer Wahrscheinlichkeit von 0,95 liegt er zwischen  $-0,31$  und  $0,05$ . Die Wahrscheinlichkeit, dass  $b$  grösser als null ist, beträgt 0,07. Die durchschnittliche Arbeitslosenquote kann mit zunehmender Flexibilität ab- oder zunehmen. Mit den vorgestellten Daten ist es daher schwierig zu behaupten, dass die Flexibilität einen Einfluss auf die durchschnittliche Arbeitslosenquote hat.

Die Prognosebänder sind sehr breit: ungefähr 20 %. Dies zeigt, dass die Kovariablen Flexibilität allein eine einzelne Arbeitslosenquote unpräzis prognostiziert. Ein Modell, dass andere Kovariablen wie Bevölkerungsstruktur, Schulbildung, Zinspolitik der Notenbank, Zusammensetzung des Arbeitsmarkts, Verhältnis des Dienstleistungssektors zum Industriesektor, etc. einbezieht, scheint daher sinnvoller zu sein. Informationen, wie man Modelle mit vielen Kovariablen aufstellen und effizient handhaben kann, findet man bei A. Raftery in [10].  $\square$

## 13.4 Prior, Likelihood, Posterior

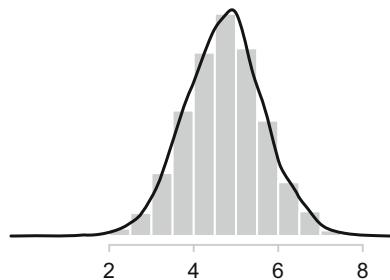
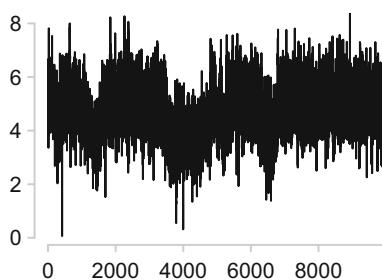
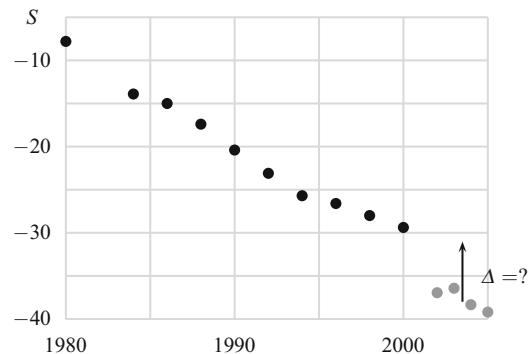
In den vorigen Abschnitten hat man neben den Daten nur minimale Vorinformation. Das folgende Beispiel zeigt, wie verfügbare Vorinformation benutzt werden kann, um die Plausibilität zu Parametern mit der Regel von Bayes besser zu aktualisieren.

**Beispiel 13.9 (Vermessung Schüttdamm)** Ab dem Jahr 2002 wurde die Setzung des Schüttdamm Châtelard-Barberine aus Beispiel 13.6 mit einem neuen Verfahren gemessen. Das Verfahren hat etwa die gleiche Präzision wie das alte Verfahren, ist aber günstiger und schneller. Abb. 13.28 zeigt die vier neuen Messungen in den Jahren 2002 bis 2005. Sie lauten:

Jahr	2002	2003	2004	2005
Setzung $S$	-36,95	-36,43	-38,34	-39,20

Die Messwerte sind deutlich tiefer als diejenigen mit dem alten Verfahren. Dies ist der Fall, weil das neue Messsystem noch nicht kalibriert ist. Die Werte des neuen Messsystems müssen daher um den Betrag  $\Delta$  nach oben versetzt oder kalibriert werden. Um  $\Delta$  zu bestimmen, hat man das schon in Beispiel 13.6 benutzte Regressionsmodell für die

**Abb. 13.28** Streudiagramm der Messwerte mit vier neuen Messungen (graue Punkte)



**Abb. 13.29** MCMC-Kette für den Posterior der Kalibrierungskonstante  $\Delta$  mit Akzeptanzrate 0,20

Setzung:

$$\mu_{S,\text{neu}}(t) = a \cdot \exp\{-b \cdot (t - 1980)\} + c - \Delta$$

Die Parameter  $a$ ,  $b$  und  $c$  sagen, wie sich der Schüttdamm setzt. Sie sind physikalische Parameter und unabhängig vom Messsystem. Die Kalibrierungskonstante  $\Delta$  kann mit der Regel von Bayes berechnet werden. Mit dem Modell der Normalverteilung mit Streuung  $\sigma$  (für das alte und neue Messsystem) ist nach Theorem 13.1

$$\text{pdf}(\Delta, a, b, c, \sigma \mid \text{Daten, Vorinfo.}) \propto \underbrace{\frac{1}{\sigma^4} \cdot \exp\{-0,5\chi^2\}}_{\text{Likelihood}} \cdot \underbrace{\text{pdf}(\Delta, a, b, c, \sigma \mid \text{Vorinfo.})}_{\text{Prior}}$$

mit

$$\chi^2 = \frac{[-36,95 - \mu_{S,\text{neu}}(2002)]^2 + \dots + [-39,20 - \mu_{S,\text{neu}}(2005)]^2}{\sigma^2}$$

In der Likelihood-Funktion steckt die Information aus den vier neuen Messungen. Als Prior dienen (1) der Posterior zu  $a$ ,  $b$ ,  $c$  und  $\sigma$  aus den alten Messungen und (2) minimale Vorinformation zum Lageparameter  $\Delta$ :

$$\text{Prior} = \underbrace{\text{pdf}(a, b, c, \sigma \mid \text{Vorinformation})}_{\text{A posteriori-Verteilung aus Beispiel 13.6}} \cdot \underbrace{1}_{\text{min. Vorinformation zu } \Delta}$$

Mit einer MCMC-Kette lässt sich daraus der Posterior für die fünf Parameter des Modells bestimmen. Bei Statistikprogrammen gibt man dazu die Messwerte ein und nennt das Daten- und das Regressionsmodell sowie die A priori-Verteilungen der Parameter:

- |                               |   |
|-------------------------------|---|
| Datenmodell alte Daten:       | $i$ -ter Messerwert $S_{\text{alt}} \sim \text{Normal}(\mu_{\text{alt},i}, \sigma)$ |
| Regressionsmodell alte Daten: | $\mu_{\text{alt},i} = a \cdot \exp\{-b \cdot (t_i - 1980)\} + c$                    |
| Datenmodell neue Daten:       | $i$ -ter Messerwert $S_{\text{neu}} \sim \text{Normal}(\mu_{\text{neu},i}, \sigma)$ |
| Regressionsmodell neue Daten: | $\mu_{\text{neu},i} = a \cdot \exp\{-b \cdot (t_i - 1980)\} + c - \Delta$           |
| Prior:                        | $a \sim \text{Uniform}(0; 100)$   |
| Prior:                        | $b \sim \text{Uniform}(0; 10)$  |
| Prior:                        | $c \sim \text{Uniform}(-200; 200)$  |
| Prior:                        | $\Delta \sim \text{Uniform}(-50; 50)$   |
| Prior:                        | $\ln \sigma \sim \text{Uniform}(\ln(10^{-3}); \ln(10^3))$                           |

Damit erhält man die A posteriori-Randverteilung von  $\Delta$  (Abb. 13.29). Der plausibelste Wert  $\Delta_0$  von  $\Delta$  ist

$$\Delta_0 = 5,20 \text{ mm}$$

Mit einer Wahrscheinlichkeit von 0,5 liegt  $\Delta$  zwischen 4,73 mm und 5,80 mm. Es besteht eine Wahrscheinlichkeit von 0,95, dass  $\Delta$  zwischen 3,75 mm und 7,16 mm ist.  $\square$

## 13.5 Verallgemeinerte lineare Modelle

Die in diesem Kapitel bisher untersuchten Beispiele handeln von kontinuierlichen Größen, deren durchschnittliche Werte mit einem Regressionsmodell in Funktion von verschiedenen Faktoren und Kovariablen beschrieben werden. Dabei spielt das Datenmodell der Normalverteilung eine wichtige Rolle. Oft nehmen Größen, die von verschiedenen Kovariablen abhängen, nur Werte Null und Eins an. Dies ist der Fall, wenn man misst, ob ein Medikament wirkt: Null bedeutet Medikament wirkt nicht, Eins heißtt Medikament wirkt. Interessant sind auch Größen, wie die Anzahl Unfälle während eines Monats oder die jährlichen Schadensfälle bei einer Versicherung. Dies sind diskrete positive Größen, die nur Werte  $0, 1, 2, \dots$  annehmen können. In den erwähnten Fällen ist das Datenmodell der Normalverteilung wenig sinnvoll. Man arbeitet daher mit anderen Verteilungen. Die folgenden Beispiele zeigen skizzenhaft, wie dies funktioniert.

**Beispiel 13.10 (Gaschromatografie)** In der Gaschromatografie spielen Eichkurven eine besondere Rolle, um Konzentrationen von BHA (Butylhydroxyanisol) im Abwasser nachzuweisen. Tab. 13.5 zeigt die Anzahl Ionen in Funktion der vorgegebenen BHA-Konzentration, die ein Massenspektrometer gezählt hat. Eine Chemikerin möchte aus der BHA-Konzentration  $K$  mit einem Regressionsmodell auf die Anzahl Ionen  $AI$  rechnen. Dies nennt man eine Eichkurve bestimmen. Die Anzahl Ionen ist eine *diskrete* und positive Größe, die  $0, 1, 2, \dots$  sein kann. Als Wissen kann die Chemikerin annehmen, dass die durchschnittliche Ionenzahl  $\mu_{AI}(K)$  endlich ist. Mit dem Prinzip von MaxEnt folgt daher nach Theorem 9.4: Die Anzahl Ionen können mit einer Poissonverteilung beschrieben werden:

$$\text{Datenmodell: } i\text{-ter Messwert } AI \sim \text{Poisson}(\lambda) \quad (13.2)$$

Der Parameter  $\lambda$  der Poissonverteilung ist der Erwartungswert, also  $\mu_{AI}(K)$ .<sup>5</sup> Er muss positiv sein. Ein einfaches Regressionsmodell der Form

$$\mu_{AI}(K) = a + b \cdot K$$

ist ungeeignet, da es negative Werte annehmen kann. Ein besseres Regressionsmodell ist:

$$\mu_{AI}(K) = \exp(a + b \cdot K)$$

Dies garantiert, dass  $\mu_{AI}(K) > 0$  ist, egal wie die Parameter  $a$  und  $b$  lauten.<sup>6</sup> Man nennt dieses Regressionsmodell ein lineares *Poisson-Regressionsmodell*. Dies ist ein Spezialfall eines *verallgemeinerten linearen Modells*.

---

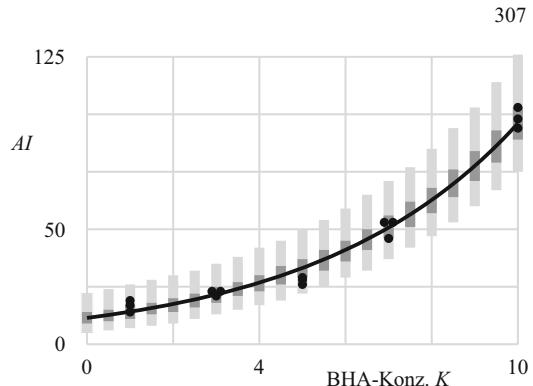
<sup>5</sup> Die Standardabweichung der Poissonverteilung ist  $\sqrt{\lambda}$ . Daher hängt bei diesem Regressionsmodell die Streuung der Zielgröße von den Kovariablen im Regressionsmodell ab!

<sup>6</sup> Anders ausgedrückt hängt der Logarithmus des Parameters linear von der erklärenden Variable ab:  $\ln\{\mu_{AI}(K)\} = a + b \cdot K$ .

**Tab. 13.5** BHA-Konzentrationen (in  $10^{-4}$  mg/L) und Anzahl Ionen (aus [2])

Nr. Messung	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
BHA-Konz.	1,0	3,0	5,0	7,0	10,0	1,0	3,0	5,0	7,0	10,0	1,0	3,0	5,0	7,0	10,0
Anzahl Ionen	17	21	26	53	98	19	23	28	46	94	14	23	29	53	103

**Abb. 13.30** Poisson-Regressionsmodell für die Anzahl Ionen in Funktion der BHA-Konzentration und Prognosebänder für weitere Messungen zum Niveau 0,5 und 0,95 (Die Messwerte sind teilweise leicht versetzt, da mehrere Datenpaare gleich sind.)



Die plausibelsten Werte von  $a$  und  $b$ , wie auch Wahrscheinlichkeitsintervalle zu  $a$  und  $b$ , kann man mit der Regel von Bayes berechnen. Dazu nennt man mit einem Statistikprogramm das Daten- und das Regressionsmodell sowie die A priori-Verteilungen der Parameter:

$$\text{Datenmodell: } i\text{-ter Messwert } AI \sim \text{Poisson}(\mu_{AI,i})$$

$$\text{Regressionsmodell: } \mu_{AI,i} = \exp(a + b \cdot K_i)$$

$$\text{Prior: } a \sim \text{Uniform}(-100; 100)$$

$$\text{Prior: } b \sim \text{Uniform}(-100; 100)$$

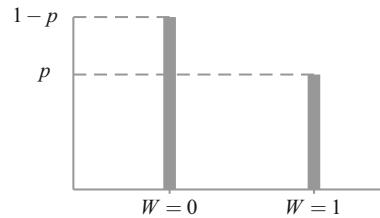
Mit einer MCMC-Simulation können daraus die A posteriori-Verteilungen von  $a$  und  $b$  berechnet werden. Die plausibelsten Werte  $a_0$  und  $b_0$  für die Parameter des Regressionsmodells lauten

$$a_0 = 2,437, \quad b_0 = 0,213$$

Sie werden nicht mit der Methode der kleinsten Quadrate berechnet. Der Grund liegt darin, dass das Datenmodell eine Poissonverteilung und nicht eine Normalverteilung ist. Mit einer Wahrscheinlichkeit von 0,95 liegt  $a$  zwischen 2,23 und 2,63. Es besteht eine Wahrscheinlichkeit von 0,95, dass  $b$  zwischen 0,19 und 0,24 liegt.

Auch Prognosebereiche für zukünftige Messungen lassen sich mit dem in diesem Kapitel vorgestellten Verfahren bestimmen. Abb. 13.30 visualisiert das Resultat. Die Prognosebänder zeigen, dass das Regressions- und das Datenmodell gut sind. Mit der Autokorrelationsfunktion der approximativen Residuen lässt sich weiter aufdecken, dass die Likelihood gut gerechnet ist: die Datenwerte sind unabhängig.

**Abb. 13.31** Datenmodell für die Wirkung  $W$  des Medikaments



Die Prognosebänder werden mit zunehmender BHA-Konzentration breiter. Dies liegt daran, dass das Poissonmodell mit Parameter  $\lambda = \mu_{AI}(K)$  eine Streuung von  $\sqrt{\lambda}$  hat. Je grösser  $\lambda$ , um so grösser ist die Streuung.  $\square$

**Beispiel 13.11 (Medikament)** Ein Arzt beschreibt, ob ein Medikament wirkt, mit der Grösse  $W$ . Dabei bedeutet  $W = 1$ , dass das Medikament wirkt.  $W = 0$  heisst, dass es wirkungslos ist. Die Wirkung  $W$  des Medikaments hänge von vielen Kovariablen  $X, Y, \dots$  ab. Mit einem Regressionsmodell kann der Arzt versuchen, die durchschnittliche Wirkung  $\mu_W(X, Y, \dots)$  in Funktion der Kovariablen zu beschreiben. Das Datenmodell für  $W$  ist einfach, da  $W$  nur Null oder Eins sein kann:

$$\mathbb{P}(W = 1 | \mu_W(X, Y, \dots)) = p, \quad \mathbb{P}(W = 0 | \mu_W(X, Y, \dots)) = 1 - p$$

Dies ist das Bernoulli-Modell. Der Graph dieses Datenmodells findet sich in Abb. 13.31. Der durchschnittlich erwartbare Wert  $\mu_W(X, Y, \dots)$  des Modells ist schnell berechnet:

$$\mu_W(X, Y, \dots) = 1 \cdot \mathbb{P}(W = 1) + 0 \cdot \mathbb{P}(W = 0) = 1 \cdot p + 0 \cdot (1 - p) = p$$

Die durchschnittliche Wirkung des Medikaments ist also gleich der Wahrscheinlichkeit  $p$ , dass  $W = 1$  ist. Da  $p$  nur Werte zwischen Null und Eins hat, ist das lineare Regressionsmodell

$$\mu_W(X, Y, \dots) = p = a + b \cdot X + c \cdot Y + \dots$$

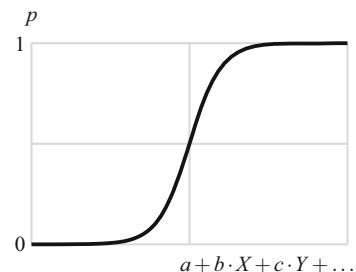
wenig sinnvoll. Man betrachtet deshalb die Chance  $\mathbb{O}$ , dass  $W = 1$  ist:  $\mathbb{O} = p/(1 - p)$ . Der Logarithmus dieser Chance kann beliebige kontinuierliche Werte annehmen. Daher wird das Regressionsmodell

$$\ln \mathbb{O} = a + b \cdot X + c \cdot Y + \dots$$

benutzt. Man nennt dies eine *logistische Regression*. Umgerechnet auf die durchschnittliche Wirkung des Medikaments, bedeutet es:

$$\mu_W(X, Y, \dots) = p = \frac{1}{1 + \exp(-(a + b \cdot X + c \cdot Y + \dots))}$$

**Abb. 13.32** Sigmoidfunktion bei der logistischen Regression



Dies ist eine Sigmoidfunktion. Ihr Graph findet sich in Abb. 13.32. Mathematisch schreibt man das Daten- und Regressionsmodell oft in kurzer Form:

$$\begin{aligned} \text{Datenmodell: } & i\text{-ter Messwert } W \sim \text{BernoulliLogit}(\mu_{W,i}) \\ \text{Regressionsmodell: } & \mu_{W,i} = a + b \cdot X_i + c \cdot Y_i + \dots \end{aligned}$$

Mit der Regel von Bayes lassen sich daraus die A posteriori-Verteilung der Regressionsparameter  $a, b, \dots$  bestimmen.  $\square$

## Reflexion

**13.1** Beim Beispiel 12.5 der Akkommodationsbreite, finden sich in Tab. 12.1 Akkommodationsbreiten des Stereo-Sehens und Alter von 19 Personen. Die durchschnittliche Akkommodationsbreite  $\mu_{\text{Stereo}}$  des Stereo-Sehens soll mit dem Regressionsmodell

$$\mu_{\text{Stereo}}(\text{Alter}) = a + b \cdot \text{Alter}$$

beschrieben werden. Als Datenmodell gilt: Hypothetische Messungen der Akkommodationsbreite streuen normalverteilt um  $\mu_{\text{Stereo}}$  mit konstanter, nicht vom Alter abhängiger Streuung. Weitere Information liegt nicht vor.

- Wie lauten die „wahrscheinlichsten“ Werte für die Parameter  $a$  und  $b$  des Regressionsmodells? Zeichnen Sie das dazugehörige Regressionsmodell in das Streudiagramm der Messwerte.
- Welche durchschnittlichen Akkommodationsbreiten prognostiziert das Regressionsmodell für die Alter 30, 40 und 100 Jahre? Welchen Werten trauen Sie?
- Bestimmen Sie Wahrscheinlichkeitsintervalle zum Niveau 0,5 und 0,95 für die beiden Parameter  $a$  und  $b$ .
- Bestimmen Sie ein Wahrscheinlichkeitsintervall zum Niveau 0,95 für die Standardabweichung  $\sigma$  des Datenmodells.

- (e) Beurteilen Sie die Qualität des Regressions- und des Datenmodells: Zeichnen Sie dazu die Prognosebänder zum Niveau 0,5 und 0,95 für weitere Messungen der Akkommodationsbreite ins Streudiagramm der Messwerte.
- (f) Wurde die Likelihood der statistischen Rechnung gut gerechnet? Überprüfen Sie dazu, ob die Messwerte als unabhängig betrachtet werden können. Arbeiten Sie mit den approximativen Residuen. (Die Reihenfolge der Messungen finden Sie in Tab. 13.3).

**13.2** Bei einem Terrassenhaus in der Westschweiz hängt der Gasverbrauch  $G$  für die Heizung hauptsächlich von der durchschnittlichen Monatsaussentemperatur  $T$  der Wintermonate ab. Einerseits möchte man mit einem Regressionsmodell von  $T$  auf den durchschnittlichen Verbrauch  $\mu_G(T)$  rechnen. Andererseits will man bei gegebener Durchschnittstemperatur  $T$  den Gasverbrauch  $G$  prognostizieren. Um dies zu bestimmen, hat man als Information die Daten in Tab. 12.3.

- (a) Visualisieren Sie die Daten in einem Streudiagramm. Wählen Sie ein sinnvolles Regressionsmodell, das den durchschnittlichen Verbrauch  $\mu_G(T)$  in Funktion der monatlichen Durchschnittstemperatur  $T$  ausdrückt.
- (b) Als Datenmodell gilt: Messungen des Gasverbrauchs  $G$  streuen normalverteilt um  $\mu_G(T)$  mit konstanter nicht von der Kovariablen  $T$  abhängiger Streuung  $\sigma$ . Wie lauten die „wahrscheinlichsten“ Werte für die Parameter des Regressionsmodells? Zeichnen Sie das zu den plausibelsten Werten der Parameter gehörende Regressionsmodell in das Streudiagramm der Messwerte.
- (c) Bestimmen Sie Wahrscheinlichkeitsintervalle zum Niveau 0,5 und 0,95 für die Parameter des Regressionsmodells.
- (d) Wie lautet ein Wahrscheinlichkeitsintervall zum Niveau 0,95 für die Standardabweichung  $\sigma$  des Datenmodells?
- (e) Beurteilen Sie die Qualität des Regressions- und Datenmodells: Berechnen Sie dazu Prognosebänder zum Niveau 0,5 und 0,95 für zukünftige Messungen des Verbrauchs  $G$  und zeichnen Sie diese ins Streudiagramm der Messwerte.
- (f) Wurde die Likelihood der statistischen Rechnung in (b) gut gerechnet? Überprüfen Sie dazu, ob die Messwerte als unabhängig betrachtet werden können. Arbeiten Sie dazu mit den approximativen Residuen.

**13.3** Ultrakurze Laserpulse mit einer Dauer von wenigen Femtosekunden können benutzt werden, um kleine, gezielte Muster in Halbleiter zu bohren. Die Fläche  $F$  der Bohrlöcher hängt stark von der Leistung  $P$  des Laserimpulses ab. Physikalische Überlegungen deuten darauf hin, dass die durchschnittliche Fläche  $\mu_F$  der Bohrlöcher mit dem linearen Regressionsmodell

$$\mu_F(P) = a + b \cdot \ln P$$

beschrieben werden kann. Um die Parameter  $a$  und  $b$  zu berechnen, wurden in [11] elf Messungen durchgeführt, die in Tab. 13.6 dargestellt sind. Das Datenmodell sei: Hypo-

**Tab. 13.6** Gemessene Flächen  $F$  der Bohrlöcher in Funktion der Leistung  $P$  des Laserimpulses (aus [11])

Nr. Messung	1	2	3	4	5	6	7	8	9	10	11
$P$ (in $\text{J}/\text{cm}^2$ )	6	8	10	12	14	16	18	20	22	24	25
$F$ (in $\mu\text{m}^2$ )	95	138	158	196	203	220	236	257	265	276	288

thetische Messungen der Flächen streuen um  $\mu_F$  mit konstanter nicht von der Leistung  $P$  abhängiger Streuung  $\sigma$ .

- (a) Zeichnen Sie das Streudiagramm der Daten.
- (b) Wie lauten die „wahrscheinlichsten“ Werte für die Parameter  $a$  und  $b$  des Regressionsmodells? Zeichnen Sie das Regressionsmodell mit den plausibelsten Werten von  $a$  und  $b$  in das Streudiagramm der Messwerte.
- (c) Bestimmen Sie Wahrscheinlichkeitsintervalle zum Niveau 0,5 und 0,95 für die beiden Parameter.
- (d) Beurteilen Sie die Qualität des Regressions- und Datenmodells: Berechnen Sie dazu Prognosebänder zum Niveau 0,5 und 0,95 für zukünftige Messungen der Fläche und zeichnen Sie diese ins Streudiagramm der Messwerte.
- (e) Von Interesse ist die Leistung  $P_0$  des Laserimpulses, bei der die Oberfläche des Halbleiters durchschnittlich nicht angebohrt wird:  $\mu_F = 0$ . Wie lautet  $P_0$  in Funktion von  $a$  und  $b$ ? Geben Sie die Plausibilität zu  $P_0$  mit einem Wahrscheinlichkeitsmodell an. Was ist der plausibelste Wert für  $P_0$ ?

**13.4** In Beispiel 3.17 ist ein Projekt vorgestellt, bei dem versucht wird, vom Fällzeitpunkt auf die Relativdichte von Holz zu rechnen. Ein Faktor, der die Relativdichte beeinflusst, ist die Jahreszeit. Mit einem Regressionsmodell soll die durchschnittliche Relativdichte  $\mu_{\text{Relativdichte}}(W)$  in Funktion der Jahreswoche  $W$  berechnet werden. Ein Regressionsmodell mit Periode von 52 Wochen ist

$$\mu_{\text{Relativdichte}}(W) = a + b \cdot \sin\left(\frac{2\pi}{52} \cdot W\right) + c \cdot \cos\left(\frac{2\pi}{52} \cdot W\right)$$

mit unbekannten Parametern  $a$ ,  $b$  und  $c$ . Um diese Parameter zu bestimmen, sollen die Daten in Tab. 13.7 benutzt werden. Das Datenmodell sei: Hypothetische Messungen der Relativdichte streuen um  $\mu$  mit konstanter nicht von der Woche  $W$  abhängiger Streuung.

- (a) Ist das Regressionsmodell linear (in den Parametern)?
- (b) Stellen Sie die Messpunkte in einem Streudiagramm dar.
- (c) Die Methode der kleinsten Quadrate bestimmt gemäß den obigen Voraussetzungen die plausibelsten Werte für die Parameter  $a$ ,  $b$  und  $c$  des Regressionsmodells. Bestimmen Sie diese mit einem Statistikprogramm. Berechnen Sie mit dem Programm zusätzlich Wahrscheinlichkeitsintervalle für diese Parameter.

**Tab. 13.7** Gemessene Relativdichten von Holz während 46 Wochen

Woche	1	2	3	4	5	6	7	8	9	10
Dichte in %	42,162	42,085	41,974	41,831	41,656	41,454	41,227	40,977	40,710	40,428
Woche	11	12	13	14	15	16	17	18	19	20
Dichte in %	40,136	39,838	39,539	39,242	38,953	38,674	38,412	38,168	37,947	37,752
Woche	21	22	23	24	25	26	27	28	29	30
Dichte in %	37,586	37,452	37,350	37,283	37,252	37,257	37,298	37,375	37,486	37,629
Woche	31	32	33	34	35	36	37	38	39	40
Dichte in %	37,803	38,006	38,233	38,482	38,750	39,032	39,324	39,621	39,921	40,218
Woche	41	42	43	44	45	46				
Dichte in %	40,507	40,785	41,048	41,292	41,513	41,707				

- (d) Beurteilen Sie die Qualität des Regressions- und Datenmodells: Berechnen Sie dazu Prognosebänder zum Niveau 0,5 und 0,95 für zukünftige Messungen der Relativdichten und zeichnen Sie diese ins Streudiagramm der Messwerte.
- (e) Zieht man von den Relativdichten jeweils die geschätzten Regressionsmodellwerte ab, so wird der Faktor ‚Jahreswoche‘ aus den Relativdichten entfernt. Sind die so *trendbereinigten* Relativdichten unter statistischer Kontrolle? Zeichnen Sie dazu eine Kontrollkarte mit geschätzten oberen und unteren Kontrollgrenzen. Überprüfen Sie dazu, ob die Messwerte als unabhängig betrachtet werden können.
- (f) Bestimmen Sie ein Wahrscheinlichkeitsintervall zum Niveau 0,95 für die Standardabweichung  $\sigma$  des Datenmodells.
- (g) Versuchen Sie, auch ein Regressionsmodell für die Relativdichten zu konstruieren, das mit dem Verfahren von Kap. 12 arbeitet. Benutzen Sie dazu gewichtete arithmetische Mittel und wählen Sie passende Bandbreiten.

**13.5** Um Glatteis auf Verkehrswegen zu verhindern, können Straßen gesalzen werden. Damit kann man die Gefriertemperatur bis auf  $-21^{\circ}\text{C}$  senken. Dies bedeutet, dass Wasser auf der Straße erst bei einer Temperatur unterhalb von  $-21^{\circ}\text{C}$  gefriert. Das auf der Straße liegende Salz wird abgetragen, wenn das Verkehrsaufkommen hoch ist. In Versuchen auf Straßen von Vimmerby und Klockrike wurde untersucht, wie stark das Salz in Funktion des Verkehrsaufkommens  $V$  abgetragen wird. Die Daten gemäß [5] für Klockrike finden sich in Tab. 13.8. Ein Regressionsmodell, das die durchschnittliche Restsalzmenge  $\mu_{\text{RS}}$  in

**Tab. 13.8** Restsalz auf der Straße in Funktion des Verkehrsaufkommens (aus [5])

$V$ (in 1000 Wagen/h)	0,5	1,8	2,4	2,8	3,4	4,2	5,8	6,0	6,4	7,4	8,9
Restsalz auf Straßen (in g/m <sup>2</sup> )	9,5	10,4	9,3	7,8	5,7	5,5	4,4	3,3	2,3	3,7	2,0

Funktion des Verkehrsaufkommens  $V$  beschreibt, ist

$$\mu_{\text{RS}}(V) = S \cdot e^{-k \cdot V}$$

In diesem physikalischen Modell sind  $S$  und  $k$  unbekannte Parameter. Als Modellannahme gilt: Hypothetische Messungen der Restsalzmenge streuen normalverteilt um  $\mu_{\text{RS}}$  mit konstanter nicht von  $V$  abhängiger Streuung  $\sigma$ .

- (a) Stellen Sie die Messwerte in einem Streudiagramm dar.
- (b) Ist das Regressionsmodell linear in den Parametern?
- (c) Bestimmen Sie die plausibelsten Werte für die Parameter  $S$  und  $k$ . Wie lauten Wahrscheinlichkeitsintervalle für die beiden Parameter? Wie stark sind die beiden Parameter korreliert (beurteilen Sie dies auch visuell).
- (d) Beurteilen Sie die Qualität des Regressions- und Datenmodells: Berechnen Sie dazu Prognosebänder zum Niveau 0,5 und 0,95 für zukünftige Messungen des Restsalzes und zeichnen Sie diese ins Streudiagramm der Messwerte.
- (e) Welche durchschnittliche Restsalzmenge prognostiziert das Regressionsmodell für Verkehrsaufkommen von 3000 Wagen/h und von 6000 Wagen/h?
- (f) Wenn auf der Strasse durchschnittlich weniger als  $4 \text{ g/m}^2$  Salz liegt, muss sie neu gesalzen werden. Ab welchen Verkehrsaufkommen ist dies der Fall?

**13.6** Gletscher, die an steilen Berghängen liegen, können abbrechen und Eislawinen produzieren, die eine Bedrohung für Bergsiedlungen bilden. Während des Sommers 1999 begann ein Gletscher oberhalb von Grindelwald stark zu fliessen. Ein Abbruch des Gletschers galt für die Wissenschaftler an der ETH Zürich als sicher. Um den Zeitpunkt des Abbruchs zu prognostizieren, wurde der Gletscher mit Stäben bestückt. Damit wurde die Verschiebung  $s$  der Gletschermasse gemessen (siehe [8]). Tab. 13.9 zeigt die Daten. Physikalische Modelle zu fliessenden Gletschern legen nahe, dass die durchschnittlich erwartbare Verschiebung  $\mu_s$  eines Gletschers, bei dem ein Abbruch droht, durch das Modell

$$\mu_s(t) = v_0 \cdot t - a_0 \left( \frac{(t_\infty - t)^{0.5} - t_\infty^{0.5}}{0.5} \right)$$

in Funktion der Zeit  $t$  beschrieben werden kann. Dabei sind  $v_0$ ,  $a_0$  und  $t_\infty$  unbekannte Parameter. Der Parameter  $t_\infty$  entspricht dem Zeitpunkt, indem der Gletscher abbricht und eine Eislawine auslöst. Als Datenmodell haben Sie: Messungen der Verschiebung  $s$  streuen um  $\mu_s$  mit konstanter nicht von der Zeit  $t$  abhängiger Streuung. Weitere Information ist nicht vorhanden.

**Tab. 13.9** Zeitliche Entwicklung der Verschiebung  $s$  eines Gletscherstücks oberhalb von Grindelwald (aus [8])

Tag	0	2	7	12	13	15	16	17	19	20	22
Versch. $s$ (in m)	0,000	0,529	1,936	3,598	3,977	4,802	5,219	5,667	6,669	7,208	8,464



**Abb. 13.33** Drei Streudiagramme

- Stellen Sie die Messpunkte in einem Streudiagramm dar.
- Ist das Regressionsmodell linear in den Parametern?
- Bestimmen Sie die plausibelsten Werte für die Parameter  $v_0$ ,  $a_0$  und  $t_\infty$  mit dem Modus der A posteriori-Verteilung. Wie lauten Wahrscheinlichkeitsintervalle für die Parameter?
- Kontrollieren Sie die Qualität des Regressions- und Datenmodells: Berechnen Sie dazu Prognosebänder zum Niveau 0,5 und 0,95 für Messungen der Verschiebung und zeichnen Sie diese ins Streudiagramm der Messwerte.
- Wurde die Likelihood der statistischen Rechnung gut gerechnet? Überprüfen Sie dazu, ob die Messwerte als unabhängig betrachtet werden können. Arbeiten Sie mit den approximativen Residuen.
- Bestimmen Sie Wahrscheinlichkeitsintervalle zum Niveau 0,5 und 0,95 für den Absturzzeitpunkt  $t_\infty$ .
- Der Absturz des Gletschers erfolgt nach  $t_\infty = 26,8$  Tagen. Vergleichen Sie dies mit dem in (c) bestimmten, plausibelsten Wert und den in (f) berechneten Wahrscheinlichkeitsintervallen. Sind Sie mit dem Resultat zufrieden?
- Bestimmen Sie Wahrscheinlichkeitsintervalle für die restlichen Parameter  $v_0$ ,  $a_0$  und die Streuung  $\sigma$ .

**13.7** Abb. 13.33 zeigt die Streudiagramme zu Werten einer Grösse  $Y$ , die vom erklärenden Faktor  $X$  abhängt bei drei verschiedenen Versuchen. Man vermutet, dass der durchschnittliche Wert von  $Y$  durch das Regressionsmodell  $\mu_Y(X = x) = a + b \cdot x$  berechnet werden kann. Dabei sind  $Y$  kontinuierliche Werte, die um  $\mu_Y(x)$  normalverteilt streuen.

- Zeichnen Sie von Hand die Regressionsgerade, ermittelt mit der Methode der kleinsten Quadrate, in die Streudiagramme ein.
- Beurteilen Sie grob, ob aussergewöhnliche Messwerte vorhanden sind.
- Welche Versuche sind schlecht geplant?

**13.8** In einer technischen Untersuchung wurde untersucht, wie Milch in Rohren verschleppt wird. Dazu wird Farbmilch durch ein Rohr gepumpt. Danach wird das Rohr mit Druckluft gereinigt und eine gewisse Zeit so belassen. Anschliessend wird das Rohr mit Milch gespült. Die Anreicherung der Milch mit den Farbmilchrückständen erlaubt es zu

**Tab. 13.10** Kalibrationswerte in Funktion der Konzentration (in  $10^{-3}$  mL/mL, aus [4])

Versuchsnummer	1	8	4	7	2	6	5	3	9
Konzentration	0,2	0,4	0,8	1,6	3,2	6,4	12,8	25,6	51,2
Kalibration Spektrometer	0,41	0,11	0,5	0,63	0,83	0,98	1,76	3,39	4,63

modellieren, wie die Milch im Rohr verbreitet wird. Tab. 13.10 zeigt das Resultat von neun Messungen. Plausible Modelle, die den durchschnittlichen Wert  $\mu_{\ln K}$  des Logarithmus der Kalibration in Funktion des Logarithmus  $\ln(\text{Konz})$  der Konzentration approximieren, sind

$$\text{Modell 1: } \mu_{\ln K}(\text{Konz}) = a + b \cdot \ln(\text{Konz})$$

$$\text{Modell 2: } \mu_{\ln K}(\text{Konz}) = \alpha + \beta \cdot \ln(\text{Konz}) + \gamma \cdot [\ln(\text{Konz})]^2$$

Das Datenmodell sein: Der Logarithmus von Messungen der Kalibration variiert um  $\mu_{\ln K}$  mit konstanter nicht von  $\ln(\text{Konz})$  abhängiger Streuung. Vor der Datensammlung bestehe nur minimale Information zu den Regressionsparametern  $a$  und  $b$ ,  $\alpha$ ,  $\beta$  und  $\gamma$ , wie auch zur Streuung  $\sigma$ .

- (a) Zeichnen Sie die Messwertpaare ( $\ln(\text{Konz}) / \ln(K)$ ) in einem Streudiagramm auf.
- (b) Schätzen Sie die Parameter  $a$  und  $b$  mit den plausibelsten Werten, wenn Sie mit dem Modell 1 arbeiten. Eignet sich das Modell? Berechnen Sie dazu Prognosebereiche für zukünftige Messwerte. Ist die Likelihood gut gerechnet?
- (c) Schätzen Sie die Parameter  $\alpha$ ,  $\beta$  und  $\gamma$  mit den plausibelsten Werten, wenn Sie mit dem Modell 2 arbeiten. Eignet sich das Modell? Ist die Likelihood gut gerechnet?
- (d) Bestimmen Sie Wahrscheinlichkeitsintervalle zum Niveau 0,5 und 0,95 für die Regressionsparameter der beiden Modelle.

---

## Literatur

1. H. A. Anderson, G. Hentz, A. Glasser, K. K. Stuebing, R. E. Manny, Minus-Lens-Stimulated Accommodative Amplitude Decreases Sigmoidally with Age: A Study of Objectively Measured Accommodative Amplitudes from Age 3, *Inv. Ophthalmology & Visual Science*, **49**(7) (2008)
2. M. Andrjuchowa, Bestimmen der Konzentration von BHA, BBP, DBP im Abwasser, Diplomarbeit, Fachhochschule beider Basel, Muttenz (1996)
3. D. Bättig, Statistische Modellierung der vertikalen Setzung des Schüttdamms Châtelard-Barbienne, Institut für Risiko- und Extremwertanalyse, Berner Fachhochschule, Bericht zuhanden Schweizerische Bundesbahnen (2005)
4. Ph. Bernhard, Prüfprozedur der Probenahme auf Milchsammelwagen, Bachelorarbeit, Berner Fachhochschule, Burgdorf (2008)
5. G. Blomqvist, M. Gustafsson, Patterns of Residual Salt on Road Surface, Case Study, vti, Note **33A** (2005)
6. R. Christensen, W. Johnson, A. Branscum, T. E. Hanson, *Bayesian Ideas and Data Analysis, An Introduction for Scientists and Statisticians* (CRC Press, Chapman & Hall Book, 2010)

7. A. Duane, Studies in Monocular and Binocular Accommodation, with Their Clinical Application, *Trans. Am. Ophthalmol. Soc.* **20**, 132–157 (1922)
8. M. Funk, H.-E. Minor, Eislawinen in den Alpen: Erfahrungen mit Schutzmassnahmen und Früherkennungsmethoden, *Wasserwirtschaft* **91**, 362–368 (2001)
9. H. J. Lang, J. Huder, P. Amann, *Bodenmechanik und Grundbau: Das Verhalten von Boden und Fels und die wichtigsten grundbaulichen Konzepte* (Springer Verlag, 7. Auflage, 2003)
10. A. E. Raftery, Bayesian Model Selection in Social Research. *Sociological Methodology*, **25**, 111–163 (1995)
11. R. Schor, Bohren von Lithium Niobat mit ps-Impulsen, Bachelorarbeit Maschinentechnik, Berliner Fachhochschule (2007)

*Nachdem sie indessen ungefähr eine halbe Stunde lang gelaufen und wieder ganz trocken geworden waren, rief der Brachvogel plötzlich: „Ende des Wettkaufs!“, und alle drängten sich, noch ganz ausser Atem, um ihn und fragten: „Aber wer ist Sieger?“*

*Dies konnte der Brachvogel nicht ohne tieferes Nachdenken beantworten, und so sass er längere Zeit hindurch da und legte den Zeigefinger an die Stirn (eine Haltung, in der ihr gewöhnlich Goethe auf den Titelbildern sitzen seht), während ringsum alles schwieg und wartete. Endlich sagte der Brachvogel: „Alle sind Sieger, und jeder muss einen Preis bekommen.“*

*Lewis Carroll, Alice im Wunderland (Insel Taschenbuch, 1973, S. 30)*

## Zusammenfassung

Nicht direkt messbare Größen werden oft aus verschiedenen Gruppen berechnet. So will man aus einer Stichprobe die durchschnittlichen Raten der Neuerkrankungen an Lungenkrebs in verschiedenen Regionen kennen. Eine Ingenieurin will die Haftkraft von Klebeetiketten (eine Größe, die wegen Messunsicherheiten und variierender Kovariablen nicht direkt bestimmbar ist) aus drei Produktionsorten oder -arten vergleichen. Es ist verbreitet und praktisch, solche Vergleiche mit der *Laplace-Approximation* und dem *Standardfehler* durchzuführen. Was dies ist und worauf man hierbei achten sollte, wird in diesem Kapitel beschrieben. Oft sind Gruppen ähnlich oder gleich strukturiert. Diese Information lässt sich in die statistischen Modelle einbauen. Damit lassen sich nicht direkt messbare Größen der verschiedenen Gruppen sehr effizient vergleichen. Dies zeigt der letzte Abschnitt in diesem Kapitel.

## 14.1 Die Methode von Laplace

Die Plausibilität und Wahrscheinlichkeitsintervalle zu einer nicht direkt messbaren Größe oder zu einem Parameter können mit der A posteriori-Verteilung beschrieben werden. Physikerinnen und Physiker, Ingenieurinnen und Ingenieure, Ökonominnen und Ökonomen schreiben Wahrscheinlichkeitsintervalle zu einem Parameter  $\theta$  meist in der Form

$$\theta = \theta_0 \pm \delta\theta = \theta_0 \pm \text{SE}(\theta)$$

Ein Beispiel mit Zahlen ist  $\theta = (341,3 \pm 0,2)$  kg. Hier ist  $\theta_0$  der plausibelste Wert von  $\theta$ . Weiter nennt man  $\delta\theta$  oder  $\text{SE}(\theta)$  den *Standardfehler* oder die *Standardunsicherheit* (engl. *error bar* oder *standard error*) der A posteriori-Verteilung von  $\theta$ . Die obige Präzisionsangabe ist ein Wahrscheinlichkeitsintervall für  $\theta$  zum Niveau von *ungefähr* 0,68. Meist basiert diese Präzisionsangabe auf der Methode von Laplace. Wie dies gemacht wird, zeigt der folgende Abschnitt.

Der Posterior pdf( $\theta$  | Daten) des Parameters  $\theta$  besitze einen Modus  $\theta_0$ . Abb. 14.1 illustriert die Situation. Man kann nun messen, wie breit die Verteilung ist. Dies gibt an, wie präzis der Modus den Parameter  $\theta$  schätzt. Dazu versucht man, die Verteilung mit einer Parabel zu approximieren. Die Verteilung kann aber sehr spitz sein. Um solche Spitzen zu glätten, ist es besser, die logarithmierte Dichtefunktion

$$L(\theta | \text{Daten}) = \ln [\text{pdf}(\theta | \text{Daten})]$$

mit einer Parabel zu approximieren. Dieses von Laplace stammende Vorgehen ist also

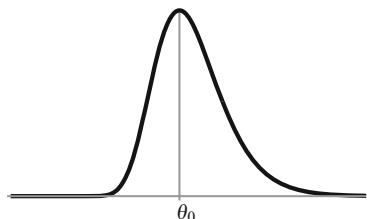
logarithmiere pdf  $\longrightarrow$  approximiere  $L$  mit Parabel  $\longrightarrow$  exponentiere

Das Taylorpolynom vom Grad 2 mit dem Modus  $\theta_0$  als Stützstelle approximiert  $L$  am besten mit einer Parabel:

$$L(\theta) \approx L(\theta_0) + L'(\theta_0) \cdot (\theta - \theta_0) + \frac{1}{2} \cdot L''(\theta_0) \cdot (\theta - \theta_0)^2$$

Die A posteriori-Dichtefunktion hat ihr Maximum beim Modus  $\theta_0$ . Damit hat auch  $L(\theta)$  dort sein Maximum. Liegt der Modus nicht am Rand der Verteilung, so ist deshalb (meist)

**Abb. 14.1** A posteriori-Verteilung der nicht direkt messbaren Größe  $\theta$



$L'(\theta_0) = 0$ . Weiter formt der Graph von  $L$  beim Maximum eine Linkskurve, also ist  $L''(\theta_0) < 0$ . Es folgt:

$$L(\theta) \approx L(\theta_0) + \frac{1}{2} \cdot L''(\theta_0) \cdot (\theta - \theta_0)^2 = L(\theta_0) - \frac{1}{2} \cdot \left( \frac{\theta - \theta_0}{1/\sqrt{-L''(\theta_0)}} \right)^2$$

Exponentiert man die Gleichung, erhält man eine Verteilung, die den Posterior des Parameters  $\theta$  approximiert:

$$\text{pdf}(\theta | \text{Daten}) \approx \text{Approx} \propto \exp \left\{ -0,5 \cdot \left( \frac{\theta - \theta_0}{1/\sqrt{-L''(\theta_0)}} \right)^2 \right\}$$

Dies ist die Dichtefunktion der Normalverteilung mit Modus  $\theta_0$  und Standardabweichung  $1/\sqrt{-L''(\theta_0)}$ . Die hier auftretende Standardabweichung nennt man den *Standardfehler* (engl. *standard error*) der A posteriori-Verteilung des Parameters  $\theta$ . Man bezeichnet weiter die zweite Ableitung

$$I(\theta | \text{Daten}) = -L''(\theta | \text{Daten}) = -\frac{d^2}{d\theta^2} \ln [\text{pdf}(\theta | \text{Daten})]$$

als die *beobachtete Information* (engl. *observed information*). Man hat also:

**Theorem 14.1 (Methode von Laplace – Präzisionsangabe mit dem Standardfehler)**  
*Approximiert man die Plausibilität zu einer nicht direkt messbaren Grösse  $\theta$  mit der Normalverteilung, so schreibt man*

$$\theta = \theta_0 \pm \delta\theta \quad \text{oder} \quad \theta = \theta_0 \pm \text{SE}(\theta)$$

*Hier ist  $\theta_0$  der Modus der A posteriori-Verteilung  $\text{pdf}(\theta | \text{Daten})$  und der Standardfehler  $\text{SE}$  wird mit Hilfe der beobachteten Information berechnet:*

$$\text{SE} = \frac{1}{\sqrt{I(\theta_0)}} \quad \text{mit } I(\theta) = -\frac{d^2}{d\theta^2} \ln [\text{pdf}(\theta | \text{Daten})]$$

Approximiert man die A posteriori-Verteilung von  $\theta$  mit der Normalverteilung, folgt gemäss Kap. 10, dass  $\theta$  mit einer Wahrscheinlichkeit von ungefähr 0,68 zwischen  $\theta_0 - \text{SE}$  und  $\theta_0 + \text{SE}$  liegt. Weiter hat man eine Wahrscheinlichkeit von *etwa* 0,95, dass  $\theta$  zwischen  $\theta_0 - 1,96 \cdot \text{SE}$  und  $\theta_0 + 1,96 \cdot \text{SE}$  ist. Diese Rechnung erlaubt es, schnell Wahrscheinlichkeitsintervalle zu berechnen. Die Rechnung über Quantile und MCMC-Simulationen entfällt.

Als *Faustregel* gilt: Ist (a) die Anzahl  $n$  der Messungen gross, sind (b) die Messwerte unabhängig und liegt (c) der Modus nicht am Rand der A posteriori-Verteilung, so ist die Laplace-Approximation gut.<sup>1</sup> Es gilt dann sogar

$$\text{SE} \propto 1/\sqrt{n}$$

Dies bedeutet, dass mit zunehmender Anzahl Messungen die Wahrscheinlichkeit wächst, dass der Modus den Parameter immer präziser schätzt. Statistikprogramme können die Laplace-Approximation meist schnell berechnen. Eine Warnung ist hier angebracht:

Die Laplace-Approximation kann verwendet werden, um die A posteriori-Verteilung eines Parameters möglichst einfach zu beschreiben. Sie sollte nicht benutzt werden, um ein Prognosemodell für weitere Messwerte oder zukünftige Beobachtungen zu approximieren.

**Beispiel 14.1 (Zeit zwischen zwischen starken Erdbeben)** In Beispiel 9.1 wird die durchschnittliche Zeit  $\mu$  zwischen zukünftigen, starken Erdbeben berechnet. Man hat für die A posteriori-Verteilung des Parameters  $\mu$

$$\text{pdf}(\mu | \text{Daten, min. Vorinformation}) \propto \underbrace{\mathbb{P}(\text{Daten} | \mu)}_{\text{Likelihood}} \cdot \underbrace{\text{pdf}(\mu | \text{min. Vor.})}_{\text{Prior}}$$

Die Likelihood bestimmt sich aus dem Datenmodell. Dieses sagt, wie Zeiten zwischen Erdbeben streuen. Wird dazu die Exponentialverteilung gewählt, erhält man für die A posteriori-Verteilung von  $\mu$

$$\text{pdf}(\mu | \text{Daten, min. Vorinformation}) \propto \mu^{-28} \cdot \exp\{-13910,55 \text{ Tage}/\mu\} \cdot \mu^{-1}$$

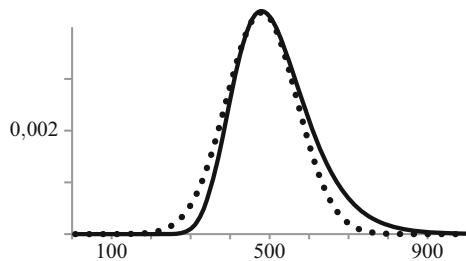
Der Modus ist  $\mu_0 = 480,1$  Tage.

<sup>1</sup> Man zeigt das Folgende:

- (1) Der Modus der A posteriori-Verteilung ist konsistent. Dies bedeutet: er schätzt die nicht direkt messbare Grösse immer besser, wenn die Anzahl Messungen zunimmt.
- (2) Die A posteriori-Verteilung konzentriert sich bei unabhängigen Messungen immer stärker um den Modus, wenn die A priori-Verteilung vernünftig gewählt ist.
- (3) Die Methode von Laplace approximiert die A posteriori-Verteilung gut, wenn der Modus nicht am Rand der Verteilung liegt.

Die Details dazu – und interessante Gegenbeispiele – findet man in [3] in Anhang B.

**Abb. 14.2** Approximation der A posteriori-Verteilung von  $\mu$  mit der Methode von Laplace durch eine Normalverteilung (gepunktet)



Die Methode von Laplace approximiert den Posterior von  $\mu$  mit einer Normalverteilung. Der Logarithmus  $L$  des Posteriors ist

$$L(\mu | \text{Daten, min. Vorinformation}) = \text{Konstante} - 29 \cdot \ln \mu - \frac{13910,55 \text{ Tage}}{\mu}$$

Damit beträgt die beobachtete Information  $I(\mu)$

$$I(\mu) = -L''(\mu) = -\frac{29}{\mu^2} + \frac{2 \cdot 13910,55 \text{ Tage}}{\mu^3}$$

Setzt man in den Ausdruck für  $\mu$  den Modus  $\mu_0 = 480,1$  Tage ein, erhält man

$$\text{SE}(\mu) = \frac{1}{\sqrt{I(\mu_0)}} = \left[ -\frac{29}{(480,1 \text{ Tage})^2} + \frac{2 \cdot 13910,55 \text{ Tage}}{(480,1 \text{ Tage})^3} \right]^{-1/2} = 89,2 \text{ Tage}$$

Ein Wahrscheinlichkeitsintervall zum Niveau von etwa 0,68 ist daher

$$\mu = \mu_0 \pm \text{SE}(\mu) = (480,1 \pm 89,2) \text{ Tage}$$

Mit einer Wahrscheinlichkeit von etwa 0,95 liegt  $\mu$  zwischen  $480,1 - 1,96 \cdot 89,2 = 305,3$  Tagen und  $480,1 + 1,96 \cdot 89,2 = 654,9$  Tagen. Abb. 14.2 zeigt die Plausibilität zu  $\mu$ , sowie gepunktet die Laplace-Approximation mit der Normalverteilung.  $\square$

Die Formeln (9.1) und (9.4) mit den Standardfehlern  $\text{SE}(\exp)$ ,  $\text{SE}(\text{poiss})$  für die Parameter der Exponential- und Poissonverteilung, sowie die Gleichung von de Moivre (10.2) entstehen mit der Methode von Laplace.<sup>2</sup>

<sup>2</sup> Ist das Datenmodell die Exponentialverteilung mit Parameter  $\mu$ , so lautet die A posteriori-Verteilung des Parameters

$\text{pdf}(\mu | \text{Daten, Vorinformation}) \propto \mu^{-n} \cdot \exp(-n \cdot \bar{x}/\mu) \cdot \text{pdf}(\mu | \text{Vorinformation})$

## 14.2 Die $\delta$ -Methode

Bei Präzisionsangaben mit Standardfehlern zu nicht direkt messbaren Größen lassen sich schnell Wahrscheinlichkeitsintervalle zu verschiedenen Niveaus berechnen. Wie pflanzen sich solche Präzisionsangaben auf transformierte Größen fort? Oder allgemeiner, wie rechnet man mit Präzisionsangaben, bei denen Standardfehler erwähnt sind? Die  $\delta$ -*Methode* oder die *Gauss'sche Fehlerfortpflanzung* antwortet auf solche Fragen.

Eine Person beschreibt, bei gegebener Information  $I$ , ihre Plausibilität zu einem Parameter  $X$  mit dem Standardfehler

$$X = X_0 \pm \delta X = X_0 \pm \text{SE}(X)$$

Gegeben sei ein zweiter Parameter  $Y$ , der von  $X$  abhängt:  $Y = g(X)$ . Dies kann  $Y$  der Kehrwert  $1/X$  von  $X$  sein. Dabei hängen  $X$  und  $Y$  in eindeutiger Weise voneinander ab. Damit ist  $X = X_0 \pm \delta X$  genau dann, wenn  $Y = g(X_0 \pm \delta X)$  ist. Mit einem Taylorpolynom 2. Ordnung lässt sich dieser Ausdruck approximieren:

$$Y = g(X_0 \pm \delta X) \approx g(X_0) \pm g'(X_0) \cdot \delta X + 0,5 \cdot g''(X_0) \cdot (\pm \delta X)^2$$

Ordnet man die Summanden um, erhält man, falls  $g'(X_0) \neq 0$ ,

$$Y \approx \underbrace{g(X_0) + 0,5 \cdot g''(X_0) \cdot (\delta X)^2}_{Y_0} \pm \underbrace{g'(X_0) \cdot \delta X}_{\delta Y}$$

Der Standardfehler  $\delta(Y)$  von  $Y$  ist also etwa  $g'(X_0) \cdot \delta X$ . Weiter ist der plausibelste Wert von  $Y$  etwa  $g(X_0) + 0,5 \cdot g''(X_0) \cdot (\delta X)^2$ . Man lässt den zweiten Summanden bei diesem Ausdruck weg, wenn  $(\delta X)^2$  klein ist:

**Theorem 14.2 (Übertragen von Präzisionsangaben mit der  $\delta$ -Methode)**

Von einer Größe  $X$  kennt man ihre Plausibilität mit dem Standardfehler:

$$X = X_0 \pm \delta X$$

Dabei ist  $\bar{x}$  das arithmetische Mittel der Daten und  $n$  ist die Anzahl Messungen. Ist die Likelihood dominant, so ist der Modus  $\mu_0$  der Verteilung  $\bar{x}$ . Die beobachtete Information lautet

$$I(\mu) = -\frac{n}{\mu^2} + 2 \cdot \frac{n \cdot \bar{x}}{\mu^3}$$

Setzt man hier den Modus  $\mu_0$  ein, erhält man  $I(\mu_0) = n/\bar{x}^2$ . Deshalb ist der Standardfehler  $\text{SE}(\exp) = \bar{x}/\sqrt{n}$ . Dies ergibt die Formel (9.1).

Hängt eine zweite Grösse  $Y$  eindeutig von  $X$  durch eine Gleichung  $Y = g(X)$  ab, so ist – falls  $g'(X_0) \neq 0$  –:

$$Y = Y_0 \pm \delta Y \approx g(X_0) \pm g'(X_0) \cdot \delta X$$

oder auch genauer

$$Y = Y_0 \pm \delta Y \approx g(X_0) + \frac{1}{2} \cdot g''(X_0) \cdot (\delta X)^2 \pm g'(X_0) \cdot \delta X$$

Diese Formeln sind approximativ. Sie können schlecht sein, insbesondere wenn der Quotient  $\delta X/X_0$  gross ist. Sie sollten also mit Vorsicht benutzt werden.<sup>3</sup>

**Beispiel 14.2 (Zeit zwischen starken Erdbeben)** In Beispiel 14.1 ist die durchschnittliche Zeit  $\mu$  zwischen zukünftigen, grossen Erdbeben angegeben:

$$\mu = \mu_0 \pm \text{SE}(\mu) = \mu_0 \pm \delta\mu = (480,1 \pm 89,2) \text{ Tage}$$

In einem Zeitfenster von zehn Jahren hat man durchschnittlich

$$N = \frac{10 \cdot 365 \text{ Tage}}{\mu}$$

große Erdbeben. Mit der Ableitung  $N'(\mu) = -3650 \text{ Tage}/\mu^2$  und der obigen ersten Formel berechnet man

$$N \approx \frac{10 \cdot 365 \text{ Tage}}{480,1 \text{ Tag}} \pm \frac{-3650 \text{ Tage}}{(480,1 \text{ Tag})^2} \cdot 89,2 \text{ Tage} = 7,6 \pm 1,4$$

Dies ist ein Wahrscheinlichkeitsintervall für  $N$  zum Niveau von etwa 0,68. Ein genaueres Resultat erhält man mit der zweiten Formel. Die zweite Ableitung  $N''(\mu)$  ist  $-7300 \text{ Tage}/\mu^3$ . Damit folgt

$$\frac{10 \cdot 365 \text{ Tage}}{480,1 \text{ Tag}} + \frac{1}{2} \cdot \frac{-7300 \text{ Tage}}{(480,1 \text{ Tag})^3} \cdot (89,2 \text{ Tage})^2 = 7,3$$

Deshalb ist  $N = N_0 \pm \delta N \approx 7,3 \pm 1,4$ .

□

Was passiert, wenn man zwei Präzisionsangaben mit Standardfehler addiert oder multipliziert? So kann man versuchen, die Präzision zur Summe  $S = X + Y$ , aus den zwei Intervallen

$$X = X_0 \pm \delta X, \quad Y = Y_0 \pm \delta Y$$

---

<sup>3</sup> Ein Beispiel, bei der die  $\delta$ -Methode schlecht ist, findet man in [7] auf den Seiten 74–77.

zu bestimmen. Das Quadrat von  $\delta X$  ist die Varianz der normalverteilten Approximation von  $X$ . Sie ist die durchschnittliche quadratische Abweichung von  $X_0$ . Variiert  $X$  um  $\Delta X$  und  $Y$  um  $\Delta Y$ , so variiert  $S$  um  $\Delta S = \Delta(X + Y) = \Delta X + \Delta Y$ . Also

$$(\Delta S)^2 = (\Delta X)^2 + (\Delta Y)^2 + 2 \cdot \Delta X \cdot \Delta Y$$

Man betrachtet den durchschnittlichen Wert dieser Gleichung. Der durchschnittliche Wert von  $(\Delta S)^2$  ist die Varianz  $(\delta S)^2$ . Sind  $X$  und  $Y$  unabhängig, so beeinflussen sich  $\Delta X$  und  $\Delta Y$  nicht. Damit ist

$$\text{Durchschnitt}(\Delta X \cdot \Delta Y) = \text{Durchschnitt}(\Delta X) \cdot \text{Durchschnitt}(\Delta Y) = 0 \cdot 0 = 0$$

Das Resultat ist  $(\delta S)^2 = (\delta X)^2 + (\delta Y)^2$ . Es ist deshalb:<sup>4</sup>

**Theorem 14.3 (Übertragen von Summen und Differenzen mit der  $\delta$ -Methode)**

Von zwei Größen  $X$  und  $Y$  kenne man Präzisionsangaben mit dem Standardfehler:

$$X = X_0 \pm \delta X, \quad Y = Y_0 \pm \delta Y$$

Sind  $X$  und  $Y$  unabhängig, so gilt

$$X \pm Y = (X_0 \pm Y_0) \pm \sqrt{(\delta X)^2 + (\delta Y)^2}$$

Will man Präzisionsangaben mit Standardfehlern addieren oder subtrahieren, so muss man also die quadrierten Unsicherheiten  $\delta X$  und  $\delta Y$  addieren. Zwei Bemerkungen dazu sind angebracht:

---

<sup>4</sup> Man findet ähnliche und andere Rechnungen in den meisten Büchern, die sich mit Wahrscheinlichkeitsmodellen befassen. Präziser kann man sogar sagen: Sind zwei Größen  $X$  und  $Y$  normalverteilt, so sind auch die Summe  $S = X + Y$  und die Differenz  $D = X - Y$  normalverteilt. Die Moden sind  $X_0 + Y_0$  und  $X_0 - Y_0$ . Die Varianz  $(\delta S)^2$  von  $S$  und  $D$  ist, falls  $X$  und  $Y$  unabhängig sind, gleich  $(\delta X)^2 + (\delta Y)^2$ . Um dies zu zeigen, braucht man die gemeinsame Verteilung  $\text{pdf}(X, Y)$  von  $X$  und  $Y$ . Sind  $X$  und  $Y$  unabhängig, so ist nach der Multiplikationsregel

$$\text{pdf}(X, Y) = \text{pdf}(X) \cdot \text{pdf}(Y) \propto \exp\{-0.5 \cdot \chi^2\} \quad \text{mit} \quad \chi^2 = \left(\frac{X - X_0}{\delta X}\right)^2 + \left(\frac{Y - Y_0}{\delta Y}\right)^2$$

Die Wahrscheinlichkeit, dass  $S = S_0 \pm \delta S$  ist, ist gleich dem Volumen unter der Fläche des Graphen der Dichtefunktion  $\text{pdf}(X, Y)$ , beschränkt durch  $X + Y = S_0 \pm \delta S$ . Man erhält ein Integral, das sich mit der Substitution  $S = X + Y$ ,  $T = Y$  und algebraischen Tricks durch die Dichte der Normalverteilung ausdrücken lässt.

- (1) Die Formeln sind falsch, falls  $X$  und  $Y$  nicht unabhängig sind. Sind  $X$  und  $Y$  positiv korreliert, so ist das Produkt  $\Delta X \cdot \Delta Y$  durchschnittlich positiv.<sup>5</sup> Damit wird der Standardfehler von  $X + Y$  grösser als in der Formel angegeben.
- (2) Auch wenn die Grössen  $X$  und  $Y$  unabhängig sind, sind es die Summen  $X + Y$  und Differenzen  $X - Y$  nicht mehr.

In ähnlicher Weise lassen sich (approximative) Formeln für Produkte und Quotienten oder komplizierte Ausdrücke herleiten. Wenn die Grössen  $X$  und  $Y$  unabhängig sind, gilt:<sup>6</sup>

$$X \cdot Y \approx X_0 \cdot Y_0 \pm X_0 \cdot Y_0 \cdot \sqrt{\left(\frac{\delta X}{X_0}\right)^2 + \left(\frac{\delta Y}{Y_0}\right)^2}$$

Analog hat man

$$\frac{X}{Y} \approx \frac{X_0}{Y_0} \pm \frac{X_0}{Y_0} \cdot \sqrt{\left(\frac{\delta X}{X_0}\right)^2 + \left(\frac{\delta Y}{Y_0}\right)^2}$$

Die Formeln besagen, dass bei Produkten und Quotienten die quadrierten relativen Präzisionen  $\delta X/X$  und  $\delta Y/Y$  addiert werden. Es folgen zwei Beispiele dazu.

**Beispiel 14.3 (Elektrischer Stromkreis)** In einem Stromkreis wird der Widerstand  $R$  nach dem ohmschen Gesetz aus der Spannung  $U$  und dem Strom  $I$  berechnet:

$$R = \frac{U}{I}$$

Ein Ingenieur misst die Spannung  $U$  und den Strom  $I$  je viermal in unabhängiger Art voneinander. Wegen Kovariablen streuen die Messwerte um die Spannung und den Strom. Die Tabelle zeigt zusammengefasst die Messungen:

	U	I
arith. Mittel	220,3 V	51,2 A
emp. Standardabw.	1,4 V	2,1 A

<sup>5</sup> Man nennt den durchschnittlichen Wert von  $\Delta X \cdot \Delta Y$  auch die *Kovarianz* (engl. *Covariance*) der gemeinsamen Verteilung von  $X$  und  $Y$ . Man hat allgemein für  $S = X + Y$ :

$$(\delta S)^2 = (\delta X)^2 + (\delta Y)^2 + 2 \cdot \text{Cov}(X, Y)$$

<sup>6</sup> Ist  $P = X \cdot Y$ , so erhält man aus  $\ln P = \ln X + \ln Y$  die erwähnte Formel. Nach der Transformationsformel ist  $\ln X \approx \ln X_0 \pm \delta X/X_0$ . Somit folgt

$$\delta \ln P \approx \sqrt{\left(\frac{\delta X}{X_0}\right)^2 + \left(\frac{\delta Y}{Y_0}\right)^2}$$

Vor den Messungen hat der Ingenieur nur minimale Information zu  $U$  und  $I$ . Weiter wählt er als Datenmodell die Normalverteilung. Mit der Gleichung von de Moivre (10.2) folgt:

$$U = U_0 \pm \delta U = \bar{U} \pm \frac{s_U}{\sqrt{10}} = 220,3 \text{ V} \pm \frac{1,4 \text{ V}}{\sqrt{10}} = (220,4 \pm 0,4) \text{ V}$$

Ebenso hat man ein Wahrscheinlichkeitsintervall für den Strom  $I$  zum Niveau von etwa 0,68:

$$I = I_0 \pm \delta I = 51,2 \text{ A} \pm \frac{2,1 \text{ A}}{\sqrt{10}} = (51,2 \pm 0,7) \text{ A}$$

Da  $U$  und  $I$  in unabhängiger Art gemessen sind, berechnet man ein approximatives Wahrscheinlichkeitsintervall für den Widerstand  $R$  mit der obigen Formel:

$$R \approx \frac{220,3 \text{ V}}{51,2 \text{ A}} \pm \frac{220,3 \text{ V}}{51,2 \text{ A}} \cdot \sqrt{\left(\frac{0,4 \text{ V}}{220,3 \text{ V}}\right)^2 + \left(\frac{0,7 \text{ A}}{51,2 \text{ A}}\right)^2}$$

Man erhält  $R \approx (4,3 \pm 0,2) \Omega$ . Dies ist ein Wahrscheinlichkeitsintervall für den Widerstand  $R$  zum Niveau von etwa 0,68.  $\square$

**Beispiel 14.4 (Waagebalken)** Dieses klassische Beispiel findet sich in [1] und [4]. Eine Ingenieurin möchte die Gewichte  $A$  und  $B$  zweier Objekte mit einem Waagebalken messen. Mit nur zwei Messungen sollen die Gewichte  $A$  und  $B$  bestimmt werden. Die Ingenieurin legt dazu einmal das Objekt  $A$  und anschliessend das Objekt  $B$  auf die Waage. Sie erhält Messungen  $A_0$  und  $B_0$ . Damit ist

$$A = A_0 \pm \text{SE}, \quad B = B_0 \pm \text{SE}$$

Präziser werden die Resultate, wenn man, wie in Abb. 14.3 dargestellt, die Grössen  $S = A + B$  und  $D = A - B$  bestimmt. Man hat  $S = S_0 \pm \text{SE}$  und  $D = D_0 \pm \text{SE}$ . Aus der Summe  $S$  und der Differenz  $D$  lassen sich die Gewichte von  $A$  und  $B$  durch Rechnung bestimmen:

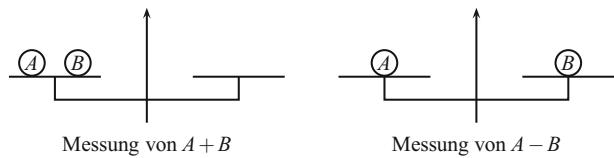
$$A = \frac{S + D}{2}, \quad B = \frac{S - D}{2}$$

Mit der  $\delta$ -Methode folgt, dass

$$A = \frac{1}{2} \left( S_0 + D_0 \pm \sqrt{\text{SE}^2 + \text{SE}^2} \right) = \frac{S_0 + D_0}{2} \pm 0,71 \cdot \text{SE}$$

Dies ist eine bessere Präzision. Der Statistiker C. Daniel kommentiert in [1] das Beispiel wie folgt:

**Abb. 14.3** Messung auf einer Waage von  $A + B$  und  $A - B$



The disadvantage of this „weighing design“ is that no information is available until the data are in. The reward for the delay is, in this case, the double precision. The moral [...] is that each observation can be made to yield information on two (or more) parameters. Indeed the number of times that each observation can be used increases steadily with the number of observations in each balanced set. What is required is *planning*.

Das Beispiel weist darauf hin, dass Überlegungen zur  $\delta$ -Methode benutzt werden können, um nicht direkt messbare Größen möglichst präzise zu bestimmen.  $\square$

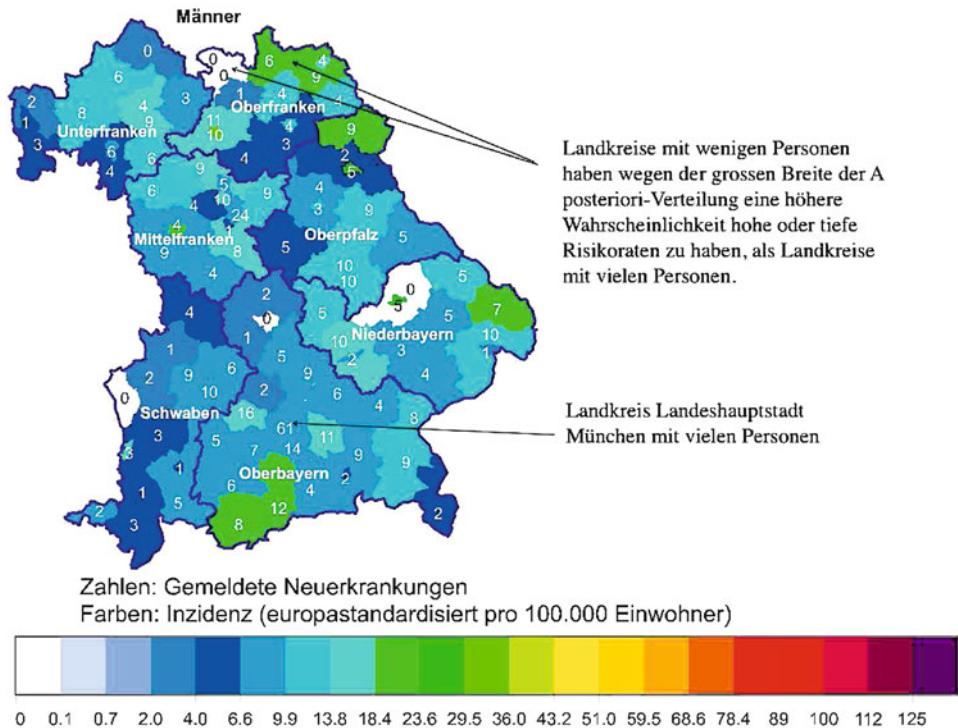
### 14.3 Eine gefahrenrächtige Gleichung

Die Laplace-Approximation der A posteriori-Verteilung einer nicht direkt messbaren Grösse  $\theta$  besagt, dass mit einer Wahrscheinlichkeit von etwa 0,68

$$\theta = \theta_0 \pm \text{SE}(\theta)$$

Bei vielen Datenmodellen ist der Modus  $\theta_0$  das arithmetische Mittel der Datenwerte. Bei unabhängigen Datenwerten ist  $\text{SE} \propto 1/\sqrt{n}$ . Je grösser der Stichprobenumfang  $n$  also ist, desto präziser bestimmt der Modus der A posteriori-Verteilung in der Regel die gesuchte Grösse. *Insbesondere ist es wahrscheinlicher, sehr hohe oder tiefe Werte von arithmetischen Mitteln bei kleinen Stichproben als bei grossen Stichproben zu beobachten!* Kostspielige Untersuchungen haben dies nicht berücksichtigt und damit Entscheidungen aus Daten getroffen, die ungenau waren. Der Statistiker H. Wainer zeigt in [8] solche Beispiele und nennt die Gleichung  $\theta = \theta_0 \pm \text{SE}(\theta)$  daher auch die *gefährlichste Gleichung der Statistik*.

**Beispiel 14.5 (Bevölkerungsbezogenes Krebsregister Bayern)** Seit 1998 registriert der Freistaat Bayern neu auftretende Krebserkrankungen. Nach einer vierjährigen Startphase mit nur der Hälfte der Landkreise und kreisfreien Städte werden seit 1.1.2002 alle bösartigen Neubildungen und ihre Frühformen flächendeckend in ganz Bayern erfasst. Abb. 14.4 zeigt gemeldete Neuerkrankungen von Hodenkrebs während eines Jahres. Die Zahlen zeigen die Anzahl der Neuerkrankungen pro Landkreis. In Farbe sind die durchschnittlichen Neuerkrankungen pro 100 000 Einwohner dargestellt. Geht man davon aus, dass das Risiko an Hodenkrebs zu erkranke überall in Bayern etwa gleich ist, ist es nicht überraschend, dass der Landkreis „München Landeshauptstadt“ mit der grössten



**Abb. 14.4** Neuerkrankungen von Hodenkrebs während eines Jahres (*Zahlen*) und durchschnittliche Neuerkrankungen pro 100 000 Einwohner (*in Farbe*), aus: Bevölkerungsbezogenes Krebsregister Bayern (Hrsg.): Krebs in Bayern im Jahr 2004, Seite 49, Erlangen 2007

Anzahl von 1 300 000 Personen in der Farbskala in der Mitte liegt. Bei den Landkreisen mit kleinen Bevölkerungszahlen besteht eine grosse Wahrscheinlichkeit, Anteile mit tiefen oder hohen durchschnittlichen Neuerkrankungen anzutreffen: Im Nordosten, bei den schwach bevölkerten Landkreisen mit 40 000–100 000 Personen, liegen hohe und tiefe durchschnittliche Neuerkrankungen nebeneinander. Ein Phänomen, das nicht lokale Gegebenheiten aufzeigt, sondern die Breite der A posteriori-Verteilungen widerspiegelt.

Wie kann man die Abbildung besser gestalten? Man kann zuerst als Vorwissen annehmen, dass das Risiko an Hodenkrebs zu erkranken, überall in Bayern etwa gleich ist. Mit Daten und Studien aus früheren Jahren, kann man den Anteil  $A$  der Leute in Bayern mit Hodenkrebs berechnen. Man hat nach Theorem 5.3

$$\text{pdf}_{\text{Bayern}}(A \mid \text{alte Daten, min. Vorinformation}) \propto A^a \cdot (1 - A)^{N-a}$$

Dabei ist  $a$  die Anzahl Erkrankungen und  $N$  die Anzahl männlicher Personen in Bayern.  $N$  ist gross. Mit den neuen Daten –  $b$  Erkrankungen in einer Region  $X$  mit  $n$  männlichen

Personen – kann man den Anteil  $A$  für die Region  $X$  berechnen:

$$\text{pdf}_{\text{Region } X}(A \mid \text{Daten, Bayern}) \propto \underbrace{A^b \cdot (1 - A)^{n-b}}_{\text{Likelihood aus Region } X} \cdot \underbrace{A^a \cdot (1 - A)^{N-a}}_{\text{Prior aus Bayern}}$$

Dies ist das gleiche Vorgehen wie in Beispiel 5.5. Der plausibelste Wert  $A_0$  von  $A$  beträgt

$$A_{0, \text{Region } X} = \frac{b}{n} \cdot z + \frac{a}{N} \cdot (1 - z) \quad \text{mit } z = \frac{n}{n + N}$$

Ist die Region  $X$  klein, so ist  $z$  klein. Der plausibelste Wert für den Anteil an Erkrankungen wird dann durch den Anteil der Erkrankungen in Bayern geprägt. Ein möglicherweise grosser Anteil an Hodenkrebs in der kleinen Region  $X$  wird damit nach unten korrigiert. Mit den so berechneten, korrigierten Anteilen lässt sich die Karte aussageadäquater gestalten. Der Effekt der Unsicherheit in den kleinen Regionen wird reduziert. Dieses Vorgehen hat aber Nachteile. So werden die Daten aus früheren Jahren bei den kleinen Regionen sehr stark gewichtet. Die Daten aus früheren Jahren könnten überholt sein. Zudem benutzt man die Informationen nicht, dass das Risiko an Hodenkrebs zu erkranken, von Faktoren abhängt, die landkreisspezifisch und die landeskreisunabhängig sind. Der Abschn. 14.4 zeigt, wie dies gemacht werden kann.  $\square$

**Beispiel 14.6 (Sicherheit in Städten)** Ein Beispiel einer fragwürdigen Rangliste erwähnt R. Wainer in [8]: Aus der durchschnittlichen Dauer zwischen Verkehrsunfällen erstellte die New York Times eine Rangliste der zehn sichersten und der zehn unsicheren Städte in den USA. Betrachtet wurden dabei 200 Städte. Die zehn grössten Städte befanden sich nicht an den Enden der Rangliste. In der Tat wäre es erstaunlich, wenn sie dort erscheinen würden. Wegen der Gleichung von de Moivre, sind es nämlich die kleinen Städte, die an den Extremitäten der Rangliste aufgelistet sind.

Es ist also schwierig, wahrscheinlichste Werte von nicht direkt messbaren Grössen oder Parametern zu vergleichen. Dazu braucht man die A posteriori-Verteilungen oder zumindest eine Präzisionsangabe mit dem Standardfehler.

**Beispiel 14.7 (Ausbildung von Lehrerinnen und Lehrern)** TEDS-M untersucht, wie gut angehende Mathematiklehrpersonen ausgebildet werden. Dabei interessiert, wie die durchschnittliche Punktzahl  $\bar{P}_{\text{Land}XY}$  aller Lehrpersonen aus einem Land bei einem standardisierten Test ausfällt. Diese Grösse kann nur mit einem Zensus genau bestimmt werden. Dies ist zu aufwändig. Mit einer Stichprobe lässt sich ihre Plausibilität bestimmen. Tab. 14.1 zeigt die Auswertung aus fünfzehn Ländern mit Stichprobengrössen zwischen 47 und 1985 Personen. In der Tabelle findet man den Stichprobenumfang  $n$ , den plausibelsten Wert  $\bar{P}_{\text{Stichprobe}}$  der A posteriori-Verteilung von  $\bar{P}_{\text{Land}XY}$  und den Standardfehler SE. Üblich ist es, den Standardfehler in Klammern zu setzen, um ihn nicht mit

**Tab. 14.1** TEDS-Untersuchung zu Lehrpersonen (aus [6])

	$n$	$\bar{P}_{\text{Stichprobe}}$	SE
Taiwan	400	623	(4,2)
Singapur	570	590	(3,1)
Norwegen	296	553	(4,3)
Deutschschweiz	1207	543	(1,9)
Russland	85	535	(9,9)
Thailand	1063	528	(2,3)
USA	283	518	(4,1)
Deutschland	945	510	(2,7)
Polen	1985	490	(2,2)
Malaysia	900	488	(1,8)
Spanien	480	481	(2,6)
Botswana	66	441	(5,9)
Philippinen	47	440	(7,6)
Chile	958	413	(2,1)
Georgien	475	345	(3,9)

dem plausibelsten Wert zu verwechseln. Man sieht, dass ein Wahrscheinlichkeitsintervall von etwa 0,68 für die durchschnittliche Punktzahl bei Lehrpersonen aus Deutschland  $510,0 \pm 2,7$  ist. Wie erwartet liegen aus kleineren Stichprobenumfängen berechnete Modelen tendenziell in der unteren und oberen Hälfte der Tabelle.

Abb. 14.5 visualisiert mit Box & Whisker Plots, wie die einzelnen Punktzahlen verteilt sind. Die Punktzahlen liegen jeweils in den hellen Balken. Die grauen Balken zeigen die empirischen Quartile. 50 % der Punktzahlen liegen in diesen Balken. Die schwarzen Streifen zeigen die Wahrscheinlichkeitsintervalle zum Niveau von 0,95 für  $\bar{P}_{\text{Land}XY}$  an:  $\bar{P}_{\text{Stichprobe}} \pm 1,96 \cdot \text{SE}$ . Man sieht, dass die Punktzahlen der Probanden aus Deutschland zwischen 380 und 650 Punkten liegen. Sie streuen also stark. Die Hälfte dieser Probanden hatten Punktzahlen zwischen 440 und 540. Vergleicht man mit den Resultaten aus Botswana, sieht man, dass  $\bar{P}_{\text{Land}XY}$  präziser bestimmt ist. Dies ist der Fall, weil der Stichprobenumfang grösser ist.

Die Standardfehler der geschätzten durchschnittlichen Punktzahlen aller Lehrpersonen sind angegeben. Damit kann man beispielsweise die Wahrscheinlichkeit berechnen, dass  $\bar{P}_{\text{Russland}}$  grösser als  $\bar{P}_{\text{USA}}$  ist. Mit der  $\delta$ -Methode ist, wenn die Punktzahlen der Probanden unabhängig sind,

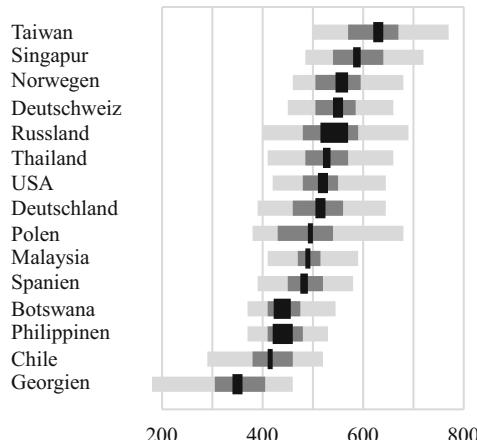
$$\bar{P}_{\text{Russland}} - \bar{P}_{\text{USA}} = (535 - 518) \pm \sqrt{9,9^2 + 4,1^2} = 17 \pm 11$$

Die A posteriori-Verteilung dieser Differenz  $D$  ist also etwa normalverteilt mit Modus 17 und Standardabweichung 11. Die Wahrscheinlichkeit, dass die Differenz grösser als Null ist, berechnet man daraus einfach mit einem Statistikprogramm:

$$\mathbb{P}(D = \bar{P}_{\text{Russland}} - \bar{P}_{\text{USA}} > 0 \mid \text{Normalverteilung}) = 0,94$$

**Abb. 14.5** TEDS-

Untersuchung zu Lehrpersonen (aus [6]): Verteilung der Punktzahlen (*hell*), Quartile (*grau*) und Wahrscheinlichkeitsintervalle zum Niveau 0,95 für den Durchschnittswert der Gesamtpopulation (*schwarz*)



Wie wahrscheinlich ist  $\bar{P}_{\text{Russland}}$  grösser als  $\bar{P}_{\text{USA}}$  und  $\bar{P}_{\text{Thailand}}$ ? Gesucht ist die Wahrscheinlichkeit  $\mathbb{P}(D_1 > 0, D_2 > 0 \mid \text{Daten})$  mit  $D_1 = \bar{P}_{\text{Russland}} - \bar{P}_{\text{USA}}$  und  $D_2 = \bar{P}_{\text{Russland}} - \bar{P}_{\text{Thailand}}$ . Die Parameter  $D_1$  und  $D_2$  sind nicht unabhängig, da  $\bar{P}_{\text{Russland}}$  in beiden Grössen vorkommt. Daher ist

$$\mathbb{P}(D_1 > 0, D_2 > 0) = \mathbb{P}(D_2 > 0 \mid D_1 > 0) \cdot \mathbb{P}(D_1 > 0)$$

Der erste Faktor auf der rechten Seite lässt sich nicht mit der  $\delta$ -Methode bestimmen. Eine Monte-Carlo-Simulation hilft hier weiter. Man erzeugt je 100 000 Werte der A posteriori-Verteilungen (Normalverteilungen) von  $\bar{P}_{\text{Russland}}$ ,  $\bar{P}_{\text{USA}}$  und  $\bar{P}_{\text{Thailand}}$ . Anschliessend zählt man, wie häufig man  $D_1 > 0$  und  $D_2 > 0$  hat. Hier die ersten fünf Werte aus einer solchen Simulation:

$\bar{P}_{\text{Russland}}$	$\bar{P}_{\text{USA}}$	$\bar{P}_{\text{Thailand}}$	$D_1$	$D_2$	$D_1 > 0 \text{ und } D_2 > 0$
534,1	522,1	526,6	12,0	7,5	Ja
535,9	512,9	527,9	23,0	8,0	Ja
513,8	517,5	527,3	-3,7	-13,5	Nein
541,1	511,2	526,1	29,9	15,0	Ja
523,6	518,9	526,9	4,7	-3,3	Nein
...	...	...	...	...	...

Die gesuchte Wahrscheinlichkeit ist

$$\mathbb{P}(D_1 > 0, D_2 > 0 \mid \text{Daten}) \approx \frac{\text{Anzahl Ja}}{100\,000} = \frac{75\,345}{100\,000} = 0,75$$

Will man Russland mit weiteren Ländern aus der Liste vergleichen, sind dazu weitere Monte-Carlo-Simulationen nötig. Dies ist aufwändig.

In der obigen Rechnung ist gezeigt, dass die durchschnittliche Punktzahl aller Lehrpersonen in Russland mit hoher Wahrscheinlichkeit höher als diejenige der USA ist. Das Resultat sollte skeptisch beurteilt werden. Die Daten stammen aus *Beobachtungen*. Es könnte daher sein, dass das Resultat nicht auf Grund des Faktors Land, sondern wegen anderen Kovariablen, wie Alter, Motivation, Einkommen oder Testart entsteht. Die Gefahr eines Simpson-Effekts ist gross. Man erhält ein aussagekräftigeres Resultat, wenn man ein Regressionsmodell in Funktion der Kovariablen aufstellt. Beispielsweise in der linearen Form ohne Interaktionen

$$\bar{P} = a + b \cdot X_1 + c \cdot X_2 + d \cdot X_3 + \dots$$

Dabei sind  $X_1, X_2, \dots$  die Kovariablen wie Land, Alter, Motivation, Einkommen, ... und  $a, b, \dots$  sind zu bestimmende Parameter.  $\square$

In vielen Zeitschriften und wissenschaftlichen Arbeiten finden sich Tabellen zu nicht direkt messbaren Größen. Dabei werden in der Regel die Moden und die zugehörigen Standardfehler der A posteriori-Verteilungen aufgelistet. Wie man solche Tabellen geschickt markieren kann, um daraus nicht voreilige Schlüsse zu ziehen, findet man in [8].

In Produktionsprozessen werden arithmetische Mittel von Größen oder von Waren verglichen. Um einen Prozess zu überwachen, können pro Zeitfenster Waren zufällig gezogen und Mittelwerte gebildet werden. Bei den pro Zeitfenster gebildeten arithmetischen Mitteln wird dann versucht, aussergewöhnlich grosse oder kleine arithmetische Mittel, die nicht auf „normalen“ Streuungen basieren, von den normalen arithmetischen Mitteln zu trennen. Dies geschieht mit geschätzten Kontrollgrenzen, analog wie im Abschn. 3.4. Sind  $\bar{X}_1, \bar{X}_2, \dots, \bar{X}_M$  die aus den  $M$  Gruppen von Stichprobenwerten berechneten arithmetischen Mittel und  $SE_1, SE_2, \dots, SE_M$  die zugehörigen Standardfehler, so betrachtet man den Mittelwert der arithmetischen Mittel und der Standardfehler:

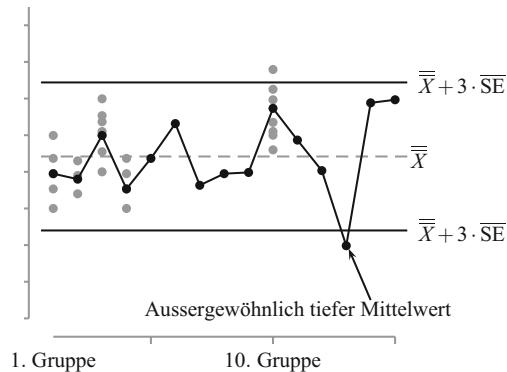
$$\bar{\bar{X}} = \frac{\bar{X}_1 + \bar{X}_2 + \dots + \bar{X}_M}{M}, \quad \bar{SE} = \frac{SE_1 + SE_2 + \dots + SE_M}{M}$$

Arithmetische Mittel, die grösser als  $\bar{\bar{X}} + 3 \cdot \bar{SE}$  oder kleiner als  $\bar{\bar{X}} - 3 \cdot \bar{SE}$  sind, werden als aussergewöhnlich bezeichnet (siehe Abb. 14.6). Die aussergewöhnlichen Mittelwerte werden als Ausdruck von nicht normaler Streuung verstanden. Mittelwerte innerhalb der Bänder streuen wegen variierender Produktionsbedingungen oder Zufälligkeiten.

Zum Schluss des Abschnitts folgt noch ein Hinweis. Angaben mit dem Standardfehler beziehen sich auf nicht direkt messbare Größen. Wie in Kap. 2 erklärt, ist es wichtig zu wissen, auf was sich eine Grösse dabei bezieht. Der Standardfehler kann je nach Bezug variieren:

**Beispiel 14.8 (Punktedurchschnitt)** Nach einer Prüfung zur Statistik, bei der maximal 44 Punkte möglich waren, ermittelt eine Lehrperson bei einer Klasse von 24 Studierenden

**Abb. 14.6** Kontrollkarte zu arithmetischen Mitteln: graue Punkte sind die einzelnen Messwerte, schwarze Punkte die daraus berechneten arithmetischen Mittel



einen Punktedurchschnitt  $\bar{P}$  von 30,5 Punkten. Dieser Schnitt ist präzis bestimmt. Es ist  $\mathbb{P}(\bar{P} = 30,5 \mid \text{Daten}) = 1$  oder

$$\bar{P} = 30,5 \pm 0,0$$

Der Wert 30,5 bezieht sich hier auf die Klasse, die die Grundgesamtheit darstellt. Die Lehrperson könnte den Punktedurchschnitt  $\bar{P}$  von 30,5 Punkten benutzen, um zukünftige Klassenschnitte  $\bar{P}_{\text{zuk}}$  zu schätzen. Dazu braucht sie ein Datenmodell, das besagt, wie Punkte um diese nicht direkt messbare Größe streuen und einen Prior. Daraus kann man die A posteriori-Verteilung von  $\bar{P}_{\text{zuk}}$  berechnen:

$$\bar{P}_{\text{zuk}} = 30,5 \pm \text{SE}(\bar{P}_{\text{zuk}})$$

Es ist unsicher, dass der Modus 30,5 und  $\text{SE} \propto 1/\sqrt{24}$  ist. Dazu müsste man garantieren, dass die 24 Stichprobenwerte unabhängig sind und keine systematischen Fehler enthalten. Systematische Fehler sind denkbar, da die Stichprobe nicht randomisiert ist. Ein ähnliches Beispiel findet sich in [2] auf Seite 556.  $\square$

## 14.4 Struktur und hierarchische Modelle

Im vorhergehenden Abschnitt wurde einführend diskutiert, wie Kennzahlen von verschiedenen Gruppen (Ländern, Regionen, Produktionsarten oder -stätten) mit Daten geschätzt und mit dem Standardfehler verglichen werden können. Geschätzt wurden die Kennzahlen jeweils nur mit Daten aus der entsprechenden Gruppe. Dies ist aber oft nicht vorteilhaft. So entstehen Phänomene, die keine gruppeninterne Gegebenheiten aufzeigen, sondern nur die Breite der A posteriori-Verteilung. Zudem werden Informationen, wie ähnlich die Gruppen sind, nicht berücksichtigt. Die folgenden Beispiele zeigen, wie solche Probleme mit *hierarchischen Modellen* beseitigt werden können.

**Tab. 14.2** Massen (in Gramm) von Stören aus der Fischzucht der Tropenhaus Frutigen AG aus sieben Becken

Becken	Messungen									
1	2000	2100	1700	1550	2050	1600	1800	1550	1300	1300
2	1050	1400	1250							
3	1250	1350	1400	1800	1600	1450	1750			
4	1300	1710	1300	1450	1550	1540	1310	1380	1510	1390
5	1610	2210								
6	1190	1720	1500	1540	1210	1300	1350	1150		
7	1230	1070	1500	1600	1280	1340	1420	1950	1480	

**Beispiel 14.9 (Fischzucht)** In Fischzuchten ist es wichtig zu wissen, wie gross die Durchschnittsmassen  $\mu$  von Fischen in Aufzuchtbecken sind. Dazu werden randomisiert Stichproben aus den Becken gezogen und die Massen der Fische gemessen. Tab. 14.2 zeigt die Stichprobenwerte aus sieben Becken der Fischzucht der Tropenhaus Frutigen AG. Um die Durchschnittsmassen  $\mu_1, \mu_2, \dots, \mu_7$  aller Fische in den sieben Becken zu berechnen, nimmt die messende Person an, dass die Daten mit der Normalverteilung beschrieben werden können. Die plausibelsten Werte für die Durchschnittsmassen sind dann die arithmetischen Mittel der jeweiligen Gruppen. Den Standardfehler SE erhält man mit der Gleichung von deMoivre (10.2). Damit hat man aus den Daten von Becken 1

$$\mu_1 \approx \frac{2000 \text{ g} + 2100 \text{ g} + \dots + 1300 \text{ g}}{10} = 1695 \text{ g}, \quad \text{SE} = \frac{s_1}{\sqrt{10}} = 92 \text{ g}$$

Mit den Daten von Becken 5 hat man

$$\mu_5 \approx \frac{1610 \text{ g} + 2210 \text{ g}}{2} = 1910 \text{ g}, \quad \text{SE} = \frac{s_5}{\sqrt{2}} = 300 \text{ g}$$

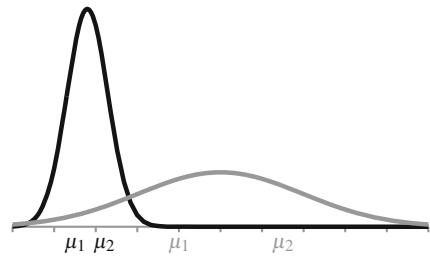
Die Schätzung ist sehr unsicher, da nur zwei Messwerte vorhanden sind. Zudem ist der grösste Messwert in der kleinsten Datengruppe. Daher der grosse Wert von 1910 g für die Schätzung von  $\mu_5$ .

Die Rechnungen lassen sich mit zusätzlicher Information verbessern. Die Fische in den sieben Becken sind alle gleich alt, wurden zum gleichen Zeitpunkt ausgesetzt und gleich gefüttert. Daher dürften die Messwerte in allen Becken gleich streuen. Als Datenmodell wählt man daher eine Normalverteilung mit einer für alle Becken gleichen Streuung  $\sigma$ :

$$\text{Datenmodell: Messwerte in Becken } i \sim \text{Normal}(\mu_i, \sigma) \quad (14.1)$$

Um den Posterior für die Durchschnittsmassen  $\mu_1, \mu_2, \dots$  in den Becken zu berechnen, braucht man einen Prior. Die Durchschnittsmassen hängen von Faktoren ab, die beckenspezifisch sind (Eigenheiten der Fische, Sonnenbestrahlung, Strömungsverhältnisse) und von anderen Faktoren, die in allen Becken gleich sind (Alter, Fütterung, Dichte

**Abb. 14.7** A priori-Verteilung für die Durchschnittsmassen  $\mu_1$  und  $\mu_2$  bei verschiedenen Werten von  $b$ : schwarze Kurve  $b$  klein (die Parameter  $\mu_1$  und  $\mu_2$  liegen nahe zusammen) und graue Kurve  $b$  gross ( $\mu_1$  und  $\mu_2$  könnten sehr verschieden sein)



der Population). Daher werden die Durchschnittsmassen in den Becken streuen. Wie stark die Streuung ist, lässt sich mit einem Wahrscheinlichkeitsmodell beschreiben. Sinnvoll scheint dazu eine Normalverteilung. Dies ergibt die A priori-Verteilungen zu den Durchschnittsmassen  $\mu_1, \mu_2, \dots, \mu_7$ :

$$\text{Prior: } \mu_1, \mu_2, \dots, \mu_7 \sim \text{Normal}(a, b) \quad (14.2)$$

Dies beschreibt, wie die Durchschnittsmassen streuen oder *kombiniert* (auch *gepoolt* genannt) sind. Sie streuen um  $a$  mit Standardabweichung  $b$ . Man erzeugt damit eine Strukturfunktion zwischen den Becken.<sup>7</sup> Die gesuchten Parameter  $\mu_1, \dots, \mu_7$  hängen von den nicht beobachtbaren Parametern  $a$  und  $b$  ab. Man nennt  $a$  und  $b$  *Hyperparameter*. Eine Person, die denkt, dass beckenspezifische Faktoren kaum Einfluss auf die einzelnen Durchschnittsmassen haben, könnte  $b$  sehr klein wählen: eine sehr schmale Normalverteilung. Es ist dann

$$\mu_1 \approx \mu_2 \approx \dots \approx \mu_7$$

Ist aber  $b$  gross, so erhält man eine breite Verteilung. Eine gute Wahl als Vorinformation für eine Person, die beckenspezifische Faktoren als wesentlich ansieht. Abb. 14.7 illustriert die Situation. Die Idee ist nun: Die Hyperparameter oder die Kombinationsstärke sollen nicht „willkürlich“ gesetzt, sondern mit der Bayes-Regel Likelihood  $\times$  Prior aus den Daten berechnet werden. Die unbekannten Größen sind die Parameter  $\mu_1, \mu_2, \dots, \mu_7, \sigma$  und die Hyperparameter  $a$  und  $b$ . Also ist

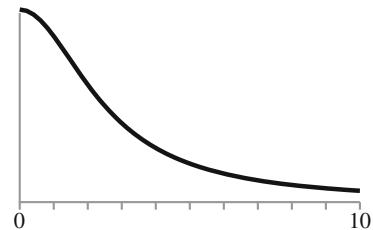
$$\text{pdf}(\mu_1, \mu_2, \dots, \sigma, a, b \mid \text{Daten}) = \text{Likelihood} \cdot \underbrace{\mathbb{P}(\mu_1, \mu_2, \dots, \sigma, a, b)}_{\text{Prior}}$$

Die Likelihood berechnet man aus dem Datenmodell der Normalverteilung, also aus der Gleichung (14.1). Für den Prior hat man –  $\sigma$  ist ein Skalierungsparameter, der zwischen 10 und 1000 g ist –:

$$\mathbb{P}(\mu_1, \mu_2, \dots, \sigma, a, b) = \mathbb{P}(\sigma) \cdot \mathbb{P}(\mu_1, \mu_2, \dots, a, b) \propto \frac{1}{\sigma} \cdot \mathbb{P}(\mu_1, \mu_2, \dots, a, b)$$

<sup>7</sup> Siehe dazu auch die Strukturfunktion (9.3).

**Abb. 14.8** A priori-Verteilung für den Hyperparameter  $b$ : eine halbe Cauchyverteilung



Der zweite Faktor wird mit der Multiplikationsregel berechnet:

$$\begin{aligned}\mathbb{P}(\mu_1, \mu_2, \dots, a, b) &= \mathbb{P}(\mu_1, \mu_2, \dots | a, b) \cdot \mathbb{P}(a, b) \\ &= \mathbb{P}(\mu_1 | a, b) \cdot \mathbb{P}(\mu_2 | a, b) \cdots \cdots \mathbb{P}(\mu_7 | a, b) \cdot \mathbb{P}(a) \cdot \mathbb{P}(b)\end{aligned}$$

Die Faktoren  $\mathbb{P}(\mu_i | a, b)$  berechnet man aus der Gleichung (14.2):  $\mu_i$  ist normalverteilt mit Modus  $a$  und Standardabweichung  $b$ . Wie soll man aber die Priors  $\mathbb{P}(a)$  und  $\mathbb{P}(b)$  zu den Hyperparametern  $a$  und  $b$  wählen? Erfahrungen der Fischzüchter zeigen, dass die Durchschnittsmassen um 1500 g liegen sollten:

$$a \sim \text{Normal}(1500, 10\,000)$$

Der Hyperparameter  $b$  ist ein Skalierungsparameter, der sehr klein aber auch sehr gross sein könnte. Eine beliebte und konservative Wahl für den Hyperparameter  $b$  ist die halbe Cauchyverteilung mit Skalierung scale = 2,5 (eine vertiefende Diskussion dazu findet man in [3]):

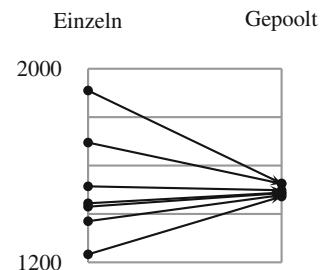
$$\text{pdf}(b) \propto \frac{1}{1 + (b/2,5)^2}$$

Abb. 14.8 zeigt die Verteilung. Mit einer MCMC-Simulation lassen sich daraus die A posteriori-Verteilungen von  $\mu_1, \mu_2, \dots$  sowie ihre plausibelsten Werte berechnen. Implementiert man das Modells mit einem Statistikprogramm, müssen das Datenmodell und die A-Priori-Verteilungen der Parameter und Hyperparameter in mathematischer Form angegeben werden:

- |                            |  |
|----------------------------|--|
| Datenmodell Becken 1:      | $i$ -ter Messwert Becken 1 $\sim \text{Normal}(\mu_1, \sigma)$ |
| Datenmodell Becken 2:      | $i$ -ter Messwert Becken 2 $\sim \text{Normal}(\mu_2, \sigma)$ |
| $\vdots$                   | $\vdots$   |
| Datenmodell Becken 7:      | $i$ -ter Messwert Becken 7 $\sim \text{Normal}(\mu_7, \sigma)$ |
| Prior Parameter $\sigma$ : | $\ln(\sigma) \sim \text{Uniform}(\ln(10); \ln(1000))$          |
| Prior Durchschnittsmassen: | $\mu_1, \mu_2, \dots, \mu_7 \sim \text{Normal}(a, b)$          |
| Prior Hyperparameter:      | $a \sim \text{Normal}(1500, 10000)$                            |
| Prior Hyperparameter:      | $b \sim \text{Halb-Cauchy}(\text{scale} = 2,5)$                |

**Tab. 14.3** Resultate Becken einzeln betrachtet und mit hierarchischem Modell kombiniert

Becken	1	2	3	4	5	6	7
Durchschnittsmasse (einzeln)	1695	1233	1514	1444	1910	1370	1430
SE (einzeln)	(92)	(101)	(78)	(42)	(300)	(71)	(84)
Durchschnittsmasse (gepoolt)	1525	1472	1498	1488	1526	1478	1486
SE (gepoolt)	(70)	(74)	(48)	(47)	(89)	(57)	(50)

**Abb. 14.9** Auswirkung des hierarchischen Modells auf die einzelnen Schätzungen der plausibelsten durchschnittlichen Massen

Wegen der Hyperparameter spricht man von einem *hierarchischen Modell* (engl. *hierarchical model*, auch *multilevel model* oder *mixed effects model*). Daraus bildet ein Statistikprogramm mit der Formel Likelihood  $\times$  Prior die A posteriori-Verteilungen der Parameter und Hyperparameter. Tab. 14.3 zeigt die Resultate. Mit dem hierarchischen Modell, das alle Becken einbezieht und kombiniert, wird der plausibelste Wert von  $\mu_5$  (ein Becken mit sehr kleinem Stichprobenumfang) nun deutlich nach unten korrigiert! Man spricht von *Shrinkage*. Abb. 14.9 illustriert dies. Alle durchschnittlichen Massen  $\mu_1, \dots, \mu_7$  sind etwa gleich. Dies ist ein Zeichen, dass sie kaum von beckenspezifischen Faktoren beeinflusst werden.<sup>8</sup> Der Standardfehler der Schätzung der Durchschnittsfischmasse des Beckens 2 ist nun deutlich kleiner. Alle Standardfehler der Schätzungen sind einheitlicher und variieren zwischen 46 und 82 g. □

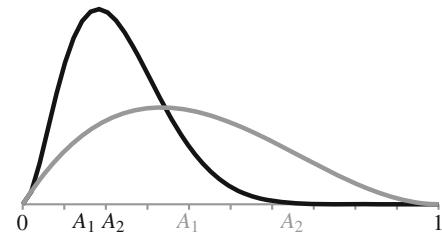
**Beispiel 14.10 (Bevölkerungsbezogenes Krebsregister Bayern)** Beim Beispiel 14.5 möchte man das Risiko von Männern in den Landkreisen von Bayern bestimmen, an Hodenkrebs zu erkranken. Verhindert werden soll, dass bei den schwach bevölkerten Landkreisen zu tiefe oder zu hohe Risiken berechnet werden. Man nimmt als Vorwissen an, dass der Anteil  $A_i$  der Männern im Landkreis  $i$  an Hodenkrebs zu erkranken, von Faktoren abhängt, die landeskreisspezifisch sind und landeskreisunabhängig sind. Eine Beta-Verteilung mit Parametern  $\alpha > 0$  und  $\beta > 0$  beschreibt, wie die  $A_1, A_2, \dots$  kombiniert sind:

$$\text{Prior: } A_1, A_2, \dots \sim \text{Beta}(\alpha, \beta)$$

Die Parameter  $\alpha$  und  $\beta$  sind die Hyperparameter des hierarchischen Modells. Abb. 14.10 illustriert die Situation. Sind  $\alpha$  oder  $\beta$  gross, so ist die Dichtefunktion schmal. Eine Person,

<sup>8</sup> Die wahrscheinlichsten Werte für den Hyperparameter  $b$  ist  $b_0 = 26,5$  g. Die durchschnittlichen Fischmassen der Becken streuen also in einem Bereich mit Breite  $2 \cdot 26,5$  g = 53 g.

**Abb. 14.10** A priori-Verteilung für die Anteile  $A_1$  und  $A_2$  bei verschiedenen Werten von  $\alpha$  und  $\beta$ :  $\alpha = 3, \beta = 10$  (die Anteile  $A_1$  und  $A_2$  liegen in der Nähe von 0,2) und  $\alpha = 2, \beta = 3$  (graue Kurve: die Anteile  $A_1$  und  $A_2$  könnten sehr verschieden sein)



die denkt, dass landeskreisspezifische Faktoren kaum Einfluss auf die einzelnen Anteile haben, könnte solche Werte für  $\alpha$  oder  $\beta$  wählen. Sind aber  $\alpha$  und  $\beta$  eins, so erhält man eine Gleichverteilung. Eine vielleicht gute Wahl für eine Person, die landeskreisspezifische Faktoren als wesentlich ansieht. Die Hyperparameter werden sinnvollerweise wie beim obigen Beispiel mit der Bayes-Regel Likelihood  $\times$  Prior berechnet. Sind  $a_i$  die Anzahl Erkrankungen und  $n_i$  die Anzahl männlicher Personen im Landkreis  $i$ , so ist bei  $N$  Landkreisen die A posteriori-Verteilung der Parameter  $A_1, A_2, \dots$  und der Hyperparameter  $\alpha, \beta$ :

$$\text{pdf}(A_1, A_2, \dots, \alpha, \beta \mid \text{Daten}) = \underbrace{\left( \prod_{i=1}^N A_i^{a_i} \cdot (1 - A_i)^{n_i - a_i} \right)}_{\text{Likelihood aus Bernoulli-Modell}} \cdot \underbrace{\mathbb{P}(A_1, A_2, \dots, \alpha, \beta)}_{\text{Prior}}$$

Der Prior lautet mit der Multiplikationsregel:

$$\begin{aligned} \mathbb{P}(A_1, A_2, \dots, \alpha, \beta) &= \mathbb{P}(A_1, A_2, \dots \mid \alpha, \beta) \cdot \mathbb{P}(\alpha, \beta) \\ &= \underbrace{\mathbb{P}(A_1 \mid \alpha, \beta)}_{\text{Beta-Verteilung}} \cdot \underbrace{\mathbb{P}(A_2 \mid \alpha, \beta)}_{\text{Beta-Verteilung}} \cdot \dots \cdot \underbrace{\mathbb{P}(A_N \mid \alpha, \beta)}_{\text{Beta-Verteilung}} \cdot \underbrace{\mathbb{P}(\alpha, \beta)}_{\text{Prior}} \end{aligned}$$

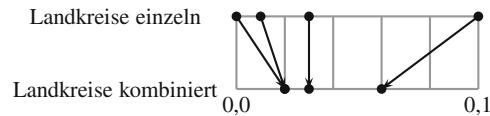
Wie soll man den Prior zu den Hyperparametern  $\alpha$  und  $\beta$  wählen? Als Vorinformation habe man: landeskreisspezifische Faktoren könnten eine Rolle spielen. Daher sind sehr grosse Werte von  $\alpha$  und  $\beta$  tendenziell weniger plausibel als kleine Werte. Eine konservative Wahl dazu ist die halbe Cauchyverteilung mit Skalierung 2,5 (eine Diskussion dazu findet man wiederum in [3]):

$$\mathbb{P}(\alpha, \beta) = \mathbb{P}(\alpha) \cdot \mathbb{P}(\beta) \propto \frac{1}{1 + (\alpha/2,5)^2} \cdot \frac{1}{1 + (\beta/2,5)^2}$$

Mit einer MCMC-Simulation werden daraus die A posteriori-Verteilungen von  $A_1, A_2, \dots$  sowie ihre plausibelsten Werte berechnet. Tab. 14.4 zeigt die Resultate an einem Beispiel mit vier Landkreisen. Betrachtet man die Landkreise einzeln, so sind die plausibelsten Werte für den Anteil der Krebsrate  $1/10 = 0,1, 0/20 = 0,0, 1/100 = 0,01$  und

**Tab. 14.4** Resultate einzeln betrachtet oder mit hierarchischem Modell kombiniert

	Landkreis 1	Landkreis 2	Landkreis 3	Landkreis 4
Anzahl Krebs	1	0	1	12
Stichprobenumfang	10	20	100	400
pl. Krebsrate (Landkreise einzeln)	0,10	0,00	0,01	0,03
pl. Krebsrate (hierarchisches Modell)	0,06	0,02	0,02	0,03

**Abb. 14.11** Auswirkung des hierarchischen Modells auf einzelne Schätzung der plausibelsten Krebsrate

$12/400 = 0,03$ . Mit dem hierarchischen Modell, das alle Landkreise kombiniert, wird der plausibelste Wert von  $A_1$  (ein Landkreis mit kleinem Stichprobenumfang) nun deutlich nach unten in Richtung des plausibelsten Werts von  $A_4$  (ein Landkreis mit grossem Stichprobenumfang) korrigiert! Dieses Shrinkage zeigt Abb. 14.11.  $\square$

Weitere Beispiele, Diskussionen und ausführliche Informationen zu hierarchischen Modellen findet man in den Büchern [3] und [5].

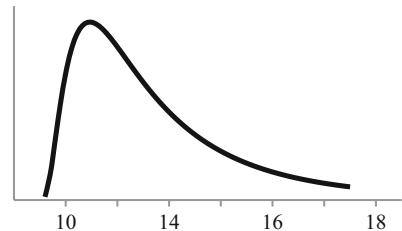
## Reflexion

**14.1** Bei gegebener Information  $\mathcal{K}$  macht eine Person Aussagen zu einer nicht direkt messbaren Grösse  $m$ , die zwischen 10 und 20 liegt, mit einem Wahrscheinlichkeitsmodell:

$$\text{pdf}(m \mid \mathcal{K}) \propto (m - 10)^3 \cdot (20 - m)^4$$

- Zeichnen Sie den Graphen der Dichtefunktion. Wie ist das Modell verteilt? Wie lautet der Modus?
- Bestimmen Sie die Wahrscheinlichkeit, dass  $m$  zwischen 15 und 16 liegt. Wie gross ist die Wahrscheinlichkeit, dass  $m$  grösser als 12,5 ist? Wie lautet die Wahrscheinlichkeit, dass  $m$  zwischen 12 und 15 liegt?
- Wie lautet die beobachtete Information des Wahrscheinlichkeitsmodells?
- Approximieren Sie das Modell mit einer Normalverteilung. Beurteilen Sie anhand der Graphen der beiden Dichtefunktionen, ob die Approximation gut ist. Wie lautet der Standardfehler SE von  $m$ ? Bestimmen Sie daraus Wahrscheinlichkeitsintervalle für  $m$  zum Niveau von etwa 0,68 und etwa 0,95.

**Abb. 14.12** Graph der A posteriori-Dichtefunktion, um die Plausibilität zu  $h$  zu beschreiben



**14.2** Die Plausibilität zu einem Parameter  $K > 0$  beschreibt eine Person mit einem Wahrscheinlichkeitsmodell mit Dichtefunktion

$$\text{pdf}(K \mid \text{Daten}) = \frac{1}{(K+1)^2} \quad \text{für } K \geq 0$$

- (a) Zeichnen Sie den Graphen der Dichtefunktion.
- (b) Berechnen Sie die beobachtete Information des Modells.
- (c) Ist es sinnvoll, das Modell mit einer Normalverteilung zu approximieren?

**14.3** Abb. 14.12 zeigt den Graphen der A posteriori-Dichtefunktion, um die Plausibilität zu einer Grösse  $h$  zu beschreiben.

- (a) Wie ist das Modell verteilt? Wie lautet der Modus? Wo ungefähr liegt der Median?
- (b) Zeichnen Sie in die Graphik die Dichtefunktion der approximierenden Normalverteilung ein. Wie gross ist der Standardfehler von  $h$  ungefähr? Kreuzen Sie die richtige Antwort an:

SE  $\approx 0,3$      SE  $\approx 1,8$      SE  $\approx 3,5$      SE  $\approx 10,4$

- (c) Bestimmen Sie eine Präzisionsangabe von  $h$  in der Form  $h \pm \text{SE}$ .

**14.4** Eine Person formuliert mit der Laplace-Approximation die Plausibilität zu einer Grösse  $\alpha$ :

$$\alpha = \alpha_0 \pm \text{SE}(\alpha) = 15,6 \pm 1,4$$

Berechnen Sie mit einem Statistikprogramm die Wahrscheinlichkeiten  $\mathbb{P}(\alpha \leq 17,5)$ ,  $\mathbb{P}(\alpha < 16,0)$ ,  $\mathbb{P}(\alpha > 14,1)$ ,  $\mathbb{P}(\alpha > 10)$  und  $\mathbb{P}(15,0 < \alpha \leq 16,0)$ .

**14.5** Von einer Grösse  $m$  hat man die Präzisionsangabe mit dem Standardfehler:

$$m = m_0 \pm \text{SE}(m) = m_0 \pm \delta m = 60,1 \pm 3,2$$

Bestimmen Sie Präzisionsangaben mit dem Standardfehler für die Grössen  $m^2$ ,  $4m^2$ ,  $1/m$ ,  $\ln m$  und  $m/(1+m)$ .

**14.6** Eine Ingenieurin hat mit der Laplace-Approximation die Plausibilität zweier Größen  $A$  und  $B$  bestimmt:

$$A = A_0 \pm \delta A = 10,0 \pm 0,2 \quad B = B_0 \pm \delta B = 2,0 \pm 0,1$$

- (a) Wie lauten Präzisionsangaben für die Größen  $A^3$  und  $1/A$ ?
- (b) Bestimmen Sie Wahrscheinlichkeitsintervalle zum Niveau von etwa 0,68 und etwa 0,95 der Größen  $A + B$ ,  $A \cdot B$ ,  $A - B$  und  $B/A$ . Nehmen Sie dazu an, dass  $A$  und  $B$  unabhängig sind.

**14.7** Eine Person misst in unabhängiger Art die Dimensionen  $L$ ,  $B$  und  $H$  einer rechteckigen Schachtel. Sie erhält:

$$L = L_0 \pm \delta L = (154,2 \pm 0,5) \text{ mm}$$

$$B = B_0 \pm \delta B = (82,3 \pm 0,5) \text{ mm}$$

$$H = H_0 \pm \delta H = (70,4 \pm 0,5) \text{ mm}$$

Bestimmen Sie Wahrscheinlichkeitsintervalle zum Niveau von etwa 0,68 und 0,95 für das Volumen und die Oberfläche der Schachtel.

**14.8** Die Plausibilität zu drei Größen  $a$ ,  $b$  und  $c$  wird mit dem Standardfehler angegeben:

$$a = a_0 \pm \delta a = 5,0 \pm 0,4 \quad b = 7,0 \pm 0,3 \quad c = 15,5 \pm 0,2$$

Die A posteriori-Verteilungen der drei Größen sind unabhängig.

- (a) Bestimmen Sie Wahrscheinlichkeitsintervalle zum Niveau von etwa 0,95 für die Größen  $a$ ,  $b$  und  $c$ .
- (b) Berechnen Sie Wahrscheinlichkeitsintervalle zum Niveau von etwa 0,68 der Größen  $a + b^2$ ,  $a + b + c$  und von  $a - b + a \cdot c^2$ .

**14.9** Ein Zitat aus der Zeitung *20 Minuten* vom 10.2.2011:

Ausgesprochen freundlich, mit Fantasie gesegnet und ein schneller Denker soll er sein, der Wassermann. Glaubt man einer Auswertung der Versicherungsgesellschaft Allianz Suisse, sind Wassermänner auch die besten Autofahrer der Schweiz. Sowohl bei Haftpflichtfällen als auch bei Vollkaskoschäden hatten die Wassermänner 2010 eine um durchschnittlich 3,5% niedrigere Schadensfrequenz als der Durchschnitt.

Das Zitat bezieht sich auf eine Studie mit einem Stichprobenumfang von 40 000 Personen mit den Resultaten in Tab. 14.5.

- (a) Bestimmen Sie die Wahrscheinlichkeit, dass die Schadensfrequenz beim Sternzeichen Wassermann kleiner ist als beim Sternzeichen Steinbock.

**Tab. 14.5** Sternzeichen und Schadensfrequenz (mit Standardfehler) (aus 20 Minuten, 10.2.2011)

Sternzeichen	Schadenfrequenz	SE
Wassermann	25,0	(0,9)
Steinbock	27,6	(0,8)
Jungfrau	27,7	(1,0)
Fisch	27,8	(0,9)
Widder	27,8	(1,1)
Krebs	28,6	(0,8)
Löwe	29,4	(0,8)
Skorpion	29,5	(1,0)
Zwilling	29,6	(0,9)
Schütze	29,8	(0,9)
Waage	30,3	(1,1)
Stier	30,4	(1,0)

- (b) Berechnen Sie mit einer Monte-Carlo-Simulation die Wahrscheinlichkeit, dass die Schadensfrequenz beim Sternzeichen Wassermann kleiner ist als bei den Sternzeichen Steinbock und Jungfrau.
- (c) Teilen Sie die Sternzeichen nach statistischen Kriterien in drei Gruppen ein: solche mit tiefer, mittlerer und hoher Schadensfrequenz.

**14.10** In einer Fabrik interessiert man sich für die täglichen Durchschnittsmassen von gefüllten Flaschen. Um die Plausibilität zu diesen Durchschnittsmassen zu bestimmen, werden pro Tag zehn Flaschen gewogen. Tab. 14.6 zeigt die plausibelsten Werte der Durchschnittsmassen (mit ihren Standardfehlern) aus dreizehn Produktionstagen. Ist die Produktion unter statistischer Kontrolle? Zeichnen Sie dazu eine Kontrollkarte mit oberen und unteren Kontrollgrenzen.

**Tab. 14.6** Durchschnittsmassen aus 13 Losen mit Angabe des Standardfehlers

Los	Durchschnittsmasse in g	SE
1	895,3	(2,2)
2	891,2	(3,0)
3	892,4	(0,3)
4	891,1	(2,1)
5	894,2	(2,2)
6	889,7	(1,8)
7	893,9	(3,2)
8	892,8	(2,5)
9	893,2	(2,2)
10	890,5	(1,8)
11	889,7	(0,8)
12	892,1	(1,9)
13	894,3	(2,7)

## Literatur

1. C. Daniel, *Applications of Statistics to Industrial Experimentation* (John Wiley & Sons, New York, 1976)
2. D. Freedman, R. Pisani, R. Purves, *Statistics* 4. Aufl. (W. W. Norton & Company, 2007)
3. A. Gelman, J. B. Carlin, H. S. Stern, D. B. Rubin, *Bayesian Data Analysis* (Chapman & Hall, 2004)
4. H. Hotelling, Some problems in weighing and other experimental techniques, *Am. Math. Stat.* **15**, 297–306 (1944)
5. R. McElrath, *Statistical Rethinking, a Bayesian Course with Examples in R and Stan* (Chapman & Hall, 2016)
6. F. Oser, H. Biedermann, M. Kopp, S. Steinmann, C. Brühwiler, S. Krattenmacher, TEDS-M: Erste Ergebnisse für die Lehrerausbildung in der Deutschschweiz, Medienmitteilung, 15.4.2010
7. D. S. Sivia, J. Skilling, *Data Analysis, a Bayesian Tutorial* 2nd ed. (Oxford Science Publications, 2010)
8. H. Wainer, *Picturing the Uncertain World, How to Understand, Communicate, and Control Uncertainty through Graphical Display* (Princeton University Press, 2009)

„Wie lautet das Urteil?“ fragte der König ungefähr zum zehntenmal.  
„Nein, nein!“ sagte die Königin, „zuerst die Strafe, dann das Urteil!“  
„Schluss mit dem Gefasel!“ sagte Alice laut. „Zuerst die Strafe, wo  
gibt denn so was!“

Lewis Carroll, *Alice im Wunderland* (Insel Taschenbuch, 1973, S. 125)

---

## Zusammenfassung

Verschiedene Regressionsmodelle können den durchschnittlichen Wert einer Zielgröße in Funktion von erklärenden Variablen  $A, B, C, \dots$  oder in Funktion der erklärenden Variablen  $X, Y, \dots$  bestimmen. Es stellt sich dann die Frage: „Welches Modell ist plausibler?“ Mit statistischen Methoden lässt sich diese Frage beantworten und Regressionsmodelle können quantitativ gegeneinander beurteilt werden. Im ersten Teil des Kapitels findet man dazu eine Einführung.

In der medizinischen Forschung und in den Sozialwissenschaften ist es üblich, Hypothesen zu nicht direkt messbaren Größen zu formulieren: „Tabakkonsum bewirkt ein durchschnittlich höheres Risiko an Lungenkrebs zu erkranken.“ oder „Personen zwischen 20 und 30 Jahren haben beim Autofahren höhere Schadensfrequenzen als andere Personen.“ Mit der A posteriori-Verteilung der nicht direkt messbaren Größen kann ausgerechnet werden, wie plausibel solche Hypothesen sind. Eine einführende und kritische Diskussion dazu führt der zweite Abschnitt des Kapitels.

---

## 15.1 Plausibilität von Modellen

Mit einem Regressionsmodell versucht man von Faktoren oder Kovariablen auf eine Zielgröße zu rechnen. So wird in Kap. 13 in Beispiel 13.1 versucht, die durchschnittliche Akkommodationsbreite  $\mu_{\text{Akko}}$  in Funktion des Alters zu bestimmen. Man könnte auch  $\mu_{\text{Akko}}$  in Funktion des Alters und des Geschlechts der Person zu modellieren versuchen. Wäre dies ein besseres Regressionsmodell? Hier ein zweites Beispiel dazu:

**Beispiel 15.1 (Holzeigenschaften)** In Beispiel 3.17 ist ein Projekt vorgestellt, das die Wirkung des Fällzeitpunkts auf die Relativdichte von Holz untersucht. Eine Kovariable, die bekannterweise die Relativdichte beeinflusst, ist die Jahreszeit. Eine Person kann daher versuchen, mit einem Regressionsmodell von der Jahreszeit auf die durchschnittliche Relativdichte zu rechnen. Es gibt Personen, die denken, dass auch der Stand des Mondes auf die Relativdichte wirkt. Solche Personen werden mit einem Regressionsmodell mit den beiden Kovariablen Jahreszeit und Mondstand arbeiten. Es stellt sich die Frage: Welches Modell ist vorzuziehen?  $\square$

Meist versucht man jenem Regressionsmodell  $\mathcal{M}$  den Vorzug zu geben, das die höchste Plausibilität hat.<sup>1</sup> Dazu misst man die Wahrscheinlichkeit  $\mathbb{P}(\mathcal{M} \mid \text{Daten})$ , die Plausibilität des Modells  $\mathcal{M}$ , gegeben die Daten.<sup>2</sup> Um sie zu berechnen, hilft die Regel von Bayes:

$$\mathbb{P}(\mathcal{M} \mid \text{Daten}) = \frac{\mathbb{P}(\text{Daten} \mid \mathcal{M}) \cdot \mathbb{P}(\mathcal{M})}{\mathbb{P}(\text{Daten})}$$

Dabei ist  $\mathbb{P}(\mathcal{M})$  die A priori-Wahrscheinlichkeit für das Modell  $\mathcal{M}$ . Der Nenner im obigen Bruch spielt keine Rolle, wenn man die Plausibilität zweier Modelle  $\mathcal{M}_1$  und  $\mathcal{M}_2$  vergleicht. Es folgt nämlich

$$\underbrace{\frac{\mathbb{P}(\mathcal{M}_1 \mid \text{Daten})}{\mathbb{P}(\mathcal{M}_2 \mid \text{Daten})}}_{\text{A posteriori}} = \underbrace{\frac{\mathbb{P}(\text{Daten} \mid \mathcal{M}_1)}{\mathbb{P}(\text{Daten} \mid \mathcal{M}_2)}}_{\text{BF}_{\mathcal{M}_1:\mathcal{M}_2}} \cdot \underbrace{\frac{\mathbb{P}(\mathcal{M}_1)}{\mathbb{P}(\mathcal{M}_2)}}_{\text{A priori}} \quad (15.1)$$

Den ersten Quotienten links des Gleichheitszeichens nennt man den *Bayes-Faktor*  $\text{BF}_{\mathcal{M}_1:\mathcal{M}_2}$ . Er bestimmt, um wie viel die Plausibilität des Modells  $\mathcal{M}_1$  gegenüber dem Modell  $\mathcal{M}_2$  gewachsen oder gefallen ist, wenn Daten vorhanden sind. Ist er nahe bei eins, hat man nicht viel aus den Daten gelernt: Die A posteriori-Plausibilitäten entsprechen etwa den A priori-Plausibilitäten der beiden Modelle. Ist er sehr gross oder nahe bei null, beeinflussen vor allem die Daten das A posteriori-Verhältnis der Plausibilitäten der beiden Modelle. Der Geophysiker und Statistiker H. Jeffreys empfiehlt die in Tab. 15.1 aufgelistete Interpretation des Bayes-Faktors.<sup>3</sup> Den Bayes-Faktor kann man wie folgt berechnen.<sup>4</sup> Hat das erste Modell  $\mathcal{M}_1$  einen Parameter  $\theta$ , so verbindet das Gesetz der Marginalisierung

<sup>1</sup> Es existieren auch andere Kriterien: Man kann das Regressionsmodell vorziehen, dass zukünftige Messungen der Zielgröße am besten prognostiziert. Siehe dazu [5].

<sup>2</sup> In der frequentistischen Statistik spricht man von der Varianz-Analyse (engl. analysis of variance [ANOVA]). Man berechnet einen *p*-Wert, der zum Teil mit der inversen Wahrscheinlichkeit der Plausibilität  $\mathbb{P}(\mathcal{M} \mid \text{Daten})$  verknüpft ist.

<sup>3</sup> Es gibt auch andere Interpretationen.

<sup>4</sup> Der folgende Text folgt den Erläuterungen von [13] auf den Seiten 78–83. Eine detaillierte Rechnung mit Angaben zu Originalartikeln von Tierney und Kadane, sowie von Jeffreys findet man in [11].

**Tab. 15.1** Interpretation des Bayes-Faktors nach H. Jeffreys (aus [3])

$\text{BF}_{\mathcal{M}_1:\mathcal{M}_2}$ (bzw. $1/\text{BF}_{\mathcal{M}_1:\mathcal{M}_2}$ )	Evidenz für das Modell $\mathcal{M}_1$ (bzw. $\mathcal{M}_2$ )
1 bis 3	Kaum wähnenswert (engl. barely worth mentioning)
3 bis 10	Positiv (engl. substantial)
10 bis 30	Stark (engl. strong)
30 bis 100	Sehr stark (engl. very strong)
Grösser als 100	Entscheidend (engl. decisive)

den Parameter mit dem Modell:

$$\mathbb{P}(\text{Daten} \mid \mathcal{M}_1) = \int \mathbb{P}(\text{Daten und } \theta \mid \mathcal{M}_1) d\theta$$

Mit dem Multiplikationsgesetz kann man den Integranden umformen. Man erhält

$$\mathbb{P}(\text{Daten} \mid \mathcal{M}_1) = \underbrace{\int \underbrace{\mathbb{P}(\text{Daten} \mid \theta)}_{\text{Likelihood}} \cdot \underbrace{\mathbb{P}(\theta)}_{\text{A priori Wissen zu } \theta} d\theta}_{\text{Marginalisierte Likelihood}}$$

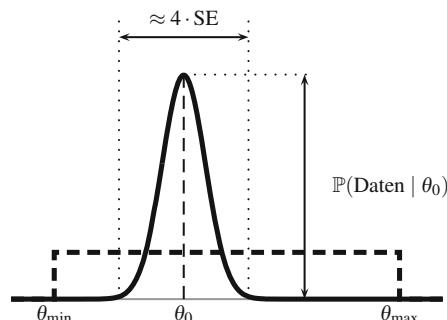
Dieses Integral nennt man die *marginalisierte Likelihood*. Um es zu bestimmen, wollen wir davon ausgehen, dass der Parameter  $\theta$  ein Lageparameter ist, der zwischen zwei Werten  $\theta_{\min}$  und  $\theta_{\max}$  liegt. Weiter habe man vor der Datensammlung minimale Information zu diesem Parameter:

$$\mathbb{P}(\theta) = \text{pdf}(\theta) = \frac{1}{\theta_{\max} - \theta_{\min}} \quad \text{für } \theta_{\min} \leq \theta \leq \theta_{\max}$$

Der Graph der Likelihood  $\mathbb{P}(\text{Daten} \mid \theta)$  ist meist glockenförmig. Abb. 15.1 zeigt die Situation: die Likelihood und der Prior des Parameters  $\theta$ . Die Likelihood-Funktion habe ihr Maximum bei  $\theta_0$ , dem plausibelsten Wert von  $\theta$ . Mit der Methode von Laplace lässt sie sich mit einer Normalverteilung mit Streuung gleich dem Standardfehler  $\text{SE}(\theta)$  approximieren:

$$\mathbb{P}(\text{Daten} \mid \theta) = \mathbb{P}(\text{Daten} \mid \theta_0) \cdot \exp \left\{ -0,5 \cdot \left( \frac{\theta - \theta_0}{\text{SE}(\theta)} \right)^2 \right\}$$

**Abb. 15.1** Plausibilität zum Parameter  $\theta$ : Gestrichelt ist die A priori-Dichtefunktion, ausgezogen die Likelihood dargestellt



Die marginalisierte Likelihood ist deshalb

$$\mathbb{P}(\text{Daten} \mid \mathcal{M}_1) \approx \frac{\mathbb{P}(\text{Daten} \mid \theta_0)}{\theta_{\max} - \theta_{\min}} \cdot \int_{-\infty}^{\infty} \exp \left\{ -0,5 \cdot \left( \frac{\theta - \theta_0}{\text{SE}(\theta)} \right)^2 \right\} d\theta$$

Das verbleibende Integral kann man explizit berechnen.<sup>5</sup> Man erhält

$$\mathbb{P}(\text{Daten} \mid \mathcal{M}_1) \approx \mathbb{P}(\text{Daten} \mid \theta_0) \cdot \frac{\sqrt{2\pi} \cdot \text{SE}(\theta)}{\theta_{\max} - \theta_{\min}} \quad (15.2)$$

Der Bayes-Faktor lässt sich daraus bestimmen. Hat man ein Regressionsmodell  $\mathcal{M}_1$  mit einem Parameter  $a$  und ein zweites Modell  $\mathcal{M}_2$  mit einem Parameter  $b$ , so ist

$$\text{BF}_{\mathcal{M}_1:\mathcal{M}_2} \approx \frac{\mathbb{P}(\text{Daten} \mid a_0)}{\mathbb{P}(\text{Daten} \mid b_0)} \cdot \frac{\text{SE}(a)}{a_{\max} - a_{\min}} \cdot \frac{b_{\max} - b_{\min}}{\text{SE}(b)}$$

Analysiert man diesen Ausdruck, ergibt sich:

- (1) Je besser die Daten zum Modell  $\mathcal{M}_1$  passen, umso höher wird die Spitze  $\mathbb{P}(\text{Daten} \mid a_0)$  der Likelihood. Um so plausibler wird damit das Modell  $\mathcal{M}_1$ . Dies bedeutet:

Ein Modell wird umso plausibler, je besser die Daten zum Modell passen.

Dies scheint sinnvoll.

- (2) Die Differenz  $a_{\max} - a_{\min}$  sagt, wie präzis man den Parameter  $a$  vor der Datensammlung kennt. Der Standardfehler  $\text{SE}(a)$  drückt aus, wie präzis der Parameter  $a$  dank Daten bestimmt ist. Der Quotient  $\text{SE}(a)/(a_{\max} - a_{\min})$  misst deshalb, wie stark man wegen der Daten an Präzision für  $a$  gewonnen hat. Je mehr man an Präzision gewonnen hat, um so kleiner wird dieser Quotient. Dies heisst:

Je grösser die gewonnene relative Präzision für den Parameter ist, um so weniger wird das Modell plausibel.

---

<sup>5</sup> Die Dichtefunktion der Normalverteilung ist

$$\text{pdf}(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -0,5 \cdot \left( \frac{x - \mu}{\sigma} \right)^2 \right\}$$

Weil  $\int_{-\infty}^{\infty} \text{pdf}(x \mid \mu, \sigma) dx = 1$  ist, folgt  $\int_{-\infty}^{\infty} \sqrt{2\pi\sigma^2} \cdot \text{pdf}(x \mid \mu, \sigma) dx = \sqrt{2\pi}\sigma$ .

Dies scheint auf den ersten Blick seltsam. Eine hohe gewonnene relative Präzision bedeutet, dass das Modell sehr „starr“ oder unflexibel wird. Es besteht die Gefahr, dass es zu sehr an die Daten angepasst ist. Die Wahrscheinlichkeit erhöht sich damit, dass zukünftige Messwerte falsch prognostiziert und das Modell abgewertet wird.

Die obige Rechnung lässt sich auch bei Regressionsmodellen mit mehreren Parametern durchführen. Wählt man ein Datenmodell, bei dem die  $n$  Messungen normalverteilt um das Regressionsmodell  $\mathcal{M}_1$  mit  $K_1$  Parametern  $a, b, \dots$  streuen, so ist

$$\mathbb{P}(\text{Daten} \mid \mathcal{M}_1) \approx \mathbb{P}(\text{Daten} \mid a_0, b_0, \dots) \cdot \left( \frac{1}{\sqrt{n}} \right)^{K_1}$$

Dies ist analog zu Gleichung (15.2). Der erste Faktor ist die Likelihood mit den plausibelsten Werten für die Parameter des Regressionsmodells. Der zweite Faktor entspricht dem Term  $\text{SE}(\theta)$  für die  $K_1$  Parameter. Der Standardfehler SE ist nämlich für  $n$  unabhängige Messwerte proportional zu  $1/\sqrt{n}$ . Es ist üblich, den obigen Ausdruck anders zu schreiben:

$$\mathbb{P}(\text{Daten} \mid \mathcal{M}_1) \approx \exp\{-\text{BIC}_1/2\}$$

Dabei ist  $\text{BIC}_1$  das *Bayes Informationskriterium* oder *Schwarz Kriterium* des Modells  $\mathcal{M}_1$  (siehe dazu [12]):

$$\text{BIC}_1 = -2 \cdot \ln[ \underbrace{\mathbb{P}(\text{Daten} \mid a_0, b_0, \dots)}_{\text{Likelihood mit plausibelsten Werten}} ] + K_1 \cdot \ln n$$

Der erste Summand in dieser Formel wird die *Deviance* des Modells genannt. Diese wird berechnet, indem man in die Likelihood die plausibelsten Werte  $a_0, b_0, \dots$  für die Parameter des Modells  $\mathcal{M}_1$  einsetzt, logarithmiert und mit  $-2$  multipliziert. Die Likelihood mit den plausibelsten Werten des Regressionsmodells lässt sich aber mit der Summe  $\text{SS}_{\text{res}}^{(1)}$  der Residuen im Quadrat ausdrücken. Man erhält

$$\text{BIC}_1 = n \cdot \ln \text{SS}_{\text{res}}^{(1)} + n \cdot [\ln(2\pi) - \ln n + 1] + (K_1 + 2) \cdot \ln n$$

Solche Ausdrücke sind in der Praxis beliebt, da Statistikprogramme sie schnell rechnen können. Die Differenz  $\Delta\text{BIC} = \text{BIC}_1 - \text{BIC}_2$  des BIC zwischen zwei *linearen* Regressionsmodellen  $\mathcal{M}_1$  und  $\mathcal{M}_2$  mit  $K_1$  und  $K_2$  Parametern bei  $n$  unabhängigen Messwerten oder Beobachtungen ist somit

$$\Delta\text{BIC}(\mathcal{M}_1 : \mathcal{M}_2) = n \cdot \ln \frac{\text{SS}_{\text{res}}^{(1)}}{\text{SS}_{\text{res}}^{(2)}} + (K_1 - K_2) \cdot \ln n$$

Dabei ist  $\text{SS}_{\text{res}}^{(i)}$  die Summe der (approximativen) Residuen im Quadrat des Regressionsmodells  $i$ . Der Bayes-Faktor  $\text{BF}_{\mathcal{M}_1:\mathcal{M}_2}$  ist dann, wenn die Likelihoods dominant sind oder

vor der Datensammlung nur minimale Information zu den Lage- und Skalenparametern vorhanden ist und mit einer Normalverteilung mit konstanter Streuung als Datenmodell gearbeitet wird:

$$BF_{\mathcal{M}_1:\mathcal{M}_2} \approx \exp \left\{ -\frac{\text{BIC}_1}{2} + \frac{\text{BIC}_2}{2} \right\} = \exp \left\{ -\frac{\Delta \text{BIC}}{2} \right\} \quad (15.3)$$

Welche Daten führen zu einem grossen Bayes-Faktor? Je negativer  $\Delta \text{BIC}$ , umso grösser ist der Bayes-Faktor und umso plausibler wird das Modell  $\mathcal{M}_1$ .

Ist die Summe der Residuen im Quadrat im Modell  $\mathcal{M}_1$  klein, so passt das Modell gut zu den Daten. Der Quotient  $SS_{\text{res}}^{(1)}/SS_{\text{res}}^{(2)}$  wird dann klein. Damit wird der erste Summand in  $\Delta \text{BIC}$  stark negativ. Hat das Modell  $\mathcal{M}_1$  jedoch viel mehr Parameter als das zweite Modell, so wird der zweite Summand in  $\Delta \text{BIC}$  gross. Damit wird ein Modell mit vielen Parametern weniger zuverlässig plausibel. Die Konsequenz aus der Gleichung (15.3) ist:

Ein lineares Regressionsmodell wird umso plausibler, je besser das Modell zu den Daten passt und je weniger Parameter es dazu benötigt.

Im Folgenden wird das Gesagte an zwei Beispielen illustriert:

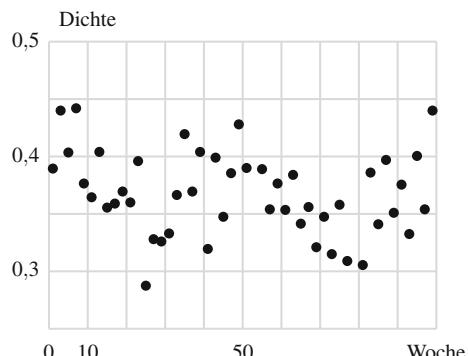
**Beispiel 15.2 (Holzeigenschaften)** In Beispiel 15.1 ist erwähnt, dass die Jahreszeit auf die Relativdichte von Holz wirkt. Abb. 15.2 zeigt gemessene 48 Werte von Relativdichten. Der Abstand zwischen zwei Messungen beträgt zwei Wochen. Man kann versuchen, mit einem Regressionsmodell vom Jahreszeitpunkt  $t$  auf die durchschnittliche Relativdichte  $\mu(t)$  zu rechnen. Ein solches Modell ist

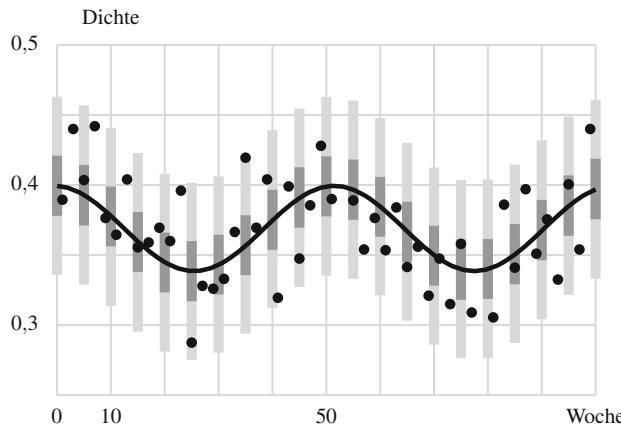
$$\mu(t) = A + b \cdot \sin(\omega_{\text{Jahr}} \cdot t + \varphi)$$

Dabei ist  $\omega_{\text{Jahr}}$  die Kreisfrequenz des Jahreszyklus:

$$\omega_{\text{Jahr}} = \frac{2\pi}{\text{Jahresperiode}} = \frac{2\pi}{52 \text{ Wochen}}$$

**Abb. 15.2** Gemessene Relativdichten einer Holzsorte, gemessen während zweier Jahre





**Abb. 15.3** Relativdichte: Regressionsmodell (Jahreszyklus) für die durchschnittliche Relativdichte und Prognosebänder zum Niveau 0,5 und 0,95 für weitere Messwerte

Das Regressionsmodell ist nicht linear im Parameter  $\varphi$ . Es ist daher besser, es umzuformulieren. Eine Sinusfunktion lässt sich als Summe einer Sinus- und Cosinusfunktion schreiben. Man hat deshalb das äquivalente lineare Regressionsmodell

$$\mu(t) = A + B \cdot \sin(\omega_{\text{Jahr}} \cdot t) + C \cdot \cos(\omega_{\text{Jahr}} \cdot t)$$

Um die Parameter  $A$ ,  $B$  und  $C$  des Regressionsmodells zu bestimmen, braucht man ein Datenmodell. Als Modellannahme gelte: Messwerte streuen normalverteilt um  $\mu(t)$  mit konstanter Streuung  $\sigma$ . Beim Jeffreys' Prior zu  $\sigma$  und bei flachen Prioren zu den Parametern des Regressionsmodells lassen sich die plausibelsten Werte von  $A$ ,  $B$  und  $C$  mit der Methode der kleinsten Quadrate rechnen. Man erhält

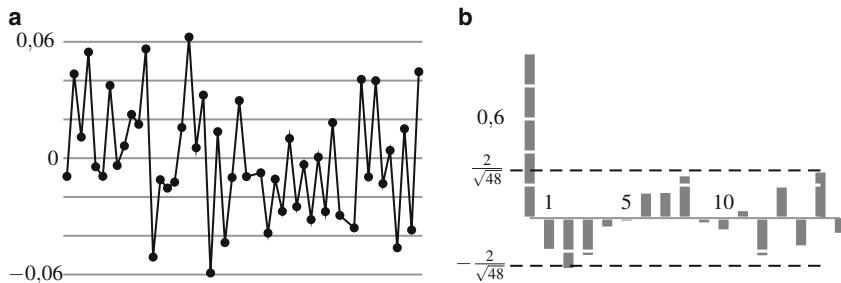
$$\mu(t) = 0,369 - 0,002 \cdot \sin(\omega_{\text{Jahr}} \cdot t) + 0,030 \cdot \cos(\omega_{\text{Jahr}} \cdot t)$$

Dabei ist, wie ein Statistikprogramm schnell liefert,

$$A = A_0 \pm \delta A = 0,369 \pm 0,004$$

Weiter ist  $B = B_0 \pm \delta B = -0,002 \pm 0,006$  und  $C = 0,030 \pm 0,007$ . Die zwei wesentlichen Parameter sind also  $A$  und  $C$ . Der Parameter  $B$  ist sehr klein. Zudem ist er wenig präzis bestimmt. Sein Standardfehler ist grösser als der plausibelste Wert von  $B$ .

Abb. 15.3 zeigt die Daten, das gerechnete Regressionsmodell und Prognosebänder für zukünftige Messwerte zum Niveau 0,5 und 0,95. Ist das Regressionsmodell sinnvoll? Alle Messwerte liegen innerhalb der Prognosebänder. Dies spricht für das Modell. Der Graph der Autokorrelationsfunktion der Residuen und der Residuenplot in Abb. 15.4 zeigen



**Abb. 15.4** Residuenplot (a) und Graph der Autokorrelationsfunktion der Residuen beim Regressionsmodell mit Jahreszyklus (b)

weiter, dass die Likelihood (Messwerte unabhängig) sinnvoll gerechnet ist. Das Regressionsmodell mit dem Jahreszyklus ist daher brauchbar. Die Prognosebänder sind breit. Dies deutet darauf hin, dass zusätzliche Kovariablen auf die Relativdichte wirken.

Es gibt Personen, die denken, dass auch der Stand des Mondes auf die Relativdichte wirkt. Solche Personen werden mit einem Regressionsmodell mit der zusätzlichen Kovariablen Mondstand arbeiten. Hier ein solches lineares Regressionsmodell:

$$\begin{aligned}\mu(t) = & A + B \cdot \sin(\omega_{\text{Jahr}} \cdot t) + C \cdot \cos(\omega_{\text{Jahr}} \cdot t) \\ & + D \cdot \sin(\omega_{\text{Mond}} \cdot t) + E \cdot \cos(\omega_{\text{Mond}} \cdot t)\end{aligned}$$

Im Modell ist

$$\omega_{\text{Mond}} = \frac{2\pi}{\text{Mondperiode}} = \frac{2\pi}{4,219 \text{ Wochen}}$$

die Kreisfrequenz des Mondes. Die Auswertung dieses Modells liefert die Wahrscheinlichkeitsintervalle zum Niveau 0,68 mit den Standardfehlern:

$$A = A_0 \pm \delta A = 0,369 \pm 0,004,$$

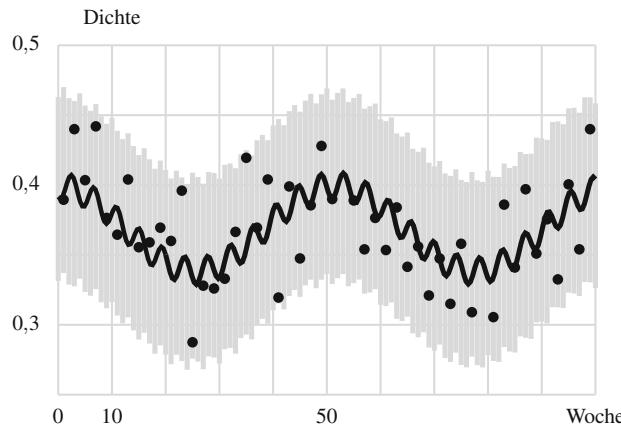
und

$$B = B_0 \pm \delta B = -0,002 \pm 0,006, \quad C = 0,031 \pm 0,006$$

sowie

$$D = -0,005 \pm 0,006, \quad E = -0,008 \pm 0,006$$

Die Parameter  $D$  und  $E$ , die den Einfluss des Mondzyklus auf die Relativdichte beschreiben, sind sehr klein. So ist der Jahreszyklus, der beim Parameter  $C$  sichtbar ist, rund fünfmal höher. Zudem sind die beiden Parameter sehr unpräzis bestimmt. Abb. 15.5 zeigt die Daten, das gerechnete Regressionsmodell und Prognosebänder für zukünftige Messwerte zum Niveau 0,95. Ist dieses Regressionsmodell sinnvoll? Wiederum liegen alle Messwerte innerhalb der Prognosebänder. Dies spricht für das Modell. Die ausgeprägten



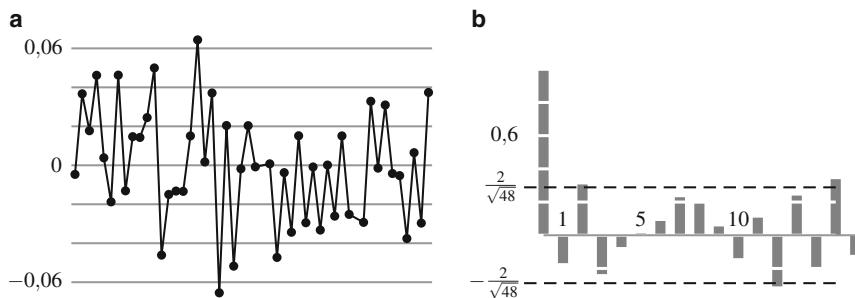
**Abb. 15.5** Relativdichte: Regressionsmodell (Jahres- und Mondzyklus) und Prognosebänder zum Niveau 0,95 für weitere Messwerte

kleinen Schwingungen könnten darauf hindeuten, dass das Regressionsmodell überangepasst ist. Der Graph der Autokorrelationsfunktion der Residuen und der Residuenplot in Abb. 15.6 zeigen, dass die Likelihood (Messwerte unabhängig) sinnvoll gerechnet ist. Das Regressionsmodell mit dem Jahres- und Mondzyklus ist daher ebenfalls sinnvoll.

Es stellt sich die Frage: Welches Modell ist plausibler? Aus den Regressionsmodellen kann man die (approximativen) Residuen berechnen. Daraus erhält man die Summe der Residuen im Quadrat für die beiden Regressionsmodelle:

$$SS_{\text{res}}^{(\text{Jahr})} = 0,04231, \quad SS_{\text{res}}^{(\text{Jahr, Mond})} = 0,04010$$

Die Summe ist kleiner beim Modell mit dem Jahres- und Mondzyklus. Dies ist nicht erstaunlich, besitzt doch dieses Modell mehr Parameter. Deshalb kann es besser an die



**Abb. 15.6** Residuenplot (a) und Graph der Autokorrelationsfunktion der Residuen beim Regressionsmodell mit Jahres- und Mondzyklus (b)

Messwerte angepasst werden. Die Differenz  $\Delta\text{BIC}$  des BIC ist

$$\Delta\text{BIC}(\mathcal{M}_{\text{Jahr}} : \mathcal{M}_{\text{Jahr, Mond}}) = 48 \cdot \ln \frac{0,04231}{0,04010} + (3 - 5) \cdot \ln 48 = -5,176$$

Der Bayes-Faktor – approximativ berechnet mit dem BIC – ist

$$\text{BF}_{\mathcal{M}_{\text{Jahr}}:\mathcal{M}_{\text{Jahr, Mond}}} \approx \exp \left\{ -\frac{-5,176}{2} \right\} = 13,30$$

Gemäss Tab. 15.1 ist dies eine starke Evidenz für das einfachere Modell ohne den Mondzyklus. Das zeigt auch die weitere Rechnung. Nach Gleichung (15.1) ist

$$\frac{\mathbb{P}(\mathcal{M}_{\text{Jahr}} \mid \text{Daten})}{\mathbb{P}(\mathcal{M}_{\text{Jahr, Mond}} \mid \text{Daten})} \approx 13,30 \cdot \frac{\mathbb{P}(\mathcal{M}_{\text{Jahr}})}{\mathbb{P}(\mathcal{M}_{\text{Jahr, Mond}})}$$

Vor der Datensammlung nimmt eine Person an, dass beide Modelle gleich plausibel sind:

$$\mathbb{P}(\mathcal{M}_{\text{Jahr}}) = \mathbb{P}(\mathcal{M}_{\text{Jahr, Mond}}) = 0,5$$

Daraus erhält man die A posteriori-Wahrscheinlichkeiten für die beiden Modelle:<sup>6</sup>

$$\mathbb{P}(\mathcal{M}_{\text{Jahr}} \mid \text{Daten}) \approx 0,93, \quad \mathbb{P}(\mathcal{M}_{\text{Jahr, Mond}} \mid \text{Daten}) \approx 0,07$$

Das Modell ohne den Mondzyklus ist klar plausibler als das Modell mit dem Mondzyklus.  $\square$

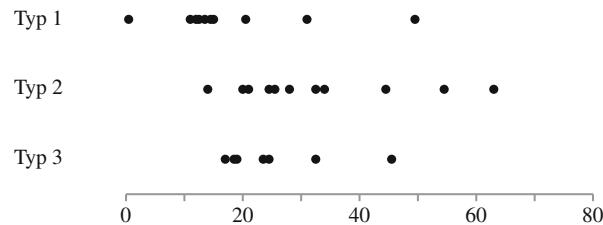
**Beispiel 15.3 (Erosion von Seeufern)** Seeufer in der Schweiz werden mit verschiedenen Vegetationstypen bepflanzt, um Erosion zu verhindern. In einer Untersuchung aus dem Jahr 2008 wurde versucht, den Erosionsfaktor am Ufer des Bielersees für drei verschiedene Vegetationstypen zu bestimmen. In Tab. 15.2 finden sich die Messwerte. Diese streuen jeweils um die typenbezogenen Erosionsfaktoren, da Kovariablen, wie Windrichtungen, Wellenstärken und Wasserstand, vorhanden sind. Am besten lassen sich die Messwerte mit Box & Whiskerplots oder mit Punkten wie in Abb. 15.7 darstellen. Die Abbildung zeigt, dass die Messwerte in allen drei Typen etwa gleich streuen. Die wichtige Frage der

**Tab. 15.2** Erosionsfaktor am Ufer des Bielersees bei drei verschiedenen Vegetationstypen

Typ 1	20,5	14,0	49,5	12,0	31,0	12,5	4,5	15,0	11,0	13,5
Typ 2	20,0	21,0	63,0	32,5	28,0	24,5	44,5	34,0	14,0	25,5
Typ 3	32,5	24,5	18,5	19,0	23,5	17,0	45,5			

<sup>6</sup> Es ist  $\mathbb{P}(\mathcal{M}_{\text{Jahr}}) + \mathbb{P}(\mathcal{M}_{\text{Jahr, Mond}}) = 1$  und  $\mathbb{P}(\mathcal{M}_{\text{Jahr}}) \approx 13,30 \cdot \mathbb{P}(\mathcal{M}_{\text{Jahr, Mond}})$ .

**Abb. 15.7** Visualisierung der Daten aus Tab. 15.2



Untersuchung ist: Haben die Vegetationstypen Einfluss auf den durchschnittlichen Erosionsfaktor?

Um diese Frage zu beantworten, braucht man Regressionsmodelle. Ein Modell, das den durchschnittlichen Erosionsfaktor  $\mu(i)$  in Funktion der Vegetationsgruppe  $i$  beschreibt, ist das lineare Regressionsmodell  $\mathcal{M}_{\text{Veg}}$

$$\mathcal{M}_{\text{Veg}} : \quad \mu(i) = \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3$$

Dabei sind  $X_1$ ,  $X_2$  und  $X_3$  kategoriale Variablen. Sie bedeuten:  $X_i = 1$  „Der Messwert stammt aus dem Typ  $i$ .“ und  $X_i = 0$  heißt „Der Messwert ist nicht aus der Gruppe  $i$ .“<sup>7</sup>

Ein zweites Modell, dass die obigen Kovariablen auslässt, ist das konstante (lineare) Regressionsmodell  $\mathcal{M}_{\text{null}}$ :

$$\mathcal{M}_{\text{null}} : \quad \mu(i) = \beta$$

Es besagt, dass alle Vegetationsgruppen den gleichen durchschnittlichen Erosionsfaktor  $\beta$  verursachen.

Nimmt man an, dass die Messwerte normalverteilt um die Regressionsmodelle streuen, lassen die plausibelsten Werte des Regressionsmodells  $\mathcal{M}_{\text{Veg}}$  mit der Methode der kleinsten Quadrate bestimmen. Man erhält

$$\beta_1 = 18,350 \quad \beta_2 = 32,864 \quad \beta_3 = 25,786$$

Daraus erhält man die Summe der Residuen im Quadrat  $\text{SS}_{\text{res}}^{(\text{Veg})} = 4449,0$ . Beim Modell  $\mathcal{M}_{\text{null}}$  ist der plausibelste Wert von  $\beta$  gleich 25,911 und man hat  $\text{SS}_{\text{res}}^{(\text{null})} = 5552,5$ . Somit ist

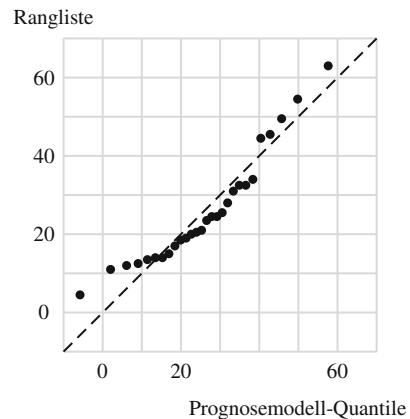
$$\Delta \text{BIC}(\mathcal{M}_{\text{Veg}} : \mathcal{M}_{\text{null}}) = 28 \cdot \ln \frac{4449,0}{5552,5} + (3 - 1) \cdot \ln 28 = 0,46$$

Der Bayes-Faktor – approximativ berechnet mit dem BIC – ist  $\exp\{-0,46/2\} = 0,79$ . Nach Gleichung (15.1) ist damit

$$\frac{\mathbb{P}(\mathcal{M}_{\text{Veg}} \mid \text{Daten})}{\mathbb{P}(\mathcal{M}_{\text{null}} \mid \text{Daten})} \approx 0,79 \cdot \frac{\mathbb{P}(\mathcal{M}_{\text{Veg}})}{\mathbb{P}(\mathcal{M}_{\text{null}})}$$

<sup>7</sup> Beliebt ist es, das Modell  $\mathcal{M}_{\text{Veg}}$  anders zu schreiben:  $\mu(i) = \alpha + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3$  mit  $\beta_1 + \beta_2 + \beta_3 = 0$ . Mit dieser Schreibweise messen die  $\beta_i$  die Abweichungen des Lageparameters der Gruppe  $i$  vom Gesamtmittel  $\alpha$ . Das Modell hat drei Parameter, da die Summe der  $\beta$ 's null ist.

**Abb. 15.8** QQ-Plot der Erosionsfaktoren beim Datenmodell der Normalverteilung



Fazit: Die Daten haben nicht viel dazu beigetragen, einen Entscheid zugunsten eines der beiden Modelle zu fällen. Nimmt man vor der Datensammlung an, dass beide Modelle gleich plausibel sind, so ist

$$\mathbb{P}(\mathcal{M}_{\text{Veg}}) = \mathbb{P}(\mathcal{M}_{\text{null}}) = 0,5$$

Damit erhält man die A posteriori-Wahrscheinlichkeiten für die beiden Modelle:

$$\mathbb{P}(\mathcal{M}_{\text{Veg}} \mid \text{Daten}) \approx 0,44, \quad \mathbb{P}(\mathcal{M}_{\text{null}} \mid \text{Daten}) \approx 0,56$$

Tendenziell ist das einfache Modell  $\mathcal{M}_{\text{null}}$  deshalb empfehlenswerter. Mit dem einfachen Modell hat der Vegetationstyp keinen Einfluss auf den durchschnittlichen Erosionsfaktor. Aus diesem Modell kann man das Prognosemodell für weitere Messwerte berechnen. Der QQ-Plot in Abb. 15.8 zeigt aber, dass das gewählte Datenmodell der Normalverteilung diskutierbar ist. Vor allem die kleinen gemessenen Erosionsfaktoren sind mit der Normalverteilung schlecht beschrieben.  $\square$

**Beispiel 15.4 (Fischzucht)** In der Fischzucht der Topenhaus Frutigen AG werden Störe aufgezogen. Dabei ist es wichtig zu wissen, wie gross die Massen der Fische in den Aufzuchtbecken sind. Störe reagieren sehr empfindlich auf menschliche Berührungen und werden daher nicht mit Waagen gewogen. Man will die Massen der Störe über Fotoaufnahmen ihrer Kontur (siehe Abb. 15.9) bestimmen. Tab. 15.3 zeigt die Massen, Kontourflächen inklusive Flossen und die Kontourflächen ohne Flossen („Miniflächen“) von 10 Stören. Zwei einfache Regressionmodelle  $\mathcal{M}_1$  und  $\mathcal{M}_2$ , um von den Flächen auf die durchschnittlich erwartbare Masse  $\mu_M$  zu rechnen, sind:

$$\mathcal{M}_1 : \quad \mu_M = a + b \cdot \text{Fläche} \quad \mathcal{M}_2 : \quad \mu_M = c + d \cdot \text{Minifläche}$$

**Tab. 15.3** Massen (in Kilogramm), Kontourflächen und Miniflächen (in dm<sup>2</sup>) von 10 Stören aus der Fischzucht des Tropenhauses Frutigen AG (KTI 9029, PFWI-IW, 2007)

Masse	3,40	2,80	2,75	2,15	3,25	2,45	1,70	2,35	1,25	3,10
Fläche	7,0822	6,6232	6,2812	5,0348	6,8419	5,5685	4,6208	5,2712	3,8865	6,4602
Minifläche	6,9628	6,1618	5,7058	4,6643	6,6207	4,8808	4,3434	4,8120	3,3245	6,2272

**Abb. 15.9** Kontourfläche eines Störs mit einer Länge von 64 cm

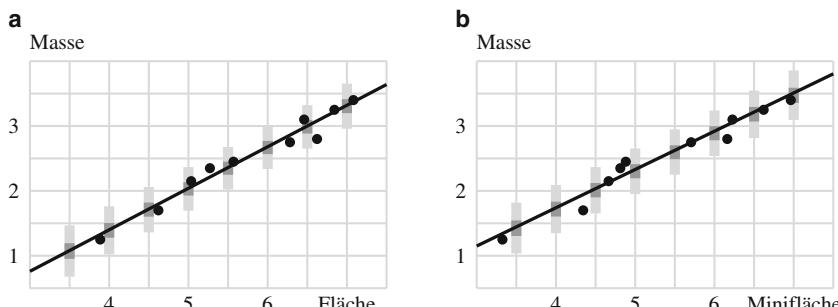


Nimmt man an, dass die Messwerte normalverteilt um die Regressionsmodelle streuen, so kann man die plausibelsten Werte für die Parameter  $a$ ,  $b$ ,  $c$  und  $d$  mit der Methode der kleinsten Quadrate berechnen. Man erhält die folgenden Wahrscheinlichkeitsintervalle zum Niveau 0,95:

$$\begin{aligned} \mathcal{M}_1 : \quad a &= -1,16 \pm 0,58, \quad b = 0,64 \pm 0,10 \\ \mathcal{M}_2 : \quad c &= -0,62 \pm 0,54, \quad d = 0,59 \pm 0,10 \end{aligned}$$

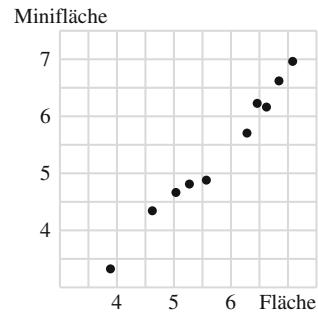
Die wichtigen Parameter  $b$  und  $d$  sind mit einem brauchbaren relativen Fehler von  $\pm 17\%$  bestimmt. Abb. 15.10 zeigt die Daten, die gerechneten Regressionsmodelle und Prognosebänder für zukünftige Messwerte zu den Niveaus 0,5 und 0,95. Beide Modelle scheinen gut und es ist möglich, mit ihnen aus den fotografierten Flächen die Masse eines Fischs mit einer Präzision von  $\pm 0,4\text{ kg}$  zu bestimmen. Die Summe der Residuen im Quadrat der Modelle lauten:

$$\text{SS}_{\text{res}}^{\mathcal{M}_1} = 0,150, \quad \text{SS}_{\text{res}}^{\mathcal{M}_2} = 0,170$$



**Abb. 15.10** Regressionsmodell und Prognosebänder zum Niveau 0,5 und 0,95 für weitere Messwerte: mit Flächen (a) und mit Miniflächen (b)

**Abb. 15.11** Streudiagramm der Flächen und Miniflächen



Somit ist

$$\Delta \text{BIC}(\mathcal{M}_1 : \mathcal{M}_2) = 10 \cdot \ln \frac{0,150}{0,170} + (2 - 2) \cdot \ln 10 = -1,26$$

Der Bayes-Faktor – approximativ berechnet mit dem BIC – ist  $\exp\{1,26/2\} = 1,88$ . Nach Tab. (15.1) unterscheiden sich die beiden Modelle kaum nennenswert.

Man könnte nun versuchen, die Massen mit beiden Flächen-Faktoren besser zu prognostizieren. Dazu kann man das Regressionsmodell

$$\mathcal{M}_3 : \mu_M = \alpha + \beta \cdot \text{Fläche} + \gamma \cdot \text{Minifläche}$$

betrachten. Es ergeben sich die folgenden Wahrscheinlichkeitsintervalle zum Niveau 0,95 für die Parameter des Modells:

$$\alpha = -0,970 \pm 0,831 \quad \beta = 0,397 \pm 0,731 \quad \gamma = 0,224 \pm 0,672$$

Das Modell ist kaum benutzbar, da die Parameter relative Fehler von über  $\pm 100\%$  besitzen. Was ist hier passiert? Die beiden Faktoren „Flächen“ und „Miniflächen“ sind stark korreliert. Man hat für den empirischen Korrelationskoeffizienten nach Pearson

$$\rho_{\text{emp}}(\text{Flächen}, \text{Miniflächen}) = 0,99$$

Dies verdeutlicht auch Abb. 15.11. Aus der Fläche (bzw. Minifläche) kann man somit die Minifläche (bzw. Fläche) approximativ berechnen. Die zwei Faktoren zusammen besitzen nicht mehr Information als ein einzelner Faktor. Die Regel von Bayes, die dazu dient, aus den Daten zu lernen, ist daher nicht in der Lage, beide Faktoren mit einer guten Präzision zu trennen. Die folgende Argumentation verdeutlicht dies: Ist  $\Delta$  die Differenz zwischen der Fläche und der Minifläche, so lautet das Modell  $\mathcal{M}_3$

$$\mu_M = \alpha + (\beta + \gamma) \cdot \text{Minifläche} + \beta \cdot \Delta$$

Die Fläche der Flossen  $\Delta$  trägt nur minimal zur Masse bei und ist vernachlässigbar. Dies entspricht somit dem Modell  $\mathcal{M}_2$  mit

$$\beta + \gamma = d = 0,59 \pm 0,10$$

Die Parameter  $\beta$  und  $\gamma$  des Regressionsmodells  $\mathcal{M}_3$  können nicht separiert werden. Nur ihre Summe ist bekannt. Sowohl mit den Flächen wie auch mit den Miniflächen lässt sich die Masse eines Störs prognostizieren. Verwendet man aber beide Faktoren gleichzeitig in einem Modell, werden die Modelparameter unsicher.  $\square$

**Eine Bemerkung** Die oben besprochenen Beispiele zeigen einführend, wie Regressionsmodelle verglichen und ausgewählt werden können. Dazu benutzt man die Gleichung (15.1) mit der marginalen Likelihood. Die Gleichung sollte jedoch mit Bedacht benutzt werden. Es kann sein, dass ein komplexes Modell plausibler ausfällt als ein einfaches, aber den Nachteil hat, schwer erklärbar zu sein. Vielleicht ist es auch überangepasst. Oder es können Kovariablen, die im komplexen Modell vorhanden sind, vermengt oder stark miteinander korreliert sein: das Modell wird unsicher, obwohl die Kovariablen die Zielgröße „erklären“. In solchen Fällen kann es vorteilhaft sein, mit verschiedenen einfacheren, vielleicht weniger plausiblern Modellen zu arbeiten. Eine vertiefte und beispielhafte Diskussion zur Modellwahl mit Hilfe des BIC findet man in [11]. Es gibt aber auch andere Informationskriterien als das BIC, um Modelle auszuwählen. Diese versuchen das Modell zu bestimmen, sodass zukünftige Messwerte der Zielgröße mit möglichst kleinem Risiko prognostiziert werden.<sup>8</sup> Eine andere beliebte Methode ist, Modelle zu mitteln. Dies verhindert die „willkürliche“ Wahl eines bestimmten Modells. Sind etwa  $\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_N$  verschiedene Regressionmodelle für eine Zielgröße, deren Plausibilitäten zu eins addieren, so folgt mit dem Gesetz der totalen Wahrscheinlichkeit und dem Multiplikationsgesetz

$$\begin{aligned}\mathbb{P}(\text{Messwert Zielgröße}) &= \sum_{k=1}^N \mathbb{P}(\text{Messwert Zielgröße und Modell } \mathcal{M}_k) \\ &= \sum_{k=1}^N \underbrace{\mathbb{P}(\text{Messwert Zielgröße } | \mathcal{M}_k)}_{\text{Prognose aus Modell } k} \cdot \underbrace{\mathbb{P}(\mathcal{M}_k)}_{\text{Plausibilität des Modells } k}\end{aligned}$$

Um einen Wert der Zielgröße zu prognostizieren, mittelt man über die Prognosen der einzelnen Modelle aus, gewichtet nach der Plausibilität der Modelle. Statistische Programme können dies meist automatisch durchführen. Details und Beispiele dazu findet man in den Büchern [5] und [6].

---

<sup>8</sup> Man spricht von möglichst gutem *out-of-sample prediction*. Im Gegensatz dazu misst das BIC die Plausibilität des Modells bezüglich der gemessenen Daten, also tendenziell das *in-sample prediction*.

## 15.2 Plausibilität von Hypothesen

In Sozial- und Wirtschaftswissenschaften, in der Chemie und in der medizinischen Forschung ist es verbreitet, mit Hypothesen zu nicht direkt messbaren Größen zu arbeiten. Man stellt zwei sich ausschliessende Hypothesen auf (auch als  $H_0$  und  $H_1$  oder als  $H_0$  und  $H_A$  für Null- und Alternativ-Hypothese geschrieben), die interessante, wissenschaftliche Postulate darstellen. Hier Beispiele dazu:

**Beispiel 15.5 (HNV-Indikator)** Beim Beispiel 5.3 interessiert der Anteil  $A$  der 1 km<sup>2</sup>-Flächen in Deutschland, die weniger als 5 % landwirtschaftliche Fläche haben. Politische Entscheidungsträger wünschen sich, dass  $A$  höchstes 0,7 ist. Damit stellt sich die Frage, welche der Hypothesen  $H_0$ : „ $A$  ist höchstens 0,7“ oder  $H_1$ : „ $A$  ist grösser als 0,7.“ wahrscheinlicher ist.  $\square$

**Beispiel 15.6 (Geburten)** Seit dem 18. Jahrhundert sind Daten zu Geburten und Todesfällen in verschiedenen Ländern Europas vorhanden. So mussten seit 1771 die Behörden in Frankreich alle Geburten und Todesfälle nach Paris melden (siehe [8]). Dabei wurde festgestellt, dass etwas mehr Jungen als Mädchen geboren wurden. Dabei blieb das Verhältnis der Jungen- und Mädchengeburten über mehrere Jahre in etwa konstant. Ist  $A$  der Anteil der Mädchengeburten, so will man wissen, ob die Hypothese  $H_0$ : „ $A$  ist kleiner als 0,5“, plausibler ist als die Hypothese  $H_1$ : „ $A$  ist grösser als 0,5.“ Die Antwort lieferte Laplace mit den vorhandenen Daten aus Frankreich (siehe [7] auf Seite 81).  $\square$

Eine Person gibt sich also zwei Hypothesen zu einer nicht direkt messbaren Größe  $\mu$  vor. Die Person kann zuerst mit ihrer vorhandenen Vorinformation  $\mathcal{I}$  setzen, wie plausibel die Hypothesen sind:

$$\text{A priori: } \pi_0 = \mathbb{P}(H_0 | \mathcal{I}), \quad \pi_1 = \mathbb{P}(H_1 | \mathcal{I})$$

Dabei sind  $\pi_0$  und  $\pi_1$  zwei Wahrscheinlichkeiten ungleich null. Mit der Regel von Bayes kann man aus Messungen oder Beobachtungen über den Parameter  $\mu$  lernen. Daraus berechnet man die A posteriori-Plausibilitäten der beiden Hypothesen:

$$\text{A posteriori: } p_0 = \mathbb{P}(H_0 | \text{Daten}, \mathcal{I}), \quad p_1 = \mathbb{P}(H_1 | \text{Daten}, \mathcal{I})$$

Ist  $p_0$  nahe bei eins, wird man die Hypothese  $H_0$  der Hypothese  $H_1$  vorziehen. Empfehlenswert ist es zusätzlich zu bestimmen, wie stark dieser Wert von den Daten und nicht von der Vorinformation  $\mathcal{I}$  abhängt. Dazu vergleicht man die beiden Quotienten  $p_0/p_1$  und  $\pi_0/\pi_1$ . Man schreibt

$$\frac{p_0}{p_1} = \text{BF}_{0:1} \cdot \frac{\pi_0}{\pi_1}$$

Die Zahl  $\text{BF}_{0:1}$  nennt man den *Bayes-Faktor*. Ist  $H_1$  die Negation der Nullhypothese, so sind  $p_1 = 1 - p_0$  und  $\pi_1 = 1 - \pi_0$ . Die Quotienten  $p_0/p_1$  und  $\pi_0/\pi_1$  sind dann die Chancen der Hypothese 0. Man hat in diesem Fall also

$$\text{A posteriori-Chance von } H_0 = \text{BF}_{0:1} \times \text{A priori-Chance von } H_0$$

Ist der Bayes-Faktor sehr gross oder nahe bei null, beeinflussen vor allem die Daten die A posteriori-Chance. Ist er nahe bei eins, hat man nicht viel aus den Daten gelernt: Die A posteriori-Chance entspricht etwa der A priori-Chance der Hypothese. Man kann Tab. 15.1 benutzen, um einen Bayes-Faktor zu interpretieren. Das vorgestellte Rechenverfahren nennt man einen *bayesschen statistischen Test*<sup>9</sup>. Hier zwei Beispiele dazu

**Beispiel 15.7 (HNV-Indikator)** Beim Beispiel 15.5 interessiert sich eine Biologin für der Anteil  $A$  der  $1 \text{ km}^2$ -Flächen in Deutschland, die weniger als 5 % landwirtschaftliche Fläche haben, und für die zwei Hypothesen  $H_0$ : „ $A$  ist höchstens 0,7“, bzw.  $H_1$ : „ $A$  ist grösser als 0,7.“ Die Biologin habe nur minimale Vorinformation  $\mathcal{I}$  zu  $A$ . Sie beschreibt daher  $A$  mit dem flachen Prior (siehe Abb. 15.12). Die Wahrscheinlichkeiten der Hypothesen  $H_0$  und  $H_1$  kann man mit den Gesetzen zur Wahrscheinlichkeit berechnen. Dazu muss man mit dem Gesetz der Marginalisierung den Parameter  $A$  und die Hypothesen miteinander verknüpfen:

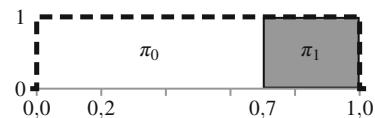
$$\pi_0 = \mathbb{P}(H_0 \mid \mathcal{I}) = \int_0^1 \mathbb{P}(H_0 \text{ und } A \mid \mathcal{I}) dA$$

Das Multiplikationsgesetz  $\mathbb{P}(H_0 \text{ und } A) = \mathbb{P}(H_0 \mid A) \cdot \mathbb{P}(A)$  hilft den Integranden auszurechnen:

$$\pi_0 = \int_0^1 \mathbb{P}(H_0 \mid A, \mathcal{I}) \cdot \mathbb{P}(A \mid \mathcal{I}) dA$$

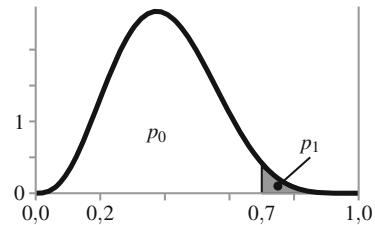
Ist  $A$  grösser als 0,7, so beträgt die Wahrscheinlichkeit der Hypothese  $H_0$  null:  $\mathbb{P}(H_0 \mid A) = 0$ . Ist  $A$  aber kleiner als 0,7, so ist diese Wahrscheinlichkeit eins. Da-

**Abb. 15.12** Prior des Parameters  $A$ , um damit A priori-Wahrscheinlichkeiten zu den Hypothesen zu berechnen



<sup>9</sup> Arbeitet man mit der frequentistischen Interpretation von Wahrscheinlichkeiten, so sind andere statistische Tests verbreitet. Sie berechnen  $p$ -Werte und arbeiten mit daraus abgeleiteten Begriffen wie die *statistische Signifikanz*. Diese messen aber nicht, wie plausibel eine Hypothese ist. Zudem ist „statistische Signifikanz“ nicht äquivalent zu wissenschaftlicher Signifikanz und weist nicht darauf hin, wie stark oder wie wichtig ein Effekt ist. Siehe dazu [1] und [2].

**Abb. 15.13** Posterior des Parameters  $A$ , um damit  $A$  posteriori-Wahrscheinlichkeiten zu den Hypothesen zu berechnen



her ist

$$\pi_0 = \mathbb{P}(H_0 \mid \mathcal{I}) = \int_0^{0,7} \underbrace{\mathbb{P}(A \mid \mathcal{I})}_{\text{Prior zu } A} dA = \int_0^{0,7} 1 dA = 0,7$$

Analog ist

$$\pi_1 = \mathbb{P}(H_1 \mid \mathcal{I}) = 0,3$$

Die Biologin wählt randomisiert acht Stichprobenflächen aus. Davon sind drei mit weniger als 5 % landwirtschaftlicher Fläche. Daher lautet die  $A$  posteriori-Verteilung des Anteils  $A$ , wie in Beispiel 5.3 im Detail gerechnet,

$$\text{pdf}(A \mid \text{Daten}) = 504 \cdot A^3 \cdot (1 - A)^5$$

Die  $A$  posteriori-Wahrscheinlichkeiten von  $H_0$  und  $H_1$  werden daraus in analoger Weise wie oben berechnet:

$$p_0 = \mathbb{P}(H_0 \mid \text{Daten}) = \int_0^{0,7} \underbrace{\mathbb{P}(A \mid \text{Daten})}_{\text{Posterior zu } A} dA = \int_0^{0,7} 504 \cdot A^3 \cdot (1 - A)^5 dA = 0,97$$

Analog ist  $p_1 = \mathbb{P}(H_1 \mid \text{Daten}) = 0,03$  (siehe Abb. 15.13). Die Biologin kann daher mit hoher Plausibilität behaupten, dass der Anteil  $A$  höchstens 0,7 beträgt. Der Bayes-Faktor ist

$$0,97/0,03 = \text{BF}_{0:1} \cdot 0,7/0,3$$

Also ist  $\text{BF}_{0:1} = 13,86$ . Die Daten haben daher die  $A$  priori-Chance von  $H_0$  um 13,86 erhöht. Dies spricht stark für die Hypothese  $H_0$ .  $\square$

**Beispiel 15.8 (Geburten)** Man interessiert sich beim Beispiel 15.6 für den Anteil  $A$  an Mädchengeburten. Dabei will man wissen, ob die Hypothese  $H_0$ : „ $A$  ist kleiner als 0,5“, plausibler ist als die Hypothese  $H_1$ : „ $A$  ist grösser als 0,5.“ Laplace benutzte die Geburtslisten von Paris zwischen 1745 und 1770 (siehe [6]), um die Plausibilitäten von  $H_0$  und  $H_1$  zu berechnen: geboren wurden 241 945 Mädchen und 251 527 Jungen. Als Vorwissen

nahm Laplace an, dass er zum Anteil  $A$  Indifferenz hat:  $\text{pdf}(A) = 1$ . Daher lauten die A priori-Wahrscheinlichkeiten der Hypothesen

$$\pi_0 = \mathbb{P}(H_0) = \int_0^{0.5} 1 \, dA = 0,5 \quad \text{und} \quad \pi_1 = \mathbb{P}(H_1) = 0,5$$

Der Posterior für den Anteil  $A$  berechnet man mit der Regel von Bayes durch das Produkt der Likelihood und des Priors. Man erhält eine Beta-Verteilung mit Parametern 241 946 und 251 528:

$$\text{pdf}(A \mid \text{Daten}) \propto A^{241\,945} \cdot (1 - A)^{251\,527}$$

Mit einem Statistikprogramm, das die Beta-Verteilung implementiert hat, erhält man

$$p_0 = \mathbb{P}(H_0 \mid \text{Daten}) = \int_0^{0.5} \text{pdf}(A \mid \text{Daten}) \, dA = 1 \quad \text{und} \quad p_1 = \mathbb{P}(H_1 \mid \text{Daten}) = 0$$

Laplace erhielt ohne Computer ein genaueres Resultat:  $p_1 \approx 1,15 \cdot 10^{-42}$ . Es ist also hochwahrscheinlich, dass Jungengeburten häufiger sind als Mädchengeburten, da  $\text{BF}_{0:1} = 1/(1,15 \cdot 10^{-42}) = 8,70 \cdot 10^{41}$ .  $\square$

Zu einer nicht direkt messbaren Grösse  $\mu$  können verschiedene Hypothesen aufgestellt werden. Oft findet man die folgende Variante:

$$H_0 : \mu = a, \quad H_1 : \mu \neq a \tag{15.4}$$

Dabei ist  $a$  eine vorgegebene Zahl. Man spricht von einer *Punkt-Nullhypothese* für den Parameter  $\mu$ . Punkt-Nullhypotesen sind heikel. So ist es unklar, wie die A priori-Wahrscheinlichkeit für  $H_0$  aus Informationskriterien oder dem Prinzip der maximalen Entropie gewählt werden soll.<sup>10</sup> Weiter kann man den Parameter  $\mu$  aus Messungen nicht exakt, sondern nur mit einer Präzision  $\pm \delta\mu$  bestimmen. Daher ist es fragwürdig, die Plausibilität von  $H_0$  zu berechnen. Zudem sind die Hypothesen

$$H_0 : \mu \leq a, \quad H_1 : \mu > a \tag{15.5}$$

informativer als die diejenigen in (15.4). Man berechnet nicht nur die Plausibilität, dass  $\mu \neq a$  ist, sondern man bestimmt aus (15.5) zusätzlich wie plausibel das Vorzeichen von  $\mu - a$  ist.<sup>11</sup> So interessiert beim obigen Beispiel zu den Geburten nicht, ob die Geburtenraten von Mädchen und Jungen gleich oder ungleich sind ( $H_0 : A = 0,5$  zu  $H_1 : A \neq 0,5$ ), sondern ob mehr Jungen ( $H_0 : A < 0,5$ ) oder mehr Mädchen ( $H_1 : A > 0,5$ ) geboren werden. Oft sind Punkt-Nullhypotesen auch wissenschaftlich oder technisch unplausibel:

<sup>10</sup> Siehe dazu [4] auf den Seiten 103–148.

<sup>11</sup> Eine detaillierte Diskussion dazu ist im Artikel „What is the question?“ in [9]. Schwierigkeiten und Paradoxa zu Punkt-Nullhypotesen, sowie weiterführende Literatur findet man in [10].

**Beispiel 15.9 (Kanalwärmetauscher)** Beim Beispiel 2.13 interessiert die durchschnittliche Verformung  $\mu_V$  von Wärmetauschern. Ein einfaches Regressionsmodell, das die Verformung in Funktion der Dicke des Siggenblechs  $S$  des Wärmetauschers beschreibt, ist

$$\mu_V = a + b \cdot S$$

Es ist für die Produzenten der Wärmetauscher klar, dass die Geometrie der Wärmetauscher – also auch die Dicke  $S$  des Siggenblechs – einen Einfluss auf die Verformung hat. Es ist daher kaum sinnvoll, die Punkt-Nullhypothese

$$H_0 : b = 0 \quad (S \text{ hat keinen Einfluss auf } \mu_V), \quad H_1 : b \neq 0$$

zu betrachten. Relevanter sind die Hypothesen „Zunehmende Dicke verkleinert die Verformung“ ( $H_0 : b < 0$ ) und „Zunehmende Dicke vergrößert die Verformung“ ( $H_1 : b > 0$ ).  $\square$

Viele Statistiker raten ab, mit Punkt-Nullhypotesen zu arbeiten. Physikerinnen und Ingenieure gehen mit der Punkt-Nullhypothese  $H_0 : \mu = a$  pragmatisch um. Sie berechnen nicht, wie plausibel  $H_0$  ist, sondern bestimmen Wahrscheinlichkeitsintervalle für  $\mu$ :  $\mu = \mu_0 \pm \delta\mu$ . Ist das Intervall mehrere Vielfache von  $\delta\mu$  vom Wert  $a$  entfernt, so wird  $H_0$  als nicht realistisch definiert (siehe dazu die Diskussion in [9]). Ganz allgemein wird im Ingenieurwesen wenig versucht, Hypothesen zu nicht direkt messbaren Größen mit statistischen Werkzeugen zu plausibilisieren. Informativer ist es, Wahrscheinlichkeitsintervalle für die nicht direkt messbaren Größen anzugeben.

Statistische Methoden, um zwischen zwei Hypothesen  $H_0$  und  $H_A$  zu entscheiden, besitzen eine Sensitivität (Richtigpositiv-Rate)  $\beta$  und eine Spezifität (Richtignegativ-Rate)  $\alpha$ . Entscheidend man sich wegen eines Wertes  $BF+$ , berechnet aus der Likelihood, für die Hypothese  $H_A$ , so ist:

$$\mathbb{P}(BF+ \mid H_A) = \beta, \quad \mathbb{P}(BF- \mid H_0) = \alpha$$

Man nennt  $\beta$  die *Macht* (engl. *Power*) und  $1 - \alpha$  den *Fehler 1. Art* (manchmal auch *p-Wert*) der statistischen Methode. Hohe Sensitivität und Spezifität garantieren aber nicht, dass die Hypothese  $H_A$  bei „positivem“ Wert  $BF+$  stark plausibel wird. Ist etwa die A priori-Wahrscheinlichkeit von  $H_A$  nur 0,01 und – wie bei vielen verwendeten Methoden in statistischen Publikationen –  $\beta = \alpha = 0,95$ , so hat man bei einem positiven Antwort der Methode nach Abschn. 1.1

$$\mathbb{O}(H_A \mid BF+) = \frac{\text{Sensitivität}}{1 - \text{Spezifität}} \cdot \mathbb{O}(H_A) = \frac{0,95}{1 - 0,95} \cdot \frac{0,01}{1 - 0,01} = 0,19 : 1$$

Die A posteriori-Wahrscheinlichkeit, dass  $H_A$  wahr ist, beträgt damit nur 16 %. Es ist daher entscheidend, dass man sich vor der Datensammlung Gedanken zum Prior  $\mathbb{P}(H_A)$  von  $H_A$  macht. Das Dilemma aber ist: Ist  $H_A$  neu und wissenschaftlich interessant, so wird ein skeptischer Wissenschaftler die Wahrscheinlichkeit  $\mathbb{P}(H_A)$  klein setzen.

## Reflexion

**15.1** Von zwei Regressionsmodellen mit je einem Parameter kennt man das Folgende:

Modell	Maximaler Wert Likelihood	rel. Gewinn Präzision
$\mathcal{M}_1$	$3,45 \cdot 10^4$	80 %
$\mathcal{M}_2$	$4,35 \cdot 10^6$	99,9 %

Der relative Gewinn der Präzision, sagt, um wie viel ein Parameter eines Modells dank Daten genauer bestimmt ist als vor der Datensammlung. So ist also beim Modell 1 mit Parameter  $a$

$$\text{SE}(a)/(a_{\max} - a_{\min}) = 0,2$$

- (a) Welches Modell passt besser zu den Daten?
- (b) Welches Modell ist plausibler?
- (c) Wie lauten die A posteriori-Wahrscheinlichkeiten der beiden Modelle, wenn vor der Datensammlung beide Modelle gleich plausibel sind?

**15.2** Die Jungfraubahn AG ist im Winterhalbjahr mit auftretenden Windstürmen des Guggiföhns konfrontiert, die eine Gefahr für ihre Zugskompositionen sind. Der Guggiföhn ist vorhanden, wenn der Luftdruckunterschied zwischen der Alpennordseite und der Alpensüdseite gross ist. Aus Messungen möchte man die durchschnittlich erwartbare Windgeschwindigkeit  $\mu_v$  aus der Druckdifferenz  $\Delta p_1$  zwischen den Standorten Cimetta und Napf und der Druckdifferenz  $\Delta p_2$  zwischen den Standorten Corvatsch und Jungfrau mit einem Regressionsmodell berechnen. Einfache Modelle sind:

$$\text{Modell 1: } \mu_v = a + b \cdot \Delta p_1$$

$$\text{Modell 2: } \mu_v = A + B \cdot \Delta p_2$$

$$\text{Modell 3: } \mu_v = \alpha + \beta \cdot \Delta p_1 + \gamma \cdot \Delta p_2$$

Als Datenmodell soll gelten: Die Geschwindigkeiten variieren normalverteilt um die Regressionmodelle mit konstanter nicht von den Kovariablen abhängiger Streuung. Um die Parameter der Regressionmodell zu bestimmen, dienen die 21 Messungen in Tab. 15.4. Vor der Datensammlung bestehe nur minimale Information zu den Parametern der Regressionsmodelle wie auch zur Streuung.

- (a) Zeichnen Sie die Messwertpaare  $(\Delta p_i | v)$  in einem Streudiagramm. Schätzen Sie die Parameter des Modells 1. Wie lauten Wahrscheinlichkeitsintervalle für  $a$  und  $b$ ? Eignet sich das Modell? Berechnen Sie dazu Prognosebereiche für zukünftige Messwerte. Ist die Likelihood gut gerechnet?

**Tab. 15.4** Druckdifferenzen (in Pa) und Windgeschwindigkeit  $v$  (in km/h) bei der kleinen Scheidegg, wo der Guggiföhn gefährlich werden kann (aus [3])

$v$	10	11	13	31	20	36	40	46	29	21	43
$\Delta p_1$	33,0	33,8	37,1	40,7	45,5	46,1	46,1	44,1	44,9	47,5	50,6
$\Delta p_2$	19,0	20,1	19,4	19,8	15,7	13,5	14,7	16,3	16,7	14,1	19,1
$v$	31	32	62	71	80	86	34	49	61	81	
$\Delta p_1$	48,7	44,8	45,3	49,6	53,7	50,5	56,2	55,7	59,2	58,5	
$\Delta p_2$	22,6	28,0	31,6	33,2	30,9	29,1	34,8	38,2	41,8	39,3	

- (b) Zeichnen Sie die Messwertpaare ( $\Delta p_2 \mid v$ ) in einem Streudiagramm. Schätzen Sie die Parameter des Modells 2. Wie lauten Wahrscheinlichkeitsintervalle für  $A$  und  $B$ ? Eignet sich das Modell? Berechnen Sie dazu Prognosebereiche für zukünftige Messwerte. Ist die Likelihood gut gerechnet?
- (c) Berechnen Sie die Quadratsumme der Residuen der Modelle 1 und 2. Beantworten Sie mit dem BIC: Welches der Modelle 1 und 2 ist plausibler?
- (d) Schätzen Sie die Parameter  $\alpha$ ,  $\beta$  und  $\gamma$  des Modells 3 mit Hilfe der Methode der kleinsten Quadrate. Wie lauten Wahrscheinlichkeitsintervalle für die Parameter des Regressionsmodells?
- (e) Wie stark sind die beiden Kovariablen  $\Delta p_1$  und  $\Delta p_2$  miteinander korreliert? Zeichnen Sie dazu ein Streudiagramm der Kovariablenwerte und berechnen Sie den empirischen Korrelationskoeffizienten.
- (f) Berechnen Sie die Quadratsumme der Residuen des Modells 3. Beantworten Sie mit dem BIC: Welches der drei Modelle ist am plausiblesten? Gehen Sie dabei aus, dass die A priori-Wahrscheinlichkeiten der drei Modelle gleich, also  $1/3$ , sind.

**15.3** In einem Tiergehege mit vielen kleinen Tieren interessiert der Anteil  $A$  der weiblichen Tiere. Jemand behauptet, dass der Anteil  $A$  grösser als 0,5 sei ( $H_0 : A > 0,5$ ). Die alternative Hypothese dazu ist „ $A$  ist kleiner als 0,5“ ( $H_1 : A < 0,5$ ). Es ist keine weitere Information zum Anteil an weiblichen Tieren im Gehege vorhanden.

- (a) Wie lauten die A priori-Wahrscheinlichkeiten der Hypothesen  $H_0$  und  $H_1$ ?
- (b) Um verbesserte Plausibilitäten zu den Hypothesen zu rechnen, werden zwölf Tiere nacheinander gefangen und ihr Geschlecht festgestellt. Gezogen wurden die Tiere durch Ziehen mit Zurücklegen, um zu garantieren, dass die Messwerte unabhängig sind. Hier das Resultat:

Gefangene Tiere: 12,    weiblich: 10,    männlich: 2

Wie lauten die A posteroiri-Wahrscheinlichkeiten der beiden Hypothesen? Ist die Nullhypothese deutlich plausibler als die Hypothese 1? Wie lautet der Bayes-Faktor des statististischen Tests?

- (c) Bestimmen Sie Wahrscheinlichkeitsintervalle zum Niveau 0,5 und 0,95 für  $A$  aus der A posteriori-Verteilung von  $A$ . Wie lautet die Fünf-Zahlen-Zusammenfassung zu  $A$ ?  
 (d) Führen Sie die Rechnung von (b) und (c) noch einmal durch, diesmal mit einer Stichprobe von 36 Tieren, bei der 30 weiblich waren.

**15.4** Eine Firma möchte, dass ihre produzierten, wiederaufladbaren Batterien bei voller Belastung durchschnittlich mindestens 300 Minuten brauchen, um vollständig entladen zu sein. Ist  $\bar{T}_{\text{Prod}}$  die durchschnittliche Entladungszeit aller produzierten Batterien – ein unbekannter Skalierungsparameter –, so interessiert sich die Firma deshalb für die Nullhypothese  $H_0 : \bar{T}_{\text{Prod}} \geq 300$ . Die Alternativhypothese dazu ist  $H_1 : \bar{T}_{\text{Prod}} < 300$ . Zudem weiss man, dass  $\bar{T}_{\text{Prod}}$  zwischen 100 und 600 Minuten liegt.

- (a) Wie lauten die A priori-Wahrscheinlichkeiten der Hypothesen  $H_0$  und  $H_1$ ?  
 (b) Berechnen Sie die A posteriori-Wahrscheinlichkeiten der beiden Hypothesen, wenn drei gemessenen Entladungszeiten: 301,2, 378,7 und 315,0 Minuten vorliegen. Nehmen Sie dazu an, dass die Entladungszeiten mit der Exponentialverteilung modelliert werden können.  
 (c) Bestimmen Sie aus dem Datenmodell auch ein 95 % Wahrscheinlichkeitsintervall für  $\bar{T}_{\text{Prod}}$ .  
 (d) Rechnen Sie die Aufgabe (b) noch einmal durch: als Information haben Sie die drei Messwerte von (b) und die zwei zusätzlichen Messwerte 350,8 und 370,1 Minuten.

**15.5** Eine Person vermutet, dass ihr Gewicht  $G$  grösser als 70 kg ist:  $H_0 : G \geq 70$  kg. Die alternative Hypothese dazu ist  $H_1 : G < 70$  kg. Die Person weiss zusätzlich, dass ihr Gewicht zwischen 60 und 90 Kilo liegt.

- (a) Wie lauten die A priori-Wahrscheinlichkeiten der Hypothesen  $H_0$  und  $H_1$ ?  
 (b) Um verbesserte Plausibilitäten zu den Hypothesen zu rechnen, misst die Person viermal ihr Gewicht mit einer Personenwaage:

$$75,5 \text{ kg} \quad 74,8 \text{ kg} \quad 75,2 \text{ kg} \quad 75,7 \text{ kg}$$

Weiter geht sie davon aus, dass die Messwerte, gegeben  $G$ , normalverteilt um  $G$  streuen. Wie lauten die A posteroiri-Wahrscheinlichkeiten der beiden Hypothesen? Ist die Nullhypothese deutlich plausibler als die Hypothese 1? Wie lautet der Bayes-Faktor des statististischen Tests?

- (c) Bestimmen Sie Wahrscheinlichkeitsintervalle zum Niveau 0,5 und 0,95 für das Gewicht  $G$ . (Arbeiten Sie dazu mit einer Monte-Carlo-Simulation oder mit Hilfe der Formeln zur  $t$ -Verteilung in Anhang A.)

## Literatur

1. American Statistical Association: The ASA's Statement on *p*-Values: Context, Process, and Purpose. *The American Statistician* **70** (2), 129–131 (2016)
2. American Statistical Association, ASA Statement on Statistical Significance and *P*-Values. *The American Statistician* **70** (2), 131–133 (2016)
3. D. Bättig, O. Mermoud, Prognostizierung und Modellierung von Windböen des Guggiföhns auf der kleinen Scheidegg. Bericht i-REX Berner Fachhochschule Burgdorf zuhanden Jungfraubahn AG (2008)
4. R. T. Cox, Of inference and inquiry. In: Proc. Maximum Entropy Formalism Conference, MIT Press, Cambridge (1979)
5. R. McElrath, *Statistical Rethinking, a Bayesian Course with Examples in R and Stan* (Chapman & Hall, 2016)
6. A. Gelman, J. B. Carlin, H. S. Stern, D. B. Rubin, *Bayesian Data Analysis* (Chapman & Hall, 2004)
7. C. C. Gillispie, *Pierre-Simon Laplace 1749–1827: a life in exact science* (Princeton University Press, 1997)
8. S. B. McGrawe, *Die Theorie, die nicht sterben wollte (Wie der englische Pastor Thomas Bayes eine Regel entdeckte, die nach 150 Jahren voller Kontroversen heute aus Wissenschaft, Technik und Gesellschaft nicht mehr wegzudenken ist)* (Springer Spektrum, Springer Verlag, Berlin, Heidelberg, 2014)
9. E. T. Jaynes, Papers on Probability, Statistics and Statistical Physics. Edited by R. D. Rosenkrantz. (Kluwer Academic Publishers, 1989)
10. P. M. Lee, *Bayesian Statistics, an Introduction* (John Wiley & Sons Ltd., 2012)
11. A. E. Raftery, Bayesian, Model Selection in Social Research. *Sociological Methodology*, **25**, 111–163 (1995)
12. G. Schwarz, Estimating the dimension of a model. *The Annals of Statistics*, **6**, No. 2, 461–464 (1978)
13. D. S. Sivia, J. Skilling, *Data Analysis, a Bayesian Tutorial* 2nd ed. (Oxford Science Publications, 2010)

„Was sein muss, muss sein“, sagte der König bedrückt.

Lewis Carroll, Alice im Wunderland (Insel Taschenbuch, 1973, S. 117)

**Zusammenfassung** In den Kap. 9 und 10 wird diskutiert, wie man Parameter berechnet und weitere Messwerte von unsicheren Größen prognostiziert, wenn das Datenmodell die Exponential-, die Poisson- oder die Gauss-Verteilung ist. Bei manchen Untersuchungen hat man außer den Daten nur minimale Vorinformation zu den Parametern der Datenmodelle. Die Verteilung des Prognosemodells kann man dann bei den erwähnten Modellen explizit berechnen. Man erhält Gamma-, Pareto-, Negativ-Binomial-,  $t$ - und Chiquadrat-Verteilungen. Weil die erhaltenen Formeln und Verteilungen weit verbreitet sind, werden sie hier vorgestellt und teilweise hergeleitet.

---

## A.1 Datenmodell Exponentialverteilung

Die Dichtefunktion der Exponentialverteilung für eine Messgröße  $X$ , die keine negativen Werte annehmen kann, ist

$$\text{pdf}(X = x \mid \mu) = \mu^{-1} \cdot \exp(-x/\mu)$$

Hat man nur minimale Vorinformation zum Skalierungsparameter  $\mu$ , so wählt man als Prior nach Jeffreys

$$\text{pdf}(\mu \mid \text{min. Vorinformation}) \propto 1/\mu$$

Man nennt dann die A posteriori-Verteilung von  $\mu$  eine *inverse Gammaverteilung* mit Form  $n$  und Skalierung  $\bar{x}$ . Dabei ist  $\bar{x}$  das arithmetische Mittel der  $n$  unabhängigen Mess-

werte oder Beobachtungen.<sup>1</sup> Der plausibelste Wert  $\mu_0$  für  $\mu$  lautet

$$\mu_0 = \frac{n}{n+1} \cdot \bar{x}$$

Die A posteriori-Verteilung sagt, dass der plausibelste Wert für  $\mu$  hier nicht das arithmetische Mittel der Messwerte ist.

Das Prognosemodell, um zukünftige Beobachtungen oder Messwerte  $X$  zu prognostizieren, ist

$$\text{pdf}(X = x \mid \text{Daten, min. Vor.}) = \int_0^{\infty} \underbrace{\mu^{-1} \cdot \exp(-x/\mu)}_{\text{Datenmodell: Exponential}} \cdot \underbrace{\text{pdf}(\mu \mid \text{Daten, m. Vor.})}_{\text{Posterior: Inverse Gamma}} d\mu$$

Das Integral kann man explizit berechnen. Man erhält die *verallgemeinerte Paretoverteilung* (engl. *generalized paretdistribution* [GPD]) mit Lage (engl. location) 0, Skalierung (engl. scale)  $\bar{x}$  und Form (engl. shape)  $1/n$ :

$$\text{pdf}(\text{Messwert} = x \mid \text{Daten, min. Vorinformation}) = \frac{1}{\bar{x}} \cdot \left(1 + \frac{x}{n \cdot \bar{x}}\right)^{-(n+1)} \quad \text{für } x \geq 0$$

Dabei ist  $\bar{x}$  wiederum das arithmetische Mittel der  $n$  Messwerte. Das  $\gamma$ -Quantil  $q_\gamma$  dieses Prognosemodells ist explizit berechenbar. Es lautet

$$q_\gamma = n \cdot \bar{x} \cdot \left[ \left( \frac{1}{1-\gamma} \right)^{(1/n)} - 1 \right]$$

## A.2 Datenmodell Poissonverteilung

Die Massenfunktion der Poissonverteilung für eine diskrete Messgrösse  $N$ , die keine negativen Werte hat, ist

$$\mathbb{P}(N = k \mid \lambda) = \frac{\lambda^k}{k!} \cdot \exp(-\lambda) \quad \text{für } k = 0, 1, 2, \dots$$

Hat man keine Vorinformation zum Skalierungsparameter  $\lambda$ , so wählt man als Prior  $\text{pdf}(\lambda \mid \mathcal{I}) \propto 1/\lambda$ . Der Posterior von  $\lambda$  ist dann eine Gammaverteilung mit Form  $n \cdot \bar{x}$  und Rate  $n$ .<sup>2</sup> Der plausibelste Wert  $\lambda_0$  von  $\lambda$  lautet, falls das arithmetische Mittel  $\bar{x}$  der  $n$

<sup>1</sup> Die Dichtefunktion der inversen Gammaverteilung mit Form  $\alpha > 0$  und Skalierung  $\beta > 0$  ist:

$$\text{pdf}(x \mid \alpha, \beta) \propto x^{-\alpha-1} \cdot \exp(-\beta/x) \quad \text{für } x \geq 0$$

<sup>2</sup> Die Dichtefunktion der Gammaverteilung mit Form  $\alpha > 0$  und Rate  $\beta > 0$  lautet:

$$\text{pdf}(x \mid \alpha, \beta) \propto x^{\alpha-1} \cdot \exp(-\beta \cdot x) \quad \text{für } x > 0$$

unabhängigen Datenwerte grösser als  $1/n$  ist,

$$\lambda_0 = \bar{x} - \frac{1}{n}$$

Der plausibelste Wert für den Parameter  $\lambda$  ist also (leicht) kleiner als das arithmetische Mittel der Datenwerte.

Die Massenfunktion, um zukünftige Beobachtungen oder Messwerte beim Poissonmodell zu prognostizieren, lautet:

$$\mathbb{P}(\text{Messwert} = k \mid \text{Daten, min. Vor.}) = \frac{(\text{size} + k - 1)!}{(\text{size} - 1)! \cdot k!} \cdot \text{prob}^{\text{size}} \cdot (1 - \text{prob})^k$$

Dies ist eine *Negativ-Binomialverteilung* mit Kennzahlen  $\text{size} = n \cdot \bar{x}$  und  $\text{prob} = n/(n + 1)$ . Dabei ist  $\bar{x}$  das arithmetische Mittel der  $n$  unabhängigen Datenwerte.

**Beispiel A.1 (Anzahl grosser Schäden)** Beim Beispiel 9.4 zu den Schadensfällen kann die Wahrscheinlichkeit berechnet werden, dass sich im nächsten Halbjahr fünf Schadensfälle ereignen. Es sind  $n = 10$  Beobachtungen vorhanden und das arithmetische Mittel der Datenwerte ist  $\bar{x} = 5,3$ . Daher ist  $\text{size} = 10 \cdot 5,3 = 53$  und  $\text{prob} = 10/11$ . Also ist

$$\mathbb{P}(N_{\text{zukunft}} = 5 \mid \text{Daten, min. Vor.}) = \frac{57!}{52! \cdot 5!} \cdot \left(\frac{10}{11}\right)^{53} \cdot \left(\frac{1}{11}\right)^5 = 0,166$$

Die Monte-Carlo-Simulation mit Theorem 9.5 ergibt ein ähnliches Resultat.  $\square$

### A.3 Datenmodell Normalverteilung

Wenn das Vorwissen zur Gauss-Verteilung mit Modus  $\mu$  und Standardabweichung  $\sigma$  minimal ist, so kann man die A posteriori-Plausibilitäten der beiden Parameter und die Prognoseverteilung explizit berechnen. Man erhält  $t$ - und inverse Chiquadrat-Verteilungen. Weil die  $t$ -Verteilung weit verbreitet ist, wird sie hier hergeleitet.

**Beispiel A.2 (Chloridgehalt)** Bei der Rechnung in Beispiel 10.4 wird angenommen, dass aus der vorhandenen Information Messwerte normalverteilt um den gesuchten Chloridgehalt Cl streuen. Bei sieben Messwerten und minimaler Vorinformation lautet die gemeinsame A posteriori-Dichtefunktion für den Lageparameter Cl und den Skalierungsparameter  $\sigma$

$$\text{pdf(Cl, } \sigma \mid \text{Daten, min. Vorinformation}) \propto \underbrace{\frac{1}{\sigma^7} \cdot e^{-0,5 \cdot \chi^2}}_{\text{Likelihood}} \cdot \underbrace{1 \cdot \frac{1}{\sigma}}_{\text{Prior}}$$

Hier ist  $\chi^2 = [(102,8 - \text{Cl})^2 + (103,3 - \text{Cl})^2 + \dots + (102,1 - \text{Cl})^2]/\sigma^2$ . Um zu berechnen, wie gross und wie plausibel der Chloridgehalt Cl ist, muss man  $\sigma$  aus der A posteriori-Verteilung gemäss Theorem 6.1 ausintegrieren:

$$\text{pdf}(\text{Cl} | \text{Daten, min. Vorinformation}) \propto \int_0^\infty \frac{1}{\sigma^7} \cdot e^{-0,5 \cdot \chi^2} \cdot 1 \cdot \frac{1}{\sigma} d\sigma$$

Man substituiert nun die Variable  $\sigma$  durch  $\chi$ . Es ist  $\chi^2 = Z/\sigma^2$  mit  $Z = (102,8 - \text{Cl})^2 + (103,3 - \text{Cl})^2 + \dots + (102,1 - \text{Cl})^2$ . Also beträgt  $\sigma = \sqrt{Z}/\chi$ . Deshalb ist  $d\sigma = -\sqrt{Z}/\chi^2 \cdot d\chi$ . Mit der Substitution wird das Integral zu

$$\begin{aligned} \text{pdf}(\text{Cl} | \text{Daten, min. Vor.}) &\propto \int_0^\infty \left( \frac{\chi}{\sqrt{Z}} \right)^7 \cdot \exp\{-0,5 \cdot \chi^2\} \cdot \frac{\chi}{\sqrt{Z}} \cdot \frac{\sqrt{Z}}{\chi^2} d\chi \\ &= \left( \frac{1}{\sqrt{Z}} \right)^7 \int_0^\infty \chi^6 \cdot \exp\{-0,5 \cdot \chi^2\} d\chi \end{aligned}$$

Das verbleibende Integral ist eine Zahl, die nicht von Cl oder anderen unbekannten Parametern abhängt. Somit ist:

$$\text{pdf}(\text{Cl} | \text{Daten, min. Vorinformation}) \propto \left( \frac{1}{\sqrt{(102,8 - \text{Cl})^2 + \dots + (102,1 - \text{Cl})^2}} \right)^7$$

Dies nennt man eine *Student-t-Verteilung* oder kurz *t-Verteilung*. Sie ist unimodal und symmetrisch. Im Gegensatz zur Normalverteilung strebt hier die Dichtefunktion nicht exponentiell gegen Null. Man hat für grosse Cl:

$$\text{pdf}(\text{Cl} | \text{Daten}) \propto \left( \sqrt{7 \cdot \text{Cl}^2} \right)^{-7} \propto \text{Cl}^{-7} \quad \text{für Cl gross}$$

Damit sind auch entfernte Werte vom Modus viel plausibler als beim Modell der Normalverteilung. Die um eins reduzierte Potenz sieben – also sechs – nennt man den Freiheitsgrad der Verteilung. Der Modus  $\text{Cl}_0$  der Verteilung ist gleich dem arithmetischen Mittel  $102,84 \text{ mol/m}^3$  der sieben Messwerte. Man nennt den Modus auch die *Lage* (engl. *location*) der *t*-Verteilung. Wie breit die Verteilung ist, hängt ebenfalls von den Messungen ab. Man spricht von der *Skalierung* (engl. *scale*) der *t*-Verteilung.<sup>3</sup> Auf halber Höhe der Dichtefunktion ist die *t*-Verteilung ungefähr 2,35-mal die Skalierung breit. Die Skalierung ist

---

<sup>3</sup> Die Dichtefunktion pdf der Student-*t*-Verteilung mit Lage  $\mu$ , Skalierung  $a > 0$  und  $n \geq 1$  Freiheitsgraden ist:

$$\text{pdf}(x | n, \mu, a) \propto \left( \frac{1}{1 + (1/n) \cdot (x - \mu)^2 / a^2} \right)^{(n+1)/2}$$

**Tab. A.1** Tabelle mit  $(1 + \gamma)/2$ -Quantilen der  $t$ -Verteilung mit Lage null und Skalierung eins

	$n - 1$	1	2	3	4	5	6	7	8	9	10	...	20	...	$\infty$
0,50	1,00	0,82	0,76	0,74	0,73	0,72	0,71	0,71	0,70	0,70	0,70	...	0,69	0,67	
0,68	1,82	1,31	1,20	1,13	1,10	1,08	1,07	1,06	1,05	1,05	1,05	...	1,02	1,00	
$\gamma$	6,31	2,92	2,35	2,13	2,02	1,94	1,89	1,86	1,83	1,81	1,81	...	1,72	1,64	
0,95	12,71	4,30	3,18	2,78	2,57	2,45	2,36	2,31	2,26	2,23	2,23	...	2,09	1,96	
0,9973	235,7	19,2	9,22	6,62	5,51	4,90	4,53	4,28	4,09	3,96	3,96	...	3,42	3,00	

hier gleich  $s/\sqrt{7}$ . Dabei ist  $s$  die empirische Standardabweichung der Messwerte. Man erhält  $0,962 \text{ mol/m}^3/\sqrt{7} = 0,36 \text{ mol/m}^3$ .  $\square$

Die  $t$ -Verteilung mit  $n - 1$  Freiheitsgraden, Lage  $\bar{x}$  und Skalierung  $s/\sqrt{n}$  beschreibt den Posterior zum Modus der Normalverteilung, wenn nur minimale Vorinformation vorliegt. Der plausibelste Wert für  $\mu$  ist das arithmetische Mittel  $\bar{x}$  der  $n$  unabhängigen Messwerte. Mit einer Wahrscheinlichkeit von  $\gamma$  ist

$$\mu = \bar{x} \pm t_{n-1}(\gamma) \cdot \frac{s}{\sqrt{n}} \quad (\text{A.1})$$

Dabei liest man den Faktor  $t_{n-1}(\gamma)$  aus Tab. A.1 ab und  $s$  ist die empirische Standardabweichung der Messwerte. So ist  $t_6(0,5) = 0,72$  oder  $t_9(0,95) = 2,26$ . In der Messtechnik schreibt man statt des Niveaus  $\gamma$  den *Erweiterungsfaktor*  $k$  hin. Der Erweiterungsfaktor entspricht dem  $t$ -Quantil, wenn  $n - 1$  unendlich gross wäre. So entspricht  $k = 1$  dem Niveau 0,68,  $k = 1,96$  dem Niveau 0,95 und  $k = 2$  dem Niveau 0,955. Ist  $n - 1 = \infty$ , so spricht man von der Gleichung von *de Moivre*. Quantile der  $t$ -Verteilung sind auch in Statistikprogrammen leicht abrufbar.

Gleichung (A.1) besagt, dass der Modus der Normalverteilung mit zunehmender Anzahl  $n$  *unabhängiger* Datenwerte immer präziser bestimmt werden kann. Die Präzision ist dabei proportional zu  $1/\sqrt{n}$ . Es ist aber meist unsicher, dass bei grosser Anzahl  $n$  die Messwerte unabhängig sind. Ein gesundes Misstrauen gegenüber den Formeln von de Moivre sollte in der Praxis daher immer präsent sein.<sup>4</sup>

Für den meist weniger wichtigen Streuungsparameter  $\sigma$  des Datenmodells erhält man bei minimaler Vorinformation den Posterior

$$\text{pdf}(\sigma \mid \text{Daten, min. Vorinformation}) \propto \frac{1}{\sigma^n} \cdot \exp \left\{ -\frac{(n - 1) \cdot s^2}{2\sigma^2} \right\} \quad \text{mit } \sigma > 0$$

Der plausibelste Wert für  $\sigma$  ist  $\sqrt{(n - 1)/n} \cdot s$ . Dabei ist  $s$  die empirische Standardabweichung der  $n$  unabhängigen Datenwerte. Auch hier nimmt die Breite des Wahrscheinlichkeitsintervalls mit zunehmendem  $n$  ab.

<sup>4</sup> Siehe dazu auch die Erläuterungen in Abschn. 3.2 und Kap. 9. Es gibt Verfahren, um Datenmodelle zu konstruieren, die Abhängigkeiten zwischen den Messwerten abbilden.

Die A posteriori-Verteilung der Standardabweichung  $\sigma$  lässt sich mit Theorem 8.2 auf die A posteriori-Verteilung für die Varianz  $\text{Var} = \sigma^2$  umrechnen.<sup>5</sup> Man erhält

$$\text{pdf}(\text{Var} \mid \text{Daten, min. Vorinformation}) \propto \frac{1}{\text{Var}^{(n+1)/2}} \cdot \exp \left\{ -\frac{(n-1) \cdot s^2}{2 \cdot \text{Var}} \right\}$$

Dies ist eine *inverse Chiquadratverteilung* mit  $n-1$  Freiheitsgraden und Skalierung  $s$ . Sie ist in Statistikprogrammen implementiert.<sup>6</sup> Daher sind Wahrscheinlichkeitsintervalle für die Varianz (und daraus solche für die Standardabweichung) schnell berechenbar.

Auch das Prognosemodell für einen Messwert  $X$  lässt sich explizit bestimmen. Dazu muss man das folgende Integral berechnen:

$$\text{pdf}(X = x \mid \text{Daten, m. Vor.}) \propto \underbrace{\int \int \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2 \right\}}_{\text{Datensetmodell}} \underbrace{\text{pdf}(\mu, \sigma) \, d\mu d\sigma}_{\text{Posterior}}$$

Man erhält eine  $t$ -Verteilung mit  $n-1$  Freiheitsgraden, Lage  $\bar{x}$  und Skalierung  $\sqrt{s^2 + s^2/n}$ . Mit einer Wahrscheinlichkeit von  $\gamma$  liegen diese im Bereich

$$\text{weiterer Messwert} = \bar{x} \pm t_{n-1}(\gamma) \cdot \sqrt{s^2 + \frac{s^2}{n}} \quad (\text{A.2})$$

Dabei ist  $\bar{x}$  das arithmetische Mittel und  $s$  die empirische Standardabweichung der  $n$  unabhängigen Datenwerte.

Beachten Sie: mit zunehmender Datenmenge  $n$  nimmt hier die Breite des Wahrscheinlichkeitsintervalls nicht ab. Für  $n$  gross erhält man  $\bar{x} \pm k \cdot s$ . Die Breite des Prognosemodells ist also nie kleiner als die Streuung der beobachteten Messwerte.

**Beispiel A.3 (Chloridgehalt)** Beim Beispiel 10.4 lautet das arithmetische Mittel der sieben Messwerte  $\bar{Cl} = 102,84 \text{ mol/m}^3$ . Die empirische Standardabweichung  $s$  beträgt  $0,96 \text{ mol/m}^3$ . Gegeben minimale Vorinformation zum Chloridgehalt, ist die A posteriori-Verteilung für den Chloridgehalt  $Cl$   $t$ -verteilt mit 6 Freiheitsgraden, mit Lage  $102,84 \text{ mol/m}^3$  und mit Skalierung  $s/\sqrt{7} = 0,36 \text{ mol/m}^3$ . Der Graph der Dichtefunktion ist in Abb. 10.11 dargestellt. Der plausibelste Wert  $Cl_0$  des Chloridgehalts ist das arithmetische Mittel der Messwerte. Er beträgt  $\bar{Cl} = 102,84 \text{ mol/m}^3$ . Mit Gleichung (A.1)

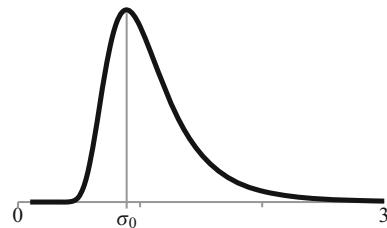
<sup>5</sup> Die Transformation ist eindeutig:  $\text{Var} = \sigma^2$  und  $\sigma = \sqrt{\text{Var}}$ .

<sup>6</sup> Die inverse Chiquadratverteilung mit  $m > 0$  Freiheitsgraden und Skalierung  $s > 0$  hat die Dichtefunktion

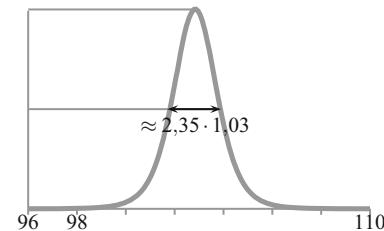
$$\text{pdf}(x \mid m, s) \propto \frac{1}{x^{m/2+1}} \cdot \exp \left\{ -\frac{m \cdot s^2}{2 \cdot x} \right\}$$

Ist  $X$  invers chiquadrat-verteilt, so ist  $1/X$  chiquadrat-verteilt.

**Abb. A.1** Plausibilität zur Streuung  $\sigma$  des Datenmodells, berechnet aus den Daten



**Abb. A.2** Prognosemodell für weitere Messungen:  $t$ -Verteilung mit sechs Freiheitsgraden, Lage  $\bar{Cl}$  und Skalierung  $1,03 \text{ mol}/\text{m}^3$



ist

$$\begin{aligned}\bar{Cl} &= \bar{Cl} \pm t_{7-1}(0,5) \cdot \frac{s}{\sqrt{7}} = \left( 102,84 \pm 0,72 \cdot \frac{0,96}{\sqrt{7}} \right) \text{ mol}/\text{m}^3 \\ &= (102,84 \pm 0,26) \text{ mol}/\text{m}^3 \quad [k = 0,67]\end{aligned}$$

ein Wahrscheinlichkeitsintervall zum Niveau 0,5 für den Chloridgehalt.

Wie gross die Streuung des Datenmodells ist und wo sie plausibel liegt, gibt der Posterior von  $\sigma > 0$  an:

$$\text{pdf}(\sigma \mid \text{Daten, min. Vorinformation}) \propto \frac{1}{\sigma^7} \cdot \exp \left\{ -\frac{6 \cdot (0,96)^2}{2\sigma^2} \right\}$$

Sein Graph ist in Abb. A.1 dargestellt. Die Verteilung ist unimodal schief. Der „wahrscheinlichste“ Wert  $\sigma_0$  für  $\sigma$  lässt sich mit der angegebenen Formel berechnen:

$$\sigma_0 = \sqrt{\frac{n-1}{n}} \cdot s = \sqrt{\frac{6}{7}} \cdot s = 0,89 \text{ mol}/\text{m}^3$$

Mit einer Wahrscheinlichkeit von 0,5 ist  $\sigma$  zwischen  $0,84 \text{ mol}/\text{m}^3$  und  $1,26 \text{ mol}/\text{m}^3$ .

Wo liegen weitere mit derselben Methode gemessene Chloridgehalte? Man benutzt gemäss dem vorigen Abschnitt dazu eine  $t$ -Verteilung mit 6 Freiheitsgraden, Lage  $102,84 \text{ mol}/\text{m}^3$  und Skalierung  $\sqrt{s^2 + s^2/7} = 1,03 \text{ mol}/\text{m}^3$ . Der Graph der Dichtefunktion ist in Abb. A.2 gezeigt. Mit Formel (A.2) beträgt die Wahrscheinlichkeit 0,5,

dass ein weiterer Messwert im Bereich

$$\begin{aligned}\text{Messwert} &= \overline{\text{Cl}} \pm t_{7-1}(0,5) \cdot \sqrt{s^2 + \frac{s^2}{n}} \\ &= \left( 102,84 \pm 0,72 \cdot \sqrt{0,96^2 + \frac{0,96^2}{7}} \right) \text{ mol/m}^3 = (102,84 \pm 0,74) \text{ mol/m}^3\end{aligned}$$

liegt. □

---

# Sachverzeichnis

## A

- A-posteriori-Verteilung, 123
- A-priori-Verteilung, 122
- acf, 146
- Akzeptanzrate, 104, 107
- ANOVA, 358
- Anteil, 126
- Ausreisser, 257
- ausschliessen, 80
- Autokorrelation, 64, 145

## B

- Bandbreite, 282
- Bayes-Faktor, 358, 373
- Bayesregel, 119
- Beobachtung, 10, 53
- Beta-Verteilung, 126
- Bias, 32, 43
- BIC, 361
- Binomialverteilung
  - negativ, 383
- Blockbildung, 56
- Burnin, 103

## C

- CE-Diagramm, 8, 33
- Chance, 3, 85
- chartchunk, 242
- Chiquadratverteilung
  - invers, 386

## D

- Daten
  - diskret, 28
  - kategorial, 28
  - stetig, 28
- Datenmodell, 4, 7, 122

## Wahl, 204

- Datenmüll, 29
- Deviance, 361
- Dichtefunktion, 91

## E

- Effekt, 257
- Eichung, 42
- Entropie, 176
  - relativ, 176
  - Shannon, 176
- Erwartungswert, 181
- Erweiterungsfaktor, 213
- Evidenz, 120
- Experiment, 9, 53
- Exponentialmodell, 193
- Exponentialverteilung, 91
- Extremwert, 257

## F

- Faktor, 31
- Faustregel
  - Akzeptanzrate, 107
  - Autokorrelation, 147
  - Korrelation, 143
- Fehler
  - 1. Art, 52
  - 2. Art, 52
  - grob, 255
  - systematisch, 32, 40
- Fehlerfortpflanzung, 334
- Fünf-Zahlen-Zusammenfassung, 99

## G

- Gammaverteilung, 382
  - invers, 381
- Genauigkeit, 52

- Gesetz  
 Addition, 78  
 Konvexität, 78  
 Marginalisierung, 7  
 Multiplikation, 79  
 totale Wahrscheinlichkeit, 80
- Gleichverteilung  
 diskret, 90  
 stetig, 95
- goldener Schnitt, 242
- Grundgesamtheit, 26  
 konzeptionell, 27
- H**
- Halbwertsbreite, 213
- Häufigkeit, 86
- Histogramm, 250
- Hyperparameter, 347
- Hypothese, 372
- I**
- Indifferenz, 174
- Information, 76  
 beobachtet, 331
- Interaktion, 37
- K**
- Kalibrierung, 42
- Kernel-Regression, 282
- Kistendiagramm, 258
- Klasse, 250
- kleinste Quadrate, 219
- Konfundierung, 10, 56, 144
- Kontingenztafel, 236
- Kontrolle  
 statistisch, 13, 58
- Kontrollgruppe, 42
- Kontrollkarte, 272
- Korrelation, 143
- Kovariablen, 31
- Kredibilität, 131
- Kreuztabelle, 81, 236
- Kreuzvalidierung, 194
- L**
- Lage, 384
- Lageparameter, 173
- Laplaceapproximation, 331
- LCL, 59
- Likelihood, 4, 5, 120
- marginalisiert, 359
- logit-Transformation, 184
- LSL, 214
- Lügenfaktor, 249
- M**
- Macht, 376
- Marginalverteilung, 138
- Massenfunktion, 89
- MaxEnt, 179  
 Exponential, 188  
 Normal, 211  
 Poisson, 197
- MCMC, 100
- Median, 96  
 empirisch, 258
- Messung  
 bivariat, 28  
 multivariat, 28  
 univariat, 28
- Messwerte  
 klassiert, 250
- MH-Algorithmus, 101
- Mischverteilung, 263
- Mittel  
 arithmetisches, 58
- Modalwert, 92
- Modell  
 diskret, 89  
 gemeinsam, 138  
 hierarchisch, 345, 349  
 Regression, 277, 291  
 robust, 261  
 stetig, 90
- Modellvergleich, 358
- Modus, 89, 92
- Monte-Carlo, 100, 238
- MSE, 211
- N**
- Niveau, 32
- Normalmodell, 224
- Normalverteilung, 212
- P**
- Paradoxon  
 Simpson, 54
- Parameter, 88
- Paretdiagramm, 36

- Paretoverteilung, 382  
Placebo, 41  
Placebo-Effekt, 41  
Plan  
    faktoriell, 36  
Plausibilität, IX, 76  
Poissonmodell, 203  
Pooling, 347  
Power, 376  
Präzision, 29, 52, 131, 149  
Prognose, 88  
    Messwert, 162  
Prognoseband, 297  
Prognosemodell, 160  
Prozessfähigkeit, 215  
Prozesskontrolle, 62  
Prozesspotential, 215  
Prozessregelung, 13  
Prozessteuerung, 13  
*p*-Wert, 376
- Q**  
QQ-Plot, 196, 204  
Quantil, 98  
Quartil, 99  
    empirisch, 257  
    oberes, 99  
    unteres, 99  
Quartilsdifferenz, 210  
    empirisch, 257  
Quintil, 99
- R**  
Randomisierung, 48  
Randverteilung, 138  
    berechnen, 140  
RCT, 54  
Regression  
    Gauss, 298  
    logistisch, 319  
    nicht parametrisch, 280  
    Poisson, 317  
    Prognoseband, 297  
Regressionsmodell, 277, 280, 291  
repräsentativ, 49  
Residuum, 302  
Risikofunktion, 216  
Risikomanagement, 98
- S**  
schätzen, 96  
Screening, 17  
Sechs Sigma, 215  
Sensitivität, 2, 118  
Sensitivitätsanalyse, 129  
Shrinkage, 349  
Skalierung, 384  
Skalierungsparameter, 173  
Spezifikationsgrenzen, 214  
Spezifität, 2  
SRS, 51  
Stabdiagramm, 247  
Stammblattdiagramm, 242  
Standardabweichung, 210  
    empirisch, 59  
Standardfehler, 225, 330, 331  
standardisiert, 274  
Standardunsicherheit, 330  
Statistik  
    Bayes, 4  
Stichprobe, 27  
Stichprobenunsicherheit, IX  
Störparameter, 218  
Streudiagramm, 59  
    kodiert, 279  
Streumodell, 4, 7, 122  
Strukturfunktion, 199  
Studentverteilung, 384  
suffizient, 224  
Sukzessionsregel, 161
- T**  
Test  
    statistisch, 372  
Toleranzgrenzen, 214  
TQM, 214  
Transformation, 172  
Typ A Unsicherheit, 38  
Typ B Unsicherheit, 29
- U**  
UCL, 14, 59  
Umfang, 27  
unabhängig, 85  
Unsicherheit, 210  
Untersuchung  
    doppelt blind, 42  
Urliste, 59

- USL, 214
- V**
- Value at Risk, 98
- VaR, 98
- Variable
- abhangig, 9
  - sekundär, 58
  - unabhängig, 9
- Varianz, 210
- verbunden, 142
- Vermengung, 10, 56, 144
- Versuchsplanung, VIII
- vertauschbar, 58
- Verzerrung, 32, 43
- Volatilität, 210
- Vollerhebung, 27
- W**
- Waagebalken, 338
- Wahrscheinlichkeit
- frequentistisch, 86
- invers, 2
- Plausibilität, 76
- Wahrscheinlichkeitsintervall, 97
- Wartezeit, 5
- Weibull-Verteilung, 93
- Wiederholung, 51
- Wirkungsraum, 34
- Z**
- Zeitreihe, 272
- Zensus, 27
- Zentralwert, 96
- Ziehen
- einfach, 51
  - mit Zurücklegen, 50
  - ohne Zurücklegen, 50
  - stratifiziert, 51
  - zufällig, 48
  - zweistufig, 51
- Zielsetzung, 26
- Zufallsgrösse, 88
- Zufallsvariable, 88



# Willkommen zu den Springer Alerts

Jetzt  
anmelden!

- Unser Neuerscheinungs-Service für Sie:  
aktuell \*\*\* kostenlos \*\*\* passgenau \*\*\* flexibel

Springer veröffentlicht mehr als 5.500 wissenschaftliche Bücher jährlich in gedruckter Form. Mehr als 2.200 englischsprachige Zeitschriften und mehr als 120.000 eBooks und Referenzwerke sind auf unserer Online Plattform SpringerLink verfügbar. Seit seiner Gründung 1842 arbeitet Springer weltweit mit den hervorragendsten und anerkanntesten Wissenschaftlern zusammen, eine Partnerschaft, die auf Offenheit und gegenseitigem Vertrauen beruht.

Die SpringerAlerts sind der beste Weg, um über Neuentwicklungen im eigenen Fachgebiet auf dem Laufenden zu sein. Sie sind der/die Erste, der/die über neu erschienene Bücher informiert ist oder das Inhaltsverzeichnis des neuesten Zeitschriftenheftes erhält. Unser Service ist kostenlos, schnell und vor allem flexibel. Passen Sie die SpringerAlerts genau an Ihre Interessen und Ihren Bedarf an, um nur diejenigen Informationen zu erhalten, die Sie wirklich benötigen.

Mehr Infos unter: [springer.com/alert](http://springer.com/alert)