1 Eine Einführung und ein Überblick

„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

D. Bättig, *Angewandte Datenanalyse*, Statistik und ihre Anwendungen,
DOI 10.1007/978-3-662-54220-0_1

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:

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

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 Information. 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 Malariaprophylaxe 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:[1] $\mathbb{P}(\text{Malaria} \mid \mathcal{K}) = 0{,}3$. 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:

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

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[2]

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

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

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

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

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

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

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

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

[1] 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.

[2] 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:

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

Man nennt dies ein Modell für Messwerte. Es bezeichnet mit einer Wahrscheinlichkeit, wie gemessene Frequenzen um eine Frequenz *streuen*. 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 bis 300 mV	3–5 Hz	0,10
	5–10 Hz	0,05
	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 technischem 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

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

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 μ 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:

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

Diese Wahrscheinlichkeit beschreibt, wie Zeiten zwischen starken Erdbeben um die durchschnittliche Wartezeit μ 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 μ 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 μ und die Information aus den Daten:

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

Wir wollen hier von einer unerfahrenen Person ausgehen, die wenig Vorwissen $\mathcal{K}$ hat. Sie nimmt an, dass μ 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 μ 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 μ nicht präzis. Man weiss ja aus (a) nur, wo μ mit hoher Wahrscheinlichkeit liegt. Daher lässt sich μ nicht einfach in die Gleichung (1.3) einsetzen. Man braucht dazu das Gesetz der Marginalisierung. Es besagt, dass man alle möglichen Werte μ in diese Gleichung einsetzt und diese nach ihrer Wahrscheinlichkeit gewichtet mittelt. In einer Formel ausgedrückt ist also

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

Die Summe kann man bestimmen.[3] 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. □

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 µm. 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. □

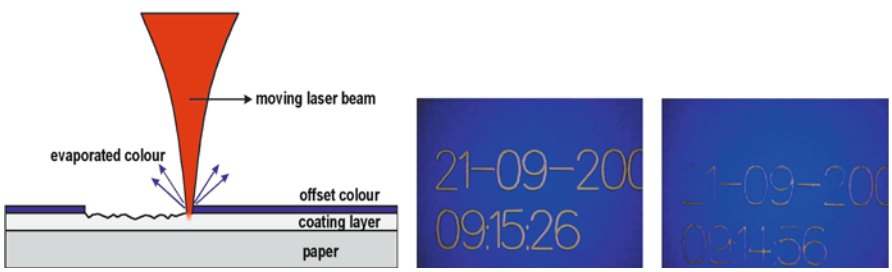

Abb. 1.1 Lasermarkierung und zwei mit Calcium-Karbonatfarbe beschichtete Papiere, markiert mit Laserstrahl

[3] Die Rechnung wird später erklärt.

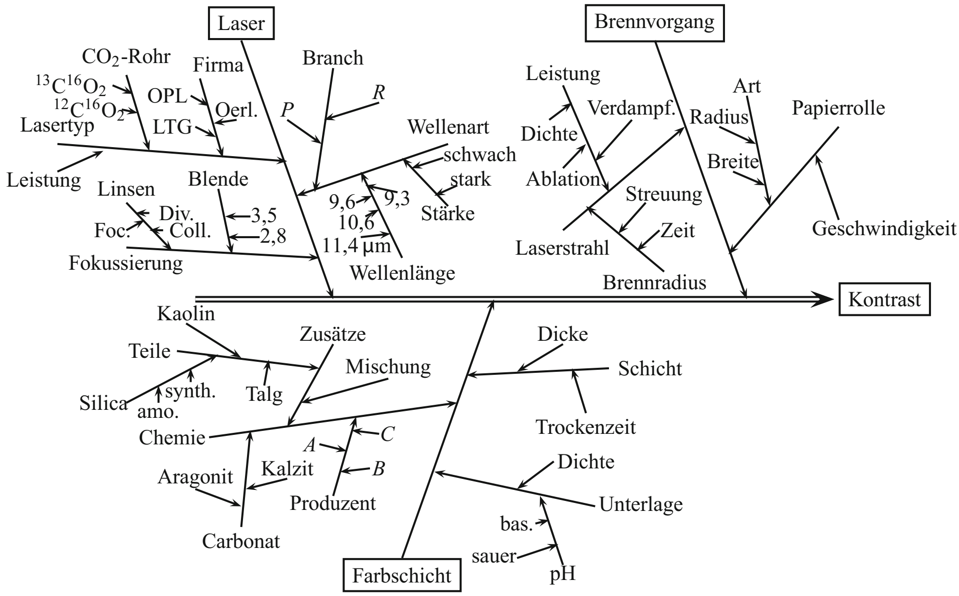

Abb. 1.2 Ein Ursache-Wirkungs-Diagramm

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-

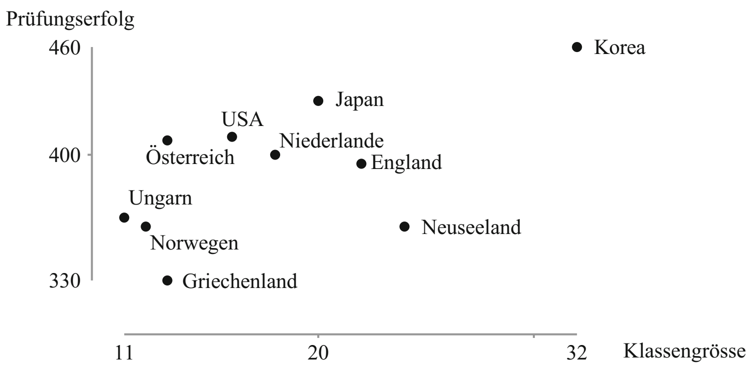

Abb. 1.3 Klassengrösse und Lernerfolg bei zehn Ländern, OECD Untersuchung *Education at Glance* aus dem Jahr 2005

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*).[4]

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:

[4] Konfundieren heisst vermengen oder durcheinander geraten.

http://www.springer.com/978-3-662-54220-0

Angewandte Datenanalyse
Der Bayes'sche Weg
Bättig, D.
2017, XVI, 393 S. 269 Abb., 11 Abb. in Farbe.,
ISBN: 978-3-662-54220-0