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 *einzigen* Werkzeug, der Regel von Bayes. Die Regel erklärt, wie man Informationen aus Daten und Zusatzinformationen verarbeiten kann. Die Zusatzinformation kann früheres 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

im Verzeichnis *Bücher* oder unter www.baettig.one verfügbar. R ist eine unter 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.

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

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