Statistik und ihre Anwendungen

Walter Zucchini
Andreas Schlegel
Oleg Nenadić
Stefan Sperlich

Statistik für Bachelor- und Masterstudenten

Eine Einführung für Wirtschafts- und Sozialwissenschaftler

Springer

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

Azizi Ghanbari, S.
Einführung in die Statistik für Sozial- und Erziehungswissenschaftler 2002

Bickeböller, H.; Fischer, C.
Einführung in die Genetische Epidemiologie 2007

Dehling, H.; Haupt, B.
Einführung in die Wahrscheinlichkeitstheorie und Statistik
2. Auflage 2004

Dümbgen, L.
Stochastik für Informatiker 2003

Falk, M.; Becker, R.; Marohn, F.
Angewandte Statistik 2004

Franke, J.; Härdle,W.; Hafner, C.
Einführung in die Statistik der Finanzmärkte
2. Auflage 2004

Greiner, M.
Serodiagnostische Tests 2003

Handl, A.
Multivariate Analysemethoden 2003

Hilgers, R.-D.; Bauer, R.; Scheiber, V.
Einführung in die Medizinische Statistik 2. Auflage 2007

Kohn, W.
Statistik Datenanalyse und Wahrscheinlichkeitsrechnung 2005

Kreiß, J.-P.; Neuhaus, G.
Einführung in die Zeitreihenanalyse 2006

Ligges, U.
Programmieren mit R 3. Auflage 2008

Meintrup, D.; Schäffler, S.
Stochastik Theorie und Anwendungen 2005

Plachky, D.
Mathematische Grundbegriffe der Stochastik 2002

Pruscha, H.
Statistisches Methodenbuch Verfahren, Fallstudien, Programmcodes 2005

Schumacher, M.; Schulgen, G.
Methodik klinischer Studien 3. Auflage 2008

Steland, A.
Mathematische Grundlagen der empirischen Forschung 2004

Zucchini, W.; Schlegel, A.; Nenadić, O.; Sperlich S.
Statistik für Bachelor- und Masterstudenten 2009

Walter Zucchini · Andreas Schlegel
Oleg Nenadić · Stefan Sperlich

Statistik für Bachelor- und Masterstudenten

Eine Einführung für Wirtschafts- und Sozialwissenschaftler

Springer

Prof. Dr. Walter Zucchini
Andreas Schlegel
Dr. Oleg Nenadić
Prof. Dr. Stefan Sperlich

Universität Göttingen
Institut für Statistik und Ökonometrie
Platz der Göttinger Sieben 5
37073 Göttingen

ISBN 978-3-540-88986-1 e-ISBN 978-3-540-88987-8
DOI 10.1007/978-3-540-88987-8
Springer Dordrecht Heidelberg London New York

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

Einbandentwurf: WMXDesign GmbH, Heidelberg

Gedruckt auf säurefreiem Papier

Springer ist Teil der Fachverlagsgruppe Springer Science+Business Media (www.springer.de)

Für Wilfried Ahlborn, mit Dank für seinen langjährigen Einsatz für unser Institut.

Vorwort

Wie der Titel bereits andeutet, richtet sich dieses Buch an Anfänger im Fach Statistik und eignet sich für Bachelor- und Masterstudenten aller Disziplinen, auch wenn viele Beispiele ihren Schwerpunkt in den Wirtschaftswissenschaften haben. Die wesentlichen Konzepte statistischer Methoden, ihre Ideen und Umsetzung werden im Detail erklärt und anhand von Beispielen erläutert. Das Buch enthält daher recht viele, wenn auch meist einfache, Beispiele aus der Praxis, die das Interesse der Leser wecken und die Anwendung der Methoden illustrieren sollen. Der Stil ist betont mathematisch informell, wenn auch mathematisch exakt, denn unser Fokus war primär das Vermitteln der zugrunde liegenden Konzepte. Aus diesem Grund handelt es sich bei diesem Buch auch nicht um ein Referenz- oder Nachschlagewerk, sondern um ein vorlesungsunterstützendes Buch, das natürlich ebenso für das Selbststudium geeignet und gedacht ist.

Das heute vorliegende Werk entstand aus den vorlesungsbegleitenden Unterlagen der Statistik-Grundvorlesungen an der Wirtschaftswissenschaftlichen Fakultät der Georg-August-Universität Göttingen. Der Umfang des vorgestellten Lehrstoffes ist auf eine Veranstaltung mit ca. 45 Stunden Vorlesung und 45 Stunden Übungen (15 in Großübung, 30 in Kleingruppen) ausgerichtet. Außerdem liegen noch ca. 30 Stunden praktische Computerübung, für gewöhnlich mit der statistischen Programmiersprache R, der Veranstaltung zugrunde. Daher sind, als begleitende Ergänzungen aber auch als eigenständige Volumina, zwei weitere Werke in Vorbereitung: Ein Übungsbuch mit Rechen- und Wissensfragen und eine deutschsprachige Einführung in die statistische Programmiersprache R.

Am viele Jahre dauernden Prozess, in dem dieses ursprüngliche Manuskript zum jetzigen Buch gewachsen ist, haben viele weitere Personen mitgewirkt. Wir danken für ihre Beiträge und Hilfe insbesondere Prof. Dr. Fred Böker, Herrn Philipp Kunze, Frau Ellen Riefling, Dr. Britta Schnoor, Frau Katja Stempel und Herrn Michael Vorfeld.

Göttingen,
November 2008

Walter Zucchini
Andreas Schlegel
Oleg Nenadić
Stefan Sperlich

Inhaltsverzeichnis

Kapitel 1
Der Zufall in unserer Welt — Einführende Beispiele und Grundbegriffe

Jeder hat eine Vorstellung davon, was man unter **Statistik** versteht, und viele denken dabei sicherlich zunächst an umfangreiche Tabellen oder grafische Darstellungen, die bestimmte Sachverhalte in komprimierter Weise verdeutlichen. Dies ist jedoch nur ein Teil der Statistik, die sogenannte beschreibende oder **deskriptive Statistik**, die dazu dient, umfangreiche Datensätze mit Hilfe von Abbildungen und Kennzahlen anschaulich darzustellen.

In den meisten Fällen geht die Statistik jedoch weit über die reine Beschreibung von Datensätzen hinaus. In der Regel sind vorliegende Daten nur eine Stichprobe aus einer sogenannten Grundgesamtheit, und man möchte aus der Stichprobe Schlussfolgerungen für die Grundgesamtheit ziehen. Dieser Teil der Statistik wird schließende oder **induktive Statistik** genannt.

Zu Beginn dieses ersten Kapitels werden zunächst einige praktische Anwendungsbeispiele statistischer Methoden vorgestellt, um einen Eindruck von den vielfältigen Anwendungsmöglichkeiten der Statistik zu vermitteln. Im hinteren Teil des Kapitels werden dann einige wichtige Grundbegriffe der Statistik, wie zum Beispiel Stichprobe und Grundgesamtheit, eingeführt.

1.1 Deterministische und stochastische Modelle

Bevor die einführenden Anwendungsbeispiele statistischer Methoden und **Modelle** vorgestellt werden, sollen zunächst die Begriffe **deterministisches Modell** und **stochastisches Modell** erläutert werden. Dazu ist zunächst der Begriff des Modells zu definieren. Ein Modell lässt sich etwa als vereinfachte Beschreibung der Realität definieren. Ein anschauliches Beispiel ist eine Landkarte, die eine bestimmte, reale Landschaft vereinfacht auf einem Blatt Papier beschreibt.

Man hat es in der Statistik immer mit Daten zu tun, d.h. mit Größen, die gemessen, gezählt oder auf andere Art und Weise quantifiziert werden können. Statistische Modelle werden durch mathematische Formeln, durch Zahlen oder als Grafik gegeben. Auf dieser Grundlage ist eine engere Definition des Modells sinnvoll:

W. Zucchini, A. Schlegel, O. Nenadić, S. Sperlich, *Statistik für Bachelor- und Masterstudenten*, © Springer-Verlag Berlin Heidelberg 2009

Ein **Modell** ist die Beschreibung eines quantitativ erfassbaren Phänomens.

Die nachfolgenden Beispiele und Bemerkungen dienen dazu, einen Eindruck von der Bedeutung stochastischer, d.h. zufallsabhängiger Sachverhalte in verschiedensten Bereichen des Lebens, sowie der Anwendung statistischer Modelle in diesem Zusammenhang zu vermitteln. Lediglich das erste Beispiel ist nicht durch zufällige Einflüsse geprägt. Es stellt eher die Ausnahmesituation als die Regel dar.

Beispiel 1.1. Schwingungsdauer eines Pendels

Zunächst wird ein Beispiel für ein deterministisches Modell betrachtet, und zwar für die Schwingungsdauer eines Pendels. Die Physik liefert dazu eine Theorie, aus der man ableiten kann, dass die Schwingungsdauer T von der Länge L des Pendels abhängt und durch die Gleichung

$$T = 2\pi\sqrt{\frac{L}{g}}$$

beschrieben wird, wobei π die Kreiszahl und g die Erdbeschleunigung bezeichnet.

Um die Schwingungsdauer T eines realen Pendels zu berechnen und somit das Modell für einen bestimmten Zweck zu verwenden, benötigt man die Länge des Pendels und die Erdbeschleunigung am Ort des Pendels. In Göttingen z.B. ist g etwa $9.81\,m/s^2$ (die Einheit ist hier *Meter dividiert durch Sekunde zum Quadrat*); dann ergibt sich beispielsweise für $L = 7.5\,m$ die Schwingungsdauer

$$T = 2\pi\sqrt{\frac{7.5\,m}{9.81\,m/s^2}} = 5.5\,s\,.$$

Dabei wird vorausgesetzt, dass der Pendelausschlag klein gegenüber der Pendellänge ist, d.h. dass das Pendel nur kleine Winkel durchläuft. Man benötigt also nur die Länge des Pendels L, um die Schwingungsdauer T zu bestimmen.

Diese Formel ist ein Beispiel für ein Modell, das den quantitativen Zusammenhang zwischen zwei Größen, der Länge eines Pendels und der Schwingungsdauer, beschreibt (bei gegebener Erdbeschleunigung). Grafisch ist dieser Zusammenhang in Abb. 1.1 dargestellt. Sowohl die Grafik als auch die Formel sind als Modell für das Pendel zu verstehen. Lediglich die Darstellungsform ist eine andere. Die Anwendungsmöglichkeit des Modells ist offensichtlich: wir können mit Hilfe dieses Modells die Schwingungsdauer T für verschiedene Werte von L bestimmen.

Falls jetzt die Frage aufkommt, aus welchem Grund man sich für die Schwingungsdauer eines Pendels interessieren sollte: bis zur Erfindung der Quarz-Uhren in den 1930er Jahren waren Pendeluhren über Jahrhunderte das genaueste Mittel der Zeitmessung, und auch heute noch hat sicher der eine oder andere eine Pendeluhr zu Hause.

Das mathematische Pendel ist ein Beispiel für eine **deterministische Beziehung** zwischen zwei Größen. Mit *deterministisch* ist gemeint, dass es für jeden Wert der

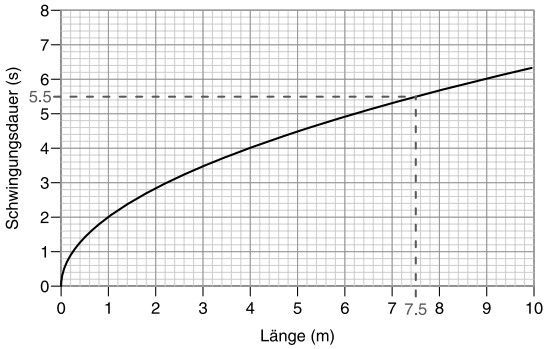

Abb. 1.1 Modell für die Schwingungsdauer eines Pendels in Abhängigkeit von der Länge

Länge L *genau einen* Wert für die Schwingungsdauer T gibt. Die Schwingungsdauer ist durch die Länge des Pendels determiniert. Es gibt hier keine Unsicherheit oder Unbestimmtheit. Wenn die Länge des Pendels bekannt ist, kennt man auch die Schwingungsdauer. Ganz anders ist die Beziehung zwischen zwei Größen in dem folgenden Beispiel.

Beispiel 1.2. Blockzeit eines Linienfluges

Wer privat oder geschäftlich mit dem Flugzeug reist, ist sicherlich nicht nur daran interessiert, sicher am Ziel anzukommen, sondern auch möglichst schnell und pünktlich. Dabei hängt die Dauer eines Linienfluges in erster Linie von der Länge der Flugstrecke ab.

Tabelle 1.1 enthält die Strecke d in nautischen Meilen (entspricht 1 852 Metern) sowie die dazugehörige Dauer t in Minuten für 100 zufällig ausgewählte inneramerikanische Flüge (mit einer Flugstrecke von maximal 1 500 Meilen) der Fluggesellschaft American Airlines im Februar 2006.[1] Die Dauer umfasst dabei die Zeit vom Losrollen eines Flugzeugs von der Start-Position bis zum Stillstand auf der Ziel-Position (*on blocks*) und wird daher auch *Blockzeit* genannt. Sie beinhaltet neben der reinen Flugzeit auch die sogenannte Taxi-Out-Zeit (Zeit vom Losrollen bis zum Abheben) sowie die Taxi-In-Zeit (Zeit vom Aufsetzen bis zum Stillstand). Die im Flugplan einer Fluggesellschaft angegebene Dauer eines Fluges stellt immer die geplante Blockzeit dar, und nicht die reine Flugzeit.

Wenn man Tabelle 1.1 betrachtet, stellt man fest, dass es den erwarteten Zusammenhang zwischen der Flugstrecke und der Blockzeit gibt; je länger die Flugstrecke, desto länger ist tendenziell die Blockzeit. Die Beziehung zwischen Flugstrecke und Blockzeit ist jedoch von anderer Art als die oben betrachtete Beziehung zwischen der Länge und der Schwingungsdauer eines Pendels. Im Falle des Pendels gehört

[1] Die Original-Daten der American Airlines Flüge im Februar 2006, die im Rahmen dieses Beispiels betrachtet werden, stammen aus der *Airline On-Time Performance Data* Datenbank, die das US-amerikanische Bureau of Transportation Statistics auf seiner Internetseite *http://www.transtats.bts.gov* zur Verfügung stellt (Stand 24. April 2008).

Tabelle 1.1 Flugstrecke d in nautischen Meilen und Blockzeit t in Minuten für 100 zufällig ausgewählte inneramerikanische American Airlines Flüge (mit einer maximalen Flugstrecke von 1 500 Meilen) im Februar 2006

d	258	1 189	1 145	258	403	612	175	733	337	761	783	468	762	1 017	888
t	64	195	178	72	78	146	46	138	70	144	100	79	175	137	150
d	748	733	416	1 437	950	888	1 121	1 235	988	1 055	583	1 217	868	1 235	1 171
t	126	105	98	220	154	143	193	193	168	174	106	207	160	170	182
d	1 145	1 062	1 389	733	1 045	1 440	190	175	1 313	175	950	868	190	1 205	551
t	204	203	197	148	158	210	67	50	182	53	147	155	63	199	115
d	1 171	1 045	236	583	1 035	1 471	867	1 162	1 017	1 055	1 171	551	1 235	641	1 068
t	173	142	65	124	179	195	126	185	172	183	196	102	181	118	168
d	569	1 431	190	733	1 464	1 235	177	190	247	786	551	1 055	592	1 182	1 213
t	89	243	49	131	199	165	62	59	82	124	96	162	115	189	166
d	551	1 302	1 372	448	190	867	762	987	678	334	964	612	1 144	177	551
t	82	182	197	86	58	167	128	164	110	86	140	142	167	59	96
d	762	612	603	1 456	1 189	861	522	1 005	733	1 438					
t	141	128	95	222	177	149	114	159	149	212					

zu jedem Wert der Länge genau ein Wert für die Schwingungsdauer. Die Schwingungsdauer ist durch die Länge eindeutig bestimmt. Im Beispiel mit der Blockzeit ist das anders. Es gibt z.B. 5 Flüge mit einer Flugstrecke von 733 Meilen und dazugehörigen Blockzeiten von 138, 105, 148, 131 und 149 Minuten. Die Blockzeit ist also nicht eindeutig durch die Flugstrecke bestimmt. Vielmehr scheint es zufällige Schwankungen zu geben. Eine solche Beziehung nennt man stochastisch. Eine grafische Darstellung der Beziehung zwischen Flugstrecke und Blockzeit gibt Abb. 1.2.

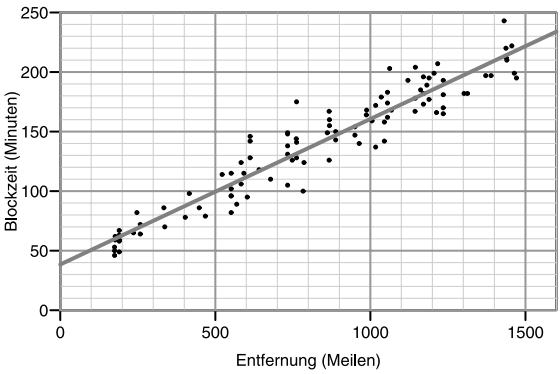

Abb. 1.2 Flugstrecke d in nautischen Meilen und Blockzeit t in Minuten für 100 zufällig ausgewählte inneramerikanische American Airlines Flüge (mit einer maximalen Flugstrecke von 1 500 Meilen) im Februar 2006 sowie angepasste Gerade