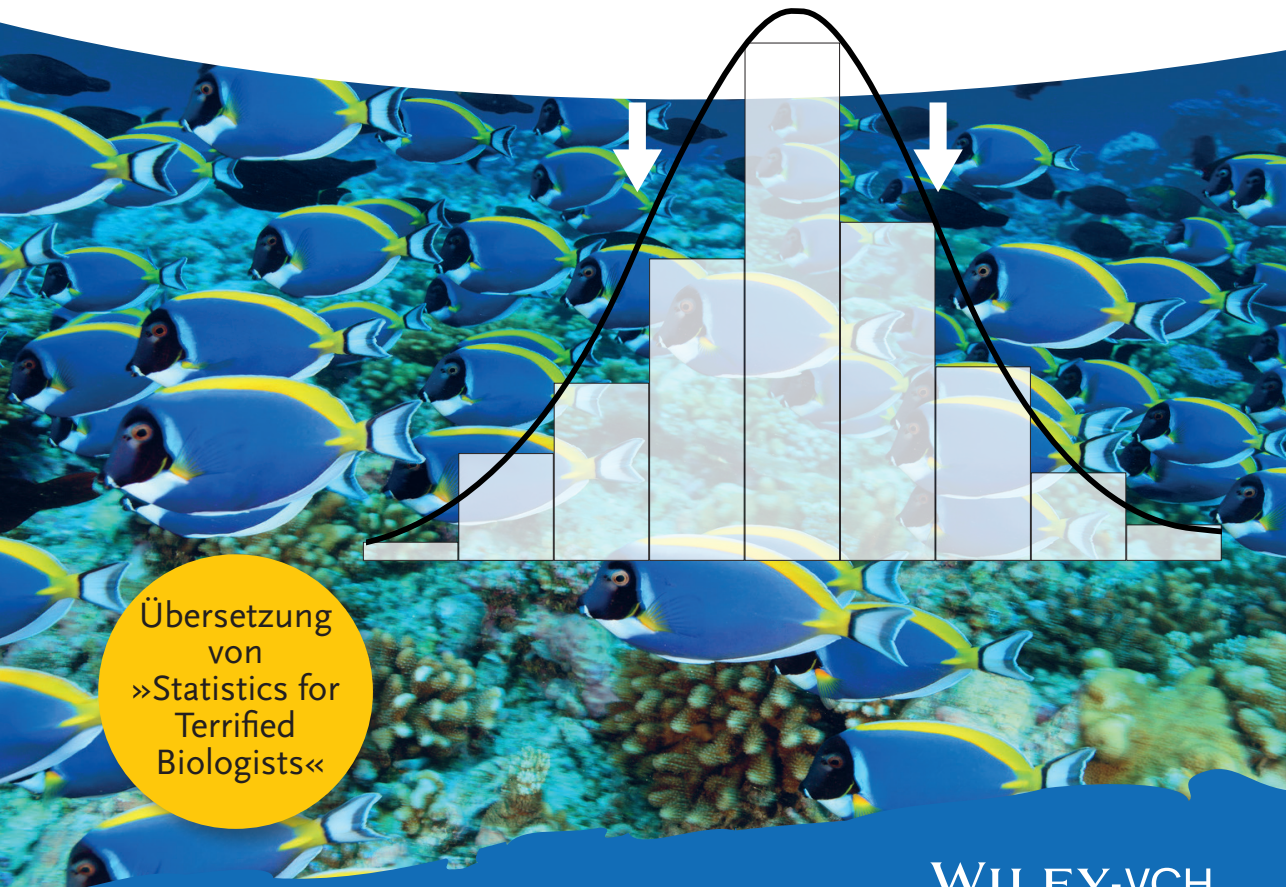




Helmut van Emden

Statistik ohne Albträume

Eine Einführung für Biowissenschaftler



Übersetzung
von
»Statistics for
Terrified
Biologists«

WILEY-VCH

Helmut F. van Emden

Statistik ohne Alpträume

**Beachten Sie bitte auch weitere interessante Titel
zu diesem Thema**

Rowe, P.

**Statistik für Mediziner und
Pharmazeuten**

2012

Print ISBN: 978-3-527-33119-2,
auch als E-Book

Fletcher, H., Hickey, I.

Genetik
für Biologen, Biochemiker,
Pharmazeuten und Mediziner

2013

Print ISBN: 978-3-527-33475-9,
auch als E-Book

McLennan, A., Bates, A., Turner, P.,
White, M.

Molekularbiologie
für Biologen, Biochemiker,
Pharmazeuten und Mediziner

2013

Print ISBN: 978-3-527-33476-6,
auch als E-Book

Alberts, B., Bray, D., Hopkin, K.,
Johnson, A., Lewis, J., Raff, M.,
Roberts, K., Walter, P.

**Lehrbuch der Molekularen
Zellbiologie**

4. Auflage

2012

Print ISBN: 978-3-527-32824-6

Lüttge, U., Kluge, M.

**Botanik – Die einführende
Biologie der Pflanzen**

6. Auflage

2012

Print ISBN: 978-3-527-33192-5

Wink, M. (Hrsg.)

Molekulare Biotechnologie
Konzepte, Methoden und Anwendungen,
2. Auflage

2011

Print ISBN: 978-3-527-32655-6

Helmut F. van Emden

Statistik ohne Albträume

Eine Einführung für Biowissenschaftler

Übersetzt von Michael Knorrenschild

WILEY-VCH
Verlag GmbH & Co. KGaA

Autor

Helmut F. van Emden

University of Reading
Department of Agriculture
Spur E, TOB1, Earley Gate
RG6 6AR Reading
United Kingdom

Übersetzer

Michael Knorrenschild

Hochschule Bochum
Fachbereich Elektrotechnik und Informatik
Lennershofstr. 140
44801 Bochum
Deutschland

© Erhan Ergin/Fotolia.com für die in der
Randspalte verwendeten Symbole

Alle Bücher von Wiley-VCH werden sorgfältig erarbeitet. Dennoch übernehmen Autoren, Herausgeber und Verlag in keinem Fall, einschließlich des vorliegenden Werkes, für die Richtigkeit von Angaben, Hinweisen und Ratschlägen sowie für eventuelle Druckfehler irgendeine Haftung.

Bibliografische Information der Deutschen Nationalbibliothek

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

© 2014 WILEY-VCH Verlag GmbH & Co. KGaA, Boschstr. 12, 69469 Weinheim, Germany

Alle Rechte, insbesondere die der Übersetzung in andere Sprachen, vorbehalten. Kein Teil dieses Buches darf ohne schriftliche Genehmigung des Verlages in irgendeiner Form – durch Photokopie, Mikroverfilmung oder irgendein anderes Verfahren – reproduziert oder in eine von Maschinen, insbesondere von Datenverarbeitungsmaschinen, verwendbare Sprache übertragen oder übersetzt werden. Die Wiedergabe von Warenbezeichnungen, Handelsnamen oder sonstigen Kennzeichen in diesem Buch berechtigt nicht zu der Annahme, dass diese von jedermann frei benutzt werden dürfen. Vielmehr kann es sich auch dann um eingetragene Warenzeichen oder sonstige gesetzlich geschützte Kennzeichen handeln, wenn sie nicht eigens als solche markiert sind.

Umschlaggestaltung Bluesea Design

Typesetting le-tex publishing services GmbH, Leipzig, Deutschland

Druck und Bindung betz-druck GmbH, Darmstadt, Deutschland

Print ISBN 978-3-527-33388-2

ePDF ISBN 978-3-527-68802-9

ePub ISBN 978-3-527-68801-2

Mobi ISBN 978-3-527-68803-6

Gedruckt auf säurefreiem Papier

Inhaltsverzeichnis

Vorwort *XI*

- 1** **Zum Gebrauch dieses Buches** *1*
 - 1.1 Einführung *1*
 - 1.2 Der Text in den Kapiteln *1*
 - 1.3 Was Sie bei auftretenden Problemen tun sollten *2*
 - 1.4 Wichtig zu wissen *3*
 - 1.5 Zahlenbeispiele im Text *3*
 - 1.6 Die Kästen *3*
 - 1.7 Wissen testen *4*
 - 1.8 Noch einmal in Kürze *4*
 - 1.9 Warum überhaupt das Ganze? *4*
 - 1.10 Mehr zum Thema *6*

- 2** **Einführung** *7*
 - 2.1 Was ist Statistik? *7*
 - 2.2 Schreibweisen *8*
 - 2.3 Schreibweisen für die Mittelwertberechnung *10*

- 3** **Streuung zusammengefasst** *11*
 - 3.1 Einführung *11*
 - 3.2 Verschiedene Größen für Streuung *12*
 - 3.2.1 Wertebereich *12*
 - 3.2.2 Gesamtabweichung *12*
 - 3.2.3 Mittlere Abweichung *13*
 - 3.2.4 Varianz *14*
 - 3.3 Warum $n - 1$? *15*
 - 3.4 Warum quadrierte Abweichungen? *16*
 - 3.5 Die Standardabweichung *17*
 - 3.6 Das nächste Kapitel *19*
 - 3.7 Wissen testen *19*

4	Summen von verschiedenen Quadraten	21
4.1	Einführung	21
4.2	Mit Rechenmaschinen geht die Berechnung der Summen von Quadraten schneller	22
4.2.1	Addierte Quadrate	22
4.2.2	Der Korrekturfaktor	22
4.3	Vorsicht vor Verwirrung mit dem Ausdruck „Summe der Quadrate“	23
4.4	Wissen testen	24
5	Die Normalverteilung	25
5.1	Einführung	25
5.2	Häufigkeitsverteilungen	25
5.3	Die Normalverteilung	26
5.4	Wie viel Prozent entsprechen einer Standardabweichungseinheit?	28
5.5	Sind die Prozentwerte immer die gleichen?	28
5.6	Andere vergleichbare Skalen aus dem Alltag	30
5.7	Die Standardabweichung als Schätzung der Häufigkeit des Auftretens einer Zahl in einer Stichprobe	30
5.8	Von Prozenten zu Wahrscheinlichkeiten	31
6	Die Relevanz der Normalverteilung bei biologischen Daten	35
6.1	Wiederholung	35
6.2	Ist unsere beobachtete Verteilung normal?	36
6.3	Was kann man tun, wenn die Verteilung zweifellos nicht normal ist?	38
6.3.1	Transformation	38
6.3.2	Gruppieren von Stichproben	40
6.3.3	Einfach nichts tun!	40
6.4	Wie viele Stichproben brauchen wir?	40
6.4.1	Faktoren, die die Anzahl der nötigen Stichproben beeinflussen	41
6.4.2	Die Berechnung der nötigen Stichprobenanzahl	41
7	Weitere Berechnungen zur Normalverteilung	43
7.1	Einführung	43
7.2	Ist „A“ größer als „B“?	44
7.3	Die Messlatte für die Entscheidung	44
7.4	Herleitung des Standardfehlers einer Differenz zweier Mittelwerte	46
7.4.1	Schritt 1: Von der Varianz der Einzelwerte zur Varianz der Mittelwerte	47
7.4.2	Schritt 2: Von der Varianz der Einzelwerte zur Varianz der Differenzen	49
7.4.3	Schritt 3: Kombination von Schritt 1 und Schritt 2; der Standardfehler der Differenz der Mittelwerte (SFDM)	51

- 7.4.4 Zusammenfassung der Berechnung des SFDM aus der Varianz der Einzelwerte 52
- 7.5 Die Bedeutung des Standardfehlers der Differenz zweier Mittelwerte 53
- 7.6 Zusammenfassung 53
- 7.7 Wissen testen 57

- 8 Der t -Test 59**
 - 8.1 Einführung 59
 - 8.2 Das Prinzip des t -Tests 60
 - 8.3 Der t -Test in statistischen Begriffen 61
 - 8.4 Warum t ? 61
 - 8.5 Tabellen für die t -Verteilung 62
 - 8.6 Der Standard- t -Test 65
 - 8.7 Der t -Test für Mittelwerte bei ungleichen Varianzen 70
 - 8.8 Der gepaarte t -Test 76
 - 8.9 Wissen testen 82

- 9 Einseitig oder zweiseitig? 83**
 - 9.1 Einführung 83
 - 9.2 Warum ist die Varianzanalyse mit dem F -Test zweiseitig? 83
 - 9.3 Der zweiseitige F -Test 84
 - 9.4 Wievielseitig ist nun der t -Test? 85
 - 9.5 Fazit zur Frage einseitig oder zweiseitig 86

- 10 Varianzanalyse – Was ist das? Wie geht das? 87**
 - 10.1 Einführung 87
 - 10.2 Summen der Abweichungsquadrate in der Varianzanalyse 88
 - 10.3 Ein fiktives Zahlenbeispiel zur Analyse mit Anova 88
 - 10.4 Die Tabelle für die Summe der Abweichungsquadrate 90
 - 10.5 Die Aufteilung der Streuung in Tabelle C mit Anova 90
 - 10.6 Die Beziehung zwischen t und F 99
 - 10.7 Einschränkungen bei der Varianzanalyse 101
 - 10.8 Vergleich zwischen Gruppenmittelwerten in der Varianzanalyse 104
 - 10.9 Der kleinste signifikante Unterschied (LSD) 106
 - 10.10 Eine Warnung zum Gebrauch des kleinsten signifikanten Unterschieds 108

- 11 Versuchsplanung zur Varianzanalyse 113**
 - 11.1 Einführung 113
 - 11.2 Volle Randomisierung 114
 - 11.3 Randomisierte Blöcke 118
 - 11.4 Unvollständige Blöcke 124
 - 11.5 Lateinische Quadrate 126

- 11.6 Split-Plot-Pläne 135
- 11.7 Wissen testen 136

- 12 Einführung in die faktorielle Versuchsplanung 139**
- 12.1 Was ist ein faktorieller Versuch? 139
- 12.2 Interaktion 141
- 12.3 Wie verändert ein faktorieller Versuch die Form der Varianzanalyse? 145
- 12.4 Summen der Abweichungsquadrate für Interaktionen 147

- 13 Zweifaktorielle Versuche 149**
- 13.1 Einführung 149
- 13.2 Ein Beispiel für einen 2-Faktor-Versuch 149
- 13.3 Analyse des 2-Faktor-Versuchs 150
- 13.4 Zwei wichtige Punkte zur Erinnerung, bevor es ans nächste Kapitel geht 157
- 13.5 Analyse von faktoriellen Versuchen mit uneinheitlicher Anzahl Wiederholungen 157
- 13.6 Wissen testen 161

- 14 Faktorielle Versuche mit mehr als zwei Faktoren (kann bei Bedarf übersprungen werden) 163**
- 14.1 Einführung 163
- 14.2 Verschiedene „Ordnungen“ von Interaktion 164
- 14.3 Beispiel für einen 4-Faktor-Versuch 165
- 14.4 Wissen testen 184

- 15 Faktorielle Versuche mit Split-Plots 187**
- 15.1 Einführung 187
- 15.2 Herleitung des Split-Plot-Plans aus dem randomisierten Versuchsplan 188
- 15.3 Freiheitsgrade in der Split-Plot-Analyse 191
- 15.4 Ein Zahlenbeispiel für einen Split-Plot-Versuch mitsamt Analyse 194
- 15.5 Vergleich von Split-Plot- und randomisierten Block-Plan 199
- 15.6 Anwendungen von Split-Plot-Plänen 202
- 15.7 Wissen testen 204

- 16 Der t-Test in der Varianzanalyse 205**
- 16.1 Einführung 205
- 16.2 Kurze Wiederholung aus relevanten früheren Abschnitten 206
- 16.3 Test auf kleinsten signifikanten Unterschied 207
- 16.4 Mehrfachreihentests 208
- 16.5 Das Testen von Differenzen zwischen Mittelwerten 213

- 16.6 Darstellung der Testergebnisse auf Unterschiede zwischen Mittelwerten 215
- 16.7 Die Analyse der Versuchsergebnisse mit Varianzanalyse in den Kapiteln 11 bis 15 216
- 16.8 Wissen testen 226

- 17 Lineare Regression und Korrelation 229**
 - 17.1 Einführung 229
 - 17.2 Ursache und Wirkung 230
 - 17.3 Weitere Fallstricke, die nur auf Sie warten 230
 - 17.4 Regression 235
 - 17.5 Unabhängige und abhängige Variablen 236
 - 17.6 Der Regressionskoeffizient b 236
 - 17.7 Berechnung des Regressionskoeffizienten b 238
 - 17.8 Die Regressionsgleichung 244
 - 17.9 Ein durchgerechnetes Beispiel mit realen Daten 245
 - 17.10 Korrelation 253
 - 17.11 Verallgemeinerungen der Regressionsanalyse 256
 - 17.11.1 Nichtlineare Regression 258
 - 17.11.2 Mehrfache lineare Regression 260
 - 17.11.3 Mehrfache nichtlineare Regression 261
 - 17.11.4 Kovarianzanalyse 262
 - 17.12 Wissen testen 265

- 18 Chi-Quadrat-Tests 267**
 - 18.1 Einführung 267
 - 18.2 Wann χ^2 und wann nicht 268
 - 18.3 Das Problem niedriger Häufigkeiten 269
 - 18.4 Yates' Kontinuitätskorrektur 269
 - 18.5 Der χ^2 -Anpassungstest 270
 - 18.6 Der χ^2 -Unabhängigkeitstest 279
 - 18.7 Wissen testen 284

- 19 Nichtparametrische Methoden – was ist das? 287**
 - 19.1 Klarstellung 287
 - 19.2 Einführung 288
 - 19.3 Vor- und Nachteile der beiden Varianten 289
 - 19.4 Einige Beispiele für die Datenorganisation in nichtparametrischen Tests 291
 - 19.5 Die wesentlichen verfügbaren nichtparametrischen Methoden 294

- Anhang A Wie viele Wiederholungen?** 297
- A.1 In diesem Kapitel 297
- A.2 Die Konzepte dahinter 297
- A.3 „Simple“ Berechnung der Anzahl der notwendigen Wiederholungen 301
- A.4 Genauere Berechnung der Anzahl der notwendigen Wiederholungen 302
- A.5 Wie man das Gegenteil beweist 304

Anhang B Statistische Tabellen 305

Richtig gelöst 315

Mehr zum Thema 333

Stichwortverzeichnis 335

Vorwort

Ich habe in meinem Fachgebiet, Insektenkunde in Landwirtschaft und Gartenbau, mehrere Bücher geschrieben und herausgegeben, aber jedesmal auf Anregung eines Verlags oder Kollegen. Dieses Buch ist auf zweierlei Weise anders. Zum einen ist es ein Buch, das ich ganz bewusst immer schon schreiben wollte, und zum anderen verlasse ich damit meinen Komfortbereich.

Die Ursprünge dieses Buches gehen bis zu meiner Berufung an die Abteilung für Gartenbau der Universität Reading unter Prof. O.V.S. Heath, FRS, zurück. Prof. Heath war die Wichtigkeit von Statistik bewusst: Zu einer Zeit, als universitätsweite Statistik-Kurse nicht vorgesehen waren, lehrte er es den Studenten im Abschlussjahrgang selbst. Nach seiner Emeritierung wurde der Kurs von seinem Nachfolger, Prof. Peter Huxley, übernommen und ich wurde gebeten begleitend dazu praktische Übungen zu veranstalten. Man kann eigentlich nichts unterrichten, was man nicht selbst versteht, aber ich habe es dennoch versucht.

Mit der Zeit übernahm ich dann den kompletten Kurs selbst. Zu der Zeit lief er im zweiten Studienjahr und wurde im dritten fortgesetzt durch einen fakultätsweiten Kurs. Die Vorlesungen habe ich nicht fortgeführt, der gesamte Kurs spielte sich in einem Labor ab, wo ich die Studierenden schrittweise durch die Berechnungen (mit Taschenrechnern) führte. Diese räumliche Situation bedeutete ständige Interaktion mit Studierenden in einer gänzlich anderen Weise als in Vorlesungen und so wurde mir schnell klar, dass viele Biologen neuronal so verdrahtet sind, dass für sie die herkömmlichen Art von Statistikunterricht ein großes Hemmnis ist.

Was meine Studierenden benötigten war Vertrauen – Vertrauen, dass statistische Ideen und Methoden nicht pure Theorie sind, sondern tatsächlich mit realen biologischen Daten arbeiten und vor allem einen logischen Hintergrund haben! Mit den Jahren begann ich zu verstehen, dass die Studierenden regelmäßig an denselben Stellen stolperten und ihr Vertrauen Schaden nahm. Jedes Jahr probierte ich neue Wege aus um ihnen über diese Krisenstellen hinweg zu helfen: Mit der Zeit war ich bei allen erfolgreich, wurde mir berichtet.

Die Wirksamkeit meiner ungewöhnlichen Lehrmethoden konnte tatsächlich gemessen werden. Nachdem meine Studierenden den fakultätsweiten Kurs, der von professionellen Statistikern unterrichtet wurde, hinter sich hatten, wurden

sie gemeinsam mit Gruppen von Studierenden aus anderen Abteilungen der Fakultät (die damals „Landwirtschaft und Ernährung“ hieß) geprüft. Diese anderen Studierenden hatten denselben Kurs „Angewandte Statistik“ im dritten Studienjahr besucht. Die meisten meiner Studierenden (etwa alle bis auf drei aus ca. 20) fanden sich in der Notenliste in einem abgesetzten Block am oberen Ende wieder, sie hatten 70 %, 80 % oder sogar 90 % der Punkte erzielt. Obwohl es auch ein bis zwei Studierende aus anderen Kursen mit hohen Punktzahlen gab, gab es dahinter doch eine größere Lücke, bis die Studierenden mit 60 % kamen und von da ab ging es kontinuierlich hinunter bis in den einstelligen Prozentbereich.

Ich bin daher zuversichtlich, dass dieses Buch mit seinen Kürzeln wie QADS und seinem „erst den Flur entlang laufen, bevor es aufwärts geht“ nützlich für Biologen sein wird. Andere bisher unbekannte Begriffe sind die „Leitlinie“ und „Supersummen“ mit ihren Indizes – nichtsdestotrotz wurde all das von meinen Studierenden über die Jahre als sehr hilfreich eingeschätzt. Eines wird auffallen, wenn man durch das Buch blättert: Wo sind denn die Gleichungen und Formelzeichen? Diese sind weitgehend durch Zahlen und Worte ersetzt. Die Biologen, die ich unterrichtete – und ich halte diese nicht für atypisch – wussten mit einer „45“ gut etwas anzufangen, aber selten mit einem „ x “. Außerdem habe ich festgestellt, dass Studierende einige statistische Prinzipien vergessen und dann mit ihren Berechnungen in Komplikationen geraten. Die betreffenden Stellen sind im Buch mit einem kleinen Schlüssel markiert – „Wichtig zu wissen ...“.



Das Buch beschränkt sich auf die traditionellen Grundlagen der parametrischen Statistik: t -Test, Varianzanalyse, lineare Regression und χ^2 -Test. Jedoch wird auch darauf hingewiesen, wenn es irgendwo wichtige Erweiterungen dieser Techniken gibt. Im letzten Kapitel finden Sie eine Einführung in nichtparametrische Tests, aber ohne Erklärungen der Rechentechniken. Das deshalb, weil die Prinzipien der Signifikanz-Tests und die zugehörigen Rechentechniken den Umfang des Buches sprengen würden. Die Rechentechniken und dazu gehörige Tabellen findet man leicht im Internet oder in umfangreicheren Lehrbüchern. In diesem Kapitel finden Sie aber ein ausgearbeitetes Beispiel zu jedem beschriebenen Test und auch eine Checkliste für nichtparametrische Methoden verbunden mit ihren parametrischen Gegenstücken.

Viele Kapitel enden mit einer kurzen Zusammenfassung, die das schnelle Nachschlagen erleichtert. Dazu gibt es Übungsaufgaben, deren Bearbeitung wesentlich für den Lernerfolg ist.

Um Formelrechnung in Grenzen zu halten werden Rechnungen anhand zahlreicher Beispiele erläutert. Diese – genauso wie die Übungsaufgaben – stammen aus vielen Quellen, die mir leider im Laufe der Zeit abhanden gekommen sind. Eine ganze Reihe von Experimenten wurden von Gartenbau-Studierenden in Reading im Rahmen ihrer Praktika im zweiten Studienjahr durchgeführt. Andere dagegen wurden vollständig erfunden, damit alles glatt aufgeht. In weiteren wurden Zahlen oder Versuche dem jeweiligen Zweck angepasst. Ich kann mich nur bei denen entschuldigen, deren Daten ich verwendet habe ohne sie zu erwähnen; es steckt keine Absicht dahinter. Bedenken Sie aber, dass die Daten oftmals geändert

oder manipuliert wurden – sehen Sie also die Ergebnisse nicht als wissenschaftlichen Nachweis für das an, was sie scheinbar zeigen!

Heutzutage übernehmen Computerprogramme für statistische Verfahren das, was früher schweißtreibende Arbeit war. Außerdem können viele Biologen auf professionelle statistische Beratung zurückgreifen. „Warum also überhaupt die lästige Beschäftigung mit statistischen Grundlagen?“ ist daher eine berechtigte Frage, ähnlich wie „warum selbst bellen, wenn man einen Hund hat?“ Die Fragen verdienen eine Antwort, aber um mich nicht zu wiederholen: Meine Antwort darauf finden Sie gegen Ende von Kapitel 1.

Ich bin Generationen von Studierenden aus Reading außerordentlich dafür dankbar, dass sie mich herausgefordert haben ihre „Hänger“ zu überwinden und die damit wesentlich zu jedem Erfolg, den dieses Buch haben mag, beigetragen haben. Neben diesen haben mich viele Studierende aus weiterführenden Studiengängen und erfahrene Gastwissenschaftler aus Übersee angeregt meinen Kurs in Buchform zu veröffentlichen. Meine Liebe und besonderer Dank geht an meine Frau Gillian, die mich vor dem Hintergrund ihrer eigenen Erfahrung mit biologischer Statistik unterstützt und ermutigt hat, dieses Buch zu schreiben. Ihr verdanke ich auch die Idee für den Originaltitel des Buches „Statistics for Terrified Biologists“.

Großen Dank schulde ich auch den sehr hilfreichen, konstruktiven und ermutigenden Kommentaren von Prof. Rob Marrs (Fachbereich Biologie der Universität Liverpool) und Dr. Steve Coad (Fachbereich Mathematik der Queen Mary Universität London). Beide haben sich auf Wunsch des Verlags der Mühe unterzogen den gesamten Text zu lesen und dafür eine Menge Zeit investiert. Mögliche Missverständnisse ihrer Ratschläge oder andere Fehler gehen natürlich nicht auf sie zurück, sondern auf mich. Schließlich möchte ich Ward Cooper vom Blackwell Verlag danken für sein Vertrauen in diesen Biologen, der weniger vor Statistik zurückschreckt als früher.

Mai 2007

Helmut van Emden

Zum Gebrauch dieses Buches

1

In diesem Kapitel ...

... geht es um folgende Themen:

- Der Text in den Kapiteln
- Was Sie bei auftretenden Problemen tun sollten
- Wichtig zu wissen
- Zahlenbeispiele im Text
- Die Kästen
- Wissen testen
- Noch einmal in Kürze
- Warum überhaupt das Ganze
- Das Literaturverzeichnis

1.1 Einführung

Damit keine Missverständnisse aufkommen: Dieses Buch ersetzt nicht Ihre eigenen Bemühungen. Alles, was es tun kann, ist diese Bemühungen effektiv zu gestalten. Der Krimi ist nur erfolgreich, weil Sie ihn zu schnell und unkonzentriert gelesen haben – auf **diesem** Weg wird auch dieses Buch ein Mysterium bleiben.

1.2 Der Text in den Kapiteln

In den Kapiteln, insbesondere Kapitel 2–8, wird eine Denkweise entwickelt, die wesentlich für die Analyse biologischer Daten ist. Sie sollten diese Kapitel daher nacheinander und recht langsam angehen. Ich kenne nur einen Weg, der Ihnen hilft, die für das Verständnis nötige Konzentration aufrecht zu erhalten: **Schreiben Sie nebenher eigene Zusammenfassungen.**

Als ich nach Reading kam, definierte mein damaliger Dekan eine Vorlesung als „eine Technik, um Information von einem Blatt Papier vor dem Dozenten auf ein Blatt Papier vor dem Studierenden zu übertragen, ohne dass diese durch die bei-

de Köpfe läuft“. Genau deshalb betone ich: **Schreiben Sie eigene Zusammenfassungen**. Das bloße Lesen des Textes bringt wenig. Sie würden nach einiger Zeit feststellen, dass Sie an etwas ganz anderes gedacht haben, aber trotzdem dabei scheinbar einige Seiten gelesen haben – das haben wir alle schon erlebt! Das obige Zitat meines Dekans meint, das bloße Abschreiben des Gelesenen ist nur wenig besser als gar nichts zu notieren. Der Trick ist das Gelesene zu verdauen und in Ihren eigenen Worten zusammengefasst wiederzugeben. Knausern Sie nicht mit Überschriften, mit Kästchen, die durch Pfeile verbunden sind, Skizzen usw. Sie können auch verschiedenfarbige Stifte für oft wiederkehrende Begriffe wie „Varianz“ und „Korrekturfaktor“ verwenden. Kurz gesagt, benutzen Sie alles, was Sie dazu zwingt meinen Text in eine vom Original so unterschiedlich wie mögliche andere Form zu übertragen. **Das** erzwingt Konzentration und Mitdenken und macht Ihnen klar, ob Sie das im Buch Gesagte wirklich verstanden haben oder nicht. Erst wenn Sie sich sicher fühlen, ist es sinnvoll, weiter voran zu schreiten.

Das eigentliche Schreiben der Notizen ist das, worauf es ankommt – später können Sie die Notizen auch wegwerfen, wenn Sie wollen. Andererseits schadet es auch nichts, wenn Sie sie für später aufheben, um nochmal etwas zu wiederholen oder nachzuschlagen.

Gehen Sie also nicht weiter, wenn Sie nicht bereit sind. Sie tun sich keinen Gefallen, wenn Sie sich überreden weiterzulesen, wenn Sie im tiefsten Innern genau wissen, dass Sie sich selbst übers Ohr hauen.

Ein Schlüsselmoment in diesem Buch ist Abb. 7.5. An dieser Stelle sollten Sie besonders gründlich sein. Wenn Sie mit diesem Diagramm **wirklich** vertraut sind, ist das ein sicheres Omen, dass Sie alle Hindernisse bewältigt haben und vor Albträumen sicher sind.

1.3 Was Sie bei auftretenden Problemen tun sollten

Die naheliegende Antwort ist: Gehen Sie zurück zu der letzten Stelle im Buch, an der Sie noch guter Dinge waren, und starten Sie von dort noch einmal.

Oft hilft es aber auch, wenn man nachschaut, wie andere denselben Sachverhalt erklären. Es ist also eine gute Idee in die entsprechenden Seiten eines anderen Statistik-Lehrbuchs zu schauen (in der Literaturliste am Ende des Buches finden Sie ein paar Vorschläge dazu, natürlich gibt es auch noch andere gute Lehrbücher).

Eine dritte Möglichkeit wäre, jemanden zu finden, der Ihnen persönlich die Dinge erklären kann. Kennen Sie so eine hilfreiche Person? Wenn Sie an einer Universität sind, könnte das ein Kommilitone sein oder ein Angestellter. Die Person, die damals versuchte meinem Jahrgang Statistik beizubringen, versagte komplett, jedenfalls was mich persönlich betrifft. In einer Eins-zu-eins-Situation konnte er mir die Dinge aber glänzend erklären.



1.4 Wichtig zu wissen

Ich habe einige Schlüsselkonzepte der Statistik mit einem Schlüsselsymbol markiert, die nach meiner Erfahrung leicht vergessen werden mit der Folge, dass Probleme auftreten ohne dass man weiß, woher sie stammen. Vertrauen Sie mir also, es ist wirklich wirklich lohnenswert diese Themen fest in Ihrem **eigenen** Gedächtnis verankert zu haben.

1.5 Zahlenbeispiele im Text

Um „Formelrechnung“ so weit wie möglich zu vermeiden, habe ich zur Veranschaulichung der Arbeitsweisen von statistischen Analysen und Tests konkrete Zahlen verwendet. Es bringt Ihnen wahrscheinlich nicht viel, wenn Sie mit dem Taschenrechner die von mir beschriebenen verschiedenen Rechenschritte nachrechnen, aber vergewissern Sie sich in jedem Schritt, dass Sie verstehen, woher jede einzelne Zahl in der Rechnung stammt und warum sie in dieser Weise verwendet wird.

Wenn Sie allerdings am Ende einer durchgerechneten Analyse oder eines Tests angekommen sind, sollten Sie zur ursprünglichen Datenquelle im Buch zurückgehen und von dort an die Rechenschritte mit dem Taschenrechner zu reproduzieren versuchen. Widerstehen Sie der Versuchung im Buch vorzublättern um die nächsten Rechenschritte nachzuschauen, außer Sie haben sich wirklich und unwiderruflich festgefahren. In diesem Fall ist es ratsam, die Zusammenfassung am Ende eines Kapitels (falls es dort eine gibt), und nicht den vorherigen erklärenden Text, zu verwenden.

1.6 Die Kästen

Vorkenntnisse und Vorerfahrungen der Leserinnen und Leser auf dem Gebiet der Statistik sind sehr unterschiedlich, und ebenso die Fähigkeit statistische Konzepte zu erfassen und zu behalten. Demzufolge werden manche Leser an einigen Stellen lieber ohne große Erklärungen meinerseits oder ohne eine weitere Wiederholung von Rechenschritten voranschreiten wollen.

Für diejenigen aber, die an diesen Stellen die Dinge nicht einfach so hinnehmen wollen, habe ich möglicherweise hilfreiche Fakten und Rechnungen in Kästen platziert, damit die anderen nicht aufgehalten oder irritiert werden. Die Rechnungen in den Kästen können besonders hilfreich sein, wenn Sie, wie oben vorgeschlagen, ein Zahlenbeispiel aus dem Text wiederholen und dabei einen Kasten benötigen um festzustellen, warum Sie steckengeblieben sind oder vielleicht auch wo Sie vom rechten Weg abgekommen sind.



1.7 Wissen testen

Am Ende mehrerer Kapitel stoßen Sie auf Übungsaufgaben, zu deren Lösung Sie, dort angekommen, auch in der Lage sein sollten.

Das ist dann der Moment, an dem Sie innehalten sollten und diese auch wirklich lösen sollten. Anders als für die Zahlenbeispiele zwischendurch dürfen Sie dabei alles aus den vorherigen Kapitel oder Zusammenfassungen benutzen, um sich Methoden und Vorgehen nochmal in Erinnerung zu rufen. Benutzen Sie einen **Taschenrechner** und denken Sie daran, Zwischenergebnisse zu notieren. Auf diese Weise fällt es Ihnen leichter Abweichungen vom Lösungsweg zu finden, wenn Ihre Endergebnisse nicht mit denen im Abschnitt *Richtig gelöst* am Ende des Buches übereinstimmen. Lesen Sie den Beginn des Abschnittes *Richtig gelöst* frühzeitig – dort ist erklärt, dass ähnliche, aber nicht identische Endergebnisse Sie nicht beunruhigen oder zu zeitaufwändigem Nachrechnen bewegen sollten. Ich kann Ihnen versichern, dass Sie erkennen werden, ob Sie den statistischen Schritten in der Übungsaufgabe richtig gefolgt sind, wenn Sie Ihre Zahlen mit denen der Lösung vergleichen. Sie werden außerdem sofort erkennen, wenn Sie es nicht getan haben.

Es ist wirklich wichtig, dass Sie diese Übungsaufgaben mit dem Taschenrechner ganz bewusst rechnen, und zwar dann, wenn Sie auf sie im Buch stoßen und nicht erst viel später. Dies ist der beste Weg, den Ihnen das Buch bieten kann, um das Thema in Ihrem Langzeitgedächtnis zu verankern und um Ihnen Zuversicht zu vermitteln, dass Sie verstehen, was Sie tun.

Die Autoren der meisten anderen Statistik-Bücher sehen das genauso und geben Ihnen Übungsaufgaben auf den Weg. Ich möchte Sie dazu anregen weitere Zuversicht und Erfahrung zu erlangen, indem Sie die Methoden aus diesem Buch auch auf die Übungsaufgaben der anderen Büchern anwenden, wenn Sie mögen.



1.8 Noch einmal in Kürze

Manche Kapitel enden mit einer Zusammenfassung, in der die Kerninhalte des Kapitels in wenig mehr als einer Seite komprimiert dargestellt sind. In diesen Zusammenfassungen können Sie Rechnungen nachschlagen, die über das vorige Kapitel verstreut dargestellt sind, mit hoffentlich genug Erklärungen, um Ihr Gedächtnis auf Trab zu bringen, was die Herleitung der Rechnungen betrifft. Daher sind sie nützlich, wenn Sie die Aufgaben angehen.

1.9 Warum überhaupt das Ganze?

Sie fragen sich vielleicht, warum man statistische Analysen mit dem Taschenrechner lernen sollte, wo man die Daten doch einfach in eine Excel-Tabelle oder ein an-

deres Computer-Programm eingeben könnte und alle Rechnungen automatisch ablaufen würden. Das mag ja früher sinnvoll gewesen sein, aber heutzutage ...

Lassen Sie sich versichern, dass Sie das nicht fragen würden, wenn Sie so viele Projektberichte und Abschlussarbeiten wie ich gelesen hätten und dabei die Konsequenzen von „einfach in eine Excel-Tabelle oder ein anderes Computer-Programm eingeben“ erlebt hätten. Nein, das Verständnis von dem, was der Computer tun sollte, hilft auf jeden Fall Probleme zu vermeiden.

Also, warum das Ganze?

- Versuchsplanung wird beträchtlich erleichtert, wenn Sie die Vor- und Nachteile der verschiedenen Versuchspläne verstehen und wie diese den „Versuchsfehler“ beeinflussen, gegen den wir unsere Differenzen zwischen den Versuchsgruppen testen. Das wird Ihnen jetzt nicht viel sagen, aber Sie müssen wirklich verstehen, wie der Versuchsplan und auch die Anzahl von Gruppen und Wiederholungen die „restlichen Freiheitsgrade“ beeinflussen und ob Sie besser in einseitige oder zweiseitige statistischen Tabellen schauen. Mein Rat an meine Studierenden war immer schon, dass sie, bevor sie einen Versuch angehen, erst einmal eine Tabelle für die Versuchsergebnisse erstellen, ein paar fiktive Werte eintragen und mit denen dann eine geeignete Analyse durchführen. Danach überdenkt man manches noch einmal.
- Ein Computer kann Ihre Rechnungen für Sie durchführen, aber mit dem entscheidenden Haken, dass er Ihre Eingabewerte akzeptiert ohne zu hinterfragen, ob das, was Sie von ihm wollen, sinnvoll ist oder nicht. Daher – und auch hier müssen Sie mir glauben, dass das kritische Punkte sind – wird auch auf dem Bildschirm kein Fenster aufgehen, in dem steht „Okay, für diese Zahlen sollten Sie eine nichtparametrische Analyse durchführen.“ oder „Kein Problem. Ich kann eine gewöhnliche faktorielle Varianzanalyse machen, aber Sie haben wohl vergessen, dass Sie einen Split-Plot-Plan verwendet haben.“ oder „Diese Zahlen sind offensichtlich Zahlenpaare. Warum ziehen Sie daraus keinen Nutzen in dem t -Test, den ich durchführen soll?“ oder „Es überrascht mich, dass ich die Statistik für eine Gerade durch eine Punkteschar rechnen soll, wo doch die Punkte offensichtlich auf einer deutlich gekrümmten Kurve liegen.“ usw.
- In Zukunft werden Sie zweifellos eher Computer-Programme als einen Taschenrechner für Ihre statistischen Berechnungen verwenden. Allerdings sind die Computer-Ausdrucke dieser Programme oft nicht besonders nutzerfreundlich. Üblicherweise unterstellen sie einiges an Hintergrundwissen über die innere Struktur der durchgeführten Computer-Analyse und die ausgedruckten Zahlen sind zudem mit Abkürzungen bezeichnet. Offensichtlich kann dabei ein Grundverständnis davon, was ein Computerprogramm tut, und Vertrautheit mit statistischen Begriffen nur hilfreich sein.
- Etwas besonders Wichtiges werden Sie mit diesem Buch auch gewinnen: Die Überzeugung, dass statistische Methoden keine „Blackbox“ irgendwo im Computer sind. Wenn es unbedingt sein müsste, könnten Sie (mit dem Buch daneben) alle Analysen und Tests mit dem Taschenrechner auf einem Blatt Papier durchführen. Außerdem werden Sie sich beim Einsatz von Computer-Pro-

grammen wahrscheinlich wohler fühlen, wenn Sie erst einmal davon überzeugt sind, dass die Methoden in diesem Buch auf Ihnen verständlichen Konzepten basieren.

- Darüber hinaus möchten Sie wahrscheinlich gerne die Methoden, die Sie kennen, um solche erweitern, die ich hier nicht abdecke. Sie erwarten, dass auch diese „logische, sinnvolle und verständliche Wege zu befriedigenden Beurteilungen biologischer Daten“ sind. Erweiterungen der Methoden, die ich hier abdecke (z. B. die am Ende von Kapitel 17 aufgezählt sind), erfordern Zahlen aus den Berechnungen, die ich hier abdecke. Sie sollten zuverlässig in der Lage sein festzustellen, welche das sind.
- Wahrscheinlich werden Sie sich in der Lage wiederfinden, Ihr Versuchsvorhaben und später die zugehörige Analyse mit einem professionellen Statistiker zu diskutieren. Es hilft ungemein dieselbe Sprache zu sprechen! Außerdem wird Ihnen der Statistiker viel mehr nützen, wenn Sie die Kompetenz haben zu sehen, wo Letzterer eine in der biologischen Realität begründete statistische Einschränkung übersehen hat.
- Und schließlich ist da die intellektuelle Befriedigung, ein Thema gemeistert zu haben, das für Biologen harte Kost sein kann. Leider werden Sie das erst zu schätzen wissen, wenn Sie auf Ihre Mühen nach erfolgreichem Abschluss zurückblicken. Ich versichere Ihnen, diese Belohnung gibt es wirklich. Ich weiß noch genau, wie es vor vielen Jahren war, als mir mitten in einer statischen Anfänger-Vorlesung dämmerte, wie simpel die Grundidee der Varianzanalyse ist und wie diese außerordentliche Einfachheit einem Biologen durch abgekürzte Rechenwege nebulös blieb. Mit anderen Worten, ich war nun bereit Kapitel 10 zu schreiben. Später wurde mir die Kluft zwischen Biologen und versierten Statistikern richtig deutlich durch einen Kommentar eines Statistikers zu einer ersten Version dieses Buches: „Ich schlage vor, Kapitel 10 zu streichen, das ist nicht so, wie wir es machen.“ Genug davon.



1.10 Mehr zum Thema

Am Ende des Buches finden Sie eine kurze Liste weiterer deutschsprachiger Statistik-Bücher. Diese Liste stellt nur eine Auswahl der Vielzahl von Büchern auf dem Markt dar, aber dabei handelt es sich um Empfehlungen deutschsprachiger Biologen-Kollegen. Natürlich müssen deren Empfehlungen nicht das sein, was gerade für Sie am hilfreichsten ist. Das Fehlen eines Titels in dieser Liste darf auch keinesfalls als Kritik an diesem Titel verstanden werden. Wenn Sie es in der Bibliothek finden, schauen Sie es ruhig genauer an – es könnte für Sie das optimale Buch sein.

Einführung

2

In diesem Kapitel ...

... geht es um folgende Themen:

- Was ist Statistik?
- Schreibweisen
- Schreibweisen für die Mittelwertberechnung

2.1 Was ist Statistik?

„Statistik“ steht für eine Zusammenfassung oder Ansammlung von Zahlen. Wenn man sagt „In meinem Freundeskreis ist die größte Person 1,73 m groß“, so liegt damit eine Statistik basierend auf der genauen Prüfung eines Haufens von Zahlen, nämlich der verschiedenen in Ihrem Freundeskreis auftretenden Körpergrößen, vor. Das Ergebnis besteht dann eben nur in der größten Zahl.

Wenn Sie sagen „Die Durchschnittsgröße in meinem Freundeskreis ist 1,58 m“, so ist das eine andere Statistik. Auch diesmal haben Sie alle Körpergrößen in Ihrem Freundeskreis gesammelt, aber nun haben Sie alle Zahlen benutzt, um eine einzige zusammengefasste Größe zu erhalten, die Durchschnittsgröße.

Wenn Ihr Freundeskreis riesengroß ist, ist es unpraktisch, von jedem einzelnen die Körpergröße zu ermitteln. Man kann aber wahrscheinlich eine gute Schätzung der Durchschnittsgröße erhalten, wenn man nicht alle misst, sondern nur eine genügend große Stichprobe von ihnen und dann die Durchschnittsgröße dieser Stichprobe ermittelt. Im Allgemeinen ist die Durchschnittsgröße dieser Stichprobe, besonders wenn sie geringen Umfang hat, nicht identisch mit der Durchschnittsgröße aller Ihrer Freunde. Und damit sind wir bei einem Grundprinzip der Statistik. Üblicherweise wollen wir einen *Parameter* (aus dem Griechischen für „jenseits der Messung“) berechnen, indem wir eine *Schätzung* vornehmen, die auf einer leicht messbaren Stichprobe beruht. Wir müssen daher immer sorgfältig zwischen *Parametern* und *Schätzungen* unterscheiden. In der Statistik verwenden wir die Bezeichnung „Mittelwert“ für die (auf einer *Stichprobe* beruhenden) Schätzung von etwas, was wir normalerweise nicht messen können. Diesen Pa-

parameter nennen wir den „Durchschnittswert“ (der gesamten zugrunde liegenden *Grundgesamtheit* von Zahlen).

2.2 Schreibweisen

„Addiere alle Zahlen in der Stichprobe und dividiere durch die Anzahl der Zahlen“, so berechnen wir doch den Mittelwert. Selbst die Beschreibung der Berechnung einer solch simplen statistischen Größe benötigt schon einen Haufen Worte. Und es kann noch viel komplizierter werden, siehe Kasten 2.1.

Kasten 2.1

„Multipliziere jede Zahl in der ersten Spalte mit der entsprechenden Zahl in der zweiten Spalte, und addiere dann diese Produkte. Nun subtrahiere von dieser Summe die Summe der Zahlen der ersten Spalte multipliziert mit der Summe der Zahlen der zweiten Spalte, aber dividiere vorher dieses Produkt durch die Anzahl der Zahlenpaare. Nun quadriere das ganze. Dividiere dies durch einen wie folgt zu berechnenden Divisor: Getrennt für jede der beiden Spalten quadriere die Zahlen und addiere sie und subtrahiere das Quadrat der Spaltensumme, nachdem vorher noch durch die Anzahl der Zahlen geteilt wurde. Dann addiere die Ergebnisse für die beiden Spalten.“

Wir brauchen unbedingt eine Kurzform, um statistische Berechnungen zu formulieren, und diese Kurzform ist eine *Formel*. Der Nachteil von Formeln aus Biologensicht ist, dass das ganze damit einen mathematischen Charakter erhält. Es gibt auch keine einheitlich verwendeten Schreibweisen, und die Unterschiede zwischen verschiedenen Lehrbüchern können für den Anfänger ganz schön verwirrend sein!

Ein vielleicht noch schlimmeres Problem ist für viele Biologen rein psychologischer Natur – beunruhigt Sie schon allein die Erwähnung von Mathematik? Konfrontiert man einen Biologen mit einem „ x “ anstelle einer Zahl wie „57“, so stehen die Chancen gut, dass das gleich zum Dichtmachen des Hirns führt. Trotzdem gehen die meisten statistischen Berechnungen nicht über die vier Grundrechenarten hinaus – obwohl ich zugeben muss, dass Sie auch noch quadrieren und Wurzel ziehen müssen. Heutzutage kann aber all das mit den einfachsten Taschenrechnern erledigt werden.

Die Mehrheit von Ihnen hat einen eigenen Computer oder wenigstens Zugang zu einem, sodass Sie Ihre Stichprobenwerte nur in ein Spreadsheet oder ähnliches eintragen müssen und alle erforderlichen Berechnungen bereits implementiert sind. Machen Computer denn damit nicht das Verständnis ihres Tuns überflüssig? Keineswegs, sage ich Ihnen! Ich habe all das schon in Kapitel 1 ausführlicher diskutiert, aber wiederhole es hier für den Fall, dass Sie dieses Kapitel übersprungen haben. Kurz gesagt, Sie müssen wissen, welche Programme die richtigen für

welche Datentypen sind und was ihre Grenzen sind. Ein Verständnis von Datenanalyse hilft Ihnen Ihre Versuche effektiver zu planen. Vergessen Sie nicht, der Computer hat keine Skrupel Ihre Daten präzise, aber sinnlos zu berechnen, wenn Sie ihm das genauso auftragen! Es ist außerdem nützlich, wenn Sie wissen, wie Sie den Output des Computers richtig interpretieren.

Zurück zum Thema Formeln. Wie ich schon sagte, werden wir uns bald in unübersichtlichem „Zahlensalat“ wiederfinden, und der ganze Sinn von Formeln liegt darin uns vor Augen zu halten, in welcher **Reihenfolge** wir das tun. Formeln mögen schrecklich aussehen, aber sie sind es nicht wirklich. Es wäre grob fahrlässig, wenn Sie als Anfänger denken, dass Sie mithilfe von Formeln Berechnungen ohne jede Vorerfahrung durchführen können. Sie können beispielsweise nicht einfach auf S. 257 eines Statistik-Buches nachschlagen und etwas wie

$$\frac{(\sum xy)^2}{\sum x^2 + \sum y^2}$$

in Angriff nehmen ohne die nötigen Vorarbeiten mit den Bezeichnungen erledigt zu haben! Übrigens, Kasten 2.1 übersetzt diese Formel für eine zweispaltige Tabelle von Zahlenpaaren (Werte x bzw. y) in deutsche Sprache. Mit fortschreitenden Statistik-Kenntnissen wird bei jedem Teil dieser Formel etwas bei Ihnen klingeln und eine weniger formelhafte Schreibweise dafür vor Ihrem geistigen Auge hervorrufen:

$$\frac{(\text{Summe der Kreuzprodukte})^2}{\text{Summe der Quadrate für } x + \text{Summe der Quadrate für } y}$$

Das wird Ihnen wahrscheinlich im Moment noch gar nichts sagen, aber in der Lage zu sein eine Summe von Quadraten einer Menge von Werten zu berechnen ist ein genauso alltägliches Verfahren wie eine Mittelwert-Berechnung.

Auch anderswo im Leben haben wir ständig mit Schreibweisen zu tun. Sagt Ihnen „03.11.92“ etwas? Ein Datum, vielleicht ein Geburtsdatum. Die Amerikaner benutzen eine andere Schreibweise: Sie würden dasselbe Geburtsdatum als 11.03.92 schreiben. Und kennen Sie C_m ? Wahrscheinlich, wenn Sie Musiker sind – es bezeichnet das Zusammenspiel der drei Noten C , E_b und G – der C -moll-Akkord (daher C_m).

In den ersten Kapiteln dieses Buches werden Sie Schreibweisen finden, die Sie daran erinnern, welche statistischen Größen sich hinter „Summen der Abweichungsquadrate“ verbergen und wie man sie berechnet. So bald wie möglich aber werden wir Begriffe wie „Summen der Quadrate“ verwenden, um mathematische Ausdrücke zu ersetzen. Das soll die Buchseiten weniger furchteinflößend aussehen lassen und den Lesefluss erleichtern – Sie können natürlich immer zu früheren Kapiteln zurückblättern, wenn Ihnen Schreibweisen entfallen sind. Es ist ein wenig wie beim Kochen: Beim Pfannkuchen backen benötigen Sie die ersten paar Male das Kochbuch um nachzuschlagen, dass Sie 300 ml Milch, 125 g Mehl, ein Ei und etwas Salz brauchen, aber eines Tages reicht allein das Stichwort „Pfannkuchenteig“, damit Sie wissen, was zu tun ist!

2.3 Schreibweisen für die Mittelwertberechnung

Niemand wird hoffentlich vor der Herausforderung einer Mittelwertberechnung aus fünf Körpergrößen – sagen wir 149, 176, 152, 180 und 146 cm – durch Summenbildung und Teilen durch 5 kapitulieren.

In der Statistik ist das Zeichen für Summe \sum , und die Reihe der Summanden wird mit einem Buchstaben bezeichnet, oft x oder y . $\sum x$ bedeutet also „addiere alle Werte der x -Reihe auf“, in unserem Beispiel die fünf Körpergrößen. n steht bei uns für die Anzahl der Werte, hier 5, d. h. die komplette Formel für den Mittelwert ist

$$\frac{\sum x}{n} .$$

Wir benötigen den Mittelwert so oft, dass sich eine noch kürzere Bezeichnung lohnt: den Buchstaben für die Zahlenreihe (z. B. x) mit einem Querstrich darüber, also \bar{x} .

Streuung zusammengefasst

3

In diesem Kapitel ...

... geht es um folgende Themen:

- Verschiedene Größen für Streuung
- Warum $n - 1$?
- Warum quadrierte Abweichungen?
- Die Standardabweichung

3.1 Einführung

Das Leben wäre sehr angenehm, wenn der Mittelwert eine geeignete statistische Größe zur Gesamtbeschreibung einer Zahlenreihe wäre. Leider ist das nicht der Fall. Stellen Sie sich vor, Sie würden öfter Streichhölzer kaufen – Ihnen ist vielleicht die Aufschrift „Inhalt durchschnittlich 48 Stück“ aufgefallen (ich frage mich immer, warum nicht 50?) Angenommen Sie haben sechs Schachteln Marke „ratsch“ gekauft, um die Inhaltsangabe zu überprüfen, und haben absolut sicher einen Inhalt von 48, 49, 49, 47, 48, und 47 gezählt. Wenn Sie zur Marke „ritsch“ wechseln, können Sie die Angabe „Inhalt durchschnittlich 48 Stück“ auch nicht widerlegen, wenn Sie sechs Schachteln mit 12, 62, 3, 50, 93 und 68 Stück Inhalt vorfinden. Würden Sie riskieren ein Packung „ritsch“ zu kaufen? Der Mittelwert sagt überhaupt nichts darüber aus, wie häufig in seiner Nähe liegende Zahlen auftreten. Wir müssen uns mit Streuung beschäftigen, um den Mittelwert einordnen zu können.

Das Beispiel der Körpergrößen aus Kapitel 2 zeigt sofort, dass wir an Streuung nicht vorbeikommen, sobald wir uns mit biologischen Messgrößen beschäftigen. So wie die Körpergröße von Menschen variiert, variiert auch das Gewicht von Salat auf einem Feld, gibt es unterschiedlich viele Bohnenläuse an benachbarten Puffbohnenpflanzen, unsere Dahlien werden auch nicht alle am selben Tag blühen, eine „Handvoll“ Rasendünger ist nur eine ungefähre Mengenangabe und Eier vom selben Hof werden auch nicht alle die gleiche Größe aufweisen. Der Umgang mit Streuung ist also alltägliches Handwerk für jeden Biologen.

Es gibt verschiedene Wege eine Maßzahl für Streuung einer Zahlenreihe zu definieren, und wir werden diese Varianten durchgehen und die zugehörigen Formeln erklären und anwenden. Parallel dazu werden wir in den Kästen (die Sie überspringen können, wenn Sie sie nicht nützlich finden) die Berechnungen an den beiden Stichproben mit den obigen Streichholzsorten durchführen (und dabei x für „ratsch“ und y für „ritsch“ verwenden, sodass \bar{x} und \bar{y} beide 48 sind, aber eben mit verschiedener Streuung).

3.2 Verschiedene Größen für Streuung

3.2.1 Wertebereich

Die „ratsch“-Reihe x lag im Bereich 47–49, dagegen hatte „ritsch“ y den wesentlich größeren Bereich von 3–93. Obwohl man beide Reihen mit demselben Mittelwert klar am Wertebereich unterscheiden kann, wurde dieser aber nur durch zwei der sechs Werte festgelegt. Wir könnten also wirklich eine Maßzahl für Streuung gebrauchen, die **alle** Werte unserer Reihe berücksichtigt.

3.2.2 Gesamtabweichung

Um das Optimum aus unserer Stichprobe herauszuholen, brauchen wir wirklich eine Maßzahl für Streuung, die (wie es auch der Mittelwert tut) alle zur Verfügung stehenden Werte berücksichtigt. Das einfache Addieren der Werte einer Reihe reicht aber nicht, diese ergibt ja in beiden Fällen 288 (6 mal den Mittelwert von 48).

Der Ausgangspunkt für einen vernünftigen Ansatz ist die Beobachtung, dass im Falle einer identischen Anzahl von Streichhölzern in allen Schachteln einer Marke diese Anzahlen alle gleich dem Mittelwert 48 wären. Wenn sie also nicht identisch sind, sondern variieren, dann hat jede der Anzahlen einen Beitrag zur Gesamtstreuung, die *Abweichung* (Differenz) vom Mittelwert. Wir könnten also alle Differenzen vom Mittelwert addieren (in Kasten 3.1 tun wir dies für unser Streichholzbeispiel). Als Formel ist das $\sum(x - \bar{x})$. Wenn wir einmal das Vorzeichen der Differenzen ignorieren, erhalten wir als Gesamtabweichung 4 für „ratsch“ und 162 für „ritsch“.

Das sieht schon mal gut aus. Die Sache hat aber einen Haken: Die Gesamtabweichung wird immer weiter anwachsen, wenn wir mehr Werte berücksichtigen, sodass wir nur Stichproben von gleichem Umfang vergleichen können. Besser wäre es, wenn unser Maß für die Streuung unabhängig vom Stichprobenumfang wäre, so wie es auch der Mittelwert ist – wir können Mittelwerte (z. B. von Körpergrößen von Männern und Körpergrößen von Frauen) vergleichen, auch wenn die beiden Gruppen eine unterschiedliche Anzahl von Testpersonen aufweisen.

3.2.3 Mittlere Abweichung

Der offensichtliche Ausweg ist die Streuung auf die verwendete Anzahl zu beziehen:

$$\frac{\sum(x - \bar{x})}{n}$$

Diese **mittlere** (durchschnittliche) Abweichung ist weitgehend unabhängig von der Anzahl der Stichproben. Die Rechnung in Kasten 3.2 ergibt eine kleine mittlere Abweichung von 0,67 für „ratsch“ und eine viel größere von 27 für „ritsch“.

Kasten 3.1

Für „ratsch“ haben wir: $\sum(x - \bar{x}) = (48 - 48) + (49 - 48) + (49 - 48) + (47 - 48) + (48 - 48) + (47 - 48) = 0 + 1 + 1 + (-1) + 0 + (-1) = 0$, aber 4, wenn wir die Vorzeichen ignorieren.

Für „ritsch“ haben wir: $\sum(y - \bar{y}) = (12 - 48) + (62 - 48) + (3 - 48) + (50 - 48) + (93 - 48) + (68 - 48) = -36 + 14 + (-45) + 2 + 45 + 20 = 0$ wie vorher, aber 162, wenn wir die Vorzeichen ignorieren.

Wenn wir die Vorzeichen ignorieren, erhalten wir für die Reihe x eine Summe der Abweichungen (Summe der Differenzen vom Mittelwert) $\sum(x - \bar{x})$ von 4, wesentlich niedriger als der Wert von 162 für $\sum(y - \bar{y})$ für die offensichtlich stärker variierende Reihe y .

Kasten 3.2

Mittlere Abweichung für das Streichholz-Beispiel:

Für „ratsch“ (Reihe x) ist die mittlere Abweichung die Gesamtstreuung/6 = $4/6 = 0,67$. Für die stärker variierende „ritsch“ (Reihe y) ist die mittlere Abweichung $162/6 = 27$.

Gegen die mittlere Abweichung als eine Maßzahl für die Streuung ist kaum etwas einzuwenden. Aus Gründen, die später in diesem Kapitel klar werden, ist jedoch die in der Statistik übliche Maßzahl für die Streuung **nicht** die mittlere Abweichung. Dennoch bringt uns die Idee der mittleren Abweichung schon sehr nahe an die Größe, die **tatsächlich** die übliche Maßzahl für die Streuung ist, nämlich die *Varianz*.

3.2.4 Varianz

Die Varianz hängt eng mit der mittleren Abweichung zusammen, die zur Erinnerung hier nochmal wiederholt sei:

In Worten:
$$\frac{\text{Summe aller Abweichungen vom Mittelwert}}{\text{Anzahl der Werte}}$$

oder als Formel:
$$\frac{\sum(x - \bar{x})}{n}.$$

Varianz ist fast dasselbe, aber mit zwei entscheidenden Änderungen (hier fett hervorgehoben):

$$\frac{\text{Summe aller **quadrirten** Abweichungen vom Mittelwert}}{\text{Anzahl der Werte **minus eins**}},$$

als Formel:
$$\frac{\sum(x - \bar{x})^2}{n-1}.$$

Varianz ist also die mittlere (bezogen auf $n - 1$) *quadrirte* Abweichung. In Kasten 3.3 finden Sie die entsprechende Berechnung. Die Varianz von nur 0,8 für „ratsch“ ist erheblich kleiner als die von 1189,2 für die stärker variierende Reihe „ritsch“.

Kasten 3.3

Für die „ratsch“-Reihe x berechnet sich die Varianz also durch Quadrieren der 6 Abweichungen vom Mittelwert (aus Kasten 3.1), das ganze addiert und durch 5 anstelle von 6 geteilt:

$$\frac{0^2 + 1^2 + 1^2 + 1^2 + 0^2 + 1^2}{5} = 0,8$$

und für die „ritsch“-Reihe y :

$$\frac{36^2 + 14^2 + 45^2 + 2^2 + 45^2 + 20^2}{5} = 1189,2$$

Die Varianzen der beiden Reihen unterscheiden sich enorm, die der Reihe y ist um fast 1500-mal so groß wie die der Reihe x !

Zwei Begriffe werden von nun an öfter auftauchen, die Bezeichnungen für Zähler und Nenner der Varianz-Formel. Der Nenner $n - 1$ wird als *Anzahl der Freiheitsgrade* bezeichnet, den Zähler $\sum(x - \bar{x})^2$, also die Summe der quadrierten Abweichungen, werden wir *Summe der Abweichungsquadrate* nennen.¹⁾

1) Im englischen Sprachgebrauch in der Statistik wird dasselbe missverständlich verkürzt nur als „sum of squares“ bezeichnet (A.d.Ü.).