
Datenanalyse für Naturwissenschaftler, Mediziner und Ingenieure

Mario Schneider

Datenanalyse für Natur- wissenschaftler, Mediziner und Ingenieure



Springer Spektrum

Mario Schneider
Oppenau, Deutschland

ISBN 978-3-662-61865-3 ISBN 978-3-662-61866-0 (eBook)
<https://doi.org/10.1007/978-3-662-61866-0>

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

© Springer-Verlag GmbH Deutschland, ein Teil von Springer Nature 2020

Das Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt. Jede Verwertung, die nicht ausdrücklich vom Urheberrechtsgesetz zugelassen ist, bedarf der vorherigen Zustimmung des Verlags. Das gilt insbesondere für Vervielfältigungen, Bearbeitungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Verarbeitung in elektronischen Systemen.

Die Wiedergabe von allgemein beschreibenden Bezeichnungen, Marken, Unternehmensnamen etc. in diesem Werk bedeutet nicht, dass diese frei durch jedermann benutzt werden dürfen. Die Berechtigung zur Benutzung unterliegt, auch ohne gesonderten Hinweis hierzu, den Regeln des Markenrechts. Die Rechte des jeweiligen Zeicheninhabers sind zu beachten.

Der Verlag, die Autoren und die Herausgeber gehen davon aus, dass die Angaben und Informationen in diesem Werk zum Zeitpunkt der Veröffentlichung vollständig und korrekt sind. Weder der Verlag, noch die Autoren oder die Herausgeber übernehmen, ausdrücklich oder implizit, Gewähr für den Inhalt des Werkes, etwaige Fehler oder Äußerungen. Der Verlag bleibt im Hinblick auf geografische Zuordnungen und Gebietsbezeichnungen in veröffentlichten Karten und Institutionsadressen neutral.

Planung/Lektorat: Stephanie Preuss

Springer Spektrum ist ein Imprint der eingetragenen Gesellschaft Springer-Verlag GmbH, DE und ist ein Teil von Springer Nature.

Die Anschrift der Gesellschaft ist: Heidelberger Platz 3, 14197 Berlin, Germany

Dieses Buch widme ich Felix und Aurelia.

Vorwort

Ich habe die Erfahrung gemacht, dass zwar Statistik in vielen wissenschaftlichen Disziplinen Teil des Lehrplans ist, diese jedoch häufig nur theoretisch behandelt wird und nicht anhand von Beispielen aus der wissenschaftlichen Praxis, wo es viel um Datenanalyse geht. Die entsprechenden Kurse werden dann häufig von Mathematikprofessoren gehalten, und dementsprechend ist der Stoff dann auch aufbereitet. Verstehen Sie mich nicht falsch, ich achte Mathematiker sehr, es hat jedoch auch seinen Grund, dass die Mathematik nicht zu den Naturwissenschaften zählt. Falls es dann praktische Übungen im Rahmen der Statistikmodule gibt, müssen sich die Studenten zunächst eine gewisse Zeit in ein mehr oder minder komplexes Softwarepaket einarbeiten, Zeit, die ihnen für das grundlegende Verständnis, wie Daten wissenschaftlich ausgewertet werden können, am Ende fehlt. Spätestens, wenn Sie selbst einmal ein wissenschaftliches Projekt eigenständig bearbeiten sollen, holt Sie schnell die Realität ein. Hier tauchen Fragen auf wie bspw. „Sollte ich meine Daten mit der Modellfunktion A oder B fitten?“ oder „Darf ich den Wert aus meinen Daten entfernen?“ oder „Welche Faktoren beeinflussen mein Experiment?“ usw. Ich habe in meiner beruflichen Praxis z. T. mit Wissenschaftlern zu tun gehabt, die nicht wussten, was ein Median oder was die Standardabweichung von Daten ist, geschweige denn, wie man diese errechnet. Für deren Berechnung gibt es nämlich mehr als eine Formel, die je nachdem zu stark unterschiedlichen Ergebnissen führen können (später dazu mehr). Vermutlich schrecken viele Wissenschaftler davor zurück, sich mit Datenanalyse zu beschäftigen, da sie denken, ihnen fehle erstens das mathematische Rüstzeug und zweitens Kenntnisse im Umgang mit statistischer Software. Meines Erachtens nach ist es sicherlich nicht verkehrt, ein gewisses mathematisches Vorwissen mitzubringen, aber es reicht in den meisten Fällen das Wissen aus dem Grundkurs Mathematik. Sie müssen für viele Analysen auch kein R-, SPSS- oder Minitab-Profi zu sein, um Ihre Daten analysieren zu können. Vieles können Sie mit Microsoft Excel (oder vergleichbaren Tabellenkalkulationsprogrammen) machen, von beschreibender Statistik bis hin zu *Machine Learning*. In diesem Buch werden Sie kein einziges Excel-Makro benötigen. Im Gegenteil, ich werde Ihnen zeigen, dass die meisten Berechnungen, sogar die der multivariaten Datenanalyse, mit verhältnismäßig einfachen Bordmitteln funktionieren. Mir ist

durchaus bewusst, dass es je nach Problemstellung besser geeignete Softwarelösungen gibt. Ich selbst bin ein großer Fan von R, Octave und MATLAB (The MathWorks). Doch kennen sich die meisten bereits relativ gut mit Excel aus, es ist beinahe auf jedem Bürorechner vorhanden und man erhält direktes Feedback, sobald der an einer Berechnung beteiligte Zellinhalt geändert wird. Dadurch verinnerlicht man die Grundlagen nochmals sehr gut. Hier kann ich aus eigener Erfahrung sprechen, denn mein damaliger Professor für Chemometrie, Dr. Ernst Cleve, hat auch Excel verwendet, um uns die Datenanalyse praktisch näherzubringen. Ihm möchte ich an dieser Stelle für die Durchsicht und die Anmerkungen zum Manuskript, sowie für die Bereitstellung diverser Unterrichtsmaterialien herzlich danken.

Ich habe das Buch so aufgebaut, dass wir uns im ersten Kapitel zunächst mit den Grundlagen im Umgang mit Microsoft Excel beschäftigen, v. a. damit, wie Sie Zellbezüge richtig setzen und Ihre Berechnungen automatisieren können. Außerdem lernen Sie, wie Sie in Excel mit Vektoren und Matrizen rechnen können, was in den darauffolgenden Kapiteln häufig angewendet wird.

Im zweiten Kapitel beschäftigen wir uns mit den Möglichkeiten zur Darstellung von Daten. Im Grunde genommen erhalten Sie dadurch einen kleinen Einblick in die Elemente der explorativen Datenanalyse, wobei ich dieses Wort bewusst nicht in einer der Überschriften verwendet habe. Sie lernen hier v. a., wie Sie bestimmte Daten(-Typen) in welcher Form darstellen können. Hier zeige ich Ihnen außerdem, wie Sie publikationsreife Abbildungen aus Excel heraus erzeugen können.

Im dritten Kapitel beschäftigen wir uns mit Wahrscheinlichkeitsverteilungen, die in der wissenschaftlichen Praxis an der ein oder anderen Stelle auftreten. Dies schließt diejenigen Verteilungen ein, die bei den sogenannten Hypothesentests eine wichtige Rolle spielen, mit denen wir uns ebenso in diesem Kapitel beschäftigen.

Im vierten Kapitel lernen wir verschiedene Methoden der Datenvorbehandlung kennen. Neben Skalierungs- und Glättungsmethoden beinhaltet diese auch Interpolationsmethoden, die in der Praxis des Öfteren verwendet werden.

Im fünften Kapitel kommen wir dann zur häufig in der Praxis angewendeten linearen bzw. der nicht-linearen Regression, mit deren Hilfe Sie eine Modellfunktion an Ihre experimentellen Daten anpassen können.

Im sechsten Kapitel beschäftigen wir uns mit verschiedenen Klassifikationsmethoden, um Muster in multivariaten Daten erkennen zu können. Hierbei gehe ich sowohl auf unüberwachte Methoden wie die Hauptkomponentenanalyse und Clusteranalyse ein, aber auch auf überwachte Methoden wie die des aktiven Lernens mittels künstlicher neuronaler Netze.

Im siebten Kapitel erfahren Sie dann, wie Sie Versuche mit mehreren Einflussgrößen statistisch sinnvoll planen und umsetzen können und wie Sie die Haupteinflussgrößen Ihrer Experimente bestimmen können.

Im achten Kapitel finden Sie ausführliche Lösungswege und Hilfen zu den Übungsaufgaben aus den zuvor genannten Kapiteln.

Im neunten Kapitel habe ich u.a. mathematische Grundlagen zusammengefasst, die zum besseren Verständnis der Inhalte der anderen Kapitel beitragen sollen. Außerdem finden Sie hier eine Sammlung statistischer Tabellen, die zum Lösen der Beispiele und der Übungsaufgaben benötigt werden.

Die Lösungen zu den Übungen finden Sie im nächsten Kapitel. Datenanalytische Konzepte lernt man meiner Erfahrung nach am besten, wenn man sie anhand von Übungen, oder noch besser, anhand von eigenen Problemstellungen (aus der Praxis) anwendet, daher kann ich Ihnen die praxisnahen Übungsaufgaben nur ans Herz legen. Ebenso wichtig wie die Übungen sind die Videos, die zu diesem Buch gehören. Hier werden Sie anhand von Fallbeispielen lernen, wie man Daten in Excel analysieren kann. In den Videos gehe ich komplette Analysen mit Ihnen von Anfang bis Ende durch. Ein Video kann durch Scannen des jeweiligen QR-Codes mittels QR-Code-Scanner-App aufgerufen werden. Sie werden hierbei auf eine YouTube-Website weitergeleitet, auf der Sie das jeweilige Video anschauen können. Zusätzlich stelle ich Ihnen die Videos auf meiner Website unter <http://dataanalysistools.de/videos-to-the-book/> zur Verfügung, auf die Sie, nach Eingabe des Passworts *DaNiM* (das steht für **D**atenanalyse für **N**aturwissenschaftler, **I**ngenieure und **M**ediziner), freien Zugriff haben. Wenn Sie Fragen zu Inhalten im Buch oder in den Videos haben, sparen Sie sich die Zeit und recherchieren Sie nicht stundenlang im Internet. Gehen Sie bitte den kurzen Weg und kontaktieren Sie mich einfach per Mail (contact@dataanalysistools.de). Ich werde versuchen, Ihnen möglichst schnell zu antworten. Nutzen Sie die Kontaktmöglichkeit bitte auch für Feedback, denn so weiß ich, was ich in Zukunft besser machen könnte. Sie können mich auch gerne über meine Website <http://www.dataanalysistools.de> kontaktieren. Dort führe ich einen Blog, lade hier regelmäßig Videos zum Thema Datenanalyse hoch und stelle kleine Software-Tools kostenfrei zur Verfügung. Ich wünsche Ihnen viel Spaß mit dem Buch.

Oppenau
30. Mai 2020

Mario Schneider

Inhaltsverzeichnis

1 Streifzug durch Excel	1
1.1 Tabellenkalkulation mit Excel	1
1.2 Rechnen mit Vektoren und Matrizen	4
1.3 Der Solver – ein starkes Werkzeug	7
1.4 Numerisches Differenzieren	11
1.5 Numerische Integration	19
Literatur	28
2 Darstellung von Daten	29
2.1 Tabellen	29
2.2 Diagramme	30
2.2.1 Grafische Darstellung mathematischer Funktionen	31
2.2.2 Streudiagramme	33
2.2.3 Box-Whisker-Plot	35
2.2.4 Histogramm	36
2.2.5 Kumulative Häufigkeitsverteilung	38
2.2.6 Violin-Plot	39
2.2.7 Fehlerbalken	40
2.2.8 Darstellung dreidimensionaler Daten	44
2.2.9 Publikationsreife Abbildungen in Excel	44
Literatur	46
3 Wahrscheinlichkeiten und Wahrscheinlichkeitsverteilungen	47
3.1 Was ist Wahrscheinlichkeit	47
3.2 Wichtige Wahrscheinlichkeitsverteilungen	50
3.2.1 Binomialverteilung	52
3.2.2 Poissonverteilung	53
3.2.3 Gauß'sche Normalverteilung	54
3.2.4 Logarithmische Normalverteilung	57
3.2.5 χ^2 -Verteilung – und die Gammafunktion	58
3.2.6 Student- t -Verteilung	59

3.2.7	F-Verteilung	61
3.2.8	Maßzahlen von Wahrscheinlichkeitsverteilungen	62
3.3	Erzeugung von Zufallszahlen.	70
3.4	Hypothesentests	73
3.4.1	Parametrische Tests – ANOVA & Co.	74
3.4.2	Nichtparametrische Hypothesentests	110
3.4.3	Das Problem der multiplen Vergleiche.	119
	Literatur	120
4	Datenvorbehandlung und Signalwertanalyse.	123
4.1	Datenskalierung	123
4.1.1	Zentrierung.	123
4.1.2	Bereichsskalierung.	124
4.1.3	Standardisierung/Autoskalierung.	125
4.1.4	Vektornormierung	126
4.2	Glätten von Daten	127
4.2.1	Gleitender Mittelwert.	127
4.2.2	Medianglättung	128
4.2.3	Savitzky-Golay-Glättung	129
4.2.4	Glättung mit Gauß'schen Gewichten	131
4.3	Konvolution und Dekonvolution	133
4.4	Korrelation	136
4.5	Interpolationsmethoden	140
4.5.1	Lineare Interpolation	140
4.5.2	Kubische Spline-Interpolation	140
4.6	Fouriertransformation	144
	Literatur	154
5	Lineare und nicht-lineare Regression	155
5.1	Lineare Regression.	155
5.1.1	Lineare Regression in analytischer Form.	155
5.1.2	Lineare Regression in Matrixform.	160
5.2	Nicht-lineare Regression	161
5.2.1	Gradienten-basierte Verfahren	163
5.2.2	Simplex-Optimierung.	166
5.2.3	Genetische Algorithmen.	171
5.3	Regression mit Replikaten	174
5.4	Globale Regression mit geteilten Parametern	176
5.5	Gewichtete Regression.	178
5.6	Robuste Regression	179

5.7	Beurteilung von Regressionsergebnissen/-analysen	182
5.7.1	Gütemaße der Regression	182
5.7.2	Vergleich von Fitmodellen	195
	Literatur	199
6	Multivariate Datenanalyse	201
6.1	Multivariate lineare Regression	201
6.2	Methoden der unüberwachten Klassifikation	204
6.2.1	Hauptkomponentenanalyse	205
6.2.2	Clusteranalyse	213
6.3	Methoden der überwachten Klassifikation	223
6.3.1	k -nächste-Nachbarn-Klassifikation	224
6.3.2	Bayes'sche Klassifikation	227
6.3.3	Support Vector Machine-Klassifizierung	231
6.3.4	Klassifizierung mittels Entscheidungsbäumen	240
6.3.5	Klassifizierung mittels künstlicher neuronaler Netze	241
6.3.6	Klassifikationsfehlerrate	250
	Literatur	254
7	Statistische Versuchsplanung	255
7.1	Versuchsplan 1. Ordnung	255
7.1.1	Vollfaktorieller Versuchsplan	258
7.1.2	Teilfaktorielle Versuchspläne	266
7.2	Versuchsplan 2. Ordnung	269
7.3	Zentral zusammengesetzter Versuchsplan	273
7.4	Box-Behnken-Versuchsplan	274
7.5	Allgemeine Hinweise zu Versuchsplänen	278
	Literatur	278
8	Lösungen zu den Übungen	279
8.1	Lösungen zu Kap. 1	279
8.2	Lösungen zu Kap. 2	282
8.3	Lösungen zu Kap. 3	285
8.4	Lösungen zu Kap. 4	292
8.5	Lösungen zu Kap. 5	294
8.6	Lösungen zu Kap. 6	302
8.7	Lösungen zu Kap. 7	306
9	Anhang	309
9.1	Ein klein wenig lineare Algebra	309
9.1.1	Rechnen mit Vektoren	310
9.1.2	Rechnen mit Matrizen	312
9.1.3	Eigenwerte und Eigenvektoren von Matrizen	316

9.2	Lagrange'sches Multiplikationsverfahren – Optimierung mit Nebenbedingungen	318
9.3	Mehr zur Fehlerfortpflanzung – für mathematisch Versierte	320
9.4	Statistische Tabellen	323
9.5	Software	349
	Literatur	352
	Stichwortverzeichnis	353