
Introduction to Statistics and Data Analysis

Christian Heumann · Michael Schomaker
Shalabh

Introduction to Statistics and Data Analysis

With Exercises, Solutions
and Applications in R

Christian Heumann
Department of Statistics
Ludwig-Maximilians-Universität München
München
Germany

Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology Kanpur
Kanpur
India

Michael Schomaker
Centre for Infectious Disease Epidemiology
and Research
University of Cape Town
Cape Town
South Africa

ISBN 978-3-319-46160-1 ISBN 978-3-319-46162-5 (eBook)
DOI 10.1007/978-3-319-46162-5

Library of Congress Control Number: 2016955516

© Springer International Publishing Switzerland 2016

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The success of the open-source statistical software “*R*” has made a significant impact on the teaching and research of statistics in the last decade. Analysing data is now easier and more affordable than ever, but choosing the most appropriate statistical methods remains a challenge for many users. To understand and interpret software output, it is necessary to engage with the fundamentals of statistics.

However, many readers do not feel comfortable with complicated mathematics. In this book, we attempt to find a healthy balance between explaining statistical concepts comprehensively and showing their application and interpretation using *R*.

This book will benefit beginners and self-learners from various backgrounds as we complement each chapter with various exercises and detailed and comprehensible solutions. The results involving mathematics and rigorous proofs are separated from the main text, where possible, and are kept in an appendix for interested readers. Our textbook covers material that is generally taught in introductory-level statistics courses to students from various backgrounds, including sociology, biology, economics, psychology, medicine, and others. Most often, we introduce the statistical concepts using examples and illustrate the calculations both manually and using *R*.

However, while we provide a gentle introduction to *R* (in the appendix), this is not a software book. Our emphasis lies on explaining statistical concepts correctly and comprehensively, using exercises and software to delve deeper into the subject matter and learn about the conceptual challenges that the methods present.

This book’s homepage, <http://chris.userweb.mwn.de/book/>, contains additional material, most notably the software codes needed to answer the software exercises, and data sets. In the remainder of this book, we will use grey boxes

```
R-command ( )
```

R

to introduce the relevant *R* commands. In many cases, the code can be directly pasted into *R* to reproduce the results and graphs presented in the book; in others, the code is abbreviated to improve readability and clarity, and the detailed code can be found online.

Many years of teaching experience, from undergraduate to postgraduate level, went into this book. The authors hope that the reader will enjoy reading it and find it a useful reference for learning. We welcome critical feedback to improve future editions of this book. Comments can be sent to `christian.heumann@stat.uni-muenchen.de`, `shalab@iitk.ac.in`, and `michael.schomaker@uct.ac.za` who contributed equally to this book.

We thank Melanie Schomaker for producing some of the figures and giving graphical advice, Alice Blanck from Springer for her continuous help and support, and Lyn Imeson for her dedicated commitment which improved the earlier versions of this book. We are grateful to our families who have supported us during the preparation of this book.

München, Germany
Cape Town, South Africa
Kanpur, India
November 2016

Christian Heumann
Michael Schomaker
Shalabh

Contents

Part I Descriptive Statistics

1	Introduction and Framework	3
1.1	Population, Sample, and Observations	3
1.2	Variables	4
1.2.1	Qualitative and Quantitative Variables	5
1.2.2	Discrete and Continuous Variables	6
1.2.3	Scales	6
1.2.4	Grouped Data	7
1.3	Data Collection	8
1.4	Creating a Data Set	9
1.4.1	Statistical Software	12
1.5	Key Points and Further Issues	13
1.6	Exercises	14
2	Frequency Measures and Graphical Representation of Data	17
2.1	Absolute and Relative Frequencies	17
2.2	Empirical Cumulative Distribution Function	19
2.2.1	ECDF for Ordinal Variables	20
2.2.2	ECDF for Continuous Variables	22
2.3	Graphical Representation of a Variable	24
2.3.1	Bar Chart	24
2.3.2	Pie Chart	26
2.3.3	Histogram	27
2.4	Kernel Density Plots	29
2.5	Key Points and Further Issues	32
2.6	Exercises	32
3	Measures of Central Tendency and Dispersion	37
3.1	Measures of Central Tendency	38
3.1.1	Arithmetic Mean	38
3.1.2	Median and Quantiles	40
3.1.3	Quantile–Quantile Plots (QQ-Plots)	44
3.1.4	Mode	45

3.1.5	Geometric Mean	46
3.1.6	Harmonic Mean	48
3.2	Measures of Dispersion	48
3.2.1	Range and Interquartile Range	49
3.2.2	Absolute Deviation, Variance, and Standard Deviation	50
3.2.3	Coefficient of Variation	55
3.3	Box Plots	56
3.4	Measures of Concentration	57
3.4.1	Lorenz Curve	58
3.4.2	Gini Coefficient	60
3.5	Key Points and Further Issues	63
3.6	Exercises	63
4	Association of Two Variables	67
4.1	Summarizing the Distribution of Two Discrete Variables	68
4.1.1	Contingency Tables for Discrete Data	68
4.1.2	Joint, Marginal, and Conditional Frequency Distributions	70
4.1.3	Graphical Representation of Two Nominal or Ordinal Variables	72
4.2	Measures of Association for Two Discrete Variables	74
4.2.1	Pearson's χ^2 Statistic	76
4.2.2	Cramer's V Statistic	77
4.2.3	Contingency Coefficient C	77
4.2.4	Relative Risks and Odds Ratios	78
4.3	Association Between Ordinal and Continuous Variables	79
4.3.1	Graphical Representation of Two Continuous Variables	79
4.3.2	Correlation Coefficient	82
4.3.3	Spearman's Rank Correlation Coefficient	84
4.3.4	Measures Using Discordant and Concordant Pairs	86
4.4	Visualization of Variables from Different Scales	88
4.5	Key Points and Further Issues	89
4.6	Exercises	90
 Part II Probability Calculus		
5	Combinatorics	97
5.1	Introduction	97
5.2	Permutations	101
5.2.1	Permutations without Replacement	101
5.2.2	Permutations with Replacement	101
5.3	Combinations	102

5.3.1	Combinations without Replacement and without Consideration of the Order	102
5.3.2	Combinations without Replacement and with Consideration of the Order	103
5.3.3	Combinations with Replacement and without Consideration of the Order	103
5.3.4	Combinations with Replacement and with Consideration of the Order	104
5.4	Key Points and Further Issues	105
5.5	Exercises	105
6	Elements of Probability Theory	109
6.1	Basic Concepts and Set Theory	109
6.2	Relative Frequency and Laplace Probability	113
6.3	The Axiomatic Definition of Probability	115
6.3.1	Corollaries Following from Kolomogorov's Axioms	116
6.3.2	Calculation Rules for Probabilities	117
6.4	Conditional Probability	117
6.4.1	Bayes' Theorem	120
6.5	Independence	121
6.6	Key Points and Further Issues	123
6.7	Exercises	123
7	Random Variables	127
7.1	Random Variables	127
7.2	Cumulative Distribution Function (CDF)	129
7.2.1	CDF of Continuous Random Variables	129
7.2.2	CDF of Discrete Random Variables	131
7.3	Expectation and Variance of a Random Variable	134
7.3.1	Expectation	134
7.3.2	Variance	135
7.3.3	Quantiles of a Distribution	137
7.3.4	Standardization	138
7.4	Tschebyshev's Inequality	139
7.5	Bivariate Random Variables	140
7.6	Calculation Rules for Expectation and Variance	144
7.6.1	Expectation and Variance of the Arithmetic Mean	145
7.7	Covariance and Correlation	146
7.7.1	Covariance	147
7.7.2	Correlation Coefficient	148
7.8	Key Points and Further Issues	149
7.9	Exercises	149

8	Probability Distributions	153
8.1	Standard Discrete Distributions	154
8.1.1	Discrete Uniform Distribution	154
8.1.2	Degenerate Distribution	156
8.1.3	Bernoulli Distribution	156
8.1.4	Binomial Distribution	157
8.1.5	Poisson Distribution	160
8.1.6	Multinomial Distribution	161
8.1.7	Geometric Distribution	163
8.1.8	Hypergeometric Distribution	163
8.2	Standard Continuous Distributions	165
8.2.1	Continuous Uniform Distribution	165
8.2.2	Normal Distribution	166
8.2.3	Exponential Distribution	170
8.3	Sampling Distributions	171
8.3.1	χ^2 -Distribution	171
8.3.2	t -Distribution	172
8.3.3	F -Distribution	173
8.4	Key Points and Further Issues	174
8.5	Exercises	175

Part III Inductive Statistics

9	Inference	181
9.1	Introduction	181
9.2	Properties of Point Estimators	183
9.2.1	Unbiasedness and Efficiency	183
9.2.2	Consistency of Estimators	189
9.2.3	Sufficiency of Estimators	190
9.3	Point Estimation	192
9.3.1	Maximum Likelihood Estimation	192
9.3.2	Method of Moments	195
9.4	Interval Estimation	195
9.4.1	Introduction	195
9.4.2	Confidence Interval for the Mean of a Normal Distribution	197
9.4.3	Confidence Interval for a Binomial Probability	199
9.4.4	Confidence Interval for the Odds Ratio	201
9.5	Sample Size Determinations	203
9.6	Key Points and Further Issues	205
9.7	Exercises	205
10	Hypothesis Testing	209
10.1	Introduction	209
10.2	Basic Definitions	210

10.2.1	One- and Two-Sample Problems	210
10.2.2	Hypotheses	210
10.2.3	One- and Two-Sided Tests	211
10.2.4	Type I and Type II Error.	213
10.2.5	How to Conduct a Statistical Test	214
10.2.6	Test Decisions Using the p -Value	215
10.2.7	Test Decisions Using Confidence Intervals	216
10.3	Parametric Tests for Location Parameters	216
10.3.1	Test for the Mean When the Variance is Known (One-Sample Gauss Test)	216
10.3.2	Test for the Mean When the Variance is Unknown (One-Sample t -Test)	219
10.3.3	Comparing the Means of Two Independent Samples	221
10.3.4	Test for Comparing the Means of Two Dependent Samples (Paired t -Test)	225
10.4	Parametric Tests for Probabilities	227
10.4.1	One-Sample Binomial Test for the Probability p	227
10.4.2	Two-Sample Binomial Test	230
10.5	Tests for Scale Parameters	232
10.6	Wilcoxon–Mann–Whitney (WMW) U-Test	232
10.7	χ^2 -Goodness-of-Fit Test	235
10.8	χ^2 -Independence Test and Other χ^2 -Tests.	238
10.9	Key Points and Further Issues	242
10.10	Exercises.	242
11	Linear Regression	249
11.1	The Linear Model.	250
11.2	Method of Least Squares	252
11.2.1	Properties of the Linear Regression Line.	255
11.3	Goodness of Fit	256
11.4	Linear Regression with a Binary Covariate.	259
11.5	Linear Regression with a Transformed Covariate	261
11.6	Linear Regression with Multiple Covariates	262
11.6.1	Matrix Notation	263
11.6.2	Categorical Covariates.	265
11.6.3	Transformations.	267
11.7	The Inductive View of Linear Regression.	269
11.7.1	Properties of Least Squares and Maximum Likelihood Estimators	273
11.7.2	The ANOVA Table.	274
11.7.3	Interactions	276
11.8	Comparing Different Models.	280
11.9	Checking Model Assumptions	285

11.10	Association Versus Causation	288
11.11	Key Points and Further Issues	289
11.12	Exercises	290
Appendix A: Introduction to <i>R</i>		297
Appendix B: Solutions to Exercises		321
Appendix C: Technical Appendix		423
Appendix D: Visual Summaries		443
References		449
Index		451

About the Authors

Prof. Christian Heumann is a professor at the Ludwig-Maximilians-Universität München, Germany, where he teaches students in Bachelor and Master programs offered by the Department of Statistics, as well as undergraduate students in the Bachelor of Science programs in business administration and economics. His research interests include statistical modeling, computational statistics and all aspects of missing data.

Dr. Michael Schomaker is a Senior Researcher and Biostatistician at the Centre for Infectious Disease Epidemiology & Research (CIDER), University of Cape Town, South Africa. He received his doctoral degree from the University of Munich. He has taught undergraduate students for many years and has written contributions for various introductory textbooks. His research focuses on missing data, causal inference, model averaging and HIV/AIDS.

Prof. Shalabh is a Professor at the Indian Institute of Technology Kanpur, India. He received his Ph.D. from the University of Lucknow (India) and completed his post-doctoral work at the University of Pittsburgh (USA) and University of Munich (Germany). He has over twenty years of experience in teaching and research. His main research areas are linear models, regression analysis, econometrics, measurement error models, missing data models and sampling theory.