Christian Heumann
Michael Schomaker
Shalabh

Introduction to Statistics and Data Analysis

With Exercises, Solutions and Applications in R

Second Edition

Springer

Introduction to Statistics and Data Analysis

Christian Heumann · Michael Schomaker · Shalabh

Introduction to Statistics and Data Analysis

With Exercises, Solutions and Applications in R

Second Edition

Springer

Christian Heumann
Department of Statistics
LMU Munich
Munich, Germany

Michael Schomaker
Department of Statistics
LMU Munich
Munich, Germany

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

ISBN 978-3-031-11832-6 ISBN 978-3-031-11833-3 (eBook)
https://doi.org/10.1007/978-3-031-11833-3

Preface to the Second Edition

We are gratified by the success of the first edition of "Introduction to Statistics and Data Analysis". Our lecture materials have been in great demand, and our readers have conveyed to us numerous good suggestions and ideas.

We believe that modern approaches to statistics can be taught early in one's statistics education. The topics need to be presented clearly and they need to be well connected to both traditional methods and statistical software. Based on this belief and the proposals received from our readers, we decided to add three new chapters to the second edition. A newly added chapter on logistic regression extends our comprehensive treatment of regression analysis. We also added a chapter on simple random sampling which interconnects classical statistical results with the computing-based inference method of bootstrapping. Lastly, we developed a chapter on causal inference: we believe that an early formal treatment of the subject helps students and researchers to express their actual hypotheses accurately, analyze them appropriately and understand the requirements needed to engage with more sophisticated literature.

In Chap. 10, we now illustrate alternatives to making binary decisions with statistical tests: we show how the concepts of compatibility and functions of p- and S-values for varying hypotheses aid the interpretation of test results by focusing on gradations of evidence rather than binary decision rules.

We rely on continuous critical feedback from our readers to learn what can be improved and refined further. We welcome suggestions, which can be sent to christian.heumann@stat.uni-muenchen.de, shalab@iitk.ac.in, and michael.schomaker@stat.uni-muenchen.de.

As in the previous edition, our philosophy is to explain the most fundamental concepts comprehensively, relegate more technical material to the appendix, illustrate computations with interpretations and use exercises and software to engage in more detail with the challenges and implications of the presented methods.

The updated book homepage is now available at https://statsbook.github.io/. It contains all solutions to the *R* exercises, additional information, and a current list of errata.

We hope that our new edition proves to be valuable for teaching introductory statistics. We thank Veronika Rosteck from Springer for her tremendous support while preparing this updated manuscript and to Barnaby Sheppard for his help in editing the new content. We deeply appreciate and acknowledge the continuous caring support of our families while completing the book.

Munich, Germany — Christian Heumann
Munich, Germany — Michael Schomaker
Kanpur, India — Shalabh
April 2022

Preface to the First Edition

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. Analyzing 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.

The book homepage, https://statsbook.github.io/, contains additional material, most notably the software codes needed to answer the software exercises, and data sets. In the remainder of the book, we will use gray boxes

```
R-command()
```

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 post-graduate 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 earlier versions of this book. We are grateful to our families who have supported us during the preparation of this book.

München, Germany — Christian Heumann
Cape Town, South Africa — Michael Schomaker
Kanpur, India — Shalabh
May 2016

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
1.3.1 Survey 8
1.3.2 Experiment 9
1.3.3 Observational Data 9
1.3.4 Primary and Secondary Data 10
1.4 Creating a Data Set 10
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 18
2.1.1 Discrete Data 18
2.1.2 Grouped Metric Data 18
2.2 Empirical Cumulative Distribution Function 20
2.2.1 ECDF for Ordinal Variables 21
2.2.2 ECDF for Metric Variables 22
2.3 Graphical Representation of a Variable 25
2.3.1 Bar Chart 25
2.3.2 Pie Chart 26
2.3.3 Histogram 28
2.4 Kernel Density Plots 30
2.5 Key Points and Further Issues 32
2.6 Exercises 33

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 46
3.1.5 Geometric Mean 46
3.1.6 Harmonic Mean 48
3.2 Measures of Dispersion 49
3.2.1 Range and Interquartile Range 50
3.2.2 Absolute Deviation, Variance and Standard Deviation 51
3.2.3 Coefficient of Variation 56
3.3 Box Plots 57
3.4 Measures of Concentration 59
3.4.1 Lorenz Curve 59
3.4.2 Gini Coefficient 61
3.5 Key Points and Further Issues 64
3.6 Exercises 64

4 Association of Two Variables 69
4.1 Summarizing the Distribution of Two Discrete Variables 70
4.1.1 Contingency Tables for Discrete Data 70
4.1.2 Joint, Marginal, and Conditional Frequency Distributions 72
4.1.3 Graphical Representation of Two Nominal or Ordinal Variables 75
4.2 Measures of Association for Two Discrete Variables 77
4.2.1 Pearson's χ^2 Statistic 78
4.2.2 Cramer's V Statistic 79
4.2.3 Contingency Coefficient C 80
4.2.4 Relative Risks and Odds Ratios 81
4.3 Association Between Ordinal and Metrical Variables 82
4.3.1 Graphical Representation of Two Metrical Variables 82
4.3.2 Correlation Coefficient 85
4.3.3 Spearman's Rank Correlation Coefficient 87
4.3.4 Measures Using Discordant and Concordant Pairs 89
4.4 Visualization of Variables from Different Scales 91
4.5 Key Points and Further Issues 93
4.6 Exercises 93

Part II Probability Calculus

5 Combinatorics 101
5.1 Introduction 101
5.2 Permutations 104
5.2.1 Permutations Without Replacement 105
5.2.2 Permutations with Replacement 105
5.3 Combinations 106
5.3.1 Combinations Without Replacement and Without Consideration of the Order 106
5.3.2 Combinations Without Replacement and with Consideration of the Order 107
5.3.3 Combinations with Replacement and Without Consideration of the Order 107
5.3.4 Combinations with Replacement and with Consideration of the Order 108
5.4 Key Points and Further Issues 109
5.5 Exercises 109

6 Elements of Probability Theory 113
6.1 Basic Concepts and Set Theory 113
6.2 Relative Frequency and Laplace Probability 117
6.3 The Axiomatic Definition of Probability 119
6.3.1 Corollaries Following from Kolomogorov's Axioms 120
6.3.2 Calculation Rules for Probabilities 121
6.4 Conditional Probability 122
6.4.1 Bayes' Theorem 124
6.5 Independence 126
6.6 Key Points and Further Issues 128
6.7 Exercises 128

7 Random Variables 131
7.1 Random Variables 131
7.2 Cumulative Distribution Function (CDF) 133
7.2.1 CDF of Continuous Random Variables 133
7.2.2 CDF of Discrete Random Variables 136
7.3 Expectation and Variance of a Random Variable 138
7.3.1 Expectation 139
7.3.2 Variance 140
7.3.3 Quantiles of a Distribution 142
7.3.4 Standardization 143
7.4 Tschebyschev's Inequality 144
7.5 Bivariate Random Variables 145
7.6 Calculation Rules for Expectation and Variance 149
7.6.1 Expectation and Variance of the Arithmetic Mean 150

7.7 Covariance and Correlation 152
7.7.1 Covariance 152
7.7.2 Correlation Coefficient 153
7.8 Key Points and Further Issues 154
7.9 Exercises 155

8 Probability Distributions 159
8.1 Standard Discrete Distributions 160
8.1.1 Discrete Uniform Distribution 160
8.1.2 Degenerate Distribution 162
8.1.3 Bernoulli Distribution 162
8.1.4 Binomial Distribution 163
8.1.5 The Poisson Distribution 166
8.1.6 The Multinomial Distribution 167
8.1.7 The Geometric Distribution 169
8.1.8 Hypergeometric Distribution 170
8.2 Standard Continuous Distributions 171
8.2.1 Continuous Uniform Distribution 172
8.2.2 The Normal Distribution 173
8.2.3 The Exponential Distribution 177
8.3 Sampling Distributions 179
8.3.1 The χ^2-Distribution 179
8.3.2 The t-Distribution 180
8.3.3 The F-Distribution 181
8.4 Key Points and Further Issues 182
8.5 Exercises 183

Part III Inductive Statistics

9 Inference 189
9.1 Introduction 189
9.2 Properties of Point Estimators 191
9.2.1 Unbiasedness and Efficiency 191
9.2.2 Consistency of Estimators 198
9.2.3 Sufficiency of Estimators 199
9.3 Point Estimation 201
9.3.1 Maximum Likelihood Estimation 201
9.3.2 Method of Moments 203
9.4 Interval Estimation 204
9.4.1 Introduction 204
9.4.2 Confidence Interval for the Mean of a Normal Distribution 206
9.4.3 Confidence Interval for a Binomial Probability 208
9.4.4 Confidence Interval for the Odds Ratio 210
9.5 Sample Size Determinations 212
9.5.1 Sample Size Calculation for μ 212

9.5.2 Sample Size Calculation for p 214
9.6 Key Points and Further Issues 215
9.7 Exercises 215

10 Hypothesis Testing 219
10.1 Introduction 219
10.2 Basic Definitions 220
10.2.1 One- and Two- Sample Problems 220
10.2.2 Hypotheses 221
10.2.3 One- and Two-Sided Tests 221
10.2.4 Type I and Type II Error 223
10.2.5 How to Conduct a Statistical Test 224
10.2.6 Test Decisions Using the p-Value 225
10.2.7 Test Decisions Using Confidence Intervals 226
10.3 Parametric Tests for Location Parameters 226
10.3.1 Test for the Mean When the Variance is Known (One-Sample Gauss-Test) 226
10.3.2 Test for the Mean When the Variance is Unknown (One-Sample t-Test) 230
10.3.3 Comparing the Means of Two Independent Samples 231
10.3.4 Test for Comparing the Means of Two Dependent Samples (Paired t-Test) 236
10.4 Parametric Tests for Probabilities 238
10.4.1 One-Sample Binomial Test for the Probability p 238
10.4.2 Two-Sample Binomial Test 241
10.5 Tests for Scale Parameters 243
10.6 Wilcoxon–Mann–Whitney (WMW) U-Test 243
10.7 χ^2-Goodness of Fit Test 246
10.8 χ^2-Independence Test and Other χ^2-Tests 249
10.9 Beyond Dichotomies 252
10.9.1 Compatibility 253
10.9.2 The S-Value 254
10.9.3 Graphs of p- and S-Values 256
10.9.4 Unconditional Interpretations 258
10.10 Key Points and Further Issues 260
10.11 Exercises 261

11 Linear Regression 267
11.1 The Linear Model 268
11.2 Method of Least Squares 270
11.2.1 Properties of the Linear Regression Line 273
11.3 Goodness of Fit 275
11.4 Linear Regression with a Binary Covariate 278
11.5 Linear Regression with a Transformed Covariate 279
11.6 Linear Regression with Multiple Covariates 280

11.6.1 Matrix Notation 281
11.6.2 Categorical Covariates 284
11.6.3 Transformations 286
11.7 The Inductive View of Linear Regression 288
11.7.1 Properties of Least Squares and Maximum Likelihood Estimators 292
11.7.2 The ANOVA Table 293
11.7.3 Interactions 295
11.8 Comparing Different Models 300
11.9 Checking Model Assumptions 304
11.10 Association Versus Causation 306
11.11 Key Points and Further Issues 307
11.12 Exercises 308

12 Logistic Regression 315
12.1 Parameter Interpretation 319
12.2 Estimation of Parameters and Predictions 320
12.3 Logistic Regression in *R* 320
12.4 Model Selection and Goodness-of-Fit 322
12.5 Key Points and Further Issues 326
12.6 Exercises 327

Part IV Additional Topics

13 Simple Random Sampling and Bootstrapping 331
13.1 Introduction 331
13.2 Methodology of Simple Random Sampling 332
13.2.1 Procedure of Selection of a Random Sample 336
13.2.2 Probabilities of Selection 337
13.3 Estimation of the Population Mean and Population Variance 342
13.3.1 Estimation of the Population Total 346
13.3.2 Confidence Interval for the Population Mean 347
13.4 Sampling for Proportions 349
13.4.1 Estimation of the Total Count 353
13.4.2 Confidence Interval Estimation of *P* 353
13.5 Bootstrap Methodology 356
13.6 Nonparametric Bootstrap Methodology 357
13.6.1 The Empirical Distribution Function 357
13.6.2 The Plug-in Principle 358
13.6.3 Steps in Applying the Bootstrap 359
13.6.4 Bootstrap Estimator and Bootstrap Variance 360
13.6.5 Bootstrap Estimate of the Bias and Standard Error 360
13.6.6 Bootstrap Confidence Intervals 365

13.7 Key Points and Further Issues 370
13.8 Exercises 370

14 Causality 373
14.1 Potential Outcomes 374
14.2 Causal Questions 374
14.3 The Causal Model: Directed Acyclic Graphs 376
14.3.1 Confounders and Confounding 377
14.3.2 Colliders 378
14.3.3 Mediators 379
14.4 Identification 380
14.4.1 Randomization 385
14.5 The Statistical Model: Estimation 387
14.5.1 The *g*-formula 387
14.5.2 Regression 391
14.6 Roadmap 393
14.7 Key Points and Further Issues 394
14.8 Exercises 395

Appendix A: Introduction to *R* 399

Appendix B: Solutions to Exercises 423

Appendix C: Technical Appendix 543

Appendix D: Visual Summaries 569

References 575

Index 577