

Understanding Statistics Using R

Randall Schumacker • Sara Tomek

Understanding Statistics Using R

 Springer

Randall Schumacker
University of Alabama
Tuscaloosa, AL, USA

Sara Tomek
University of Alabama
Tuscaloosa, AL, USA

ISBN 978-1-4614-6226-2 ISBN 978-1-4614-6227-9 (eBook)
DOI 10.1007/978-1-4614-6227-9
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2012956055

© Springer Science+Business Media New York 2013

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. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

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.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

*Dedicated to our children,
Rachel and Jamie
Daphne*

Preface

This book was written as a supplemental text for use with introductory or intermediate statistics books. The content of each chapter is appropriate for any undergraduate or graduate level statistics course. The chapters are ordered along the lines of many popular statistics books so it should be easy to supplement the chapter content and exercises with your statistics book and lecture materials. The content of each chapter was written to enrich a students' understanding of statistics using R simulation programs. The chapter exercises reinforce an understanding of the statistical concepts presented in the chapters.

Computational skills are kept to a minimum in the book by including R script programs that can be run for the exercises in the chapters. Students are not required to master the writing of R script programs, but explanations of how the programs work and program output are included in each chapter. R is a statistical package with an extensive library of functions that offers flexibility in writing customized statistical routines. The R script commands are run in the R Studio software which is a graphical user interface for Windows. The R Studio software makes accessing R programs, viewing output from the exercises, and graph displays easier for the student.

Organization of the Text

The first chapter of the book covers fundamentals of R. This includes installation of R and R Studio, accessing R packages and libraries of functions. The chapter also covers how to access manuals and technical documentation, as well as, basic R commands used in the R script programs in the chapters. This chapter is important for the instructor to master so that the software can be installed and the R script programs run. The R software is free permitting students to install the software and run the R script programs for the chapter exercises.

The second chapter offers a rich insight into how probability has shaped statistics in the behavioral sciences. This chapter begins with an understanding of finite and

infinite probability. Key probability concepts related to joint, addition, multiplication, and conditional probability are covered with associated exercises. Finally, the all important combination and permutation concepts help to understand the seven fundamental rules of probability theory which impact statistics.

Chapter 3 covers statistical theory as it relates to taking random samples from a population. The R script program is run to demonstrate sampling error. Basically, sampling error is expected to be reduced as size of the random sample increases. Another important concept is the generation of random numbers. Random numbers should not repeat or be correlated when sampling without replacement.

Chapter 4 covers histograms and ogives, population distributions, and stem and leaf graphs. The frequency distribution of cumulative percents is an ogive, represented by a characteristic S-shaped curve. In contrast, a data distribution can be unimodal or bimodal, increasing or decreasing in value. A stem and leaf graph further helps to visualize the data distribution, middle value and range or spread of the data. Graphical display of data is reinforced by the chapter exercises.

Chapter 5 covers measures of central tendency and dispersion. The concept of mean and median are presented in the chapter exercises, as well as the concept of dispersion or variance. Sample size effects are then presented to better understand how small versus large samples impact central tendency and dispersion. The Tchebysheff Inequality Theorem is presented to introduce the idea of capturing scores within certain standard deviations of the frequency distribution of data, especially when it is not normally distributed. The normal distribution is presented next followed by the Central Limit Theorem, which provides an understanding that sampling distributions will be normally distributed regardless of the shape of the population from which the random sample was drawn.

Chapter 6 covers an understanding of statistical distributions. Binomial distributions formed from the probability or frequency of dichotomous data are covered. The normal distribution is discussed both as a mathematical formula and as probability under the normal distribution. The shape and properties of the chi-square distribution, t-distribution, and F-distribution are also presented. Some basic tests of variance are introduced in the chapter exercises.

Chapter 7 discusses hypothesis testing by expressing the notion that “*A statistic is to a sample as a parameter is to a population*”. The concept of a sampling distribution is explained as a function of sample size. Confidence intervals are introduced for different probability areas of the sampling distribution that capture the population parameter. The R program demonstrates the confidence interval around the sample statistic is computed by using the standard error of the statistic. The statistical hypothesis with null and alternative expressions for percents, ranks, means, and correlation are introduced. The basic idea of testing whether a sample statistic falls outside the null area of probability is demonstrated in the R program. Finally TYPE I and TYPE II error are discussed and illustrated in the chapter exercises using R programs.

Chapters 8–13 cover the statistics taught in an elementary to intermediate statistics course. The statistics covered are chi-square, z, t, F, correlation, and regression. The respective chapters discuss hypothesis testing steps using these statistics. The R

programs further calculate the statistics and related output for interpretation of results. These chapters form the core content of the book whereas the earlier chapters lay the foundation and groundwork for understanding the statistics. A real benefit of using the R programs for these statistics is that students have free access at home and school. An instructor can also use the included R functions for the statistics in class thereby greatly reducing any programming or computational time by students.

Chapter 14 is included to present the concept that research should be replicated to validate findings. In the absence of being able to replicate a research study, the idea of cross validation, jackknife, and bootstrap are commonly used methods. These methods are important to understand and use when conducting research. The R programs make these efforts easy to conduct. Students gain further insight in Chap. 15 where a synthesis of research findings help to understand overall what research results indicate on a specific topic. It further illustrates how the statistics covered in the book can be converted to a common scale so that effect size measures can be calculated, which permits the quantitative synthesis of statistics reported in research studies. The chapter concludes by pointing out that statistical significance testing, i.e., $p < 0.05$, is not necessarily sufficient evidence of the practical importance of research results. It highlights the importance of reporting the sample statistic, significance level, confidence interval, and effect size. Reporting of these values extends the students' thinking beyond significance testing.

R Programs

The chapters contain one or more R programs that produce computer output for the chapter exercises. The R script programs enhance the basic understanding and concepts in the chapters. The R programs in each chapter are labeled for easy identification. A benefit of using the R programs is that the R software is free for home or school use. After mastering the concepts in the book, the R software can be used for data analysis and graphics using pull-down menus. The use of R functions becomes a simple cut-n-paste activity, supplying the required information in the argument statements.

There are several Internet web sites that offer information, resources, and assistance with R, R programs, and examples. These can be located by entering "R software" in the search engines accessible from any Internet browser software. The main Internet URL (Uniform Resource Locator) address for R is: <http://www.r-project.org>. A second URL is: <http://lib.stat.cmu.edu/R/CRAN>. There are also many websites offering R information, statistics, and graphing, for example, Quick-R at <http://www.statmethods.net>.

Contents

1 R Fundamentals	1
Install R	1
Install R Studio	3
Getting Help	4
Load R Packages	5
Running R Programs	7
Accessing Data and R Script Programs	8
Summary	9
** WARNING **	10
R Fundamentals Exercises	10
True or False Questions	10
2 Probability	11
Finite and Infinite Probability	11
PROBABILITY R Program	12
PROBABILITY R Program Output	13
Finite and Infinite Exercises	14
Joint Probability	18
JOINT PROBABILITY Exercises	21
Addition Law of Probability	23
ADDITION Program Output	24
ADDITION Law Exercises	25
Multiplication Law of Probability	26
Multiplication Law Exercises	28
Conditional Probability	29
CONDITIONAL Probability Exercises	32
Combinations and Permutations	34
Combination and Permutation Exercises	38
True or False Questions	40
Finite and Infinite Probability	40

Joint Probability	40
Addition Law of Probability	40
Multiplication Law of Probability	41
Conditional Probability	41
Combination and Permutation	41
3 Statistical Theory	43
Sample Versus Population.....	43
STATISTICS R Program.....	44
STATISTICS Program Output	45
Statistics Exercises.....	46
Generating Random Numbers.....	48
RANDOM R Program	49
RANDOM Program Output.....	50
Random Exercises.....	51
True and False Questions.....	53
Sample versus Population.....	53
Generating Random Numbers.....	53
4 Frequency Distributions	55
Histograms and Ogives	55
FREQUENCY R Program	56
FREQUENCY Program Output.....	57
Histogram and Ogive Exercises	58
Population Distributions	62
COMBINATION Exercises	65
Stem and Leaf Graph	66
STEM-LEAF Exercises	70
True or False Questions	72
Histograms and Ogives	72
Population Distributions	73
Stem and Leaf Graphs.....	73
5 Central Tendency and Dispersion	75
Central Tendency	75
MEAN-MEDIAN R Program	76
MEAN-MEDIAN Program Output.....	76
MEAN-MEDIAN Exercises	77
Dispersion	79
DISPERSION Exercises	81
Sample Size Effects	83
SAMPLE Exercises	84
Tchebysheff Inequality Theorem.....	86
TCHEBYSHEFF Exercises	90
Normal Distribution	91

- Normal Distribution Exercises 93
 - Central Limit Theorem 95
- Central Limit Theorem Exercises 101
- True or False Questions 103
 - Central Tendency 103
 - Dispersion 104
 - Sample Size Effects 104
 - Tchebysheff Inequality Theorem 104
 - Normal Distribution 105
 - Central Limit Theorem 105
- 6 Statistical Distributions 107**
 - Binomial..... 107
 - BINOMIAL R Program 109
 - BINOMIAL Program Output..... 110
 - BINOMIAL Exercises 110
 - Normal Distribution 112
 - NORMAL R Program..... 114
 - NORMAL Program Output 114
 - NORMAL Distribution Exercises..... 115
 - Chi-Square Distribution 116
 - CHISQUARE R Program 117
 - CHISQUARE Program Output 118
 - CHISQUARE Exercises..... 119
 - t-Distribution..... 122
 - t-DISTRIBUTION R Program..... 124
 - t-DISTRIBUTION Program Output 124
 - t-DISTRIBUTION Exercises 125
 - F-Distribution..... 128
 - F-DISTRIBUTION R Programs 132
 - F-Curve Program Output 132
 - F-Ratio Program Output 133
 - F-DISTRIBUTION Exercises..... 133
 - True or False Questions 135
 - Binomial Distribution 135
 - Normal Distribution 135
 - Chi-Square Distribution 136
 - t-Distribution 136
 - F-Distribution..... 136
- 7 Hypothesis Testing 137**
 - Sampling Distribution..... 137
 - DEVIATION R Program..... 139
 - DEVIATION Program Output 140

Deviation Exercises.....	141
Confidence Intervals	142
CONFIDENCE R Program.....	144
CONFIDENCE Program Output	144
Confidence Interval Exercises.....	145
Statistical Hypothesis.....	146
HYPOTHESIS TEST R Program	150
HYPOTHESIS TEST Program Output.....	151
Hypothesis Testing Exercises.....	152
TYPE I Error.....	154
TYPE I ERROR R Program.....	157
TYPE I ERROR Program Output	158
TYPE I Error Exercises.....	158
TYPE II Error	160
TYPE II ERROR R Program	163
TYPE II ERROR Program Output.....	164
TYPE II Error Exercises	164
True or False Questions	166
Sampling Distributions	166
Confidence Interval.....	166
Statistical Hypothesis.....	167
TYPE I Error.....	167
TYPE II Error	168
8 Chi-Square Test.....	169
CROSSTAB R Program.....	172
CROSSTAB Program Output.....	173
Example 1	173
Example 2	173
Chi-Square Exercises	174
True or False Questions	175
Chi-Square	175
9 z-Test	177
Independent Samples	177
Dependent Samples.....	180
ZTEST R Programs.....	184
ZTEST-IND Program Output.....	184
ZTEST-DEP Program Output	184
z Exercises	185
True or False Questions	186
z-Test.....	186
10 t-Test.....	187
One Sample t-Test.....	187
Independent t-Test.....	189

- Dependent t-Test 190
 - STUDENT R Program 192
 - STUDENT Program Output 192
- t Exercises 193
- True or False Questions 194
 - t-Test 194
- 11 F-Test** 197
 - Analysis of Variance 197
 - One-Way Analysis of Variance 198
 - Multiple Comparison Tests 200
 - Repeated Measures Analysis of Variance 201
 - Analysis of Variance R Programs 203
 - ONEWAY Program 203
 - ONEWAY Program Output 204
 - Scheffe Program Output 205
 - REPEATED Program Output 205
 - F Exercises 206
 - True or False Questions 207
 - F Test 207
- 12 Correlation** 209
 - Pearson Correlation 209
 - Interpretation of Pearson Correlation 211
 - CORRELATION R Program 214
 - CORRELATION Program Output 214
 - Correlation Exercises 215
 - True or False Questions 218
 - Pearson Correlation 218
- 13 Linear Regression** 219
 - Regression Equation 220
 - Regression Line and Errors of Prediction 221
 - Standard Scores 224
 - REGRESSION R Program 225
 - REGRESSION Program Output 226
 - REGRESSION Exercises 227
 - True or False Questions 228
 - Linear Regression 228
- 14 Replication of Results** 229
 - Cross Validation 230
 - CROSS VALIDATION Programs 230
 - CROSS VALIDATION Program Output 231
 - Cross Validation Exercises 232

- Jackknife 234
 - JACKKNIFE R Program..... 236
 - JACKKNIFE Program Output 237
- Jackknife Exercises 237
- Bootstrap 239
 - BOOTSTRAP R Program..... 242
 - BOOTSTRAP Program Output 242
- Bootstrap Exercises..... 242
- True or False Questions 244
 - Cross Validation 244
 - Jackknife 244
 - Bootstrap..... 245
- 15 Synthesis of Findings 247**
 - Meta-Analysis 247
 - A Comparison of Fisher and Gordon Chi-Square Approaches 248
 - Converting Various Statistics to a Common Metric..... 249
 - Converting Various Statistics to Effect Size Measures 249
 - Comparison and Interpretation of Effect Size Measures 250
 - Sample Size Considerations in Meta-Analysis 252
 - META-ANALYSIS R Programs 253
 - Meta-Analysis Program Output 254
 - Effect Size Program Output..... 254
 - Meta-Analysis Exercises..... 254
 - Statistical Versus Practical Significance 256
 - PRACTICAL R Program 259
 - PRACTICAL Program Output..... 260
 - PRACTICAL Exercises 260
 - True or False Questions 261
 - Meta-Analysis 261
 - Statistical Versus Practical Significance 261
- Glossary of Terms 263**
- Appendix 271**
- Author Index..... 279**
- Subject Index..... 281**