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

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.

Tuscaloosa, AL, USA

Randall Schumacker
Sara Tomek

Contents