

A Guide to Doing Statistics in Second Language Research Using R

Jenifer Larson-Hall

Contents

Introduction

List of R Packages Used

1 Getting Started with R and R Commander

1.1 Downloading and Opening

1.1.1 Downloading and Installing R

1.1.2 Customizing R

1.1.3 Loading Packages and R Commander

1.2 Working with Data

1.2.1 Entering Your Own Data

1.2.2 Importing Files into R through R Commander

1.2.3 Viewing Entered Data

1.2.4 Saving Data and Reading It Back In

1.2.5 Saving Graphics Files

1.2.6 Closing R and R Commander

1.3 Application Activity in Practicing Entering Data into R

1.4 Introduction to R's Workspace

1.4.1 Specifying Variables within a Data Set, and Attaching and Detaching Data Sets

1.5 Missing Data

1.6 Application Activity: Practicing Saving Data, Recognizing Missing Data, and Attaching and Detaching Data Sets

1.7 Getting Help with R

1.8 Using R as a Calculator

1.9 Application Activity with Using R as a Calculator

1.10 Objects

1.11 Application Activity in Creating Objects

1.12 Types of Data in R

1.13 Application Activity with Types of Data

1.14 Functions in R

1.15 Application Activity with Functions

1.16 Manipulating Variables (Advanced Topic)

1.16.1 Combining or Recalculating Variables

1.16.2 Creating Categorical Groups

1.16.3 Deleting Parts of a Data Set

1.16.4 Getting Your Data in the Correct Form for Statistical Tests

1.17 Application Activity for Manipulating Variables

1.17.1 Combining or Recalculating Variables

1.17.2 Creating Categorical Groups

1.17.3 Deleting Parts of a Data Set

1.17.4 Getting Data in the Correct Form for Statistical Tests

1.18 Random Number Generation

THERE IS NO CHAPTER 2

3 Describing Data Numerically and Graphically

3.1 Obtaining Numerical Summaries

3.1.1 Skewness, Kurtosis, and Normality Tests with R

3.2 Application Activity with Numerical Summaries

3.3 Generating Histograms, Stem and Leaf Plots, and Q-Q Plots

- 3.3.1 *Creating Histograms with R*
- 3.3.2 *Creating Stem and Leaf Plots with R*
- 3.3.3. *Creating Q-Q Plots with R*
- 3.4 Application Activity for Exploring Assumptions
- 3.5 Imputing Missing Data
- 3.6 Transformations
- 3.7 Application Activity for Transformations

THERE IS NO CHAPTER 4 or CHAPTER 5

6 Correlation

- 6.1 Creating Scatterplots
 - 6.1.1 *Modifying a Scatterplot in R Console*
 - 6.1.2 *Viewing Simple Scatterplot Data by Categories*
 - 6.1.3 *Multiple Scatterplots*
- 6.2 Application Activity with Creating Scatterplots
- 6.3 Calculating Correlation Coefficients
 - 6.3.1 *Robust Correlation*
- 6.4 Application Activity with Calculating Correlation Coefficients
- 6.5 Partial Correlation
- 6.6 Point-Biserial Correlations and Inter-rater Reliability
 - 6.6.1 *Point-Biserial Correlations and Test Analysis*
 - 6.6.2 *Inter-rater Reliability*

7 Multiple Regression

- 7.1 Graphs for Understanding Complex Relationships
 - 7.1.1 *Coplots*
 - 7.1.2 *3D Graphs*
 - 7.1.3 *Tree Models*
- 7.2 Application Activity with Graphs for Understanding Complex Relationships
- 7.3 Doing the Same Type of Regression as SPSS
 - 7.3.1 *Reporting the Results of a Regression Analysis*
 - 7.3.2 *Reporting the Results of a Standard Regression*
 - 7.3.3 *Reporting the Results of a Sequential Regression*
- 7.4 Application Activity with Multiple Regression
- 7.5 Finding the Best Fit
 - 7.5.1 *First Steps to Finding the Minimal Adequate Model in R*
 - 7.5.2 *Reporting the Results of Regression in R*
- 7.6 Further Steps to Finding the Best Fit: Overparameterization and Polynomial Regression
- 7.7 Examining Regression Assumptions
- 7.8 Application Activity for Finding the Best (Minimally Adequate) Fit
- 7.9 Robust Regression
 - 7.9.1 *Visualizing Robust Regression*
 - 7.9.2 *Robust Regression Methods*
- 7.10 Application Activity with Robust Regression

8 Chi-Square

- 8.1 Summarizing and Visualizing Data
 - 8.1.1 *Summary Tables for Goodness-of-Fit Data*
 - 8.1.2 *Summaries of Group Comparison Data (Crosstabs)*
 - 8.1.3 *Visualizing Categorical Data*
 - 8.1.4 *Barplots in R*

8.1.5 *New Techniques for Visualizing Categorical Data*

8.1.6 *Association Plots*

8.1.7 *Mosaic Plots*

8.1.8 *Doubledecker Plots*

8.2 Application Activity for Summarizing and Visualizing Data

8.3 One-Way Goodness-of-Fit Test

8.4 Two-Way Group-Independence Test

8.5 Application Activity for Calculating One-Way Goodness-of-Fit and Two-Way Group-Independence Tests

9 T-Tests

9.1 Creating Boxplots

9.1.1 *Boxplots for One Dependent Variable Separated by Groups (an Independent-Samples T-Test)*

9.1.2 *Boxplots for a Series of Dependent Variables (Paired-Samples T-Test)*

9.1.3 *Boxplots of a Series of Dependent Variables Split into Groups*

9.2 Application Activity with Creating Boxplots

9.3 Performing an Independent-Samples T-Test

9.4 Performing a Robust Independent-Samples T-Test

9.5 Application Activity for the Independent-Samples T-Test

9.6 Performing a Paired-Samples T-Test

9.7 Performing a Robust Paired-Samples T-Test

9.8 Application Activity for the Paired-Samples T-Test

9.9 Performing a One-Sample T-Test

9.10 Performing a Robust One-Sample T-Test

9.11 Application Activity for the One-Sample T-Test

10 One-Way ANOVA

10.1 Numerical and Visual Inspection of the Data, Including Boxplots Overlaid with Dotcharts

10.1.1 *Boxplots with Overlaid Dotcharts*

10.2 Application Activity for Boxplots with Overlaid Dotcharts

10.3 One-Way ANOVA Test

10.3.1 *Conducting a One-Way ANOVA Using Planned Comparisons*

10.4 Performing a Robust One-Way ANOVA Test

10.5 Application Activity for One-Way ANOVAs

11 Factorial ANOVA

11.1 Numerical and Visual Summary of the Data, Including Means Plots

11.1.1 *Means Plots*

11.2 Putting Data in the Correct Format for a Factorial ANOVA

11.3 Performing a Factorial ANOVA

11.3.1 *Performing a Factorial ANOVA Using R Commander*

11.3.2 *Performing a Factorial ANOVA Using R*

11.4 Performing Comparisons in a Factorial ANOVA

11.5 Application Activity with Factorial ANOVA

11.6 Performing a Robust ANOVA

12 Repeated-Measures ANOVA

12.1 Visualizing Data with Interaction (Means) Plots and Parallel Coordinate Plots

12.1.1 *Creating Interaction (Means) Plots in R with More Than One Response Variable*

12.1.2 *Parallel Coordinate Plots*

- 12.2 Application Activity for Interaction (Means) Plots and Parallel Coordinate Plots
- 12.3 Putting Data in the Correct Format for RM ANOVA
- 12.4 Performing an RM ANOVA the Fixed-Effects Way
- 12.5 Performing an RM ANOVA the Mixed-Effects Way
 - 12.5.1 *Performing an RM ANOVA*
 - 12.5.2 *Fixed versus Random Effects*
 - 12.5.3 *Mixed Model Syntax*
 - 12.5.4 *Understanding the Output from a Mixed-Effects Model*
 - 12.5.5 *Searching for the Minimal Adequate Model*
 - 12.5.6 *Testing the Assumptions of the Model*
 - 12.5.7 *Reporting the Results of a Mixed-Effects Model*
- 12.6 Application Activity for Mixed-Effects Models

13 ANCOVA

- 13.1 Performing a One-Way ANCOVA with One Covariate
 - 13.1.1 *Visual Inspection of the Data*
 - 13.1.2 *Checking the Assumptions for the Lyster (2004) Data*
 - 13.1.3 *Performing a One-Way ANCOVA with One Covariate*
- 13.2 Performing a Two-Way ANCOVA with Two Covariates
 - 13.2.1 *Checking the Assumptions for the Larson-Hall (2008) Data*
 - 13.2.2 *Performing a Two-Way ANCOVA with Two Covariates*
- 13.3 Performing a Robust ANCOVA in R
- 13.4 Application Activity for ANCOVA

Appendix A: Doing Things in R

A collection of ways to do things in R gathered into one place. Some are found in various places in the text while others are not, but they are collected here. Examples are “Finding Out Names of a Data Set,” “Changing Data from One Type to Another” and “Ordering Data in a Data Frame.” Ideas for troubleshooting are also included.

Appendix B: Calculating Benjamini and Hochberg’s FDR using R

Calculate p-value cut-offs for adjusting for multiple tests (the FDR algorithm is much more powerful than conventional tests like Tukey’s HSD or Scheffe’s).

Appendix C: Using Wilcox’s R Library

How to get commands for robust tests using the Wilcox WRS library into R.

References

Introduction

These online R files are a supplement to my SPSS book *A Guide to Doing Statistics in Second Language Research Using SPSS*. They will give the reader the ability to use the free statistical program R to perform all of the functions that the book shows how to do in SPSS. However, basic statistical information is not replicated in these files, and where necessary I refer the reader to the pertinent pages in my book.

R can perform all of the basic functions of statistics as well as the more widely known and used statistical program of SPSS can. In fact, I would claim that, in most areas, R is superior to SPSS, and continues every day to be so, as new packages can easily be added to R. If you have been using SPSS, I believe that switching to R will result in the ability to use advanced, customizable graphics, better understanding of the statistics behind the functions, and the ability to incorporate robust statistics into analyses. Such robust procedures are only now beginning to be more widely known among non-statisticians, and the methods for using them are varied, but in the book I will illustrate at least some methods that are available now for doing robust statistics.

In fact, I feel so strongly about robust statistics that I will not include any information about how to use R to do non-parametric statistics. I believe that with robust statistical methods available there is no reason to use non-parametric statistics. The reason parametric and non-parametric statistics were the ones that became well known was not because they were the ones that statisticians thought would be the best tests, but because their computing requirements were small enough that people could compute them by hand. Robust statistics have become possible only since the advent of strong computing power (Larson-Hall & Herrington, 2010), and authors like Howell (2002) have been predicting that robust methods will shortly “overtake what are now the most common nonparametric tests, and may eventually overtake the traditional parametric tests” (p. 692). Well, I believe the day that non-parametric statistics are unnecessary has come, so I have thrown those out. I do not think robust methods have yet overtaken traditional parametric tests, though, so I’ve kept those in. The great thing is that, with R, readers anywhere in the world, with or without access to a university, can keep up on modern developments in applied statistics.

R is an exciting new development in this road to progress in statistics. If you are like I was several years ago, you may have heard about R and even tried downloading it and taking a look at it, but you may feel confused at how to get started using it. I think of these R excerpts found on this website fondly as “R for dummies,” because I feel that is what I am in many respects when it comes to R. Most of the early chapters and more simple types of statistics will be illustrated by using both the command-line interface of R and a more user-friendly drop-down menu interface called R Commander. Some people who have written books about using R feel that using graphical interfaces is a crutch that one should not rely on, but I suspect these people use R much more often than me. I’m happy to provide a crutch for my readers and myself! I know R much better now but still open up R Commander almost every time I use R. If you work through the commands in the book using both R Commander and R, you will gradually come to feel much more comfortable about what you can do in the command-line interface. Eventually, by the end of the book, we will have come to more complicated statistics and graphics that aren’t found in R Commander’s menus. I hope by that point you will feel comfortable using only the command-line interface of R.

One thing that I’ve done in the book to try to make R more accessible is to break down complicated R commands term by term, like this:

```
write.csv(dekeyser, file="dekeyser.csv", row.names=FALSE)
```

write.csv(x)	Writes a comma-delimited file in Excel format; by default will keep column names; it saves the data as a data frame (unless it is a matrix).
--------------	--

file="dekeyser.csv"	Names file; don't forget the quotation marks around it!
---------------------	---

row.names=FALSE	If set to FALSE, names for rows are not expected. If you do want row names, just don't include this argument.
-----------------	---

The command above the line gives the command with illustration from the data set I'm using, but then I explain what each part of the command means. Try playing around with these terms, taking them out or tweaking them to see what happens.

You'll also find that all of the commands for R, as well as the names of libraries, data sets, and variables of data sets are in the Arial font. This is a way to set these apart and help you recognize that they are actual names of things in R. Some books show R commands with a command prompt, like this:

```
>write.csv
```

I have chosen to remove the command prompt, but the commands will be set apart from the rest of the text and will be in Arial so you should be able to recognize them as such.

The SPSS book is written as a way for those totally new to statistics to understand some of the most basic statistical tests that are widely used in the field of second language acquisition and applied linguistics. As you work through the information, whether using SPSS or R, I suggest that you open the data sets that are found on the website and work along with me. You will not get very much out of these sections if you just read them! You basically have to sit down at a computer and copy what I do while reading in order to learn. And since R is free and downloadable anywhere, this is easy to do as long as you have a computer and (at least an initial) internet connection! The data used in this R online version can be found on the website for the SPSS book (*A Guide to Doing Statistics in Second Language Research Using SPSS*) under the "SPSS Data Sets" link (the website is <http://cw.routledge.com/textbooks/9780805861853/>). The application activities will then help you move on to trying the analysis by yourself, but I have included detailed answers to all activities so you can check what you are doing. My ultimate goal, of course, is that eventually you will be able to apply the techniques to your own data.

Almost all of the data sets analyzed in the book and the application activities are real data, gathered from published articles or theses. I'm deeply grateful to these authors for allowing me to use their data, and I feel strongly that those who publish work based on a statistical analysis of the data should make that data available to others. The statistical analysis one does can affect the types of hypotheses and theories that go forward in our field, but we can all recognize that most of us are not statistical experts and we may make some mistakes in our analyses. Providing the raw data will serve as a way to check that analyses are done properly, and, if provided at the submission stage, errors can be caught early. I do want to note, however, that authors who have made their data sets available for this book have done so in order for readers to follow along with the analyses in the book and do application activities. If you wanted to do any other kind of analysis on your own that would be published using this data, you should contact the authors and ask for permission to either co-author or use their data (depending on your purposes).

This book was updated with the most recent version of R available at the time of writing, which was R version 2.10.1 (released 12/2009) and R Commander version 1.5.4 (released 12/2009). There is no doubt these programs will continue to change but, at least for R, most of the changes will not affect the way commands are executed. I also tested out all of the syntax in the book on a PC. I have tried R out on a Mac and it works a little differently, so I cannot say that what's in the book will work on a Mac, unfortunately. Because you have so much freedom in R, you also have to be somewhat more of a problem-solver than you do in commercial software programs like SPSS. You'll get more errors and you'll need to try to figure out what went wrong. But probably 70% of the time my errors result simply from not typing the command correctly, so look to that first. There is a section for help and troubleshooting in Appendix A that you might also want to consult.

In many cases I have used packages that are additions to R. In a later section I explain how to download packages, and I will tell the reader which library I am using if I need a special one. However, I wanted to make a list, all in one place, of all of the R libraries which I used, so that if a person wanted to download all of them at one time they'd be able to. These are listed in alphabetic order (be careful to keep the capitalization exactly the same as well):

bootStepAIC	lme4	pwr
car	MASS	Rcmdr
coin	mice	relaimpo
dprep	multcomp	robust
epitools	mvoutlier	robustbase
fBasics	nlme	TeachingDemos
HH	norm	tree
Hmisc	nortest	vcd
lattice	psych	WRS (but not directly available through R)

Writing this book has given me a great appreciation for the vast world of statistics, which I do not pretend to be a master of. Like our own field of second language research, the field of statistics is always evolving and progressing, and there are controversies throughout it as well. If I have advocated an approach which is not embraced by the statistical community wholeheartedly, I have tried to provide reasons for my choice for readers and additional reading that can be done on the topic. Although I have tried my hardest to make my book and the R online sections accessible and helpful for readers, I would appreciate being informed of any typos or errors, or any areas where my explanations have not been clear enough.

Last of all, I would like to express my gratitude to the many individuals who have helped me in my statistical journey. First of all is Richard Herrington, who has guided me along in my statistical thinking. He provided many articles, many answers, and just fun philosophizing along the way. I also want to thank those researchers whose efforts have made R, R Commander, and all of the packages for R available. They have done an incredible amount of work, all for the benefit of the community of people who want to use R. Last I would like to thank my family for giving me the support to do this.

Jenifer Larson-Hall

List of R Packages Used

Additional R packages used in online sections:

bootStepAIC	lme4	pwr
car	MASS	Rcmdr
coin	mice	relaimpo
dprep	multcomp	robust
epitools	mvoutlier	robustbase
fBasics	nlme	TeachingDemos
HH	norm	tree
Hmisc	nortest	vcd
lattice	psych	WRS (but not directly available through R)

The data used in this R online version can be found on the website for the SPSS book (*A Guide to Doing Statistics in Second Language Research Using SPSS*) under the “SPSS Data Sets” link (the website is <http://cw.routledge.com/textbooks/9780805861853/>).