A Guide to Doing Statistics in Second Language Research Using R

Jenifer Larson-Hall

Contents

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/).