An introduction to R

William Revelle
Swift 315
email: revelle@northwestern.edu

March 24, 2009

Contents

This short course will meet in Swift Hall 107 from 5-7 on Monday, Tuesday, Wednesday (March 30-April 1) and Monday, Tuesday (April 6-7).

1 Objectives

There are many possible statistical programs that can be used in psychological research. They differ in multiple ways, at least some of which are ease of use, generality, and cost. Some of the more common programs used are SAS, SPSS, and Systat. These programs have GUIs (Graphical User Interfaces) that are relatively easy to use but that are unique to each package. These programs are also very expensive and limited in what they can do. Although convenient to use, GUI based operations are difficult to discuss in written form. When teaching statistics or communicating results, it is helpful to use examples that others may use, perhaps in other computing environments. This course describes an alternative approach that is widely used by practicing statisticians, the statistical environment R.

R is used in various courses here at NU and has been adopted as the primary stats program for teaching at the University of Virginia and the University of Colorado (among others). I use it in teaching Psych 205, 371, 405, and 454.

The objective of this short course is very simple: to have you learn enough about R to start using it to facilitate your teaching and research. By the end of the course you should be wondering why you ever used SPSS or SAS.

2 Requirements and readings

A willingness to learn and to ask questions. Bringing a personal computer to class would not be a bad idea.

Handouts of the lecture notes will be linked from this outline. Most of the handouts will be either pdfs of slides or pdfs of example code.

There are a number of tutorials on learning R, ranging from the short to the extensive. For those who are familiar with SPSS or SAS, the book, R for SAS and SPSS Users by Muenchen (2009), is a good introduction. (See his webpage at http://rforsasandspssusers.com/). For psychologists, my tutorial Using R for psychological research:A simple guide to an elegant package is not a bad beginning. See also the short and very short versions of that for undergraduates. As an example of what a bright undergraduate can do to help other undergraduates use R, see K. Funkhouser's Using R to analyze a simple data set.

There are a number of other very good tutorials on the web. An essential aid is the R reference card and the search engines R seek: a search engine for R and Jonathan Baron's search engine of the R help archives.

3 Outline

3.1 Day 1: What is R? An introduction

3.1.1 What is it?

3.1.2 How to get it: CRAN

3.1.3 Packages and Task Views

3.1.4 Help and Guidance

3.1.5 Package Vignettes

3.2 Day 2: Graphical data displays and Exploratory Data Analysis

3.3 Day 3: The general linear model and its special cases

3.3.1 Regression

3.3.2 Analysis of Variance

3.3.3 Multi-level models as an alternative to repeated measures ANOVA

3.4 Day 4: Multivariate analysis

3.4.1 Factor analysis and Principal Components Analysis

3.4.2 Cluster Analysis, Multidimensional Scaling

3.4.3 Structural Equation Modeling

3.5 Day 5: R as a programming language

3.5.1 R in the lab

3.5.2 R in the classroom

3.5.3 Using R and Latex or OpenOffice to prepare documents

4 Day 1: What is R? An introduction

4.1 What is it?

The R Development Core Team (2008) has developed an extremely powerful "language and environment for statistical computing and graphics" and a set of **packages** that operate within this programming environment (R). The R program is an open source version of the statistical program S and is very similar to the statistical program based upon S, S-PLUS (also known as S+). Although described as merely "an effective data handling and storage facility [with] a suite of operators for calculations on arrays, in particular, matrices" R is, in fact, a very useful interactive package for data analysis. When compared to most other stats packages used by psychologists, R has at least three compelling advantages: it is free, it runs on multiple platforms (e.g., Windows, Unix, Linux, and Mac OS X and Classic), and combines many of the most useful statistical programs into one quasi integrated environment. R is free[1], open source software as part of the GNU[2] Project. That is, users are free to use, modify, and distribute the program, within the limits of the GNU non-license). The program itself and detailed installation instructions for Linux, Unix, Windows, and Macs are available through CRAN (Comprehensive R Archive Network) at `http://www.r-project.org`[3] Although many run R as a language and text oriented programming environment, there are GUIs available for PCs, Linux and Macs. See for example, *R Commander* by John Fox or *R-app* for the Macintosh developed by Stefano Iacus and Simon Urbanek. Compared to the basic PC environment, the Mac GUI is to be preferred.

R is an integrated, interactive environment for data manipulation and analysis that includes functions for standard descriptive statistics (means, variances, ranges) and also includes useful graphical tools for Exploratory Data Analysis. In terms of inferential statistics R has many varieties of the General Linear Model including the conventional special cases of Analysis of Variance, MANOVA, and linear regression. Statisticians and statistically minded people around the world have contributed **packages** to the R Group and maintain a very active news group offering suggestions and help. The growing collection of **packages** and the ease with which they interact with each other and the core R is perhaps the greatest advantage of R. Advanced features include correlational **packages** for multivariate analyses including Factor and Principal Components Analysis, and cluster analysis. Advanced multivariate analyses **packages** that have been contributed to the R-project include one for Structural Equation Modeling (**sem**, Hierarchical Linear Modeling (referred to as non linear mixed effects in the **nlme4** package) and taxometric analysis. All of these are available in the (>1400) free **packages** distributed by the R group at CRAN. Many of the functions described in this book are incorporated into the **psych** package. Other **packages** useful for psychometrics are described in a task-view at CRAN. In addition to be a environment of prepackaged routines, R is a interpreted programming language that allows one to create specific functions when needed.

[1] Free as in speech rather than as in beer. See `http://www.gnu.org`

[2] GNU's Not Unix

[3] The R Development Core Team (2008) releases an updated version of R about every six months. That is, as of March, 2009, the current version of 2.8.1 will be replaced with 2.9.0 sometime in April. Bug fixes are then added with a sub version number (e.g. 2.8.1 fixed minor problems with 2.8.0). It is recommended to use the most up to date version, as it will incorporate various improvements and operating efficiencies.

R is also an amazing program for producing statistical graphics. A collection of some of the best graphics is available at the webpage with a complete gallery of thumbnail of figures.

4.2 How to get it: CRAN (Comprehensive R Archive Network)

Although it is possible that your local computer lab already has R, it is most useful to do analyses on your own machine. In this case you will need to download the R program from the R project and install it yourself. Go to the R home page at `http://www.r-project.org` and then choose the Download from CRAN (Comprehensive R Archive Network) option. This will take you to list of mirror sites around the world. You may download the Windows, Linux, or Mac versions at this site. For most users, downloading the binary image is easiest and does not require compiling the program.

4.3 Packages and Task Views

One of the advantages of R is that it can be supplemented with additional programs that are included as *packages* using the `package manager.` (e.g., *sem* does structural equation modeling) or that can be added using the `source` command. Most packages are directly available through the CRAN repository. Others are available at the BioConductor `http://www.bioconductor.org` repository. Yet others are available at "other" repositories. The *psych* package may be downloaded from CRAN or from the `http://personality-project.org/r` repository. The concept of a "task view" has made downloading relevant packages very easy. For instance, the `install.views("psychometrics")` command will download over 20 packages that do various types of psychometrics.

For any other than the default packages to work, you must activate it by either using the Package Manager or the `library` command:

- e.g., `library(`**psych**`)` or `library(`**sem**`)`
- entering ?**psych** will give a list of the functions available in the **psych** package as well as an overview of their funtionality.
- `objects(package:psych)` will list the functions available in a package (in this case, **psych**).

4.4 Help and Guidance

R is case sensitive and does not give overly useful diagnostic messages. If you get an error message, don't be flustered but rather be patient and try the command again using the correct spelling for the command.

When in doubt, use the `help(somefunction)` function. This is identical to `?` somefunction where some function is what you want to know about. e.g.,
`?read.table` #ask for help in using the read.table function – see the answer in the `help` window, or
`help(read.table)` #another way of asking for help. - see the `help` window

`RSiteSearch`("keyword") will open a browser window and return a search for "keyword" in all functions available in Rand the associated packages as well (if desired) the R-Help News groups.

4.5 Package vignettes

All packages have help pages for each function in the package. These are meant to help you use a function that you already know about, but not to introduce you to new functions. An increasing number of packages have a package "vignettes" that give more of an overview of the program than a detailed description of any one function. These vignettes are accessible from the help window and sometimes as part of the help index for the program. The two vignettes for the **psych** package are also available from the personality project web page. (An overview of the psych package and Using the psych package as a front end to the sem package).

5 Basic R commands and syntax

5.1 R is just a fancy calculator

One can think of R as a fancy graphics calculator. Enter a command and look at the output. Thus,

2 + 2
4

At the abstract level, almost all operations in R consists of executing a function on an object. The result is a new object. This very simple idea allows the output of any operation to be operated on by another function.

Command syntax tends to be of the form:
`variable = function (parameters)` or
`variable <- function (parameters)`

The = and the <- symbol imply replacement, not equality. The preferred style is to use the <- symbol to avoid confusion with the test for equality (==).

The result of an operation will not necessarily appear unless you ask for it. The command
`m <- mean(x)`
will find the mean of x but will not print anything on the console without the additional request
`m`.
however, just asking `mean(x)`
will find the mean and print it.

6 Day 2: Graphical data displays and Exploratory Data Analysis

6.1 Day 3: The general linear model and its special cases

6.2 Regression

6.3 Analysis of Variance

6.4 Multi-level models as an alternative to repeated measures ANOVA

6.5 Day 4: Multivariate analysis

6.6 Factor analysis and Principal Components Analysis

6.7 Cluster Analysis, Multidimensional Scaling

6.8 Structural Equation Modeling

6.9 Day 5: R as a programming language

6.10 R in the lab

6.11 R in the classroom

6.12 Using R and Latex or OpenOffice to prepare documents

LaTeXis a text processing and formating language that can be combined with the `Sweave` function in `R` to integrate statistics within a manuscript. This is also possible to do with OpenOffice.

7 Various web resources

`http://www.rseek.org/`R seek: a search engine for R

`http://artsweb.uwaterloo.ca/~jalockli/R_exp_psy.pdf`A psychology graduate students learns R

Draft of March 24, 2009.

References

Muenchen, R. A. (2009). *R for SAS and SPSS Users.* Springer.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.