USING R IN AN INTRODUCTORY STATISTICS COURSE

Alan T. Arnholt
Department of Mathematical Sciences
Appalachian State University
Boone, NC 28608
arnholt@math.appstate.edu

1 What is R?

R is a statistical programming language not unlike the non-GUI commands in S-PLUS. R is a derivative of the original S System, an innovative software program that helps users to manage and extract useful information from data. John Chambers, the developer of the S System, is now a core member of the R development team. In fact, in the February, 2005, issue of the *Journal of Statistical Software*, Jan de Leeuw, said "It is obvious now, and it was obvious then, that S was rapidly becoming the ***lingua franca*** of statistics."

Statistical Problems Many statistical problems can be broken down into three components:

1. Collecting data
2. Analyzing / summarizing the collected data
3. Interpreting the analyzed data

The course I am teaching assumes data is collected correctly. This article will focus on the analyzing and summarizing of collected data, for which my class uses R.

My Position Regarding Statistical Learning

- If you cannot implement a concept, you do not understand the concept. (Example: Given some data, find the percent of values that fall within plus or minus two standard deviations of the mean.)
- Computing is integrally related to statistics.
- Any programming language or software program is simply a tool to implement a concept.
- Working with large data sets by hand is not time effective.
- Simulations are even more effective when the students can code them versus when the students use the instructor's code or an applet.

2 Audience and Class

My Audience: To whom am I teaching?

- Appalachian State University is a comprehensive state university.
- Students come from a variety of majors (50% psychology, 25% biology, and 25% many disciplines).
- Incoming freshman average 1100 on the SAT.
- The majority of students are sophomores.

How does my class work?

- Students install R and Tinn-R (a text editor) on machines in class — first class period.
- Students have a pen drive on which they install R and Tinn-R — second class period.
- Students use R!
- Students use scripts that mirror R on slides — this allows students to take notes (Tinn-R) during class.
- Students are in front of computers **every class**.

3 Course Objectives

Students should be able to:

1. Organize and summarize univariate data.
2. Organize and summarize multivariate data.
3. Solve problems involving the binomial and normal distributions.

Students should understand:

1. The ideas behind a sampling distribution.
2. The logic behind the creation as well as be able to compute and interpret confidence intervals for unknown parameters.
3. The logic behind hypothesis testing and be able to implement that logic with practical scenarios.

4 R and Tinn-R

4.1 My History

How did I get to R? I Used Minitab™ from 1993-2002. Then, I used S-PLUS in 2002/2003. For the fall of 2002, students used S-PLUS by typing commands. This made many of them frustrated, so I wrote GUI front end for everything students used in course. Positively, students' frustration went down. Negatively, the students' learning went down as well. In the fall of 2004, I started using R in the statistics service course. Because R is free and so similar in functionality to S-PLUS, I could not justify the cost of renewing the S-PLUS license. Also, I wanted to create robust course materials. Since GUI materials hard to keep current, R was an excellent solution.

4.2 R Features

What is great about R? The first wonderful aspect of R is that it is free. This means that students have no excuse not to get it. Secondly, it can be installed and run from a pen drive. This makes it extremely portable so that students use it more. Also, students do not have to go to a lab if they have a home computer. Next, there exists an incredible amount of documentation for this extremely powerful and difficult to misuse program. The commands do not change as GUI menus do, so class materials do not require semester by semester revisions in quite the extent necessary for a GUI program. For teaching, simulations are easy and the program itself is easily extensible. The entire program is open source, which means that if one knows what one is doing, anything can be customized. The analyses done with R are reproducible, and have a seamless integration into reports/slides using Sweave. Finally, R has become the programming language of choice for research statisticians, so there are always up-to-date packages available for virtually every statistical technique an instructor would want to use.

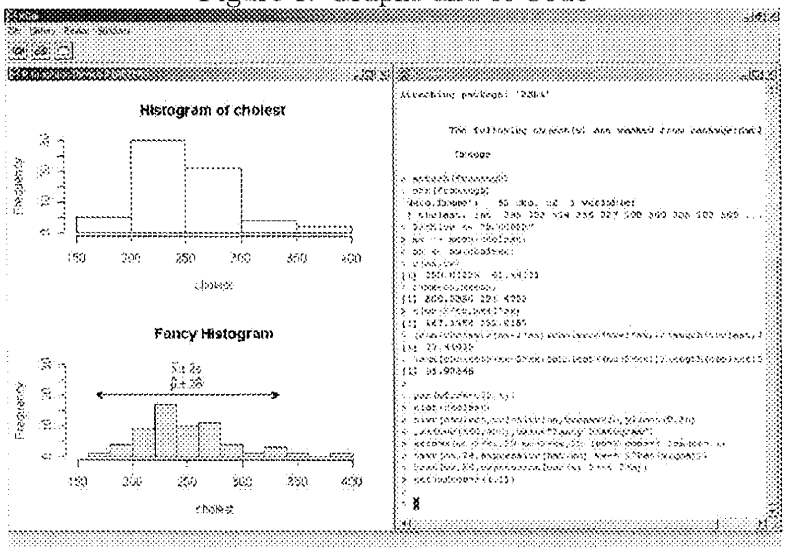
Example: The power of a programming language The cholesterol levels of 62 subjects in the Framingham Heart Study are stored in the variable `cholest` of the data frame `Framingh` found in the `BSDA` package.

- What percent of the actual data falls within one standard deviation of the mean?
- What percent of the actual data falls within two standard deviations of the mean?
- Create a histogram of the cholesterol values and depict the range of values for the mean $\pm$ two standard deviations with a double arrow.

5 Installing R and Using Packages

How can you get R working for you and your students? The directions in what follows apply to R under Windows. For other operating systems, please follow

Figure 1: Graphs and R Code



the directions provided in the FAQ - (2.5 How can R be installed?)

- Go to your nearest CRAN site to download R http://cran.r-project.org/mirrors.html.
- In the Precompiled Binary Distributions click on Windows (95 and later)
- Next click on base
- Download the current version of R by clicking on the file R-2.2.1-win32.exe. When the file download prompt appears, select `save`. Make sure you note where you save the download!

Installing R

- Navigate to the folder where the file `R-2.2.1-win32.exe` was saved.
- Double click on the file `R-2.2.1-win32.exe` and answer the Setup questions.
- Note: You may not have permission to install R on your Lab computers. However, you can always install R to a pen drive (provided your pen drive has at least 100 megs of free space - this may take 15-20 minutes) and subsequently run R from your pen drive. If you are installing R on a pen drive, make sure to specify the location where you would like R to be installed. For example, if your pen drive is in the F drive, you might specify `F:/Program Files/R/R-2.2.1` as your install folder.
- Use the default values for your installation unless you know what you are doing.

Launching R

- You should have a shortcut R icon appear on the machine where you downloaded R provided you choose the default installation values. However, if you installed

R to a pen drive on a University/Lab computer, the shortcut icon will more than likely disappear when the machine is shut down.

- To launch R, either click on the R shortcut icon on the desktop or navigate to the bin folder (`Program Files/R/R-2.2.1/bin`) and click on the `Rgui.exe` file.

Downloading Packages (`BSDA`) My class relies heavily on the `BSDA` package which needs to be both installed and loaded. To install `BSDA`,

- Click on *Packages* > *Install Package(s).*
- Select an appropriate mirror.
- Select the packages you want to install (`BSDA`).
- Click on *OK* and `BSDA` and six additional packages required by `BSDA` will be downloaded and installed.
- To load `BSDA`, click on *Packages* > *Load Package* > *BSDA*. Note: You only install a package once. However, to use the package, you must load it each time you launch R.

Using an Editor (Optional)

- Although you can type commands directly in the R console, the use of an editor is strongly recommended. There are several editors to choose from.
- `Tinn-R` is an excellent choice for students who will only use the editor to interact with R. The most recent stable version of `Tinn-R` can be found at
  http://www.sciviews.org/Tinn-R/index.html.
- If you have installed R on your pen drive, you will also want to install `Tinn-R` on your pen drive.
- To launch `Tinn-R`, click on `Tinn-R.exe` which is in the `Tinn-R/bin` folder provided the default options were selected while installing `Tinn-R`.
- To use `Tinn-R`, type your commands in the `Tinn-R` window. Select *R* > *Send to R* > *All* to send all of the typed commands to R.

6 Resources for the introductory course

- CRAN contributed Documentation — Contributed materials not all in English.
- Of particular interest to beginning students are the works "Using R for Data Analysis and Graphics — Introduction, Examples and Commentary" by John Maindonald and "Simple R" by John Verzani.
- Statistics and R — A collection of slides and R scripts I have created.