Tested Studies for Laboratory Teaching
Proceedings of the Association for Biology Laboratory Education
Vol. 35, 405-407, 2014

Biostatistics in the Classroom: Teaching Introductory Biology Students How to Use the Statistical Software 'R' Effectively

Mark A. Sarvary

Cornell University, Investigative Biology Laboratory, Comstock Hall, Ithaca NY 14853 USA
(**mas245@cornell.edu**)

Extended Abstract

Background

The Investigative Biology Course at Cornell University is designed for biology majors to provide lab experience with emphasis on processes of scientific investigation and students gain expertise in scientific methods and instrumentation. The course modules follow the "crawl, walk, run" approach to develop the capacity of students to solve increasingly challenging problems with greater independence. During the "crawl" phase students fill their scientific toolbox, learn instrumentation and skills to be able to design and conduct their own experiments and analyze and communicate their results. For multiple years students analyzed the results of their experiments by hand, or by the assistance of a Laboratory Instructor who may or may not had been proficient in using statistical software. To "even out the playing field" and to provide an additional tool to the students, we started to teach freshmen and sophomores how to use the statistical program 'R'.

'R' is a comprehensive, programming, graphical and statistical software that will provide significant help with the statistical analysis needed for the research projects conducted in biology labs. It became one of the most popular statistical software in the biological sciences in the past few years. 'R' is an implementation of the statistical programming language "S", which was developed at the Bell Laboratories. Every semester, over 400 students learn how to use 'R' in our investigative biology laboratories. After a one-semester biology lab our students do not only walk away with laboratory skills, but by using 'R', they improve their statistical, mathematical and analytical skills as well. This new addition to our labs has been very successful, and the feedback from the students, TAs and other faculty has been very positive. 'R' is a freeware, and compatible with Unix, Apple and Windows operating systems. The freeware can be downloaded from **www.r-project.org**. In addition, we also teach students how to use R-studio (**www.rstudio.com**) that organizes graphs, scripts and variables in a user friendly way. R-studio and R-project give the same results; both software use the same commands; and students can use either of them for their data analysis.

Organization

Students or groups should bring their own laptops to class, and download the software. After collecting simple data, students either enter their data in an excel file, and save it as "tab delimited text" in their working directory, or read their data directly into 'R'. It is very important that students become familiar with data entry right at the beginning. Students will explore and understand the program on their own, but struggle with data entry. Working in groups helps students solve problems and troubleshoot faster, with less help from the instructor. In the first lab students start analyzing their data by calculating mean, standard deviation, median and IQR and learn how to draw a boxplot.

Command lines can be entered and executed in the console window. Students can also type a command in the script window, highlight it, and execute it from there. The script window functions as a word processor with editable command lines, while in the console window each command needs to be retyped or recalled by using the arrows on the keyboard. During the first lab only the console window and r-project is being used. This helps students understand the program better. Students are introduced to scripts and R-studio only during their third week.

Pedagogy

'R' is a "new language" for the students, especially for those who never used programming and statistics before, and therefore it should be taught in the crawl phase, simultaneously with basic statistical methods. This helps students make the immediate connection between the statistical methods and the software. If both statistics and 'R' are taught during the first lab section, it becomes a tool that students are familiar with and will more likely use. There is a simple worksheet due at the beginning of their second lab, which ensures that they understand the first basic steps (data entry, descriptive statistics and graphing). Students may draw a boxplot or calculate means and standard deviations by hand first, and do the same immediately after that in 'R'. This exercise gives students the "aha" moment, and convinces even the most skeptical students that is it worth it to learn this new statistical programming language.

In order to have students become familiar with many aspects of the software, we have them hand type all the commands and datasets. Later, students can receive scripts from their instructors or share scripts among themselves. Instructors do not need to be statisticians or experts in 'R' in order to be able to teach the basics. This generation of students only need help with the first few steps, and after that they go and explore this software on their own.

Assessment

A worksheet or assignment is very important during the first labs; otherwise the students will not spend time with the software outside of the classroom. Simple questions on prelims and exams are also asked; however, students are not required to memorize any of the codes. The real assessment is using the software for their data analysis, providing the code in the appendix of their papers and using 'R' to create the figures for their lab reports and posters.

Examples

There are several great statistical books and online resources describing how to use 'R'. Since this is a freeware, many developers freely share their scripts online. Unfortunately, many of these books and online resources are too complex for first and second year students; therefore we write our own lab manual chapter with 'R' codes targeting the level of understanding of the students in our course.

In addition, students can use hash tags (#) to make notes for every command they type in the script window. These notes remind them what each code stands for, and therefore they create their own teaching material.

The following is a simple example we provide to our students. They can type these lines word by word (including #):

```
#Enter data manually into R. You have 5 data points for kanamycin and another 5 data points for ampicillin:
count<-c(5,7,8,19,22,11,22,33,44,55)
antibiotic<- c(rep ("kanamycin", 5), rep ("ampicillin", 5))

#Combine the two variables into one data table (data frame), and call it Example1:
Example1<-data.frame(count,antibiotic)
Example1

#Identify variables, calling the first 5 data points kan and the other 5 amp:
kan=(count[1:5])
amp=(count[6:10])

#Explore the data with descriptive statistics, first calculating the mean and standard deviation:
tapply(count,antibiotic, mean)
tapply(count,antibiotic, sd)

#Graphing. Draw two boxplots and color them orange and red:
boxplot(count~antibiotic, col=c("Orange", "Red"), ylab="number of cells", main="Serratia cell count")
```

Keywords: biostatistics, data analysis, graphing

Mission, Review Process & Disclaimer

The Association for Biology Laboratory Education (ABLE) was founded in 1979 to promote information exchange among university and college educators actively concerned with teaching biology in a laboratory setting. The focus of ABLE is to improve the undergraduate biology laboratory experience by promoting the development and dissemination of interesting, innovative, and reliable laboratory exercises. For more information about ABLE, please visit **http://www.ableweb.org/.**

Papers published in *Tested Studies for Laboratory Teaching: Peer-Reviewed Proceedings of the Conference of the Association for Biology Laboratory Education* are evaluated and selected by a committee prior to presentation at the conference, peer-reviewed by participants at the conference, and edited by members of the ABLE Editorial Board.

Citing This Article

Sarvary, M.A. 2014. Biostatistics in the Classroom: Teaching Introductory Biology Student How to Use the Statistical Software 'R' Effectively. Pages 405-407 in *Tested Studies for Laboratory Teaching,* Volume 35 (K. McMahon, Editor). Proceedings of the 35th Conference of the Association for Biology Laboratory Education (ABLE), 477 pages.
http://www.ableweb.org/volumes/vol-35/?art=42