Introducing R to Non-STEM Undergraduates in a Second Semester Statistics Course

Darlene Olsen*

Abstract

To encourage members of all disciplines to think with data, students need to analyze real data in general education statistics courses with software that is available to them after graduation. In the past, our institution has used licensed software packages that are not accessible to students once the course is over. Recently, R has been used in a second semester course since it is open-source, free software. However, coding is a daunting task for students that have no experience with programming. The purpose of this paper is to provide an example of R statistics lab on sensitivity, specificity, positive predictive value, and negative predictive value that gently introduce students to programming while exploring statistical concepts using real life data.

Key Words: Statistics education, Health Sciences, R, screening tests, sensitivity, specificity

1. Background

The statistics education community has emphasized the need to reform statistical literacy particularly at the introductory level (Hassad, 2014). The key points in the American Statistical Association (ASA) 2014 guidelines are the importance of students having extensive computing skills, the ability to work with real data, an understanding of design and limitations of the data, and the ability to communicate the results. While the intent of the ASA guidelines are intended for undergraduate statistics majors, educators emphasize the importance that students develop data-related capacities, beginning with the introductory courses (Horton et. al, 2015).

In health science, the emergence of evidence-based medicine where practitioners use research literature in making decisions about patient care has highlighted the importance of improving statistics education for these majors. Students in health related disciplines need to understand the statistical work flow used in this research which includes programming used in the analysis. Using R in introductory courses is a teaching strategy that encourages investigative inquiry and conceptual understanding that are being encouraged in the statistical education literature (Grimshaw, 2015; Horton et. al, 2014; Tishkovskaya and Lancaster, 2012).

Norwich University offers a second semester statistics course that aims to expose health science students the statistical methods often used in evidence-based practice. It is challenging to motivate students to explore and appreciate the concepts of these statistical practices (Olsen, 2016). The aims of incorporating computer labs using the statistical software R to analyze the real data was to help motivate students learn the course material, expose students the the software practioners use, and demonstrate the process of statistical analysis.

Currently, the introductory statistics courses for non-STEM majors at the university use web applications or Minitab to analyze small scale, curated data that is much easier to work with than "messy" data found in real applications. The use of technology in the classroom should give students a sense of the realistic workflow encountered in research

*Department of Mathematics, Norwich University, 158 Harmon Drive, Northfield, VT 05663

and familiarize them with the technology used by data scientists (Horton, 2015). Appropriately designing computer labs give students practice working with the type of data they may encounter in their careers. Based on an assessment performed by Grimshaw (2015), students felt empowered to ask more questions when working with real data. Infusing real data examples within a course allowed students to make connections between methods, theory, and results (Grimshaw, 2015).

Typically health science majors at our institution do not have experience with programming so the labs were crafted in a way that the students gain an understanding of R to set a foundation for further study. Each lab focused on the statistical methodology used to analyze the data but also allow students to practice data cleaning and derivation as recommended in the literature (Horton, 2015; Horton et.al, 2014). An assessment of students' perceptions of using R over Minitab was performed after implementation. A survey was developed to determine if students found the computer programming in R was useful in learning statistics.

2. Methods

2.1 Example Problem

The first topic covered in the course is the accuracy of screening tests. It is important to have genuine examples that illustrate the factors that influence the sensitivity, specificity, positive predictive value, and negative predictive value of a screening test (Olsen, 2016). Resch et al. (2016), studied a battery of tests used to screen for concussions. The premise is that using three tests is better than one. This provides a great case study for students to think about both the statistical and non-statistical issues of screening exams.

2.1.1 Simulated Data

It is recommended that instructors use realistic, cutting-edge case studies from collaborations with scientists in statistics courses (Nolan and Lang, 2015). However, this is a difficult task for many that do not have access to scientists in health related fields. As such, simulating data that matches the results of the journal article is an alternative. Using the results of Resch et al (2016), data was simulated for 100 patients. Figure 1. provides a snapshot of the data. The results of the three screening test (ImPACT, SOT, HIS-r) for each patient is given, and the true result based on a more comprehensive examination of the patient.

athlete	ImPACT	SOT	HIS-r	concussion
1	0	0	0	concussed
2	1	1	1	not_concussed
3	0	1	0	concussed
4	1	0	1	concussed
5	1	0	0	not_concussed
6	0	0	0	concussed

Figure 1. Simulated data: 0 = positive screen, 1=negative screen.

2.1.2 Lab Objectives and Code

The main focus of each lab was the statistical methodology with two or three more aspects of R programming. This particular lab focused on 1.) creating a table of results, 2.) using R packages, and 3.) creating new variables. This particular lab used the epiR package

(Stevenson et al., 2018). Figure 2. demonstrated the basic code used for this lab. Students were able to investigate the impact on sensitivity, specificity, positive predictive value, and negative predictive value for each screening test alone and for any combination of the three tests.

```
concuss <- read.table("concussion.txt", header = TRUE)

table < xtabs( ~ ImPACT+ concussion , data=concuss)

library(epiR)
epi.tests(table)

concuss$battery <- ifelse(
(concuss$ImPACT == 0) |
(concuss$SOT == 0) |
(concuss$HIS.r == 0), c("0_+"),c("1_-"))
```

Figure 2. Basic R code.

2.1.3 Assessment

The redesigned course was implemented in the fall semester of 2016. There was a total of 43 students enrolled in two sections of the course. Overall, there were eight different R labs used throughout the course.

A survey assessing students perceptions of the course, with seven questions that specifically addressed the R computer labs, was given at the end of the semester. Responses to each question were given on a 5 point Likert scale (1 – strongly disagree, 2 – disagree, 3 – neutral, 4 – agree, 5– strongly agree). The survey was approved by the Institutional Review Board. A total of 36 out of 43 completed the survey. In particular, 28 had used Minitab in their previous course.

A majority of the respondents agreed with the statement "The use of computer labs was a beneficial component of this course". Of the respondents, 71% were either neutral or disagreed with the statement "The software package R was more difficult to use than Minitab". In fact, 43% of the respondents agreed or strongly agreed with the statement "I enjoyed learning to write code using the software package R' (See Figure 3.)

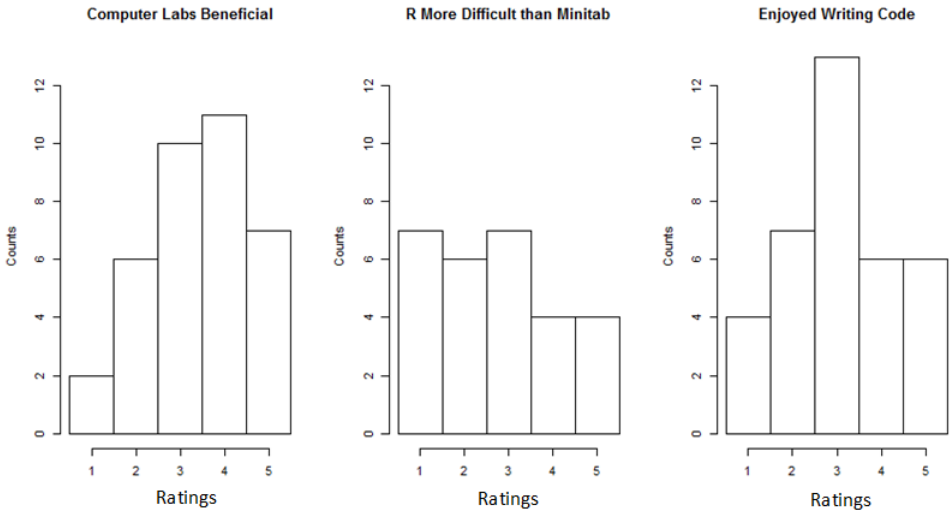

Figure 3. Results of survey.

3. Conclusions

The intent of including R labs was for the students' to realize the importance of understanding the statistical approaches used in research to better aid their critical evaluation of journal articles in evidence based practice. Using a statistical software that is used by practitioners helps students understand the statistical process. Exposing them to the nuances of data gives them a first hand look at what research may entail.

Teaching non-STEM majors programming may seem to be a daunting task. However, when approached with establishing a few small objectives for each lab, students begin to see the power of programming. Overall, teaching R in this course was successful. The survey supports that idea that R can be used in introductory statistics classes and perhaps should be considered for use by all instructors.

For code or data, please feel free to contact Darlene Olsen, dolsen1@norwich.edu.

Acknowledgements This work was supported by a Curriculum Development grant from Norwich University.

REFERENCES

American Statistical Association Undergraduate Guidelines Workgroup. 2014. 2014 curriculum guidelines for undergraduate programs in statistical science. Alexandria, VA: American Statistical Association. http://www.amstat.org/asa/files/pdfs/EDU-guidelines2014-11-15.pdf

Grimshaw, S. (2015) A Framework for Infusing Authentic Data Experiences Within Statistics Courses. *The American Statistician* 69:4, 307-314.

Hassad, R. A. (2014). The status of reform in statistics education: A focus on the introductory course. Retrieved from http://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_C196_HASSAD.pdf

Horton, J. (2015.) Challenges and Opportunities for Statistics and Statistical Education: Looking Back, Looking Forward, *The American Statistician*, 69:2, 138-145.

Horton, N., Baumer, B., and Wickham, H. (2014). Teaching precursors to data science in introductory and second courses in statistics. *arXiv preprint arXiv:1401.3269*.

Horton, N. J., Baumer, B. S., and Wickham, H. (2015). Setting the Stage for Data Science: Integration of Data Management Skills in Introductory and Second Courses in Statistics. *CHANCE*, 28, 40–50.

Nolan, D., and Temple Lang, D. (2015). Explorations in Statistics Research: An Approach to Expose Undergraduates to Authentic Data Analysis. *The American Statistician*, 69(4), 292-299.

Olsen, D. (2016). Engaging Undergraduate Health Science Students in Advanced Statistics. *Joint Statistical Meeting (JSM) Proceedings, Section on Teaching Statistics in the Health Sciences, Alexandria, VA: American Statistical Association*, 3492 – 3496.

Resch, J. E., Brown, C. N., Schmidt, J., Macciocchi, S. N., Blueitt, D., Cullum, C. M., and Ferrara, M. S. (2016). The sensitivity and specificity of clinical measures of sport concussion: three tests are better than one. *BMJ open sport and exercise medicine*, 2(1), e000012.

Stevenson, M., Stevenson, M. M., and BiasedUrn, I. (2018). Package 'epiR'.

Tishkovskaya, S., and Lancaster, G. (2012). Statistical education in the 21st century: a review of challenges, teaching innovations and strategies for reform. *Journal of Statistics Education*, 20(2), 1-55