FEATURE ARTICLE

Early Undergraduate Biostatistics & Data Science Introduction Using R, R Studio & the Tidyverse

Israel Del Toro, Kimberly Dickson, Alyssa S. Hakes, Shannon L. Newman

**Abstract**

Increasingly, students training in the biological sciences depend on a proper grounding in biological statistics, data science and experimental design. As biological datasets increase in size and complexity, transparent data management and analytical methods are essential skills for undergraduate biologists. We propose that using the software R and R Studio are effective tools to train first- and second-year undergraduate students in data visualization and foundational statistical analyses. Here, we present the redesigned laboratory curriculum for our Experimental Design and Statistics course, a required course for all first- or second-year biology majors at Lawrence University, a small liberal arts institution in northeast Wisconsin. We include an example 10-week syllabus and eight laboratory exercises (as supplementary materials) for undergraduate institutions that aim to introduce and guide students through acquiring a basic understanding of biostatistical analyses and skills using R and R Studio. We also provide a flexible framework and examples that are easily modifiable and cover the essential biostatistics and data science skills needed for biology undergraduates. Finally, we discuss the potential pitfalls and obstacles as well as the intrinsic benefits and expected outcomes of our laboratories.

**Key Words:** statistical programming; biological data analysis; data management tools; reproducible analyses; data ethics.

○ Introduction

Biological datasets are continuously growing in size and complexity (Marx, 2013); as a result, there exists an urgent need to integrate the biological sciences with data science, a field with an emphasis on using data to describe the world (DeVeaux et al., 2017; Porter & Smith, 2019). By design, data science is interdisciplinary, linking statistics, mathematics, computer science (DeVeaux et al., 2017), and in this case, the biological sciences. More and more, educators are tasked with providing training with the data management, analysis, and visualization toolkits that can increase students’ opportunities in the biological sciences and related fields, thus making this an applied interdisciplinary training opportunity. Challenges associated with creating an equitable learning environment for biology students entering with a range of interests, experiences, and preparation are often exacerbated when integrating mathematical and statistical tools (Feser et al., 2013). Students therefore require a framework of education that is flexible enough to be applied to every biological discipline ranging from biomedical to environmental sciences, but not so daunting that student interest and buy-in are discouraged.

A central theme in biological data science is that analysis must be reproducible (Fanelli, 2018) and transparent (Parker et al., 2016). The open access and freely available statistical software R (R Development Core Team, 2020) and R Studio (R Studio Team, 2015) are appropriate tools for managing, analyzing, and visualizing biological datasets in a transparent and reproducible way. Compared to the traditional use of point-and-click software, the R framework can empower students with a marketable skillset, reduce coding stress, and increase motivation when using biological data science (Guzman et al., 2019). We also suggest that introducing undergraduate students to these skills in their first or second year will afford them opportunities to further develop these abilities and computational competencies in later undergraduate years. Coding at earlier stages of learning has been shown to also enhance students’ problem-solving, critical-thinking, social, and self-management academic skills (Popat & Starkey, 2019), which are arguably useful throughout the students’ undergraduate curriculum.

As part of a grant to Lawrence University from the Howard Hughes Medical Institute through the Science Education Program, our team is in the process of redesigning the Experimental Design and Statistics course for biology and environmental studies majors. Here we present the redesigned laboratory materials that are part of this effort.

“I had never coded before and now I really love it! My confidence in designing a biological study grew significantly.” —Student
Biological Data Science Learning Objectives & Key Competencies

Two of the learning objectives established by the Biology Department at Lawrence University can be achieved with the use of R-based laboratory exercises in a statistical methods course: (1) to manage and analyze biological data using appropriate tools and statistical methods and (2) to communicate and interpret data presented in figures, tables, and text. These competencies overlap with the three key competencies suggested for undergraduate data science students, which include (A) developing students’ computational and statistical thinking, (B) developing the mathematical foundations of model building, and (C) developing a software foundation for data curation, knowledge transference, communication, and responsibility (DeVeaux et al., 2017).

We developed laboratory exercises to focus on data management and visualization in the Tidyverse (Wickham et al., 2019), a collection of R packages that are designed for solving data-science challenges by sharing a high-level design philosophy and minimizing grammar and data structures. The Tidyverse itself is a dialect of the R coding language with commonalities across multiple packages (e.g., dplyr and ggplot) in syntax, data structure, and data management philosophy. The Tidyverse was used in laboratory exercises throughout the term to encourage student practice and skill development and ultimately facilitate the learning of the R programming language. Data analyses were completed with the base R language.

Table 1: Sample Course Schedule.

| Week | Key Statistical Concept | Laboratory Theme | Data Science Competency | R Packages Required | Associated Primary Literature | Dataset Biological Classification |
|------|--------------------------|------------------|-------------------------|---------------------|-------------------------------|-----------------------------------|
| 1    | Data curation & management | Spreadsheet & Tidy data management | A, C | — | — | — |
| 2    | Reproducibility & transparency in biological data science | Introduction to the R computing environment | C | Tidyverse datasets | (Beall, 1942) | Agricultural sciences & morphometrics |
| 3    | Communicating using data, graphs & tables | Data distributions & visualization in the Tidyverse | A, C | Tidyverse datasets e1071 | (Cushny & Peebles, 1905) | Physiology |
| 4    | Test of independence | Statistical independence & power | A, C | Tidyverse pwr | (Beall et al., 2002) | Biomedical, physiology |
| 5    | Introduction to categorical data | Review & synthesis | NA | NA | NA | — |
| 6    | Two-category comparisons of center | Student t-test & related alternatives | A, B, C | Tidyverse | (Hasselquist et al., 1999; LaMunyon & Ward, 1998) | Animal behavior & physiology; sexual selection |
| 7    | Analyses of complex categorical data | ANOVA & related alternatives | A, B, C | Tidyverse | Dummy Zombie Dataset (unpublished) | Epidemiology & spatial biology |
| 8    | Ordinary least squares regression | Regression & correlation | A, B, C | Tidyverse | (Müller et al., 2011; Whitman et al., 2004) | Animal ecology, conservation biology |
| 9    | Complex multiple & nonlinear regression | Complex multiple & nonlinear regression | A, B, C | Tidyverse MASS corrgram | (Sunda & Hunteman, 1997; Broadbent et al., 2004) | Freshwater ecology; neurobiology |
| 10   | Conceptual synthesis | Review & synthesis | NA | NA | NA | — |

Table 1: Course laboratory syllabus linking specific labs with the key learning objectives of this course. Departmental learning objectives are (1) to manage and analyze biological data using appropriate tools and statistical methods, and (2) to communicate and interpret data generated in figures, tables, and text. Data science competencies are (A) developing students’ computational and statistical thinking, (B) developing the mathematical foundations of model building, and (C) developing a software foundation for data curation, knowledge transference, communication, and responsibility. Each lab is designed to be completed in one 2-hour lab section.
and required very few additional dependent packages (defined in Table 1). Furthermore, we tried to use datasets from a plethora of biological science subdisciplines, ranging from biomedical/physiological to ecological/environmental examples, with the goal of enhancing student interest and demonstrating broad applicability, both of which were intended to bolster student receptiveness and buy-in.

Laboratory Exercises and Assessment

Labs were completed in class sizes of 24 students or smaller, and each lab lasted 2 hours. In some cases, students did not complete the labs within the allotted time but were successful in finishing the assignments within the 48-hour assignment deadline. We found that students often sought help from peer teaching assistants and tutors. We begin our term with an introduction to data and metadata (a dataset that describes the main analytical data) management. Here, students learn basic “Tidy” data management and design appropriate metadata structures. By working within the Tidy data management framework, students are trained from early in their careers in a standardized way of organizing data that facilitates subsequent analysis and visualization (Wickham, 2014). Understanding proper data and metadata management is an essential skill not only for subsequent labs in this course but also for the larger biology curriculum (Qin & D’ignazio, 2010; Piorun et al., 2012).

As an example, our course is taught for three 70-minute meetings per week and one 120-minute laboratory meeting. During the first meeting of the week, we apply a traditional lecture component introducing students to the general course objectives and content. In our second weekly meeting as a class, we practice using code as a class group and address any lingering questions on the content. Part of our third weekly meeting is devoted to coding and analysis troubleshooting. In all cases we suggest that your classroom has peer teaching assistants (TAs) available. We have found that peer-to-peer education can be more effective in troubleshooting the questions and problems encountered by students in this course, this is especially true for the 120-minute laboratory section. We have also used TAs to help in after-hours group tutoring sessions, which run twice a week for one-hour intervals.

Lab 2 slowly introduces students to the R working environment using R Studio, where students will learn the basics of loading, summarizing, and graphing data in the Tidyverse. Labs 3 and 4 focus on the graphing conceptual basis of statistical distributions, tests for variable independence (chi-square), hypothesis testing, and calculating statistical power (Figure 1A). The following two labs address categorical data analyses using t-tests and ANOVA and the

Figure 1: Examples of data visualization achieved in the labs are (A) a density plot of the distribution of sleep study data (specific details can be found in Lab 3) — two treatment groups of the study and a standard normal distribution are plotted, (B) a point plot showing paired t-test visualization of the “Blackbird” testosterone manipulation dataset (Hasselquist et al., 1999) (specific details can be found in Lab 5), (C) an ANOVA visualization of means and standard error of viral outbreak rates in four demographic groups from the “Zombie” hypothetical dataset (specific details can be found in Lab 6), and (D) an ordinary least squares regression with best fit line in predicting the proportion of black coloration on lion noses based on lion age (Whitman et al., 2004) (specific details can be found in Lab 7).
Obstacles, Benefits & Outcomes

Mathematics and statistics in the sciences can be intimidating for many undergraduate students (Arnett & Van Horn, 2009). In addition to teaching new mathematical principles, our redesigned laboratory framework introduces students to an entirely new skill set, programming in the R coding language, which in itself can cause student stress and anxiety (Rogerson & Scott, 2010). Even with the existing fear of mathematics in the biological sciences, students tend to understand the value of statistical (and data science) integration within the field (Bowyer & Darlington, 2018). Therefore, we suggest that the slow and scaffolded integration of statistics and coding might be a successful way of introducing these skills so that coding can be a tool for learning statistical concepts and not a stressor or distraction. Ultimately, a steady scaffolded approach can also maximize student inclusion and level the learning curve in the classroom. Students in introductory biology courses have different levels of preparation from their high school curriculum (Thiry, 2019). Some students arrive with well-developed computational and mathematical skills, while some high schools cannot provide such training, causing a disparity in student training. To attempt to overcome this issue, we have crafted the laboratory exercises assuming that students are arriving in our classroom without any prior training in programming, data science, or statistics. We also ensured that the materials were always placed within the appropriate biological context and

Figure 2: Examples of analytical outputs achieved in the labs are (A) the t-test output of a Welch two-sample t-test of sperm size from two sexes (LaMunyon & Ward, 1998) (specific details can be found in Lab 3), (B) ANOVA table output testing differences in means of infection percentages between four demographic groups in the hypothetical Zombie dataset (see Lab 6 for specific details), and (C) ordinary least squares regression and linear model building and output considering the linear relationship between lion age and the percentage of black pigmentation on individual noses (Whitman et al., 2004) (see Lab 7 for specific details).
that proper experimental and study design was discussed such that students remained engaged and challenged by the material at multiple applied levels.

There is a need to integrate quantitative learning, including data-science skills, to the introductory biology curriculum (Speth et al., 2010) and throughout the biological disciplines (Porter & Smith, 2019). Here we have provided the laboratory materials for a course in introductory Experimental Design and Analyses aimed at biology first- and second-year undergraduate students with the main goal of fostering and developing undergraduate quantitative skill building in the biological sciences. We have placed strong emphasis on data management, visualization, and basic statistical analysis, which are the basic foundations of an undergraduate data science education (Baumer, 2015). Several cohorts of students have gone through versions of the modified computer labs, and we have found that our students are positively responding to the changes. Among our informal course surveys, students have responded with comments like:

- “I had never coded before and now I really love it! My confidence in designing a biological study grew significantly throughout the course and was immediately useful in my other biology courses. This was in large part due to the hands-on nature of the labs when it comes to finding the right analytical tool and using it.”

- “I enjoyed working with R, as it felt like a universal tool kit. I enjoyed the lab assignments, as they revealed how much I knew, or what I needed improvement upon.”

The laboratory materials that we have designed offer incoming biology majors the opportunity to learn the importance of data science, proper study design, and applied toolkits needed to succeed in subsequent upper-level biology courses. Additionally, the coding skill set developed by the students offers a unique insight into the interdisciplinary nature of the biological sciences with fields like data science. While few of these students will become active biological data scientists, general data science literacy is an urgent need (Dichev & Dicheva, 2017). With these materials, we expose our classes to the basic quantitative tools needed to succeed as a biologist in a new data-driven era.

The datasets and examples used in this lab cater to a wide array of biological science subdisciplines and are relevant for students with interests ranging from the biomedical to the environmental sciences. We have introduced our students to the R programming language, which is a marketable and applied skill set that trains students to develop independent critical thinking skills, research problems when encountering coding obstacles, think carefully about data and metadata management, and work collaboratively with peers to overcome challenges. We encourage our students to work in small teams and help each other troubleshoot whenever a coding syntax or analytical problem is encountered, much like in the professional technology and software development sector. This material engages our students’ ability to identify a problem and come up with a creative and innovative solution rather than simply memorizing and reciting foundational biological principles. Students will encounter challenges in completing these exercises, but we aim to maintain and disseminate changes to the materials as these labs continue to evolve. Ultimately, we aim to have our students leave our labs with a sense of independence and accomplishment for having learned a new foundational and applied skill set.

Acknowledgements

This manuscript was supported in part by a grant to Lawrence University from the Howard Hughes Medical Institute through the Science Education Program (GT11057). We also thank our colleagues Andrew Sage, Scott Corry, Stefan Debbert, and Rose Theisen, for their comments and reviews of earlier manuscript drafts.

References

Arnett, A. & Van Horn, D. 2009. Connecting mathematics and science: A learning community that helps math-phobic students. *Journal of College Science Teaching*, 38(6), 30–34. https://eric.ed.gov/?id=EJ848598.

Baumer, B. 2015. A data science course for undergraduates: Thinking with data. *American Statistician*, 69(4), 334–42. https://doi.org/10.1080/00031305.2015.1081105.

Baumer, B., Çetinkaya-Rundel, M., Bray, A., Loi, L. & Horton, N. 2014. R Markdown: Integrating a reproducible analysis tool. *Technology Innovations in Statistics Education*, 8(1). https://arxiv.org/abs/1402.1894.

Beall, C.M., Decker, M.J., Brittenham, G.M., Kushner, I., Gebremedhin, A. & Strohl, K.P. 2002. An Ethiopian pattern of human adaptation to high-altitude hypoxia. *Proceedings of the National Academy of Sciences of the United States of America*, 99(26), 17215–18. https://doi.org/10.1073/pnas.252649199.

Beall, G. 1942. The transformation of data from entomological field experiments. *Biometrika*, 32(3–4), 243–62. https://doi.org/10.1093/biomet/32.3-4.243.

Bowyer, J. & Darlington, E. 2018. Mathematical struggles and ensuring success: Post-compulsory mathematics as preparation for undergraduate bioscience. *Journal of Biological Education*, 52(1), 54–65. https://doi.org/10.1080/00219266.2017.1285803.

Broadbent, N.J., Squire, L.R. & Clark, R.E. 2004. Spatial memory, recognition memory, and the hippocampus. *Proceedings of the National Academy of Sciences of the United States of America*, 101(40), 14515–20. https://doi.org/10.1073/pnas.0406344101.

Cunshy, A.R. & Peebles, A.R. 1905. The action of optical isomers: II. Hyoscines. *The Journal of Physiology*, 32(5–6), 501–10. https://doi.org/10.1113/jphysiol.1905.sp001097.

De Veaux, R.D., Agrawal, M., Averett, M., Baumer, B.S., Bray, A., et al. 2017. Curriculum guidelines for undergraduate programs in data science. *Annual Review of Statistics and Its Application*, 4, 15–30. https://doi.org/10.1146/annurev-statistics-060116-053930.

Dichev, C. & Dicheva, D. 2017. Towards data science literacy. *Procedia Computer Science*, 108, 2151–60. https://doi.org/10.1016/j.procs.2017.05.240.

Fanelli, D. 2018. Is science really facing a reproducibility crisis, and do we need it to? *Proceedings of the National Academy of Sciences of the United States of America*, 115(11), 2628–31. https://doi.org/10.1073/pnas.1708272119.

Feser, J., Vasaly, H. & Herrera, J. 2013. On the edge of mathematics and biology integration: Improving quantitative skills in undergraduate biology education. *CBE Life Sciences Education*, 12(2). https://doi.org/10.1187/cbe.13-03-0057.

Guzman, L.M., Pennell, M.W., Nikelski, E. & Srivastava, D.S. 2019. Successful integration of data science in undergraduate biostatistics courses using cognitive load theory. *CBE Life Sciences Education*, 18(4). https://doi.org/10.1187/cbe.19-02-0041.

Hasselquist, D., Marsh, J.A., Sherman, P.W. & Wingfield, J.C. 1999. Is avian humoral immunocompetence suppressed by testosterone? *Behavioral Ecology and Sociobiology*, 45, 167–75. https://doi.org/10.1007/s002650050550.

LaMunyon, C.W. & Ward, S. 1998. Larger sperm outcompete smaller sperm in the nematode *Caenorhabditis elegans*. *Proceedings of the Royal
THE AMERICAN BIOLOGY TEACHER  EARLY UNDERGRADUATE BIOSTATISTICS & DATA SCIENCE  INTRODUCTION  USING R, R STUDIO  & THE TIDYVERSE

Society B: Biological Sciences, 265(1409). https://doi.org/10.1098/rspb.1998.0531.
Marx, V. 2013. The big challenges of big data. Nature, 498, 255–60. https://doi.org/10.1038/nature12415.
Müller, M.S., Porter, E.T., Grace, J.K., Awkerman, J.A., Birchler, K.T., et al. 2011. Maltreated nestlings exhibit correlated maltreatment as adults: Evidence of a “cycle of violence” in nazca boobies (Sula granti). The Auk, 128(4), 615–19. https://doi.org/10.1525/auk.2011.11008.
Parker, T.H., Nakagawa, S. & Gurevitch, J. 2016. Promoting transparency in evolutionary biology and ecology. Ecology Letters, 19(7), 726–28. https://doi.org/10.1111/ele.12610.
Piorun, M., Kafel, D., Leger-Hornby, T., Najafi, S., Martin, E., et al. 2012. Teaching research data management: an undergraduate/graduate curriculum. Journal of EScience Librarianship. https://doi.org/10.7191/jeslib.2012.1003.
Shahira, P. & Starkey, L. 2019. Learning to code or coding to learn? A systematic review. Computers and Education, 128(2019), 365–76. https://doi.org/10.1016/j.compedu.2018.10.005.
Porter, S.G. & Smith, T.M. 2019. Bioinformatics for the masses: The need for practical data science in undergraduate biology. Omics: A Journal of Integrative Biology, 23(6), 297–99. https://doi.org/10.1089/omi.2019.0080.
Qin, J. & D’ignazio, J. 2010. "The Central Role of Metadata in a Science Data Literacy Course." Journal of Library Metadata, 10(2–3), 188–204. https://doi.org/10.1080/19386389.2010.506379.
The R Foundation. 2020. The R Project for Statistical Computing. https://www.r-project.org.
Rogerson, C. & Scott, E. 2010. The fear factor: How it affects students learning to program in a tertiary environment. Journal of Information Technology Education: Research, 9, 147–71. https://doi.org/10.28945/1183.
RSstudio. 2015. Rstudio Team. https://rstudio.com.
Speth, E.B., Momsen, J.L., Moyer-Herreel, G.A., Ebert-May, D., Long, T.M., et al. 2010. 1, 2, 3, 4: Infusing quantitative literacy into introductory biology. CBE Life Sciences Education, 9(3). https://doi.org/10.1187/cbe.10-03-0033.
Sunda, W.G. & Huntaman, S.A. 1997. Interrelated influence of iron, light and cell size on marine phytoplankton growth. Nature, 90, 389–92. https://doi.org/10.1038/37093.
Thiry, H. 2019. Issues with high school preparation and transition to college BT—talking about leaving revisited: persistence, relocation, and loss in undergraduate STEM education. In Elaine Seymour & Anne-Barrie Hunter (Eds.), Talking about Leaving Revisited (pp. 137–47). Cham: Springer International. https://doi.org/10.1007/978-3-030-25304-2_5.
Whitman, K., Starfield, A.M., Quadling, H.S. & Packer, C. 2004. Sustainable trophy hunting of African lions. Nature, 428, 175–178. https://doi.org/10.1038/nature02395.
Wickham, H. 2014. Tidy data. Journal of Statistical Software, 59(10), 1–23. https://doi.org/10.18637/jss.v059.i10.
Wickham, H., Averick, M., Bryan, J., Chang, W., McGowen, L., et al. 2019. Welcome to the Tidyverse. Journal of Open Source Software, 4(43), 1686. https://doi.org/10.21105/joss.01686.

ISRAEL DEL TORO is an assistant professor of biology at Lawrence University and teaches in the biology, environmental science, and data science programs. KIMBERLY DICKSON and ALYSSA S. HAKES are associate professors of biology at Lawrence. SHANNON L. NEWMAN is a visiting professor at Lawrence, Appleton, WI, USA 54911. Email: Israel.DelToro@lawrence.edu.

SUSTAINING MEMBERS!

PROGRAMMATIC PARTNER
HHMI BioInteractive

PLATINUM LEVEL
minIPCR
Pivot Interactives

GOLD LEVEL
Bedford, Freeman & Worth High School Publishers
Bio-Rad Laboratories
BSCS Science Learning
Carolina Biological Supply Company
CGH & InnovATEBIO
Education Projects & Partnerships, LLC
Hudson Alpha Institute for Biotechnology
Lab Aids
Visible Boyy

SILVER LEVEL
ADInstruments, Inc.
Course Hero

NABT
National Association of Biology Teachers

Sustaining Members share NABT’s mission to promote biology and life science education. Learn more at www.NABT.org.