Opinion Page

Teaching research integrity – Using history and philosophy of science to introduce ideas about the ambiguity of research practice

Dhyaneswaran Palanichamy & Bruce V. Lewenstein School of Biology, Cornell University, Ithaca, NY 14853, USA dp429@cornell.edu

Statistics in Biology

Using quantitative methods in biology goes back to the experiments of Van Helmont in 1648 when he studied the effects of water and soil on the growth of willow trees. He concluded that willow trees consumed water rather than soil (Pagel 2002). Since then, biologists have had a difficult relationship with analyzing quantitative data. When Mendel used mathematical principles to analyze his data to understand heredity, no biologist at his time appreciated his work. It took decades for biologists to realize Mendel's genius (von Tschermak-Seysenegg 1951; Samuels et al. 2012). Ironically in the 21st century biologists are buried in data; knowing how to analyze them is unavoidable (Marx 2013). Despite this reality, statistics education in an introductory biology class has been mediocre at best. Here, we discuss reasons for this disparity and the advantages of using science history – specifically, the history of statistical methods – in an introductory biology course.

Statistics is a branch of mathematics that deals with the science of collecting, organizing and interpreting data. After Van Helmont in the 17th century, the next notable usage of quantitative analytical methods in biology was in the 1830s when Adolph Quetelet, a Belgian astronomer, showed that human traits such as height and chest size were distributed in a Gaussian curve (Quetelet 1835).

However, the real surge in the use of statistical methods in biological sciences took place at the beginning of the 20th century. English statisticians Francis Galton, Karl Pearson and Walter F.R. Weldon urged biologists to use statistical methods in their research. In 1901, they founded the first journal for statistical methods, named *Biometrika* (Magnello 2009). Since then statisticians have revolutionized the way biologists analyze data, resulting in scientific achievements that would have never been possible without the use of advanced statistical techniques (Keiding 2005).

For example, the yield of corn in the United States has increased by more than ten times within the last century using advanced statistical methods and breeding techniques (NASS 2016). Public health accomplishments such as successful immunizations for diseases like polio have built on statistical methods such as randomized controlled trials (Meldrum 2000). Jerome Cornfield, a statistician, designed the Framingham heart study in the 1960s that led scientists to narrow down the causes of heart diseases and strokes to dietary factors such as cholesterol, fat and salt (Truett et al. 1967). Since then heart diseases have decreased by 56% and stokes by 70% in the United States (Thom et al. 2006). The use of statistics in biological research has led to other significant achievements, including decreased infant mortality rate, increased motor vehicle safety, and better nutrition.

Despite this widespread use of statistics in biological research, learning statistical methods in an introductory biology classroom has not changed much. The pedagogy for biology and for statistics differ significantly in an introductory biology course. Biology often draws on history, referring to Darwin, Mendel, and Watson and Crick to explain key ideas. But unlike biology, teaching statistics does not usually draw on the history of a certain concept. Statistics is taught as certain set of rules that students can apply in data analysis of a biological experiment. Unfortunately, this approach means that students don't understand the context for using statistics and often fail to understand how to best use statistics in biology. Here, we discuss the rationale for changing the mode of statistics instruction, showing how science history could play an influential role in improving statistics instruction in biology courses.

Need for Better Statistics Education in Biology

The amount of data generated in biology is at an all-time high. Low cost, high throughput genome sequencing, automation, information technology and robotics in data collection have all contributed to the overabundance of data generated in the biological sciences. For example, public genome repositories such as the National Center for Biotechnology Information (NCBI) already store petabytes of data (Singer 2013). Recently a study compared the amount of data generated in four prolific domains: Astronomy, YouTube, Twitter and Genomics. By 2025, genomic data is projected by the study to need more storage requirements than all the other fields (Stephens et al. 2015). This reality has pushed biological researchers to need more data analytics skills than ever before (Feser et al. 2013). However, most biologists are not trained to have data skills – knowing how to store, integrate, move and analyze large amounts of data (National Research Council 2003).

Even though several statistical packages and software are designed for biologists and are available to perform data analysis, certain steps in biological experiments – like designing an experiment and choosing the tests based on the context of an experiment – require significant knowledge and experience in statistical analysis (Friedman 2001). Performing biological experiments using big data without the best analytical training can lead to spurious results due to biased experimental designs or incorrect interpretation of results (Mertz 2008). Since several high impact factor journals accept only studies with statistically significant results, some biologists are known to selectively publish only data that is statistically significant and this has led to a toxic habit of data falsification or p-hacking. Some suggest that this is one of the major reasons that biological researchers struggle with reproducibility of results (Head et al. 2015).

Even though students are encouraged to attend the statistics courses offered by the biometrics department of a university, many biology students do not enjoy taking these classes . This is because many of them lack context and are designed for a different audience. Teaching statistical methods as a set of rules to be applied to a certain biological problem removes context. And removing context will prevent students from gaining deep understanding. Starting with Aristotle in the 4th century B.C.E., scholars have emphasized the importance of context in learning (Weibell C. J. 2011). Students will better appreciate a statistical method if they understand why a certain statistical method is used rather than just how a method is used. This in turn, will enhance their understanding of statistical methods while simultaneously improving their confidence and creativity in data analysis.

Science History in Pedagogy

One of the first to suggest the use of history of science in general education was the renowned Harvard chemist (and later president), Dr. James B. Conant. He believed that learning science history would help students understand the strategies and "tactics" of science. Under his leadership a book titled *Harvard case histories in experimental sciences* was published (Conant 1957). The book contained eight specific cases of scientific innovation that described the process of science through case studies. Following this model, case studies such as "Davy's visit to France and the investigation of Iodine" were used in chemistry education to describe the process of science to students (Klopfer 1969).

Later, physicists came to understand the value of science history and have used it in various undergraduate physics classrooms (Gooday et al. 2008). For example, Demirci et al (2017) have implemented science history as a successful tool to facilitate deep understanding of neutrinos in a physics classroom. History of science can help especially in teaching commonly unclear concepts to students. For example Coelho (2010) uses the history of science to teach the complicated definition and understanding of "force" in physics.

However, some educators have argued that the history of certain scientific concepts might provide a poor model for responsible conduct of research and that beginners should be shielded from knowing the history of those concepts (Brush 1974). Considering that the person who coined the term "regression" in statistics is also the person who coined the term "eugenics," some argue that teaching science history might not be the best idea in statistics.

However, teaching the history of statistics is not only helpful but vital because students need to understand how usage of statistics in science has "evolved" over time. Student do need to know that Galton invented statistics for purposes that today we would call racist. That way, they can understand both the power of statistics and the care that needs to be used in applying and interpreting statistical tools. Unfortunately, almost all statistical textbooks oversimplify data analysis methods and represent an image of comprehensive certainty. This culture in the pedagogy of statistical training in biology has the potential to be disastrous, especially considering that we live in a time where data analysis is the most challenging problem of the field (Diggle 2015). With little or no information on how or why statistical methods were developed, students may become frustrated and leave the field; even more critical, not knowing the importance of context could lead to incorrect interpretation of experimental results.

Data generated in biological sciences requires unique statistical methods that can be significantly different from the methods used in other disciplines such as astronomy. In the past, when biologists were challenged with this problem, innovative individuals like Frank Wilcoxon and Charles Spearman were brave enough to invent appropriate statistical methods that worked for their field. We need the current learners of statistics to have the same mindset of innovation in analyzing data, science history can play a major role in inspiring them to do that.

Incorporating science history into undergraduate biology courses

Designing an introductory biology class is a massive challenge by itself. It is hard to incorporate the essentials of the broad field of biology in a single class. Even though a wide range of materials are covered in these courses, instructors also somehow manage to introduce basic statistical analysis. Students are exposed to these statistical methods and are asked to perform quick data analyses to get to some results and conclusions. Due to this limited exposure of statistics, most students do not consider themselves as experts in these statistical methods after an introductory biology course. Despite this, many of them start their research carriers by working in research labs where statistics appear all the time (National Research Council 2003).

Generating and evaluating scientific evidence is considered a fundamental requirement for literacy in science by the US National Academies of Science (National Research Council 2009). Despite the importance of this issue, very little research has been done on integrating statistics in undergraduate biology courses (Bialek & Botstein 2004; Colon-Berlingeri et al. 2011; Metz et al. 2008).

Before teaching students the history of statistical methods, it is important to know the student perspective on learning statistics. We conducted a survey about learning the history of statistical methods in a large introductory biology class (~400 students) at Cornell University. The course "Investigative Biology Labs" (BioG1500) introduces students to college level statistics for the first time. One of the major conclusions from the study was that the majority of the students felt learning science history would improve their understanding of statistical methods (Palanichamy et al. 2018).

Since students are exposed to only a few basic statistical methods in an introductory biology course, adding additional science history to the course material is feasible. Developing case studies related to specific statistical methods and adding them to the statistics course material could be a viable teaching strategy. This could be followed by classroom discussions, reflection papers etc.

Providing students with supplemental reading materials such as biographical articles or books related to the development of statistical methods is another strategy that one could consider (Klopper 1969). Another method is to develop short science history videos and share them in an educational online platform. For more advanced classes, one might consider providing students with original research articles or conference proceedings by scientists and discuss them in class. Science history could also be taught in a lecture format by strategically incorporating relevant history before introducing a certain statistical method. Although the educational efficacy of science history in teaching statistics needs to be experimentally shown, it is worthwhile to experiment with novel pedagogical approaches to improve statistics education in biology.

Conclusion

Developing a biology course with statistics and science history of statistical methods can be a challenging task. However, without a sound data analysis curriculum in biology we might not be training competent graduates for the current era of big data biology. Instructors are aware of this problem and are trying to improve their courses. One of the major concerns in biological research is the failure of identification of causations in big data projects. Some skeptics believe that without meaningful biological results the current trend of funding big data projects in biology could be short lived, which in turn might slow down the progress of research and development (Singer 2013). These reasons justify the urgent need for better pedagogical research on teaching statistical methods in undergraduate biology courses. Inspiring innovation in data analysis for aspiring biologists is no longer an option but a necessity. The traditional statistics instruction in biology fails in this area. Learning statistics with context using science history could be the missing link in statistics instruction of undergraduate biology courses.

References

Bialek W, Botstein D. 2004. Introductory science and mathematics education for 21st-century biologists. *Science* (80). 303:788–790.

Brush, S. G. (1974). Should the history of science be rated X? *Science*, 183(4130), 1164-1172. doi:10.1126/science.183.4130.1164.

Coelho, R. L. (2010). On the concept of force: How understanding its history can improve physics teaching. *Science & education*, 19(1), 91.

Colon-Berlingeri M, Burrowes PA. 2011. Teaching biology through statistics: application of statistical methods in genetics and zoology courses. *CBE-Life Sciences Education* 10:259–267.

Conant, J. B., & Nash, L. K. (1957). *Harvard case histories in experimental science*. Cambridge, Mass.: Harvard University Press.

Demirci N. 2016. Teaching the history of science in physics classrooms - The story of the neutrino. *Physical Education* 51:43003.

Diggle PJ. 2015. Statistics: a data science for the 21st century. *Journal of the Royal Statistical Society - Series A*. 178:793–813.

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* 12:124–128.

Friedman JH. 2001. The role of statistics in the data revolution? *International Statistical Review* 69:5–10.

Gooday G, Lynch JM, Wilson KG, Barsky CK. 2008. Does science education need the history of science? *Isis* 99:322–330.

Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD. 2015. The extent and consequences of p-hacking in science. *PLoS Biology* 13:e1002106. doi:10.1371/journal.pbio.1002106.

Keiding N. 2005. Roles of statistics in the life sciences. *International Statistical Review* 73:255–258.

Klopfer, L. E. (1969). The teaching of science and the history of science. *Journal of research in science teaching*, 6(1), 87-95.

Magnello ME. 2009. Karl Pearson and the establishment of mathematical statistics. *International Statistical Review* 77:3–29.

Marx V. 2013. Biology: The big challenges of big data. *Nature* 498:255–260.

Meldrum, M. L. (2000). A brief history of the randomized controlled trial: From oranges and lemons to the gold standard. *Hematology/oncology clinics of North America*, 14(4), 745-760.

Metz AM. 2008. Teaching statistics in biology: using inquiry-based learning to strengthen understanding of statistical analysis in biology laboratory courses. *CBE-Life Sciences Education* 7:317–326.

NASS U. 2017. Crop Production: 2016 Summary NASS U. 2017. Crop Production: 2016 Summary. *National Agricultural Statistical Service*.

National Research Council. 2003. BIO2010: Transforming undergraduate education for future research biologists. *National Academies Press*.

National Research Council. 2009. Learning science in informal environments: People, places, and pursuits. *National Academies Press*.

Palanichamy D, Sarvary MA, Williams K. 2018. Augmenting Statistics Education with Science History in Introductory Biology Courses, research paper, *Teaching as research (TAR) national conference*, Center for Teaching Innovation, Cornell University, Ithaca, NY.

Pagel, W. (2002). *Joan Baptista van Helmont: reformer of science and medicine*. Cambridge University Press.

Quetelet A. 1835. *Sur l'homme et le développement de ses facultés ou essai de physique sociale*. Bachelier.

Samuels ML, Witmer JA, Schaffner A. 2012. *Statistics for the life sciences*. Pearson education.

von Tschermak-Seysenegg E. 1951. The rediscovery of Gregor Mendel's work: An historical retrospect. *Journal of Heredity* 42:163–171.

Singer E., Wired M. 2013 Oct. Biology's Big Problem: There's Too Much Data to Handle. *Wired Magazine*.

Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, Iyer R, Schatz MC, Sinha S, Robinson GE. 2015. Big data: astronomical or genomical? *PLoS Biology* 13:e1002195.

Thom, T., Haase, N., Rosamond, W., Howard, V. J., Rumsfeld, J., ... & Kittner, S. (2006). Heart disease and stroke statistics–2006 update: a report from the American Heart Association Statistics Committee and Stroke Statistics Subcommittee. *Circulation*, 113(6), e85-e151.

Truett, J., Cornfield, J., & Kannel, W. (1967). A multivariate analysis of the risk of coronary heart disease in Framingham. *Journal of chronic diseases*, 20(7), 511-524.

von Tschermak-Seysenegg E. 1951. The rediscovery of Gregor Mendel's work: An historical retrospect. *Journal of Heredity*. 42:163–171.

Weibell CJ. 2011. *Principles of learning: 7 principles to guide personalized, student-centered learning in the technology-enhanced, blended learning environment*. Retrieved July 4:2011.