

PREPARING FUTURE LIFE SCIENCE RESEARCHERS TO ENGAGE WITH STATISTICS IN RESEARCH

Bethany White and Jastaranpreet Singh

Department of Statistical Sciences, University of Toronto, Canada

Department of Immunology, University of Toronto, Canada

bethany.white@utoronto.ca

Life science students who take undergraduate statistics courses will engage with statistics as consumers, and perhaps even producers, of quantitative research throughout their studies and professions. Several years ago, a study was conducted in an undergraduate statistics course at our institution that was designed to improve the quantitative preparation of life sciences students. Although we observed improvements in students' confidence in their abilities to choose appropriate statistical procedures and to interpret results from the beginning to the end of the course, there were still noticeable gaps in their abilities to do both at the end of the course. A follow-up survey study was recently conducted in the same course (n=164). This paper will highlight results on students' self-efficacy and abilities to recognize dependency in data and select appropriate statistical procedures. Implications for future quantitative life science course offerings will also be discussed.

INTRODUCTION

Statistical errors have been pervasive in life sciences research and are even found in journals with high-impact factors (Ocaña-Riola, 2016). Inappropriate use of statistics in research negatively impacts the quality and validity of results (Allen, 2015). This interferes with scientific progress and contributes to reproducibility concerns in research. For instance, a common error is pseudoreplication, or the use of statistical methods that assume independence on dependent measurements, which can lead to misleading findings (Lazic, 2022). As awareness of the prevalence of statistical errors in research has grown, reporting standards have been introduced by scientific journals (e.g., see Enhancing the Quality and Transparency of Health Research Network at www.equator-network.org/). Yet, statistical errors continue to be problematic in research (e.g., Weissgerber et al., 2016) and are being reinforced within life sciences disciplines by field experts, journals, and peers. This can be at least partially attributed to the insufficient statistical training and expertise of researchers and peer reviewers, and the scientific community recognizes this. A *Nature* survey conducted to explore scientific researchers' perceptions on the reproducibility crisis found that close to 90% of the 1,576 respondents agreed that “More robust experimental design,” “better statistics,” and “better mentorship” would help improve reproducibility (Baker, 2016). There have been many other calls for improvements to statistics training in life sciences (e.g., Gardenier & Resnik, 2002; Weissgerber et al., 2016). However, the required quantitative training for life sciences students is quite limited (Tong et al., 2022). In their review of the quantitative requirements in undergraduate life sciences programs (e.g. biology, biomedical sciences, physiology, pharmacology) offered by a group of 15 research-intensive universities in Canada (i.e., U15), Tong et al. (2022) found approximately *two-thirds* of the programs required at most one statistics course, while 21% required none. Since life sciences students' required statistics training is often limited, it is important for instructors of statistics courses for life scientists to carefully prioritize the statistical knowledge and skills their students learn and think carefully about how to support their students' learning. Statistics is its own scientific discipline (ASA, 2018); one that takes significant time in which to build expertise. This implies our students cannot be statistical experts after completing a course, or even a handful of courses, in statistics. Another complication is that statistical ideas are not intuitive and are challenging to learn. In fact, “inappropriate reasoning about statistical ideas is widespread and persistent, similar at all age levels (even among some experienced researchers), and quite difficult to change” (Garfield & Ben-Zvi, 2007).

These persistent issues served as the motivation for the development of a collaborative course offered by the Department of Statistical Sciences and the Human Biology Program at the University of Toronto, and for the research the authors have been conducting in this course over the last several years.

THE COURSE

The course, *STA288H1: Statistics and Scientific Inquiry in the Life Sciences*, was proposed in 2017, and co-developed and co-taught for the first time in 2018. As discussed in Tong et al. (2022), this is a non-traditional introductory statistics course that was designed to teach students about the use of statistics at all stages of scientific inquiry (Wild & Pfannkuch, 1999). STA288H1 focuses on critical thinking and decision-making, and conceptual understanding rather than calculations. Students gain experience using R and RMarkdown to analyze scientific data and create reports. There are no university mathematics prerequisites for this course. However, students must have taken a second-year university-level biology course to ensure that they are familiar with foundational biology concepts and vocabulary and have had exposure to life sciences research. STA288H1 is one of the statistics course options for various life sciences programs including Human Biology Program students, and it is the required statistics course for Pharmacology & Toxicology and Immunology undergraduate programs at the University of Toronto. Course learning outcomes were developed in consultation with faculty from these life sciences programs and all course activities and pedagogical decisions are informed by evidence-based best practices (e.g., GAISE, 2016). The multidisciplinary teaching partnership has been an essential feature of this course. The authors (i.e., a statistician and an immunologist) closely collaborate on all aspects of the course. Information on this course, as well as our rationale behind all design decisions are described in Tong et al. (2022).

To explore the impact of this course, students' perceptions about statistical practice in life sciences and their preparedness to engage with statistics in research, two survey studies have been conducted in this course to date. In the original study, although we observed that students were more confident in their abilities to choose correct statistical procedures and interpret results by the end of the course, there were still major gaps in their abilities to do both at the end of the course. In particular, students did not seem to recognize dependency in data or the inappropriateness of standard methods in this situation (White & Singh, 2021). Therefore, we conducted a follow-up study early in 2023. Several of the survey questions we posed to students in the latest study are included in the next section.

METHODS

All students enrolled in STA288H1 were invited to complete a survey via an online form that was available between March 1-17, 2023 (i.e. shortly after the midpoint in the course). Since there was pedagogical value completing the survey as a reflective exercise about statistical practice in life sciences research and on their quantitative training, survey completion was worth 1% of each student's STA288H1 course grade. Study participation, however, was completely optional.

The survey for the original study was developed in Summer 2018 to explore student perceptions about statistics and statistical practice in the context of life sciences research. That fall, the survey was closely reviewed by seven experts from different disciplines (i.e., biology, statistics, immunology, and higher education) who provided valuable feedback to improve its validity and optimize the clarity and ordering of questions. The follow-up study survey discussed in this paper was designed based on the original survey. Most of the questions were identical, but some improvements were made based on student responses in the first study, and several questions about quantitative training were added. This follow-up survey consisted of 20 questions such as demographic questions and questions to explore student self-efficacy and statistical knowledge and skills, perceptions about statistics and quantitative training, and a question asking students to indicate their consent to use their anonymized responses for research purposes. The questions discussed in this paper are presented in more detail here.

Two items from the *current statistics self-efficacy (CSSE)* instrument (Finney & Schraw, 2003) were included in the survey. Specifically, students were asked to rate their confidence in their abilities to "Select the correct statistical procedure to be used to answer a research question" and "Interpret the results of a statistical procedure in terms of the research question". Survey questions were also developed to assess students' abilities to do both in dependent data contexts to observe if students would notice the dependency and avoid pseudoreplication. The survey question in Figure 1 was meant to serve as a realistic experiment students may encounter in a lab course, and the survey question in Figure 2 was adapted from a published research article with pseudoreplication (Sato et al., 2008; Lazic, 2010).

You are in a laboratory course where you are working with a group of students to collect data on the effects of different drugs on blood vessel constriction. The experimental apparatus is set up so that an artery is mounted on a contraction measuring device in a bath. Different drug solutions can be added to the bath. The baseline contraction is measured, then the contraction is measured after adding Drug A, and again after B. After administering Drug A, the bath is emptied and washed out before Drug B is added. Each group of students uses one artery specimen. The class results are summarized in the following table:

Group #	Contraction force (g) at baseline - No drug	Contraction force (g) after Drug A	Contraction force (g) after Drug B
1	2.9	2.3	4.1
2	2.8	3.1	3.8
3	3.2	3.5	4.7
4	3.0	2.6	4.2
5	2.9	2.9	3.9
6	2.6	3.8	5.2
7	2.6	3.2	4.3
8	2.9	2.5	4.0
9	2.8	3.1	4.6
10	2.7	3.0	4.7

If you are trying to compare the efficacy (i.e. contraction-inducing capability) of these treatments, what statistical technique would be most appropriate?

- chi-squared (χ^2) test
- independent samples t-test
- one-way analysis of variance (ANOVA)
- a paired t-test
- random effects model
- repeated measures ANOVA
- simple linear regression
- Kruskal–Wallis test
- I do not know

Figure 1. Selecting statistical procedure in hypothetical lab scenario


RESULTS

164 STA288H1 students, or approximately 80% of the class, completed the survey and consented to have their anonymized responses used for research purposes. As shown in Figure 3, the vast majority of students reported being at least “slightly confident” about their abilities to interpret results and select appropriate statistical procedures. 13 students (7.9%) indicated they were “not at all confident” in their ability to select appropriate statistical techniques, while only 3 students (1.8%) rated their confidence to interpret statistical results this low. 72 of the students (44%) reported being “confident” or “very confident” in their abilities to interpret results; while only 25 students (25%) rated their confidence in their ability to select the appropriate procedure this high. Interestingly, most students (91%) tended to rate their confidence in interpreting statistical results the same or higher than their confidence in selecting appropriate procedures. This suggests that students seemed to be more comfortable interpreting statistical results than selecting appropriate statistical procedures given a research question.

In a paper published in *Nature Neuroscience* (Fig.1a), researchers classified rod terminals in the retina as either bipolar (+) or not bipolar (-). Using a total of six mice (three for each genotype, either “wild-type (+/+)” or “Pikachurin knock-out (-/-)”), they examined whether the proportions of the two rod terminals differ between wild-type (+/+) and Pikachurin (i.e., a protein involved in photoreceptor formation) knock-out (-/-) mice.

What can we conclude from their Chi-Square (χ^2) test (Fig. 1b)?

Fig. 1a



Pikachurin, a dystroglycan ligand, is essential for photoreceptor ribbon synapse formation

Shigeru Sato^{1,2}, Yoshihiro Omori¹, Kimiko Katoh¹, Mineo Kondo⁴, Motoi Kanagawa³, Kentaro Miyata⁴, Kazuo Funabiki⁵, Toshiyuki Koyasu⁶, Naoko Kajimura⁷, Tomomitsu Miyoshi⁸, Hajime Sawa¹, Kazuhiro Kobayashi¹, Aiko Tani¹, Tatsuhi Toda¹, Jiro Uekura⁹, Yasuo Tano², Takashi Fujikado^{3,7} & Takahisa Furukawa¹

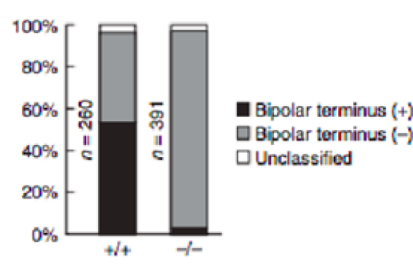


Fig 1.b: Quantitative analysis of bipolar dendrites in the wild-type(+/+) and *Pikachurin*(-/-) mouse retina. 260 and 391 measurements were taken from the 3 mice in the wild-type and knock-out groups, respectively. χ^2 Test; P-value<0.001

- Mice with the Pikachurin knock-out (-/-) tend to have a smaller proportion of bipolar terminus (+) than wild-type mice, so this proportion seems to depend on genotype.
- There is not a statistically significant difference in the proportions of bipolar terminus (+) for wild-type (+/+) and Pikachurin knock-out (-/-) mice, so the proportion does not seem to vary based on genotype.
- There is evidence against equality of the proportions of bipolar terminus (+) in wild-type (+/+) and Pikachurin knock-out (-/-) mice, suggesting this proportion differs based on genotype.
- We cannot conclude anything from this statistical test because the measurements are not independent.
- I do not know

Figure 2. Interpreting results of a statistical procedure for given research question and data

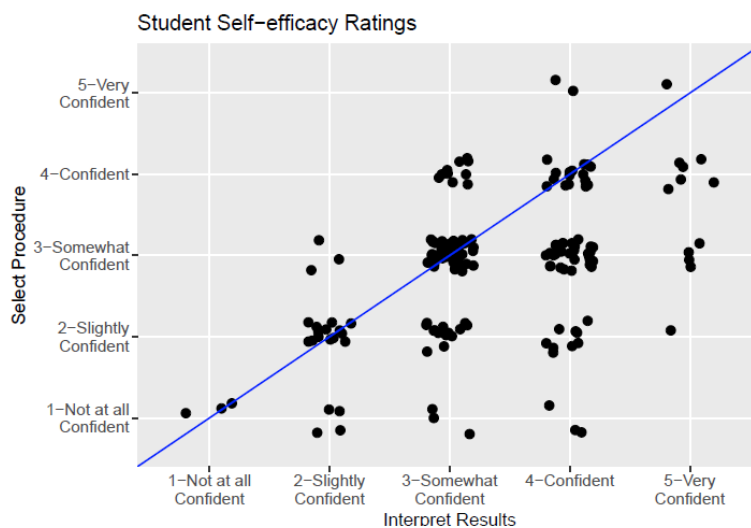


Figure 3. Reported student self-efficacy in statistics (n=164)

Next, we were interested in whether students were able to select appropriate statistical procedures and interpret results. As shown in Figure 4A, when students were asked to select the statistical procedure for the data described in Figure 1, “I do not know” (31%) and “a paired t-test” (29%) were the most popular responses. Given that students did not learn methods for dependent samples beyond the two-sample problem in the course, “I do not know” was the most appropriate

response assuming students were responding based only on the methods learned in the course. Although “A paired t-test” recognizes dependent samples, there are more than two dependent samples in the study described in Figure 1, so this is not appropriate. Some of the students who selected this option may have been thinking about conducting multiple paired t-tests and adjusting for multiple testing, but this was not presented as an option in the survey. At the end of the course in the original study, many students selected a one-way ANOVA for this question. This is an inappropriate procedure since the same artery specimens were exposed to all three treatments resulting in dependent samples. Therefore, it was encouraging that only 9% chose “one-way ANOVA” in this follow-up study.

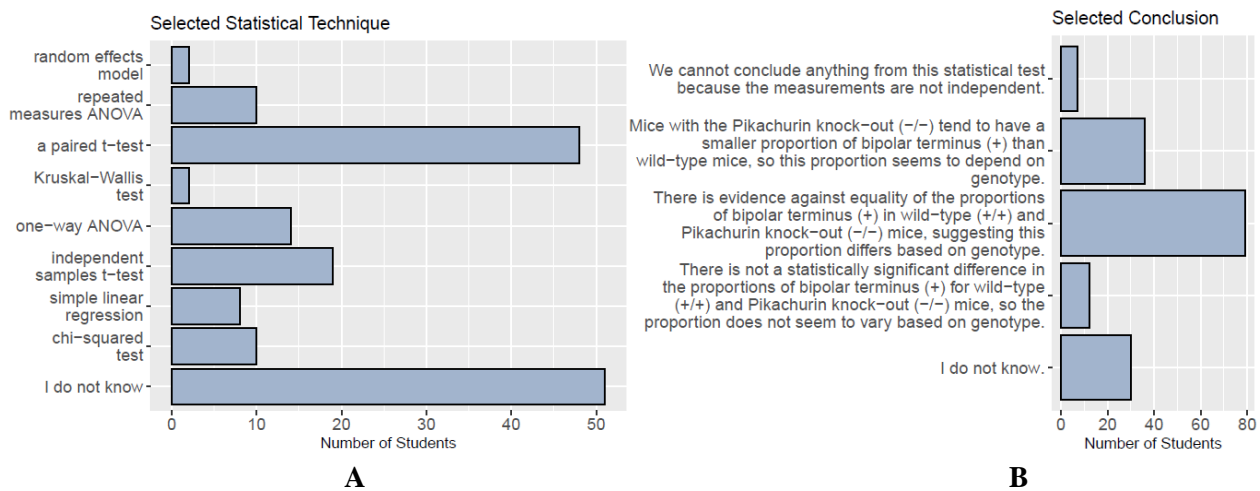


Figure 4. Students' (A) most appropriate statistical technique for data in Figure 1 and (B) conclusions based on Chi-squared test in Figure 2 (n=164)

Based on Figure 4B, students appeared to be using the reported Chi-square p-value ($p < 0.001$) to draw their conclusion to the question in Figure 2 since 48% selected “There is evidence against equality of the proportions of bipolar terminus (+) in wild-type (+/+) and Pikachurin knock-out (-/-) mice, suggesting this proportion differs based on genotype.” In fact, this was the most popular choice. Since there were only three mice per genotype studied, multiple measurements were made on each animal to give the sample sizes of 260 and 391 reported in the paper. The Chi-squared test is not appropriate for these data because it assumes there are independent measurements for each genotype. This is an example of pseudoreplication. The most appropriate response in this situation would be “We cannot conclude anything from this statistical test because the measurements are not independent.”, and only 4% of the students selected this option.

DISCUSSION

While many students reported being confident in their abilities to interpret results and select appropriate statistical procedures toward the end of STA288H1, there were still quite a few students who were unable to select techniques and correctly interpret results for the research questions described in Figures 1 and 2, both involving dependent data. Although some students selected “I do not know”, quite a few proceeded to select (32%), or interpret results from (48%), procedures assuming independent data for dependent measurements. This demonstrates how easy it was for them to miss pseudoreplication and is consistent with Lazic’s findings in his review of research studies published in a 2008 issue of *Nature Neuroscience* (Lazic, 2010). There are several limitations to this study. The survey was run shortly after the midpoint of the course, so students had not successfully completed the course when they responded. Response bias may have also been an issue as students may have misinterpreted the questions in a systematic way (e.g., in Figure 4B, students may have arrived at the most popular answer based on the data rather than the p-value because the answer did not mention statistical significance). Also, although there was an incentive for survey completion, students did not need to answer questions correctly to earn credit for survey completion.

These survey results, along with the fact that this is the only statistics course many of these students will need to take, suggests that the most important course learning outcome is to “Recognize when standard statistical procedures are not appropriate and know to seek statistical expertise early in the research process,” and reinforces the need to improve statistics training in life sciences (e.g., Tong, 2022; Gardenier & Resnik, 2002; Weissgerber et al., 2016). In the future, we plan to develop and evaluate additional course activities and assessments that are designed based on data from authentic life science research studies to target gaps, including those that push students’ statistical knowledge boundaries. If students are consumers or procedures of quantitative research in the future, they will inevitably encounter situations that are beyond the scope of their statistical knowledge and skills. So, students need to gain experience with unfamiliar data situations and help-seeking in their course. It may also help to integrate more activities involving simulation (e.g., Shiny apps) to explore consequences of model misspecification. Future research looking into student reasoning about data, including hierarchical data, and students’ perceptions about statistical practice in research and quantitative research behaviors later in their studies and research would be interesting. Multidisciplinary collaboration has been extremely valuable in this course, and the related research, so we will continue to take a multidisciplinary approach to our future work.

REFERENCES

- Allen, B. (2015). Healthy And Unhealthy Statistics: Examining The Impact Of Erroneous Statistical Analyses In Health-Related Research, *Electronic Thesis and Dissertation Repository*, The University of Western Ontario. 3119. Available at <https://ir.lib.uwo.ca/etd/3119/>.
- ASA. (2018). Overview of Statistics as a Scientific Discipline and Practical Implications for the Evaluation of Faculty Excellence [Position paper]. <https://www.amstat.org/asa/files/pdfs/POL-Statistics-as-a-Scientific-Discipline.pdf>.
- Baker, M. (2016). Reproducibility crisis. *Nature*, 533(26), 353-66.
- Finney, S. J., & Schraw, G. (2003). Self-efficacy beliefs in college statistics courses. *Contemporary educational psychology*, 28(2), 161-186.
- GAISE College Report ASA Revision Committee (2016). Guidelines for Assessment and Instruction in Statistics Education College Report 2016. [https://www.amstat.org/education/guidelines-for-assessment-and-instruction-in-statistics-education-\(gaise\)-reports](https://www.amstat.org/education/guidelines-for-assessment-and-instruction-in-statistics-education-(gaise)-reports).
- Gardenier, J. and Resnik, D. (2002). The Misuse of Statistics: Concepts, Tools, and a Research Agenda, *Accountability in Research: Policies and Quality Assurance*, 9(2), 65-74.
- Garfield, J. & Ben-Zvi, D. (2007). How Students Learn Statistics Revisited: A Current Review of Research on Teaching and Learning Statistics, *International Statistical Review*. 75, 3, 372-396.
- Lazic, S. E. (2022). Genuine replication and pseudoreplication. *Nature Reviews Methods Primers*, 2(1), 1-2.
- Lazic, S. E. (2010). The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis?. *BMC neuroscience*, 11, 1-17.
- Ocaña-Riola, R. (2016). The Use of Statistics in Health Sciences: Situation Analysis and Perspective, *Statistics in Biosciences*, 8, 204-219.
- Sato, S., Omori, Y., Katoh, K., Kondo, M., Kanagawa, M., Miyata, K., Funabiki, K., Koyasu, T., Kajimura, N., Miyosh, T., Sawai, H., Kobayashi, K., Tani, A., Toda, T., Usukura, J., Tano, Y. & Furukawa, T. (2008). Pikachurin, a dystroglycan ligand, is essential for photoreceptor ribbon synapse formation”. *Nature neuroscience*, 11(8), 923-931.
- Tong, L., White, B. J. G., & Singh, J. (2022). Bridging statistics and life sciences undergraduate education. *Journal of Biological Education*, 1-13.
- Weissgerber T. L., Garovic V. D., Milin-Lazovic J. S., Winham S. J., Obradovic Z., Trzeciakowski J. P., Milic, N. M. (2016). Reinventing Biostatistics Education for Basic Scientists. *PLoS Biol*, 14(4): e1002430.
- White, B. J. G., & Singh, J. (2021). Partnering to Prepare Tomorrow’s Life Sciences Researchers [Poster Presentation]. *United States Conference on Teaching Statistics (USCOTS)*. <https://www.causeweb.org/cause/uscots/uscots21/th-03-partnering-prepare-tomorrow%E2%80%99s-life-sciences-researchers>.
- Wild, C. & Pfannkuch, M. (1999). Statistical Thinking in Empirical Enquiry. *International Statistical Review*, 67(3), 223-265.