FINDING MEANING IN A MULTIVARIABLE WORLD: A CONCEPTUAL APPROACH TO AN ALGEBRA-BASED SECOND COURSE IN STATISTICS

Karen McGaughey[1], Beth Chance[1], Nathan Tintle[2], Soma Roy[1], Todd Swanson[3], Jill VanderStoep[3]
[1]Department of Statistics, Cal Poly, San Luis Obispo, CA
[2]Department of Math, Computer Science & Statistics, Dordt College, Sioux Center, IA
[3]Department of Mathematics, Hope College, Holland, MI
kmcgaugh@calpoly.edu

Although the teaching of the first course in statistics has improved dramatically in recent years, there has been less focus on a similarly conceptual-based second course aimed at non-majors. We present a curriculum for the second course, designed to expand statistical literacy across disciplines, which focuses on conceptual understanding of multivariable relationships through data visualization, study design, the role of confounding variables, reduction of unexplained variation, and simulation-based inference, rather than the mathematically-based discourse often used in the second course. Our curriculum uses a student-centered pedagogical approach, utilizing guided-discovery activities based on real-world case studies, facilitated by student-focused technology tools. Highlights of the curriculum and student assessment will be shared.

INTRODUCTION

The last few decades have witnessed a dramatic improvement in teaching of introductory statistics at the second and tertiary levels. New recommendations in pedagogy (e.g., GAISE Guidelines, 2005; 2016) have emphasized statistical thinking and the statistical investigative process, effective use of technology, use of genuine research studies, and active learning. In turn, recent content changes have included more integration of computationally-intensive methods such as bootstrapping and randomization tests. Much of this has been facilitated by new technology tools such as applets with a more student-centered focus, as well as educationally-focused enhancements to statistical software. These approaches appear to be more appealing to a more diverse student body than ever before, enrollments have been rising dramatically, and national standards at lower grade levels have also been increasing the use of statistics. More and more students are arriving to the university courses already with more than minimal background in statistics. Despite nearly one million students per year taking introductory statistics courses in the U.S. alone (American Mathematical Society, 2013; College Board, 2014), very few go on to take a second course in statistics, leaving a dearth of qualified practitioners who can understand and utilize multivariable statistical methods (Switzer & Horton, 2007). Hence, it is now more important than ever to offer a second course in statistics accessible to a wider audience; students without the calculus background required of the usual Stat 2 course. Our goal was to develop and implement a novel curriculum for an algebra-based second course in statistics, aimed at a wide audience across many disciplines. In this paper we provide an overview of the curriculum, highlight some of the distinctive features, and provide initial assessment results from the first full implementation at one university.

RECURRENT THEMES

Complementing the recent changes in the introductory course, attention has been shifting to the follow-up course. Historically, the second course made heavy use of calculus and linear algebra to explain the underpinnings of the procedures. Increasingly, the focus has been shifting to reflect how these methods are being used in statistical practice and to make the course accessible to a more general audience (e.g., Cannon et al, 2011; Kuiper & Sklar, 2013, Ramsey & Schafer, 2013). Our curriculum has focused on the same general topics (e.g., regression and ANOVA) but with more emphasis on the following ideas:

- Explaining variation in the response and minimizing unexplained variation
- Identifying and adjusting for possible confounding variables
- Using visualizations to tell the story of the data

In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

- Using simulation-based methods to understand the logic of inference
- The role of study design in the choice of analysis and the scope of conclusions
- Use of model equations and variable relationship diagrams to describe associations, with correspondence to regression equations and partitioning of sums of squares

All of these ideas are discussed in the context of real case studies, starting with visualizations of the data and consideration of variable relationships. The analyses, whether simulation-based or theory-based, leverage student intuition and emphasize conceptual understanding, with minimal use of formulas.

PEDAGAOGY

In particular, we wanted to follow the "spiral" approach of our first course, changing one thing at a time and revisiting key ideas (e.g., adjusting, interaction) at multiple points in the course. We wanted these changes to be motivated by the data structure (e.g., comparing two means, comparing more than two means), and we wanted to build as much of the course as possible around genuine research studies (e.g., wage discrimination, modified air storage of strawberries). The content of each section is presented through both a worked-out example, with several "think about it" reflection questions sprinkled in, and a hands-on exploration. Free on-line student-centered applets are used in conjunction with standard statistical packages (R or JMP) to explore these ideas and with a focus on interpreting output. The spiral approach is exemplified by an initial "Preliminaries" chapter that gives students three diverse contexts that will recur in the book, focusing on the issues of visualization and multivariable relationships in each setting.

COURSE OUTLINE

After the Preliminaries chapter, Chapter 1 focuses on comparing two groups on a quantitative response, reviewing many ideas from the first course, but also establishing the simulation-based inference framework that will be used throughout the course. Unlike the first course, we introduce both randomization tests and bootstrapping. We want students to always be thinking about how the study design impacts the simulation model and how the choice of simulation model impacts the analysis results. In Chapter 2, we move to comparing more than two groups (randomization tests, one-way ANOVA, multiple comparisons, randomized block designs, and adjusting for confounding variables). In this chapter, we establish the idea of a variable relationship model and dividing variation in the response variable into variation explained by the explanatory variable and unexplained variation, using pie charts to help visualize the partitioning. We also re-visit the idea of multivariable thinking addressing how we deal with confounding variables, either through the design of study in an experiment (random assignment and blocking) or through the analysis of the study (adjusting for a confounder) in an observational study. Chapter 3 focuses on designing and analyzing studies using two experimental variables (factorial designs, interactions, issues with observational data). Chapter 4 transitions to quantitative explanatory variables, reviewing ideas of regression (again simulation-based and theory-based inference), as well as coding categorical explanatory variables. Chapter 5 focuses on multiple quantitative variables, with ideas of collinearity, R^2 adjusted, and added variable plots). Chapter 6 focuses on nonlinear relationships and investigates transformations and polynomial models. Chapter 7 focuses on a categorical responses variable (logistic regression). A final chapter focuses on cases studies that involve messier data collection and analysis issues.

NEW APPLETS

Another potential advantage to continued used of simulation-based methods is more opportunities for active learning and student exploration (e.g., Chance, Wong, & Tintle, 2017). We have expanded our applet collection to include applets dealing with bootstrapping and multivariable relationships (e.g., randomized block designs and two-way ANOVA). For Randomized Block Designs, students can explore restricted randomization as well as adjusting for block effects before re-randomizing. We have also expanded some of the existing applets, giving students more opportunities to explore concepts like power, the null distribution of the p-value, and alternative statistics like the mean absolute deviation. These free applets help students focus on statistical concepts and visualizations throughout the course and provide a complement to standard statistical packages.

EXAMPLE

To give a sense for how the curriculum progresses, we highlight the first section of Chapter 3. The overall goal of the chapter is factorial designs and interactions. Students are first asked to consider different study designs involving two variables, where the goal of the study is to find the settings that optimize a response (free throw shooting points; Orn, 2017). The purpose here is to motivate the need for a factorial study looking at treatment combinations, as opposed to two one-variable-at-a-time studies. Students then look at interaction plots to summarize the results and think about different choices for statistics to summarize the interaction through one number, ultimately leading to discussion of how the "difference in differences" can measure the statistical interaction. They then use the new two-way ANOVA web applet to explore the behavior of this intuitive statistic under repeated random assignment of the observed scores to the original treatments. By the end of the first class period, they can address the original research question with a simulation-based p-value for the interaction.

ASSESSMENT

To assess gain in student understanding within the course, we developed a pre-test that was given to students at the beginning of the course. This multiple-choice test was modelled after the assessment we have used for many years in the introductory courses. Most students completed the pre-assessment outside of class, but were given homework credit for completion. Some of the questions concerned big ideas from the introductory course, but several were aimed specifically at the content in the second course (e.g., interactions). Some of the questions reappeared as quiz or exam questions or on the final exam, to track improvement. We have expanded these assessment efforts to several class testers for Spring 2018.

We are still analyzing the assessment data but have a few preliminary observations. Table 1 shows the percentage correct on the pre-test for the questions that were also given on the in-class post-test as part of the final exam for a class of undergraduates, Fall 2017. In this sample, about 59%

Table 1. Pre and post test results for questions used on both the pre-test and in-class final exam

Question	Pre % correct (n = 30)	Post % correct (n = 26)
3. Which 2 of age, sex, and year use to predict GPA	27.6%	65.4%
4. PI vs. CI (prediction intervals not discussed in course)	41.4%	38.5%
5. Larger slope means larger impact	79.3%	100%
7. How phrase conclusion with significant ANOVA	69%	100%
18. Use p-value for slope from regression table to assess significance of association	75.9%	92.3%
19. Change in regression output when add a variable implies explanatory variables are associated	79.3%	88.5%
20. Can't compare predictions in multiple regression without adjusting for other variables	60.7%	88.5%
Given description of a randomization test simulation		
26. Reason behind simulation process	44.8%	96.2%
27. Assumption behind null distribution	48.3%	88.5%
29. conclusion about research question	24.1%	61.5%
30. correct interpretation of p-value	48.3%	80.8%
31. Which dotplots will have smaller p-value (differ in n)	55.2%	88.5%
32. Which dotplots will have smaller p-value (differ in SD)	51.7%	88.5%
33. Identifying an interaction from a table of means	69%	84.6%

were female, and respondents were equally distributed between having taken their introductory statistics course in the past 3-6 months, in the past 6-12 months, and more than one year ago. About 48% received an A or A- in their first course. The results of Questions 3, 19, 20 and 33 suggest that students are generally leaving the course with an improved understanding of multivariable relationships. In addition, the students appear comfortable with simulation and the role of study design in the design of the simulation. Most students could distinguish observational studies from designed experiments and understood the implications for causation, as well as the purpose for inclusion criteria and the limits on scope of generalization. Areas where students still seemed to struggle included the purpose of random assignment, and the distinctions between confounding, collinearity, and interaction.

From the end of course evaluations in the fall, 47.4% felt the mathematical level was appropriate. The most confusing topics cited were leverage plots and those coming at the end of the course (transformations and polynomial models). Many self-reported comments were positive about the interactive, learn-by-doing nature of the course, mixing lab sessions with lecture sessions.

SUMMARY

Tremendous growth in the algebra-based first course in statistics and increasing exposure to statistics in primary and secondary schools suggest that an algebra-based second course may be appealing and necessary to a growing number of students. However, the vast majority of second courses require substantial mathematical understanding (e.g., linear algebra, calculus). Here we present a second course that assumes only an algebra-based first course, while emphasizing active learning, simulation, and real data. Key distinctions from the few other curricula we have seen for an algebra-based second course are the spiral approach (introducing key ideas earlier and spiraling over them throughout the book), and the use of visualization, simulation, and intuitive guided discovery to help motivate and reinforce student learning. Preliminary work on the curriculum is nearly complete and initial assessment results are positive but also informative of future changes to the curriculum. For example, we have seen a need to help students more clearly differentiate confounding and interaction between variables. Some of our next goals include

- Refining and disseminating our assessment tool for a second course in statistics
- Develop two more applets: one focused on visualizing Simpson's Paradox and one focused on added variable plots for multiple regression allowing students to drag variables into the model and compare the adjusted and unadjusted association
- Further reduction, replacement of technical language (e.g., orthogonality, additive, parameterization)
- More focus on technical writing skills

ACKNOLWEDGEMENTS

This research was supported by NSF grant NSF/TUES/DUE # 1612201.

REFERENCES

American Mathematical Society (2013). *Conference Board on Mathematical Sciences*. Washington, D.C..

Chance, B., Wong, J., & Tintle, N. (2017). Student performance in curricula centered on simulation-based inference: A preliminary report, *Journal of Statistics Education*, *24*(3), 114-116.

Cannon, A., Cobb, G.W., Hartlaub, B.A., Legler, J.M., Lock, R.H., Moore, T.L., Rossman, A.J. & Witmer, J. (2011). *Stat 2: Building models for a world of data.* W.H. Freeman.

College Board (2014). *Score Distribution for AP Statistics, 2014*.

Kuiper, S., & Sklar, J. (2013). *Practicing Statistics: Guided investigations for the second course*. Pearson.

Orn, A. (2017). *Effects of Pressure and Free Throw Routine in Basketball Kinematics and Sports Performance*, Master's Thesis, Arizona State University.

Ramsey, F., & Schafer, D. (2013). *The Statistical Sleuth: A course in methods of data analysis* (3rd edition). Brooks/Cole Cengage Learning.

Switzer S., & Horton N.J. (2007). What your doctor should know about statistics (but perhaps doesn't), *Chance*, 20:17–21.