ENHANCING CONCEPTUAL UNDERSTANDING WITH DATA DRIVEN LABS

Robert Gould, Gretchen Davis, Rakhee Patel and Mahtash Esfandiari
Department of Statistics, University of California, Los Angeles, United States of America
rgould@stat.ucla.edu

Teaching introductory statistics with a data-driven curriculum presents many challenges for the instructor. One challenge is to provide students with opportunities to work with data in a realistic context. If not done carefully, students spend their time struggling to learn the software, not engaging with the data. Students might be able to follow step-by-step instructions to "see" how data analysis is done, but still fail to connect this to important concepts. We report on a project to create a set of data analysis activities that use Fathom to engage students in exercises that emphasize the challenges of statistical inference beginning in the very first week of the course; involve students with real data and real research questions; and require students to discover analysis procedures on their own. The resulting set of labs emphasizes simulation and randomization-based inference procedures while working in the context of real data.

INTRODUCTION

In the 1980's, dissatisfaction with the introductory statistics curriculum in the United States came to a head. This dissatisfaction had been brewing for quite some time, but the 1980's and '90's produced a long list of publications from statisticians, education researchers, and statistics teachers calling for change and suggesting that many problems could be fixed by centering the curriculum on real data. (For example: Cobb, 1991; Marquardt, 1987; Singer & Willet, 1990.) These papers propose that using real data would better motivate students, would better teach the application of concepts and would better teach statistical thinking. Perhaps most importantly, centering on data would help shift the emphasis from rote application of mathematical formulae to a deeper conceptual understanding of the fundamentals of statistics, such as the omnipresence of variability, data collection methods, and the relation of models to reality.

One important consequence of teaching with real data is that it becomes essential that technology be included in the classroom. (“Technology”, in this paper, refers both to hardware and software, although in this discussion we are mostly concerned with software.) This inclusion of technology creates its own set of challenges. In our own introductory statistics course, these challenges are (i) the complexity and accessibility of the data we can consider depends on our choice of technology and the technology the students have available; (ii) the pedagogy to teach the use of technology within a statistics curriculum is in its infancy, which means there are many unanswered questions about how to most efficiently teach students to use software to solve statistical problems; (iii) real data are messy, and so the need to teach data cleaning and management and “data wrangling” nudge into the curriculum.

Challenge (i) is particularly important for any course in which data are used to motivate students and convince them of the relevance of statistics. One difference between our current students and students 10 years ago is that the current students actually interact with data on a daily basis, although they might not be aware of it. For example, a student's mp3 player can be thought of as a data analysis device that converts an mp3-format dataset into sound. And with the right statistical technology, this mp3 file can be evaluated using the same statistical thinking approaches we hope students will apply to more traditional data-based problems. Less exotically but perhaps more importantly, students can find their own datasets on the internet, although often these are in awkward formats or might involve complex data structures. With the right technology, students can learn to access and analyze a dataset to answer a question of personal interest, and not limit their enquiries to the formal contexts provided in most statistics texts.

Our past attempts to teach students to use professional caliber statistical software (Stata) with relatively messy datasets produced frustrating results (Gould, Kreuter & Palmer, 2007). These labs were intended to teach economics students how to use a package that they would likely use professionally, however we found that students treated the instructions as recipes to be memorized, and not as an intellectual process. Some students treated the instructions so rigidly that they blindly repeated the same steps on every set of data, regardless of the structure of the data or the error

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php [© 2010 ISI/IASE]

messages produced. The lack of a pedagogy for teaching technology is therefore strongly felt. While the literature contains numerous examples of innovative and effective uses of particular technology applications (Chance, Ben-Zvi, Garfield & Medina, 2007), there is to our knowledge no comprehensive pedagogy that sees technology as an integral part of statistical thinking. This is changing: Nolan and Temple Lane (2009) challenge instructors to make computing central to the statistics curriculum and report on several efforts to develop instructional materials. But as our experience shows, if not done properly, rote application of computation provides no more understanding does rote application of mathematical formulae.

This paper describes an on-going attempt to address these challenges at one particular university for one particular introductory statistics class. The attempt takes the form of a Data Analysis Lab Manual that consists of seven data-based and technology driven investigations. As of this writing, formal evaluation has not yet begun, but will be completed by June 2010. This paper is therefore written in the spirit of a case study, and the authors hope that by supplying details, other instructors will have some useful materials for their own students and generate their own ideas about how to include real data in their class.

This paper focuses on a trio of labs that teach informal statistical inference through the use of randomization and permutation tests. We emphasize simulation approaches because they expose students to the use of computer and software technology for solving statistical problems, and do so in a "realistic" fashion (that is, they are tools and approaches that statisticians use). The literature identifies many reasons for why simulation-based methods can be effective teaching tools. Simulation-based approaches readily lend themselves to a constructivist approach so that students' learning is based on their own experiences with randomness and they form a coherent system of reasoning that is, in some ways, more coherent than are Normal probability-based procedures (Garfield & Ben-Zvi, 2007; Mills, 2002; Cobb, 2007).

THE CONTEXT

Statistics 10, Introduction to Statistical Reasoning, emphasizes statistical literacy but was redesigned in 2005 to include a "statistical thinking" component (Garfield & Ben-Zvi, 2007). Stats 10 is taken by about 480 students per term. (A term is 10 weeks.) Students attend lecture by the professor three times each week, in groups of about 160. Twice a week they meet in smaller groups of 40 with a graduate-student Teaching Assistant (TA) once in a discussion session, and once in the computer lab. Therefore, the students perform these lab exercises under the direction of the TA.

The lab manuals are intended to give students experience working with data and strengthening their understanding of the concepts covered in lecture. Specifically, we designed these labs so that:

- students would learn that statistics was applicable to problems that matter and are relevant.
- students would reinforce their understanding of statistical concepts, particularly those concerning inference.
- students would engage intellectually with the labs, and learn to see statistical analysis as a process of inquiry and discovery, and not as a recipe.

Absent from these goals is the desire for students to learn any particular statistical software. One reason for this is that we feel there are some basic concepts in data analysis that are independent of software, and so the choice of actual software is less important for an introductory course than for a course that is preparing future professionals. We therefore chose Fathom, because we felt that, compared to other statistical software, it gets "out of the way" so that students focus on the data, and not on the software's syntax (Finzer et al., 2007).) Fathom also provides fairly sophisticated tools for accessing data. For example, in many cases, students can easily (i.e., with no programming) access data stored on the internet in the form of html tables.

BRIEF DESCRIPTIONS OF LABS

The first two labs use data from the North Carolina birth registry (provided by John Holcomb, Cleveland State University). The first lab is designed to get them acquainted with Fathom and to teach them to explore a set of data to determine whether associations between

variables exist. Specifically, students are asked to describe the effect of mother's smoking during pregnancy on the baby's health. Instructions for basic Fathom operations, such as creating summary statistics or graphs are provided in a part of the document separated from the primary research questions. Computer instructions are deliberately general, for example, "To make a graph, drag the graph object from the toolbar. Drop variable names onto the axes of the empty graph or onto the graph itself. Notice that a selector appears in the upper-right side of the graph that allows you to toggle between different types of graphs."

In the second lab, students address the affect of mothers' smoking by examining the weight of the babies at birth. In addition to the original data, students are given a "scrambled" set in which baby weights have been randomly assigned to mothers. They are asked to explain why this scrambled "chance model" is consistent with the hypothesis that smoking has no effect on babies' weight, and are asked to generate multiple instances this chance model. Finally, they are asked to write a paragraph comparing the real data to the multiple chance model instances, and say what this tells them about the effect of smoking on birth weight. This happens in the second week of the course, well before the topic of inferential statistics has been taken up in the book or lecture.

The penultimate lab, which occurs in the 8th week of the course, explores a historic data set: an early randomized study of the effectiveness of an antibiotic to cure tuberculosis. Again, students are asked to compare the actual outcome (best summarized by a two-way table) of the study to a chance model. This time they are asked to do a large number or repetitions and to estimate the probability that a test statistic as extreme or more extreme than what they observed could occur by chance. They are asked to explain the relation between this estimated probability and the concept of "p-value", which they have recently studied in class (but not in the context of randomization tests.) This lab presents several challenges. Students are able to choose any test statistic they wish, but are encouraged to consider using the value of one of the cells, for example, the number of people who died and received the antibiotic. They must then carefully explain what is meant by "as extreme or more extreme." Essentially, students are being asked to map their understanding of hypothesis tests learned in the context of proportions and means to the context of permutation tests.

PEDAGOGY

A key feature of the labs is that they often introduce concepts before these have been formally presented in class. As a simple example, students are asked to choose between a ribbon chart and a bar chart for displaying an association between two particular categorical variables. (Ribbon charts are essentially segmented bar charts, in which the width of the bars is proportional to the proportion of cases of the conditional variable.) Ribbon charts are not explained, but students must teach themselves to read the charts through comparison with bar charts. For simulation-based inference labs, students are not given a set of procedures, but instead are shown the general framework and asked to describe their own reasons for their conclusions. The goal is to get students to use their intuition to develop approaches and methodologies on their own. Particularly since there is some evidence that students often build intuition for more complex concepts based on their understanding of simpler concepts (Zieffler & Garfield, 2009), we wished to challenge students very early to think of statistics as a "natural" approach to solving some unnatural (at least from a naive point of view) problems.

Following our own experience and others reported in the literature, we hope that computer-based simulations will be more effective if preceded by more tactile experience (Lane-Getaz & Zieffler, 2006). Therefore, before doing permutations in the lab, students do a similar activity in lecture by physically shuffling cards to understand how two-way tables might appear if two categorical variables were independent.

DISCUSSION

To date, we have done only informal, formative evaluations, with the exception of a review by our external advisors. The advisors felt the labs, as a whole, were appropriate in statistical content and pedagogy, and sufficiently challenging for introductory students. A formal review by an external evaluator is scheduled for Spring, 2010. Lacking the results of the formal evaluation, our remarks will address some of the continuing challenges of this project.

Our informal evaluations and weekly discussions with TAs have shown that considerable challenges remain. Despite the fact that some researchers have labeled the current generation of students as "digital natives" (Bennett et al., 2008), we find that students still struggle over basic computer operations such as finding downloaded files. We find that even small things, such as presenting categorical variables in the form of their original numerical codes -- cause great frustration for students and distract them from the primary goal of the lab. This might indicate that basic data cleaning procedures are sufficiently challenging to deserve special attention and argues that if real data are to be the center of the curriculum, then computing must be included as an important "statistical concept" (Holcomb & Spalsbury, 2005; Nolan & Temple Lang, 2009).

A common misconception students reveal in the labs is that the null distribution of the test statistic is seen as the "real" distribution, and students reason that because the distribution is centered at 0, the null hypothesis cannot be rejected. We will investigate this further in the formal evaluation, and in particular are interested in seeing if this misunderstanding lessens after students' exposure to formal hypothesis testing.

REFERENCES

Bennet, S., Maton, K., & Kervin, L, (2008). The 'digital natives' debate: A critical review of the evidence. *British Journal of Educational Technology*, *39*(5) 775-786.

Chance, B., Ben-Zvi, D., Garfield, J., & Medina, E., (2007). The Role of Technology in Improving Student Learning of Statistics. *Technology Innovations in Statistics Education*, *1*(1). Online: http://escholarship.org/uc/item/8sd2t4rr.

Cobb, G. W. (1991). Teaching Statistics: More Data, Less Lecturing, *AMSTAT News*, No. 182.

Cobb, G. W. (2007). The Introductory Statistics Course: A Ptolemaic Curriculum? *Technology Innovations in Statistics Education*, *1*(1). Online: http://escholarship.org/uc/item/6hb3k0nz.

Finzer, W., Erickson, T., Swenson, K., & Litwin, M., (2007). On Getting More and Better Data Into the Classroom. *Technology Innovations in Statistics Education*: *1*(1). Online: repositories.cdlib.org/uclastat/cts/tise/vol1/iss1/art3.

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.

Gould, R., Kreuter, F., & Palmer, C., (2006). Towards Statistical Thinking: Making Real Data Real. In *Proceedings, International Conference on Teaching Statistics 7*, Salvador, Brazil.

Holcomb, J., & Salbury, A., (2005). Teaching Students to Use Summary Statistics and Graphics to Clean and Analyze Data. *Journal of Statistics Education*, *13*(3). Online: www.amstat.org/publications/jse/v13n3/datasets.holcomb.html.

Lane-Getaz, S. J., & Zieffler, A. S. (2006). Using simulation to introduce inference. In *Proceedings of the 2006 Joint Statistical Meetings*, Alexandria, VA. American Statistical Association.

Marquardt, D. (1987). The Importance of Statisticians. *Journal of the American Statistical Association*, *82*(397), 1-7.

Mills, J., (2002), Using Computer Simulation Methods to Teach Statistics: A Review of the Literature. *Journal of Statistics Education*, *10*(1).

Nolan, D., & Temple Lang, D., (2009). Approaches to Broadening the Statistics Curricula, Chapter 18 (pp. 357-381). In M. C. Shelley II, L. D. Yore, & B. Hand (Eds.), *Quality Research in Literacy and Science Education: International Perspectives and Gold Standards*, 2009. Dordrecht, The Netherlands, Springer.

Joiner, B. (1988). Let's Change How We Teach Statistics. *Chance*, *1*(1), 53-54.

Singer, J., & Willet, J. (1990). Improving the Teaching of Applied Statistics: Putting the Data Back into Data Analysis. *The American Statistician*, *44*(3), 223-230.

Zieffler, A., & Garfield, J., (2009). Modeling the Growth of Students' Covariational Reasoning During an Introductory Statistics Course. *Statistics Education Research Journal*, *8*(1) 7-31.