TOWARDS STATISTICAL THINKING: MAKING REAL DATA REAL

Robert Gould, Frauke Kreuter, and Christina Palmer
University of California - Los Angeles, United States
rgould@stat.ucla.edu

Although the Statistics Education community has advocated using real data to teach introductory statistics for quite some time, often these data sets are not recognizably real to statisticians since the students' limited experience with "real" statistical software and data management techniques precludes the use of truly messy data. But grappling with messy and complex data sets is important for teaching Statistical Thinking (broadly defined as "thinking like a statistician") and is appropriate for an introductory statistics course. We describe our experience collecting rich data sets and developing computer lab assignments using Stata *to teach statistical thinking to first-year university students using these data sets. Collecting useable, real, data sets turns out to be fairly difficult for several reasons, and teaching data management and analysis without resorting to rote-based rules is quite challenging.*

INTRODUCTION

Using "real" data to teach introductory statistics has become an accepted tenet of Statistics education. Advocates of real data (e.g., Cobb, 1991; Singer and Willet, 1990) propose that the inclusion of real data would improve the introductory statistics course on several important dimensions. These dimensions include exposing students to real and important questions, teaching data exploration, and improving student motivation (Singer and Willet, 1990).

Many of these dimensions address characteristics relevant to the goal of teaching *statistical literacy* (http://www.gen.umn.edu/artist/glossary.html#statistical_literacy). We are in agreement with H.G. Wells in believing that all citizens need to be statistically literate, whether they are consumers or producers of Statistics (Wells, 1903). However, statistical literacy should not be the only goal of a statistics course (GAISE College Report, 2005). *Statistical thinking* is another quality educators strive to develop in their students. Statistical thinking has been described as the complex process that statisticians go through when solving statistical problems (Ben-Zvi and Garfield, 2004). Statistical thinking is statistical literacy made active.

Just as carefully chosen data sets are a necessary component of a statistical literacy course, certain types of data sets are better suited to teaching statistical thinking. If statistical literacy requires real data, then teaching statistical thinking requires *very* real data. It is our experience that, while the "real data" revolution has produced textbooks with richer data sets to analyze, these data sets are often too simple for the purposes of teaching statistical thinking. Part of the reason for this is that statistical thinking requires some practical, computer-based skills that are often beyond the scope of an introductory class.

In this paper we address the qualities of data that we think make them "very real." We do so in the context of a data analysis lab offered as part of the introductory statistics courses we have been teaching for Economics and Biology undergraduate students. To set the stage, we begin with a plea for first courses to spend more time with one of the earliest stages of the statistical investigation cycle: data management. We then propose a schema for measuring the quality and structure of a data set as a means to determine its level of "realness," followed by a description of the approach we took at UCLA to offer very real data for analysis with statistical software to students in an introductory statistics course.

GETTING ONE'S HANDS ON THE DATA

We believe a first course in Statistics should provide some training for how to manage difficult data so that upon completing the class, students have the skills to solve some real-life problems. The popularity of spreadsheet packages, which allow columns of data, comments, graphs, and unusual character objects to peacefully co-exist makes these skills particularly important. While we cannot prepare our students for an arbitrarily messy data set, by using "very real" data in our classes, we can better prepare them for the real world of data analysis.

One motivation for providing these skills for our introductory students is that a student

must be able to "get her hands on" the data before she can demonstrate her statistical thinking skills. This means that she must be able to upload the data file into a software package, determine the format the data are stored in, e.g., byte, string, floating, change the format if necessary, and assess the quality and structure of the data. This important first stage of the statistical investigative process is often the most time consuming and frustrating, and arguably the most difficult to learn.

WHAT MAKES REAL DATA REAL? MEASURING QUALITY AND STRUCTURE OF DATA SETS

We identified several characteristics that operationalize the concepts of quality and structure of data sets and that can be used to distinguish "reality" levels in data sets.

1. *Length* (# observations). Real data sets should be large enough to make the need for computers, rather than calculators, obvious. They should be large enough that graphical and numerical summary techniques reveal structure that is otherwise not apparent.
2. *Width* (# variables). Real data sets should have enough variables that students have room for exploration, for testing alternative hypotheses, and for performing residual diagnostics.
3. *Form*. Real data include missing values, un-coded values, and sometimes misspelled values; real data can include both numerical and character-valued variables.
4. *Structure*. Real data typically have complex structure. For example, linear relations are rare and even then high correlations are unusual. Distributions can be highly skewed, or multi-modal. Students should also see that some forms of data, for example longitudinal, can be stored in different "shapes," depending on whether rows represent an observation made on a subject at a particular time, or contain all observations for a particular subject.

COLLECTING REAL DATA SETS AND DEVELOPING COMPUTER LABS TO PROMOTE STATISTICAL THINKING: THE UCLA EXPERIENCE

To accommodate requests from several campus departments, the UCLA Statistics Department developed undergraduate lower-division courses in the "Introduction to Statistical Methods in X" series, where X is a specific course for economics majors, a specific course for majors in the social sciences, a specific course for geology and environmental science majors, or a specific course for majors in health and biological sciences. These introductory statistics courses were developed to emphasize data-analytic problem solving on a foundation of statistical literacy. Over a 10-week period, the courses involve students through three hours of didactic lectures, one hour of discussion, and a one-hour computer laboratory per week. Although the courses cover nearly the same statistical concepts and techniques, they differ in terms of the real-world discipline-specific data sets used to teach statistical literacy and thinking, and promote understanding of 'what makes real data real' while developing computing and analytic skills. In order to achieve these goals, instructional development funding was awarded to Professors Gould and Palmer over the course of 3 years to (1) collect real data sets arising from the study of discipline-specific problems, and (2) develop computer labs that adroitly used the data sets.

Data sets. We decided to focus our efforts on collecting "local" data sets, under the hypothesis that students might take greater interest in data collected by faculty with whom they might have taken or with whom they might someday take classes. We hoped that these data would be from recent papers so that students would be motivated by sharing in a sense of discovery. We also hoped these data would be more real, and hence more complex, than those typically used in introductory courses. For this reason, we hired two graduate students to collect data sets from UCLA faculty and to interview those who donated data sets so that we could provide students with a detailed context of the data. We were ultimately able to obtain 12 local data sets contributed by UCLA faculty that were either the result of their own research or were data sets that they referred to in their work (Table 1).

Due to challenges we encountered in obtaining and working with local data sets (described below), we ended up supplementing the 12 local data sets with 14 non-local data sets from the literature or the web (Table 2). The local data sets represented local research and research that local researchers in client departments felt to be important. The non-local data sets represented data that our own faculty "scraped" off the internet (for example, census data, stock

prices), and data sets from previously published resources.

Table 1: Characteristics of Local Data Sets Used in UCLA Introductory Statistics Courses

Data Set	Characteristic		
	Length (# obs)	Width (# var)	Form*
Thatch Ants	1199	6	Char, cat, quant; missing; uncoded
Seed Ants	577	6	Char, cat, quant; uncoded; ranges
Risk Perception	13,442	7	Char, cat, quant; missing
Ashe Wellness Survey	640	5	Cat, quant; missing
CA Dept of Corrections	3922	5	Cat, quant; missing; unlabelled
Fast Food	410	46	Cat, quant; missing
Birds	40	9	Char, quant; missing
Cardiac	558	42	Char, quant
Guppies	119	27	Cat, quant; missing; unusual characters
Seaslugs	47	2	Quant
Students	82	8	Cat, quant
Twins Study	183	16	Cat, quant; missing
	Mean=1768.3 Median=484	Mean=14.9 Median=7.5	

* char=character, cat=categorical, quant=quantitative

Table 2: Characteristics of Non-Local Data Sets Used in UCLA Introductory Statistics Courses

Data Set	Characteristic		
	Length (# obs)	Width (# var)	Form*
Baseball	44	12	Char, cat, quant; uncoded
Census	2494	21	Cat, quant; missing; uncoded
Stock Market	500	30	Char, quant; missing; unlabelled
Broadway	18	7	Char, cat, quant
U.S. Professor Salary	1160	16	Char, quant; missing; uncoded
NY Parking Meter Collections	47	4	Cat, quant; missing
CA Missions	21	6	Char, quant; unlabelled
Birthdays	365	2	Char, quant; unlabelled; dates
SAT Scores	50	8	Char, quant
1970 Draft Lottery	366	2	Char, quant; unlabelled; dates
Body Temp	130	3	Cat, quant
Vietnam Deaths & Draft #s	12	4	Cat, quant
Captopril	15	2	Quant; paired
Green Vehicles (EPA)	798	15	Cat, quant
	Mean=430 Median=90	Mean=9.4 Median=6.5	

* char=character, cat=categorical, quant=quantitative

We compared the quality and structure of the local and non-local data sets in terms of length, width, and form as described above. Consistent with our intuition, local data sets tended to be "more real" than non-local data sets, and as such should enhance the development of statistical thinking skills. As shown by comparing Tables 1 and 2, our local data sets have greater length and width and have more complicated form than our non-local data sets (as an example, 67% of local data sets contain missing values vs. 28.5% of the non-local data sets).

We do not mean to imply that non-local data sets cannot be as complex as our local data sets, because certainly one can find very complex data sets publicly available on the web. We do feel, though, that one must hunt with complexity in mind, and one method for hunting is to search for data that are important to local researchers.

Computer lab manual. We also used our funding to hire graduate students to assist in writing and developing a computer lab manual that would teach the introductory students to use statistical software to analyze and explore the wide variety of local and non-local data sets. Our basic format was that each lab would introduce a real problem and a real data set, and then would illustrate useful statistical techniques with their software commands for approaching the problem. These commands would be executed on a data set in the lab, and students would record their thoughts and observations in a lab book. A summary activity would ask them to perform an additional "take home" analysis, often on a different data set. The motivating research question behind this take-home exercise and the accompanying data were chosen so that the techniques demonstrated in the "in class" section would be useful, but not necessarily required and not necessarily sufficient. Altogether, our lab manual consists of a set of lab exercises that currently use 18 of the 26 data sets in Tables 1 and 2. Of note, 41.6% (5/12) of the local data sets and 85.7% (12/14) of the non-local data sets are currently in active use. The lab manuals are often substantially altered by individual faculty and hence continue to evolve.

CHOOSING REAL SOFTWARE

Real data require real software. Teachers of statistics have a wide variety of software from which to choose. *Fathom*, *DataDesk*, *Tinkerplots* are examples of software that are tied to particular markets of statistics learners (*Tinkerplots*, for example, targets K-8 students) and make a strong and often successful use of graphics to allow students to immediately get their hands on data with little need to spend time learning the idiosyncrasies of the software. On the other end of the user-friendly spectrum are professional packages such as *Stata*, *R*, *Splus*, and *SAS*, with *Minitab*, *SPSS* and *JMP* occupying some middle ground. These professional packages often require the user to develop a large vocabulary of commands and in return offer great flexibility and analytic power.

We chose *Stata* for our introductory statistics computer lab for a number of reasons. *Stata* is a professional package that can handle complex data forms and allows easy access to data-format information. *Stata* is used by a growing number of businesses and organizations and is used extensively in upper-division coursework by a number of our client departments. Through a combination of command-line interfaces, "do" files and "logs," *Stata* encourages good organizational practices which, in turn, foster scientific replicability (Kohler and Kreuter, 2005; Gentleman and Lang, 2004). *Stata* teaches computer skills that, in our opinion, extend to other packages such as *SAS*. *Stata* is supported on multiple platforms, including Mac, PC, and Unix. Finally, while more expensive than the freely available *R*, it is substantially less expensive than *SAS* and has an inexpensive student version that is only slightly hampered.

EXAMPLES

We provide two examples. The first focuses on data handling and comes from a non-local data set (Berresford, 1980). The second focuses on understanding a research question, and comes from a local data set contributed by Palmer.

Birthday Lab: Dates are a common data type that can provide challenges to students. Most statistical software packages have particular data formats for handling dates, and learning to manipulate data within this construct is important for interrogating the data. This lab asks questions that are easy to answer with numerical data (e.g., How does the distribution of births vary from month to month? From day to day?) but require recoding for date-format data. For the

"take home" question, students are asked to examine a famous data set from the U.S. war in Vietnam to see if there is evidence that the draft lottery was unfair. These data use a slightly different format for dates, and so students need to further explore data structuring tools to coerce the data into a useful form.

Risk Perception Lab: In this lab, students are exposed to research on risk perception. The data came from a study in which 611 participants completed a survey in which they were asked to provide a numerical value of risk on 22 financial and health-related activities using a scale from 0-100 (100 being high risk). Examples of questionnaire items include "How risky is it to invest 80% of savings in a new medical research firm?" and "How risky is it to fly on commercial airplanes every month?" Participants were also asked questions to identify their age, race/ethnicity, gender, and worldview (hierarchist, individualist, egalitarian). From this data set, it is possible to compare perceived risk between/among groups. The statistical tests and analyses that are appropriate to address questions about group comparisons include *t*-tests of perceived risk between males and females; analysis of variance of perceived risk as a function of ethnicity or worldview; and chi-square analysis of ethnicity and worldview, or gender and worldview. A complicating aspect of these data is that they are stored longitudinally: for each observation there is only one activity for which there is a risk judgment. Therefore, for 611 participants, there are a total of 13,442 observations (611×22).

DISCUSSION

Are locally collected, and hence, more "real," data sets valuable? We found the process of collecting local data sets to be much more expensive in terms of time and effort than we had planned, and the final result was not exactly what we had imagined. Quite often the data sets we received were over-processed or from another source, or already widely available and used in teaching. Sometimes we got more than we bargained for, and the data's primary research question required techniques beyond the introductory level (for example the cardiac data set requires logistic regression and survival analysis), and so it was difficult to find "teaching moments" within these data. These difficulties likely explain the lower rate of current use of local data sets compared to the non-local data sets which tend to be less complex and hence less "real." Given these difficulties, it seems reasonable to ask whether the use of *very* real data, as embodied in our local data sets is substantially better at instilling statistical thinking than less real data as embodied by our non-local data sets.

We are unable to answer that question; however, in general, we feel that the bulk of the data sets and lab exercises are valuable. An examination of students' attitudes, performed very early in our data-collecting experience by an external evaluator required by the NSF grant used to fund the purchase of the computers, found that generally students felt the labs helped them to analyze data graphically and understand graphics and improved their confidence at solving statistical problems. Perhaps more surprisingly, students tended to "enjoy" the computer use and found it much less frightening than expected (Kreft, 2002). Anecdotally, our faculty feel that the students often are more likely to engage with the labs at a higher level than with the homework. The faculty also perceive that the labs help the students to develop statistical thinking skills as witnessed by the students' ability to complete projects involving computer-based statistical analysis and interpretation of non-lab data sets, and good performance on exams with computer-generated statistical output or "lab type" open-ended questions.

As of this writing, we have begun storing our data sets and labs in a searchable database. We believe this will encourage faculty to contribute more data and, in the spirit of "open source" development, improve our current labs. The database tracks usage of labs and enables instructors to rate the labs in terms of the effectiveness of the lab with the course so that frequently used labs that receive favorable evaluations can "rise to the top" and, conversely, less used, unsuccessful labs may sink from sight.

ACKNOWLEDGEMENTS

Development and evaluation of data sets and statistics lab manuals was supported, in part, through funding from the University of California, Los Angeles Office of Instructional Development (IIP#99-51, IIP#00-50, IIP#00-52, IIP#00-50-02) and the United States National Science Foundation (#DUE-9981172).

REFERENCES

Ben-Zvi, D. and Garfield, J. (2004). Statistical literacy, reasoning, and thinking: Goals, definitions, and challenges. In D. Ben-Zvi and J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning and Thinking*, Dordrecht: Kluwer Academic Publishers.

Berresford, G. (1980). The uniformity assumption in the birthday problem. *Mathematical Magazine*, 53(5), 286-288.

Cobb, G. (1992). Teaching statistics: More data, less lecturing. *AMSTAT News*, No. 182.

GAISE College Report. (2005). Guidelines for assessment and instruction in statistics education. The American Statistical Association, http://www.amstat.org/education/gaise/.

Gentleman, R. and Temple Lang, D. (2004). Statistical analysis and reproducible research. accepted for *Journal of Computational Graphics and Statistics*, 2005. http://www.bepress.com/bioconductor/paper2.

Kohler, U. and Kreuter, F. (2005). *Data Analysis Using Stata*. College Station: Stata Press.

Kreft, I. (2002). Students' attitudes towards statistics and computer labs. Unpublished report to the UCLA Department of Statistics and the National Science Foundation (NSF award #DUE-9981172).

Singer, J. D. and Willett, J. B. (1990). Improving the teaching of applied statistics: Putting the data back into data analysis. *The American Statistician*, 44(3), 223-230.

Wells, H. G. (1903). *Mankind in the Making*. London: Chapman and Hall.