Modern introductory statistics using simulation and data analysis

K. Larry Weldon
Department of Statistics and Actuarial Science
Simon Fraser University
Vancouver, Canada. V5A 1S6
weldon@sfu.ca

1. Introduction

Many students are required to take one course in statistics, and many of those choose to take only one. This one course is often a "service" course, aimed at satisfying the perceived needs of various majors like social science or bioscience. The tradition in these courses is to include a fairly heavy dose of inference: confidence intervals, significance testing, p-values etc. After many years by many creative instructors of trying to make this an interesting and useful course, some degree of failure must be admitted. Perhaps it is time for a radical revision of the first course so that more students will take a second course. In this paper I will describe a way in which even the very first course can be made interesting and useful for students. Then I will suggest some further data analytic techniques that could be included in a second course. Both courses emphasize data analysis rather than the usual introductory inference procedures. The combination of simulation, resampling and graphical methods provide tools which allow instructors to describe variability and probability tools without involving mathematical development. While this approach will delay instruction of the more traditional inference material, it may be more useful than the traditional material for students who only take one or two statistics courses.

Introductory service courses are generally crammed with techniques that user departments, like bioscience or sociology, feel their students must know. This requirement is likely based on the need of students to understand the use of statistics in application-area research. The tradition of journal editors is to require a fairly rigid adherence to traditional inference methods. But inference on means and proportions for large and small samples and regression inference include a bewildering array of calculation rituals accompanied by confusing jargon and difficult logic. Few students really master these techniques well enough to be able to use them on their own. Moreover, few one-course students will appreciate when these techniques are really appropriate, and so the opportunity for their making use of statistical inference would often be lost.

University politics may require that the traditional service course, with its emphasis on methods, regardless of understanding, may have to continue for some time. A possible long-term remedy to this is to provide a statistics-appreciation course that does not focus on conveying a large catalogue of techniques for inference on means and proportions, but rather illustrates some simple aspects of a wide variety of useful statistical strategies. They aim to describe what statistics is about, without necessarily arming the student with tools they can use to analyze data by themselves. Once this "service" aspect of the first course is abandoned, a wide-ranging variety of interesting examples can be presented, leaving the student with a realistic and stimulating portrayal of our discipline. For the students that still do not want a second statistics course, at least they will have some useful ideas and a view of the discipline more favorable than with the traditional course. And for the students that do choose a second course, we have a positive attitude that seems lacking in some of our more traditional service

courses. A further development in the direction of data analysis and away from traditional inference is to provide more advanced data analysis in the second course. Students who have already expressed an interest in a first course such as has been proposed may well like a continuation of the data-analytic techniques. In section 3, I suggest some data analysis techniques that can be taught with almost no prerequisites, but with the help of computer software, of course.

Traditions change very slowly in statistics education. Moore's Concepts and Controversies text originally published almost thirty years ago, and now in its fifth edition Moore(2005), responded to a need for an alternative to the traditional technique-oriented text. And Cleveland(1993) started a revolution in display of multivariate data. It is my opinion that both movements are still in their infancy in 2005.

2. Experience with a first course

A new course has been developed at Simon Fraser University with the aim of convincing students that statistics as a discipline is interesting and useful. The technique used has been to present techniques embodied in case-studies of intrinsic interest to a wide variety of students. The philosophy behind a case-oriented course has been clearly described by Cobb (1993): "In teaching, one wants data sets to illustrate the methods, of course, but ultimately the correct emphasis should be that a set of methods is used to illuminate each data set, not that the data sets are there to serve the methods. An effective way to instill this attitude is to organize the course as a series of applied problems."

This new course "Chance and Data Analysis" was first offered in 2002. The idea of the simulations was to show that certain interesting and useful real-world applications of statistics could be illustrated using simulations no more complicated than the equivalent of a fair coin toss, or picking a digit from a uniform distribution on the integers 0,1,…9. In addition, some data-based expositions used straightforward descriptive techniques to extract information expected to interest students. To give the flavor of the course, I describe a few examples of topics covered.

Simulation Demonstrations:

Sports Leagues: The idea here was that, in a competitive league, the assumption that every game was 50-50 would produce league scores surprisingly close to those seen in some actual leagues. The implications for gambling or sports commentary were explored. Students could actually produce the phenomenon using a small scale league and a fair coin. The idea of using simulation to study a complex phenomenon is incidentally introduced. The result in this case (an English soccer league) was that the observed variability in the team points (from 20 to 60) was only slightly greater than would have often been observed in a league of teams of equal winning ability (50-50 for every game). This counter-intuitive result suggested a practical betting strategy.

Portfolio Diversification: A "risky company" is simulated using two fair coins, producing each of the following alternatives 25 percent of the time, as the return to a $1 investment after 1 year. $0, $0.50, $1, $4. The simulation shows what would happen to a portfolio of independent companies like this. The illustration shows the power of diversification of investment as well as the meaning of independence and dependence. The many students opting for an undergraduate degree in business or commerce seem to have a thirst for information about investment, even though few such students would likely be active investors as students. The next example builds on this interest.

Stock Markets: The symmetric random walk with step sizes 1 demonstrates the illusion of patterns that apparently trend up or down, even though these trends offer no advantage for prediction. The artificial aspect of the ±1 step sizes can be removed by using, for example, a normal step size (positive or negative); most introductory students would be familiar with the normal model. The resulting path over 250 days could be compared with the actual stock market index over a calendar year – the nature of the graphs will appear very similar. This demonstration not only teaches students about the illusions caused by randomness but also some very practical information about interpretation of time series.

Data Collection/Manipulation Demonstrations:

Driving Risk: Each student is asked to provide the date of their first driver's license and whether on not they have been involved in an accident, anonymously, on a scrap of paper. These scraps can be summarized, making certain homogeneity assumptions, to estimate the chance that a student that has not had an accident will have one in the next month. (The calculation was based on first principles – e.g. the proportion of students with exposure 6-12 months who had been involved in an accident was easily computed from the raw data.) The chance of an accident in the next month for a student so far accident-free turned out to be about 1 percent. The ability to get this kind of information from such innocuous and minimal data was a surprise to some. The idea of survival analysis as an area of statistical expertise is simultaneously conveyed.

Marijuana Use: The randomized response technique for obtaining an unbiased answer to a sensitive question can be effectively demonstrated. One asks if the respondent uses marijuana at least once a week. Again, the toss of a fair coin determines whether the respondent answers the sensitive question, or the question whose response is known probabilistically. The logic of correcting for the answers to the insensitive question is quite transparent, requiring no formula of any kind. The estimate in one class was 20% using marijuana once per week or more.

The handouts for the first offering of the course are available at www.stat.sfu.ca/~weldon.

Some Findings and Discussion of the First Course

Examples like these are designed to accomplish the following goals: They show that statistical tools can extract information from data that is not obtainable using common sense. They show that the information that statistical tools expose can be interesting and important even for the lay population. They show that at least some of the useful concepts of statistics are understandable without a heavy mathematical preparation. They suggest that statistics is a discipline that can be useful for almost anyone, like reading, writing and arithmetic, and also that an understanding of variability broadens one's world view.

Some feedback data was obtained from students to determine what they understood, what they found confusing, and what they wanted more information about. Although the case studies were presented without emphasizing the statistical tools, nevertheless students learned which tools were important. For example, without prompting, 36 percent chose random sampling and the square root law as "most important concepts" during a mid-course survey. At another stage, after the discussion of sports leagues and portfolio diversification, 32 percent realized that simulation was the common tool that enabled the study of such phenomena without advanced statistical training – other students chose other important ideas, like seasonal adjustment and time series smoothing, as "most important". The small

point here is that the general tools were not overlooked because of the focus on interesting applications.

Student acceptance of the course was also fairly high. In anonymous student evaluations, submitted by about 70% of the students, only 18% said the course content was "not valuable". 66% of the students rated the course as a whole in the top 2 categories of a 5 point likert scale, and 92% in the top 3 categories. It was not the easy grades that put the students in an appreciative frame of mind – averages on the midterms and final exam were lower than traditional norms. These tests were designed to test understanding rather than calculation rituals, and the grades were as usual for this kind of test. Although this was the first offering for a course of this particular style and content, the content was accepted as valuable and interesting by most students.

3. Suggestions for the Second Course

Of course, a second course could be initiated to build on the data analysis approach of the first course. This has not been done at my university yet but in the following I suggest some outline of content that such a course might include. Details are available in a full paper available by e-mail (weldon@sfu.ca) In that full version, I argue that simple kinds of kernel estimation, nonparametric smoothing, multivariate data displays, resampling and the bootstrap, can be taught to students with only a background such as the first course just described.

4. Conclusion

I have tried to make the case for a much greater emphasis on motivational examples in the first course, with a minimum of formal inference, and for a follow-up course which emphasizes useful strategies of data analysis for the second course. The extent to which these suggestions replace the traditional inference is an issue that needs debate in local contexts. The examples here, and in the longer version of this paper, are intended as a resource for that debate.

REFERENCES

Cleveland, W.S. (1993) *Visualizing Data*. Hobart Press, Summit, NJ.

Cobb, G.W.(1993) Reconsidering Statistics Education: A National Science Foundation Conference. Journal of Statistical Education v.1, n.1.

Moore, David S. (2005) *Statistics: Concepts and Controversies*. 5th Edition. W.H. Freeman. New York.