TEACHING STATISTICS THROUGH DATA ANALYSIS

Thomas Piazza
Survey Research Center and Department of Sociology
University of California
Berkeley, California

The question of how best to teach statistics is sometimes phrased in terms of learning by theory versus learning by doing. Unfortunately each approach, by itself, has serious drawbacks. We all know that statistical theory and proofs tend to be forgotten soon unless put to some use. On the other hand, to plunge a student into data analysis does not guarantee that he or she will understand the rationale for, and the limits of, the procedures that can now be so easily carried out by using statistical software packages.

The real problem, as I see it, is how to guide the use of data analysis as a learning tool so that it leads students to an understanding of the statistical principles on which the various analytic techniques are based. In what follows I will summarize the pros and cons of three approaches to the use of data analysis in teaching statistics. Then I will try to draw a few conclusions about the matter. Note that the students we will be concerned with here are those who are at a beginning level and who are interested primarily in application. At a more advanced level, especially for those wishing to become professional statisticians, the balance can shift more in the direction of theoretical exposition.

1. Large Batch-Oriented Statistical Packages

The most widely diffused data analysis tools are the large statistical packages such as SAS, BMDP, SPSS-X, and other sets of programs that offer a wide variety of data management and analysis procedures. These packages can be used to carry out analyses in conjunction with a statistics course A data file can be set up for students to use, and they can carry out assignments or work on projects using an available statistical package. A partial step in this direction is the textbook by Mosteller, Fienberg, and Rourke (1988), which supplements the discussion of multiple regression by displaying and interpreting computer output; a few problems are even provided specifically for students with access to a computer with a multiple regression program.

The use of large packages for instruction has both advantages and disadvantages. The principal advantage of using statistical packages is their comprehensive nature. Although the multiplicity of available procedures can cause confusion at first, students profit by being exposed to the variety of possible methods that can be applied to the analysis of data. They also learn a great deal in the process of figuring out the meaning of the computer output. Furthermore, since students will most likely be using statistical packages in their work later on, the use of these programs can make the courses seem more realistic and practical.

It is also worth noting that these packages are readily available in most universities. If adequate computer time can be obtained for instruction, a set of exercises can be developed and quickly incorporated into a course.

A disadvantage of using statistical packages for instruction is the false sense of precision that students can acquire. The computer programs rarely complain when basic assumptions are being violated or when the form of the data is inappropriate for the procedure used. The computer printout can look good but still be nonsense.

Another problem is that students can get bogged down with the mechanics of producing the runs. The correct specification of control language and options can be so distracting that students implicitly come to think that the successful production of computer printout is the same as the solution to a statistical problem. It is important for students to remain aware that the purpose of the course is to teach statistical principles and sound analytic technique – not just to develop facility with a particular statistical package.

A related problem is that turn-around time can still be slow at many institutions. Ideally the computer should give results immediately for the problems we submit. A delay of a few hours may be tolerable; however, if the results are not available for days, it is easy for students to lose track of what the whole point of the analysis was. Once again, it is important that students not get bogged down with mechanical details.

On balance, consequently, the use of large statistical packages for teaching statistics can be helpful, but they are certainly not the ideal solution. It is worth considering other approaches as well.

2. Interactive Analysis Programs

Computer programs for minicomputers and microcomputers have begun to appear which are designed for interactive data analysis. These programs are somewhat more promising for teaching statistics than are the large batch-oriented packages. Examples are the S package from Bell Laboratories, and ISP from Berkeley; the microcomputer versions of SPSS and SAS can also be considered semi-interactive.

The distinguishing characteristic of interactive data analysis programs is that the results of one stage of analysis can immediately be used to specify the next stage. At the simplest level this means that the results can be examined on the screen of a computer terminal and used as the basis for an additional analysis specification. At a more sophisticated level this means that intermediate results are available for examination and can guide further analytic steps without requiring the user to issue a whole new set of commands.

The main advantage of interactive programs is this immediate feedback. This approach encourages exploration of the data, and it facilitates the development of good problem-solving strategies. Some of the disadvantages

of batch-oriented programs can thus be overcome. The text by McNeil (1977) illustrates some of these possibilities.

On the other hand, interactive statistical programs still have some of the same problems as batch programs for purposes of teaching statistics. There are still the problems of having to spend a lot of time learning to use the programs and of getting distracted by the mechanics of producing results. Furthermore, interactive programs produce results in the same uncritical manner as any other statistical computer program, and those results need not make much sense.

All in all, however, an interactive statistical package is likely to be more appropriate for teaching statistics than a batch-oriented package. If such a set of programs is available for student use, it is well worth investigating.

3. Simulations Based on Spreadsheets

There is one other computer application that I wish to mention in this context, and that is the use of spreadsheet programs that are widely available for microcomputers. These programs set up a large matrix in which some cells can be defined as a function of other cells. When the latter cells are modified, then the cells defined as functions also change. These functions can be as simple as the sum of two cells or as complex as a chi-square statistic based on a contingency table.

The instructional value of these spreadsheet programs is due to their ability to make clear and transparent how a set of observations generates a particular statistic. A student can see immediately how a change in observed values leads to a change in the statistic (Kelman et al., 1983). If set up appropriately, a spreadsheet application can allow the student to simulate easily the effects of different distributions or patterns and to examine the behavior of a statistic under many conditions.

Let me give a relatively simple example of a spreadsheet that I have used in my own courses. In a course on log-linear analytic methods I wanted the students to develop a good intuitive sense of the behavior of the odds-ratio in a 2x2 contingency table. So I set up a spreadsheet with a contingency table and the computed odds-ratio (plus Yule's Q and the logarithm of the odds-ratio). The students could change the cell counts of the contingency table and see immediately how those changes affected the odds-ratio (and the other associated statistics). In a very short period of time a student could explore quite thoroughly the behavior of this statistic under various conditions.

A more interesting and complex example from the same course was a spreadsheet set up to generate the model of independence and the associated chi-square statistic in a bivariate contingency table. A starting table was also included so that some of the cells could be defined as structural zeros. The student could manipulate both the cell frequencies and the starting table in order to see how well the models of independence and

quasi-independence fit the cell frequencies. I found this spreadsheet application to be particularly helpful to the students.

These spreadsheet applications can focus the students' attention on certain important principles or on the behavior of a specific statistic. Once set up, these spreadsheets are very easy for students to use, and they can provide real insights into the subject matter of the course.

The limitations of these spreadsheets, of course, are very clear. Not every calculation can be set up conveniently in such a format, and the functions available (logs, square roots, trigonometric functions, and so forth) vary from one program to the other. Another problem is that the data set for calculations must be relatively small, if students are to modify it easily and intelligibly.

All in all, however, I believe that this approach to using computers holds great promise for teaching statistical principles. At the present time it is a little cumbersome for the professor to have to set up spreadsheets for a course, but I suspect that in due time some appropriately designed software will facilitate the task or provide some equivalent methodology.

4. Conclusion

The basic question we need to address is how people learn – more specifically, how they learn statistical concepts and techniques. Most of us have probably come to the conclusion that it is not enough to teach statistical theorems and derivations, with a few exercises thrown in. We need more emphasis on the analysis of data, on putting statistics to work.

Nevertheless, it is not always easy to develop an approach that integrates theory, data analysis, and the requisite computer skills. As discussed above, statistical packages and programs can be helpful, but they are not a substitute for understanding statistical theory. As we also noted, students can spend so much time getting the computer to do something that they lose sight of the statistical principles that are supposedly being exercised.

One use of computers that I consider especially promising is the use of spreadsheet programs, which are available for practically any microcomputer. Once set up appropriately, these programs allow students to change a set of data and to see immediately the impact of those changes on a statistic. This method can help students to understand the meaning of a certain statistic or statistical principle and to acquire an intuitive grasp of how robust a statistic is under different conditions. This approach, although relatively undeveloped, is a good example of how computers and data analysis can further a theoretical understanding of statistics.

References

Kelman, P. et al. (1983). Computers in teaching mathematics. Reading, Mass.: Addison-Wesley.

McNeil, D.R. (1977). Interactive data analysis: A practical primer. New York: John Wiley.

Mosteller, F., Fienberg, S.E., & Rourke, R.E.K. (1983). Beginning statistics with data analysis. Reading, Mass.: Addison-Wesley.