A STUDY OF THE INTERACTION BETWEEN THE USE OF STATISTICAL SOFTWARE AND THE DATA ANALYSIS PROCESS

John D. McKenzie, Jr.
William H. Rybolt, and
David P. Kopcso
Babson College
Wellesley, MA, U.S.A.

Motivation

Those familiar with a statistical package are often surprised at how long it takes a novice to accomplish even a simple task. In a college environment, students often complain about the countless hours spent on a simple assignment. An examination of their computer printouts and written work often fails to reveal where the students spent their hours. Unfortunately, it is possible for students to understand conceptually what to do but still not be able to do it.

In an effort to learn where students actually ran into difficulty and discover more about the types of mistakes they were making, we undertook a series of experiments at Babson College. These experiments involved two instructors each teaching two sections of an applied statistics course. There were 154 undergraduate students in these four sections. An extensive set of references dealing with this area may be found in Kopcso, McKenzie, and Rybolt (1985). References to our past work and another recent article are present at the end of this paper.

At the beginning of the semester, the students were given CA1. For this project, they were asked to describe, with the help of the 82.1 release of Minitab, the data present in a table giving the State and Local Per Capita Tax Burdens in Fiscal 1982-1983 for the fifty states. At the end of the semester, they were asked to obtain a prediction interval for one variable given a specific value of another variable in CA5. There were six different types of data sets. Within each data set students were given randomly generated data. Again, they were expected to use Minitab to carry out the analysis. Of course, during the semester they were assigned several other assignments using Minitab.

What made CA1 and CA5 different from the other assignments was our ability to capture an electronic record of the students interaction with Minitab. Using a VAX-command procedure, the students entered Minitab with a journal-like feature enabled. This procedure created a file containing all the commands and data, which the students typed into Minitab. These files also contained Minitab error responses, and a limited number of other responses. These files do not contain the results of a requested statistical analysis. The students turned in hard copies of their computer runs so that the journal files could be compared with the hard copies. None of this information was used to grade the students' performances.

By comparing the submitted runs with the actual runs, we expected to discover where students were actually spending their time. Perhaps the real difficulties were deemed so trivial or embarrassing that they were not reported to the instructor. This knowledge could increase teaching effectiveness by focusing attention on the actual difficulties. It would also permit designers of statistical packages to improve those areas causing the greatest difficulty.

By examining the pattern of student usage for CA1 with that for CA5, we hope to discover how experience and learning influences a students ability to use a typical statistical package. If the problems of novices are very different from seasoned users, then an appropriate choice of learning material would aid both groups.

Results

The results of CA1 have been discussed elsewhere. Here we focus our attention on CA5 and some of the differences between CA1 and CA5. During CA5, 142 students ran Minitab 483 times for an average of 3.4 runs per student. One student had 19 different runs. The majority of the students, 77%, used four runs or less to complete the assignment. About 5% of the students used nine or more runs.

For CA1, 142 students ran Minitab 572 times for an average of 3.8 runs per student. The maximum number of runs was 12 by two different students. The majority, 67%, used four runs or less to complete the assignment. About 6% of the students used nine or more runs. Because of a few drops and adds, the students at the beginning and end of the semester differed by a few individuals. A total of 154 students participated in either CAI or CA5 while 138 students participated in both experiments. The median number of runs in both experiments was 3.

Was the reduction of 0.4 in the average number runs per student statistically significant? The students who made 12 more runs in the second experiment than the first experiment were clearly outliers. If these data points are eliminated, then several paired difference tests yield the same outcome at a 5% level of significance. The decrease is statistically significant.

Altogether students entered about 17,000 lines of information into Minitab. About 66% were command lines while the other 34% were data lines. The data lines are generally entered as sets of lines. There were about 770 sets of data lines indicating an average of just under eight lines per data set.

The distribution of data set sizes is bimodal. About 64% of the data sets contain 1 to 5 lines of data while 31% of the data sets contain 18 to 22 lines of data. In CA5, students were given 20 pairs of values. Thus two distinct styles of data entry emerge. Most students, 44%, preferred to enter the data one line at a time. The most popular data entry command was SET which is used to enter values for a single variable. The second most popular style, 23%, was to enter 20 lines of data at a time. The single most popular command, 12%, was LET. It was used to perform transformations on the data.

HELP represented 1% of the commands, and INFO was used 1.5% of the time. HELP was followed by HELP 37% of the time. The most common command preceding INFO was RETRIEVE 44% of the time and INFO was followed by PRINT 18% of the time.

In CA5, about 6% of the commands were followed by data sets. For CA1, only about 4% of the command lines were followed by data sets. Just over 1,300 command lines contained errors; thus 11% of the command lines in CA5 had errors associated with them. This represented an improvement over CA1 when about 15% of the command lines contained errors.

In the above quick comparison between CA1 and CA5, we find many similar patterns. As one might expect, familiarity with a package seems to reduce both the number of runs needed to complete the analysis and the percent of time an error occurs. The reduction is perhaps not as much as might be expected. One conjecture is that even though the students acquire familiarity with using Minitab, the instructors add new features at such a pace that the challenge on the part of the students remains roughly constant.

An expert running CA5 entered the following sequence of commands: READ, END, NAME, PRINT, SAVE, PLOT, BRIEF, REGRESS, LET, REGRESS, and STOP. Of course, depending upon the exact data set being analyzed, any user will generate slightly different sequences of commands. Still most experts would generate a similar sequence of commands. Twelve commands and 20 data lines were used by this expert. In CA5, 31 students used over 100 command lines, 6 students used over 300 command lines, and one student used over 700 command lines. If we compare the performance of the average student with that of the expert, then we find that students entered as least twice as many data lines as the expert and six times as many commands.

By examining the number of times, a command occurs before or after another command, we can construct a modal run. This modal run shows the most common sequence followed by a typical student. One such modal run is the sequence SET, END, SET, END, PRINT, Y, Y, PLOT, PLOT REGRESS, Y. The reason some commands occur more than once is that they are used in that chronological sequence. For example, SET, END, SET, END means that first the value of one variable was entered into a column, and then the value of the other variable was entered into the second column. The important feature is that students choose roughly the same commands as the expert, but that they use the commands far more often. The students seem to get lost in the details of the analysis. It is encouraging to see that students seem to make use of the interactive aspects of Minitab. For example, a PLOT is followed by another PLOT 23% of the time and by REGRESS 15% of the time.

The more common commands seem to be associated with errors less often than the commands new to this assignment. For example, LET, READ, and SET are associated with errors 5%, 10%, and 5% of the time, respectively. On the other hand, REGRESS was associated with an error 21% of the time. It was encouraging to see the use of file commands used to save and retrieve data sets. It was discouraging that these commands often generated errors. The commands SAVE and RETRIEVE were used incorrectly 20% and

33% of the time, respectively. Clearly students were entering their data sets from the keyboard multiple times. This is evident both from the number of data lines used as well as the inability to use data storage and retrieval commands correctly.

Recommendations

Several suggestions emerge from this preliminary analysis of CA5. Some choices of command names such as N, the number of nonmissing values, are confusing and serve no useful purpose. NONMISS would be a far better choice. Almost all Minitab commands have four letters. Commands such as DELETE, which many students tried to use, should be added to the command set. The treatment of three and four character commands should be uniform. READC1 works correctly, but SETC1 generates an error. It would be far better to require a space between the command arguments.

Many students know what they want to do but get lost in the details of the command syntax. It is time for software packages to make extensive use of screen menus for both data entry and analysis. The modal sequences of commands may be an aid in the placement of parts of statistical programs on floppy disks so as to minimize the need for disk swapping. Some unexpected side effects of Minitab were also discovered. For example, the information following NEWS is placed in the journal file as if the student had typed it from the keyboard. This is different than the treatment of other commands such as INFO and HELP.

Acknowledgments

The authors wish to thank Minitab, Inc., the Babson College Computer Center, and Gordon Prichett for their assistance.

Selected References

Kopcso, David P., McKenzie, John D., Jr., and Rybolt, William H. (1984). Pedagogical Impact of AI on Statistical Software. Paper presented at the ASA-IASC-SIAM Conference on the Frontiers in Computational Statistics, Boston.

Kopcso, David P., McKenzie, John D., Jr., and Rybolt, William H. (1985). An Interactive Approach to Improving Data Analysis in the Classroom. Proceedings of the Computer Science and Statistics: 17th Symposium on the Interface, in press.

Rybolt, William H., Kopcso, David P., and McKenzie, John D., Jr. (1986). A Study of Students' Cognitive Styles When Using a Data Analysis package. Paper to be presented at the 7th Symposium on Computational Statistics, Rome, Italy.

Thisted, Ronald A. (1986). Computing Environments for Data Analysis. Statistical Science, 1, 2, 259-275.