The Stata Journal

The *Stata Journal*, electronic version (ISSN 1536-8734) is a publication of Stata Press, and Stata is a registered trademark of StataCorp LP.

The Stata Journal (2005)
5, Number 2, pp. 248–258

Teaching statistics to physicians using Stata

Susan M. Hailpern
Department of Epidemiology and Population Health
Albert Einstein College of Medicine
Bronx, NY
shailper@aecom.yu.edu

Abstract. The Clinical Research Training Program (CRTP) at the Albert Einstein College of Medicine at Yeshiva University is a two-year program for physicians leading to a Master of Science degree in Clinical Research Methods. Beginning in July 2004, the program began teaching data analysis using Stata 8 in order to better meet the advanced statistical needs of the students. This paper details the structure and content of the course, how Stata was introduced, and the problems we encountered. Student comments and suggestions on future enhancements to Stata are included. Although challenging, our first semester teaching Stata was a success: the students all learned Stata and, more importantly, continued to use it for the analysis of their own research data after the course was complete.

Keywords: gn0027, teaching statistics to physicians, menu-driven interaction style

1 Introduction

The Clinical Research Training Program (CRTP) at the Albert Einstein College of Medicine at Yeshiva University began in 1998. It is a two-year program leading to a Master of Science degree in Clinical Research Methods. Students in the program must have a doctoral degree (most have an M.D.) with a strong interest in clinical research. The CRTP has two complementary components: a didactic curriculum, with emphasis on epidemiology, biostatistics, study design, computer methods, and research ethics; and a mentored clinical research thesis project. The U.S. National Institutes of Health, under a Clinical Research Curriculum Award, funds the program.

From its beginning, the CRTP has taught biostatistics and data analysis through classroom methodology and hands-on computer software training. Beginning in July 2004, the program changed the statistical software taught to Stata 8 to address the limitations of our previous software package, which did not have the ability to analyze many of the study designs or apply many of the methods used by our students, such as matched case–control studies, multinomial and ordinal logistic regression, frailty models for survival analysis, and analysis of complex sampling survey data. Stata was chosen because of its excellent reputation, ease of use, the addition of menu-driven interaction (new to version 8), and the wide range of procedures and options available. It was perceived as an optimal choice: the students were not scientific novices, but professionals in fields other than statistics, and they needed a tool that was both easy to use for standard analyses and powerful enough to perform advanced statistical methods.

Teaching Stata using menu-driven statistical analysis to the CRTP students provided an interesting challenge. The senior instructor had extensive experience using and teaching other statistical software but was a newcomer to Stata. This author had extensive experience using Stata but was unfamiliar with Stata's new menu-driven interaction style.

This article reports on the course structure and content and evaluates its success. The experience was positive, but detailed comments on what the students found difficult or lacking are included.

2 Course structure and content

The first semester of the CRTP includes two parallel six-week intensive courses in statistics and data analysis (table 1). Each course consists of a 3.5-hour class, taught once a week, resulting in two interrelated classes per week. We employed a class structure similar to that used in our medical school to teach epidemiology and biostatistics (Marantz, Burton, and Steiner-Grossman 2003). The statistics class is lecture-style. It discusses statistical theory, statistical tests for categorical variables (χ^2 and Fisher's exact), continuous variables (t tests, ANOVA, and correlation), and nonparametric statistical tests. The data analysis class is a hands-on computer lab, with a syllabus that parallels the statistics course (table 1). The data analysis class teaches database management, data cleaning, and data analysis.

All students were required to have access to Intercooled Stata 8. The required texts were Altman (1991) for the statistical course and Hamilton (2004) for the data analysis course. Weekly homework assignments were planned to reinforce statistical theory using the Stata software. A brief written introduction to the Stata software was distributed to students before the first data analysis class. All students had some computer experience, although most had no experience using statistical software. Students were asked to install Stata before the first class. All were successful: they found it easy to install on both Windows machines and Apple Macs.

2.1 The first session

The first data analysis class was primarily an introduction to the standard set of Stata windows. Students were introduced to the Data Editor, Data Browser, Results window, Review window, Variables window, Command window, and Viewer. Although students were taught to use the menu-driven interaction style, the distinction between menus and syntax was discussed. Variable elements, such as variable name, type, format, variable labels, and value labels, were introduced. Students were taught how to record their Stata session in a log file. We used a problem-based learning framework (Dyke, Jamrozik, and Plant 2001) in which students are given tasks that reflect situations they are likely to face in their future professions as clinical researchers. The dataset used, for the duration of the semester, was a modified subset of 1,000 observations and 40 variables from a U.S. public-use health-related database. To tie together the statistical theory taught during the first week with the data analysis material, students were taught descriptive statistics. Dialog boxes for `summarize` and `describe` were used, and their options were explored.

The first session was wrought with confusion and frustration. We had expected that there would be a steep learning curve this first week, but having been a Stata user for many years, this author underestimated how steep the curve would be. Emotions ran high as students struggled to find the appropriate dialog boxes, attempted to run statistical tests with the Data Editor open, and learned the difference between **OK** and **Submit** in the dialog boxes. The largest problem encountered at this early point was the students' desire to move Stata windows around the computer screen, resulting in some students losing windows behind other windows! If it were not for the **Window** pull-down menu on the tool bar, some of the students might still be searching for their Results window.

2.2 The second session

The second data analysis class continued to focus on becoming familiar with the Stata software. Variable manipulation and subset analysis were discussed in depth. Most, but not all, of the dialog boxes contained in the **Data** pull-down menu on the toolbar were introduced. Students were taught to generate new variables, recode variables, change variables from string to numeric and numeric to string, and change the contents of a variable. Qualifiers were discussed, as was sorting the data and dropping or keeping of variables or observations.

Student errors progressed from simple window manipulation to a fundamental misunderstanding of Stata numeric ranges. Recoding variables was the largest problem encountered in this session. Students did not understand that a missing value was treated by the system as a number. This misunderstanding came to light when they attempted to `recode` a range using the `else` syntax. While students expected a certain number of occurrences to be captured by the `else`, Stata unexpectedly recoded all the missing values into that range as well. The students were mystified as to why the number of observations changed.

Using numeric ranges in recoding caused another problem. If one range was to end at a value x and the next range was to start immediately after x, there was no convenient syntax to express these contiguous, but disjoint, ranges. Many students searched for a simple way to denote the "next value" to begin the second range, but Stata does not provide a syntax to easily denote $x + \epsilon$. Using x as the first end value and $x + 1$ as the next start value would miss any fractional values in between the cutpoints. When it was clear that an integer next value would not work, students then attempted to use a fractional difference. The difference between the precision (number of digits after the decimal point) shown in the Stata Results window and the actual value exacerbated the confusion. Students attempted to use x as one end value and $x + 0.00001$ as the next start value because they assumed that there were no values in between those cutpoints. The precision of the Data Editor window reinforced this misconception. The result was that values between the end of the first range and the beginning of the second range were still missed. The precision of the Data Editor remained a problem until students noted the full value of variables shown one at a time at the top of the window.

While some students were becoming more familiar with the software, others felt overwhelmed and requested additional weekly hands-on help sessions. An additional (optional) weekly lab session was added to the curriculum. It was used mainly for review. However, extra lab time with the students provided a wonderful opportunity to teach some additional features of Stata that were not taught as part of the regular syllabus. Immediate commands, date functions, do-files, `reshape`, and graphing were among the extra topics discussed. The slower pace of the additional lab resolved many of the problems students were having. It gave them the confidence needed to explore the software on their own. Perhaps the most satisfying unexpected outcome of this lab was the opportunity to demonstrate multiple approaches to expressing the same data manipulation (e.g., many ways to recode, many ways to subset, `~=` versus `!=`).

2.3 The third session and beyond

The third class session spent less time on gaining familiarity with the software and more time on data analysis. Combining datasets using `append` or `merge`, collapsing data, and updating the software were taught. Statistical analysis moved from descriptive statistics to comparison of means and medians for one and two samples using one- and two-sample t tests and one-way ANOVA and Kruskal–Wallis tests for the comparison of means and medians for more than two groups. We also taught χ^2, relative risk, and odds ratio for the comparison of categorical data. Student confidence was building by this time, and several students began to do analyses with their own research data.

The final three weeks of the semester were focused primarily on data analysis. Correlation and an introduction to linear regression were taught, followed by a brief introduction to multiple linear regression and two-way ANOVA. In addition, statistical power and sample size were introduced using another statistical software package. By the end of the first semester, all students were successful in learning the fundamentals of the Stata software. All students were able to complete their homework assignments using Stata. Thankfully, all passed a challenging final exam.

(*Continued on next page*)

Table 1: First semester syllabus

Week	**Biostatistics I** Text: *Practical Statistics for Medical Research* by D. G. Altman	**Data Analysis with Stata** Text: *Statistics with Stata (Updated for Version 8)* by L. C. Hamilton
1	Types of data Describing data Preparing data for analysis	Stata structure: windows, menu and syntax, data file, data structure, variable elements Descriptive statistics: `describe`, `tabulate`, `summarize`, `list`

2	Foundations of analysis:
	sampling distributions
	estimation
	hypothesis testing
	comparing groups (continuous data)
		Modifying variables, subsets, and conditions:
		`if`, `in`, `sort`, `drop`, `keep`, `save`
		Creating new variables
		`gen`, `replace`
		Functions
		Recoding
		`recode`, `encode`, `decode`

3	Comparing groups continuous data: nonparametric tests one-way ANOVA Categorical data: 2×2 tables and $2 \times k$ tables ordered categories relative risks and odds ratios	Comparison of means and medians: one and two sample Comparison of means and medians: > 2 groups one-way ANOVA and Kruskal–Wallis Comparing categorical data 2×2 tables and $2 \times k$ tables ordered categories relative risks and odds ratios
4	Relationship between two continuous variables Correlation Regression	File handling: `use`, `clear`, importing files `merge`, `update`, `append`, `collapse`, `xpose`, `reshape` Correlation and linear regression Cohort studies
5	Relationship among several variables introduction to multiple regression two-way ANOVA	Immediate commands and display Multivariate analysis: multiple regression two-way ANOVA Interpretation of outputs Confounding an interaction Case–control and cross-sectional studies
6	Summary/review	Sample size and power analysis Causal inference

3 Evaluation

Although challenging, our first semester teaching Stata was an unqualified success. For the senior instructor, the challenge was keeping ahead of the students in learning the details of the software. For this author, the challenge was to learn the menu-driven interaction style. To facilitate this learning for both instructors, we created a flow chart showing the pull-down menus needed to reach the desired dialog box required for analysis. It categorized data analysis tasks into numerical and categorical data, and further subdivided according to the number of groups to be analyzed (see appendix).

We can honestly say that there were no failures: all students were successful in learning the software and, more importantly, successful in learning to use the software to analyze data. In fact, many of the students embraced the software and basic knowledge of statistics. While menu-driven interactions were taught, a handful of students learned to write Stata syntax. Several students felt so empowered that they began to analyze their own research data, limited only by their rudimentary knowledge of statistics. It was not uncommon for this author to receive emails sent at all hours of the day and night asking for help with Stata commands on advanced analysis topics that had not yet been taught.

Since the change in statistical software was made out of necessity and not part of a study to determine which software was more appropriate for our students, we have very little hard data to document our success. We did, however, see an improvement in our end-of-the-semester course evaluations. Four questions, answered on a 4-point scale (4 = strongly agree, 3 = agree, 2 = disagree, and 1 = strongly disagree), were specifically related to the introduction to statistics and data analysis classes. The questions were

These sessions have

- Enhanced my understanding of biostatistical principles
- Increased my comfort with using the computer
- Facilitated my use of the statistical package
- Increased my understanding of the application of statistical tests and procedures to datasets

Our mean score increased on three of the four questions. Scores for the year before teaching Stata were 3.72, 3.56, 3.61, and 3.67, respectively. Scores while teaching Stata were 3.92, 3.54, 3.77, and 3.85, respectively. Of particular note is the increase in understanding of the application of statistical tests. The slight decline in our second score is a result of one student who disagreed that the sessions increased his comfort with using the computer. The student justified the response by stating that he was very comfortable with the use of computers before this course.

Students were also encouraged to write comments on the evaluation forms. In the year before teaching with Stata, we had no positive comments specific to statistics and data analysis. In contrast, there were a large number of comments when we taught with Stata. Typical comments included “overall excellent introduction to biostatistics”, “my knowledge has increased dramatically”, and “I no longer fear reading the ‘methods’ section of a paper—in fact, I now enjoy it”. It would be unscientific to say that improvement in our course evaluations was a direct result of our switch to Stata. We can, however, say that our teaching faculty and textbooks did not change in the years under consideration. In other words, the only change was the statistical software.

Our success in teaching Stata went beyond the students in this first class. Second-year students, who had been taught data analysis using another statistical package, heard of Stata’s capabilities and ease of use. They strongly requested a hands-on computer lab of their own so that they could learn to use Stata as well. Many of the students had determined that the analysis of data for their Master’s theses could not be handled in the package they had originally learned. Stata, however, included the functionality they needed, which enabled them to complete their research.

4 Student comments and suggestions

The request heard most often was the desire for an “undo” or “back” button! (This request was usually preceded by an expletive.) This gave the instructors the opportunity to remind students that the Stata software is fast and powerful: students needed to know what they were doing at all times. The menu-driven interaction style diminished the need to teach the students about do-files. The lack of an ongoing do-file made the missing undo operation a bigger problem than it need be.

Dialog boxes were a mixed blessing. On the one hand, they made it quick and easy for students to learn Stata and perform statistical tests. On the other hand, the students had a number of complaints. One of the most frequently heard complaints was the inability to find the right dialog box to perform the test desired. The `db` (dialog box) command was useful only if one knew the precise name of the procedure, which cannot be assumed for those who rely on menu-driven interaction. Students felt that menus were not intuitively placed on the tool bar. While they sought a change in organization, no consensus could be reached as to the best organizational form. Several students suggested that dialog boxes contain a brief description of the statistical test in a mouse-over “tool tip”, so they can be assured that they have chosen the correct test. Students also felt that Stata manuals were of little help for those users relying on the menu-driven interaction style. Stata manuals were written to assist those using syntax-driven commands.

5 Conclusion

In summary, Stata worked and worked well for the CRTP program. It was easy to learn and was powerful enough to handle complex data. Most importantly, the students chose to use Stata when they needed to analyze their own research data.

6 Appendix: Stata flow chart

6.1 Numerical data

- 1 group
  - One-sample *t* test
    - Statistics → Summaries, tables, and tests → Classic tests of hypothesis → One-sample mean comparison test
- 2 groups
  - Paired
    - Paired *t* test
      - Statistics → Summaries, tables, and tests → Classic tests of hypothesis → Two-sample mean comparison test (assumes data are paired unless checked as unpaired)
    - Wilcoxon sign-rank test
      - Statistics → Summaries, tables, and tests → Nonparametric test of hypotheses → Wilcoxon matched-pairs sign-rank test
      - Statistics → Summaries, tables, and tests → Nonparametric test of hypotheses → Test equality of matched pairs
  - Independent
    - Unpaired *t* test
      - Statistics → Summaries, tables, and tests → Classic tests of hypothesis → Two-sample mean comparison test (checked as unpaired)
    - Wilcoxon rank-sum test
      - Statistics → Summaries, tables, and tests → Nonparametric test of hypotheses → Mann–Whitney two-sample rank-sum test
- $>$ 2 groups
  - One-way ANOVA
    - Statistics → ANOVA/MANOVA → Analysis of variance and covariance
  - Kruskal–Wallis test
    - Statistics → Summaries, tables, and tests → Nonparametric test of hypothesis → Kruskal–Wallis rank test

6.2 Categorical data

- 2 categories investigating proportions
  - 1 group
    - z test for proportion
      - Statistics → Summaries, tables, and tests → Classic tests of hypothesis → One-sample proportion test
  - 2 groups
    - Paired
      - McNemar's test
        - Statistics → Observation/Epi. analysis → Tables for epidemiologists → Matched case–control studies
    - Independent
      - Chi-squared test
        - Statistics → Summaries, tables, and tests → Tables → Two-way tables with measures of association (check off Pearson's chi-squared test)
      - Fisher's exact test
        - Statistics → Summaries, tables, and tests → Tables → Two-way tables with measures of association (check off Fisher's exact test)
- >2 categories
  - Chi-squared test
    - Statistics → Summaries, tables, and tests → Tables → Two-way tables with measure of association (check off Pearson's chi-squared test)

6.3 Further analysis

- Regression
  - Correlation
    - Correlation coefficients
      - Pearson's
      - Spearman's
        - Statistics → Summaries, tables, and tests → Summary statistics → Correlations and covariances
        - Summaries, tables, and tests → Summary statistics → Pairwise correlations (check box for significance level for each entry)
        - Statistics → Summaries, tables, and tests → Nonparametric test of hypotheses → Spearman's rank correlation
        - Statistics → Summaries, tables, and tests → Nonparametric test of hypotheses → Kendall's rank correlation
  - Regression
    - Simple/Multiple
      - Statistics → Linear regression and related → Linear regression
      - Statistics → General postestimation → Obtain predictions, residuals, etc., after estimation
    - Logistic regression
      - Statistics → Binary outcomes → Logistic regression → (or logistic regression (reporting OR))
      - Statistics → General postestimation → Obtain predictions, residuals, etc., after estimation
      - Statistics → Binary postestimation
    - Cox regression
      - Statistics → Survival → Setup and utilities → Declare data to be survival time (you must set your data before you do any other procedure)
- Statistics → Survival → Summaries, tables, and tests

  Statistics → Survival → Regression models

7 Acknowledgments

The author would like to thank Drs. Paul Marantz and Hillel Cohen for giving me the opportunity to teach Stata to the CRTP students and the students themselves for allowing me to share my love for Stata.

8 References

Altman, D. G. 1991. *Practical Statistics for Medical Research*. London: Chapman & Hall.

Dyke, P., K. Jamrozik, and A. J. Plant. 2001. A randomized trial of a problem-based learning approach for teaching epidemiology. *Academic Medicine* 76: 373–379.

Hamilton, L. C. 2004. *Statistics with Stata (Updated for Version 8)*. Belmont, CA: Brooks/Cole.

Marantz, P. R., W. Burton, and P. Steiner-Grossman. 2003. Using the case-discussion method to teach epidemiology and biostatistics. *Academic Medicine* 78: 365–371.

About the Author

Susan M. Hailpern is a faculty associate in the Department of Epidemiology and Population Health at Albert Einstein College of Medicine. Her research interests are in the areas of heart and kidney disease, and smoking uptake among young adults. She is pursuing a doctorate in public health in the field of epidemiology at New York Medical College.