USING SAS® TO TEACH A UNIVERSITY COURSE IN STATISTICAL COMPUTING AND DATABASE MANAGEMENT

John L. Wasik, North Carolina State University

Background

One of the capabilities that employers have come to expect of recent university graduates is substantial experience with, if not competence in, working with computers. While many employers are focusing on their employees' skills in working with individual microcomputers (PC), university graduates seeking jobs in scientific fields may also be expected to have additional expertise in working with "mini", "mainframe", and/or "workstation" computers.

The Department of Statistics at the North Carolina State University (NCSU) has been providing statistical education at the undergraduate level for approximately forty-five years. A Bachelor of Science degree in Statistics has been offered during much of this period, but it has only been within the last ten years that the program has attracted substantial numbers of undergraduate majors. At the present time, a total of 98 undergraduates are majoring in statistics at North Carolina State University.

Development of a Course in Statistical Computing

A survey of employers of North Carolina State University B. S. in Statistics degree holders conducted five years ago revealed that our graduates were often hired as scientific programmers. However, their assessment was that these our graduates were not "up-to-speed" in their ability to use SAS to perform database management and/or data analysis tasks. Since the survey was conducted to provide supporting information for a five year review of programs of the Department of Statistics, the findings were utilized to argue for a major revision of the B. S. in Statistics curriculum. A total of four new courses to be required of statistics majors were developed and subsequently approved by the university. One of the new courses, ST445, Introduction to Statistical Computing and Database Management, was specifically designed to meet the expressed concerns of employers about student level of expertise in the use of SAS. The present undergraduate major in statistic curriculum at North Carolina State University requires minimums of 33 hours in statistics including the ST445 course, two courses in computer science, and 18 hours in mathematics.

Course Philosophy

The philosophy underlying the development of ST445 was based upon the observation that scientific programmers typically serve on a project team to complete a set of tasks. Development of the ST445 course also coincided with a university initiative to introduce more writing opportunities in courses offered by scientific curricula such as statistics. Thus, ST445 was also given the designation of the writing project course which meant that students had to write a major report to demonstrate their ability to communicate findings of research in a written format. The course introduces the student to various tasks undertaken by a project team during the processing and analysis of data. The goal of ST445 was to ensure that students upon completion of the course would be able: (1) define variables and provide codes for the variable values; (2) construct one or more databases containing study variables; (3) input data into constructed database(s); (4) write programs to handle one or more data sets written in raw data and/or stored SAS formats; (5) identify and correct program errors; (6) perform statistical analyses; (7) construct graphical presentations; and (8) prepare written reports using graphical output and results from statistical analyses.

It is noteworthy that not all of the above skills are typically associated with SAS programming. Indeed, it should be noted that the SAS system serves as a background for helping students develop an overview of the different types of tasks that need to be accomplished in completing a project. Essentially, the instruction in SAS usage could be viewed as developing skill in using another research tool.
Previous Experiences in Teaching with SAS

The Department of Computer Science at North Carolina State University had previously presented a course in statistical packages with a focus on the SAS system and SPSS statistical computer packages to perform statistical calculations. This course was discontinued after the Department of Computer Science decided their role was to teach computer programming languages and principles and not usage of applications computer programs.

Statistics undergraduate majors taking third and fourth year statistics methods courses typically are introduced to rudimentary SAS programming principles. These courses use SAS in either a "mainframe" or a networked workstation laboratory setting. SAS is used in these course to work homework problems so as to provide practice of realistic problem solving in statistical analysis situations. Such usage of the SAS system for supporting statistics methods courses has been reported by several individuals responsible for teaching such courses. For example, some of the earliest reports of the use of SAS as an instructional tool have been by Thisted (1979) in the American Statistician, Bobulz and Gentile (1979) at the SUGI4 conference, and Hobbs (1982) at the 1980 Ohio State University Conference on the Teaching of Statistics and Statistical Consulting. Each of these papers advocated the use of SAS as an excellent tool to perform the calculations required for working statistical problems such as encountered in homework problem sets. In addition, the papers point out the flexibility and power of the SAS programming language for use writing programs that can be used by students to demonstrate important statistical concepts such as variability and power associated with statistical hypothesis testing.

Other reports present the use of SAS in teaching the use of statistical program in a manner like the discontinued North Carolina State University Department of Computer Science statistical package course. For example, Kraemer, Kheah, and Wildman reported to 1988 SUGI11 attendees on their experiences in teaching an undergraduate statistics course on computer statistical packages at the University of Central Florida. Their course presented instruction on the use of BMDP, SAS, and SPSSX statistical software package but spent a substantial amount of time teaching data management principles using SAS.

The literature on the use of SAS in educational instructional settings turned up no cases where a course was designed exclusively to teach SAS data management principle. However, the principles espoused by Gallagher in his 1979 SUGI4 presentation of a course taught to statistically unsophisticated graduate students at the University of North Carolina School of Public Health were much in line with those identified as important in our proposed ST445 course. He identified 10 skills to be acquired by his students: (1) design of forms for computer analysis; (2) design of questions for computer analysis; (3) "punching of data" into a media that was machine readable; (4) range and consistency checking of data; (5) record linkage; (6) updating of data; (7) backups of data; (8) file documentation; (9) statistical analysis of data; and (10) customized output. His goal was to have "students to have attained sufficient background in course to be able to start and complete their computer-aided analysis of research data with minimum of professional programming support (p.175)." A review of the SAS training literature also revealed one report of the teaching of SAS with a focus on the development of database management skills. This course was taught by London (1990) in a pharmaceutical firm to both experience and non-experienced computer programmers. In this 12 week course, the students converted topics that focused on database management concepts, SAS programming principles including macro programming, and statistical analysis and graphical procedures.

Course Features

The instruction in ST445 focused on introducing SAS as a tool for computer management and processing of data. While students taking ST445 typically have had formal coursework in PASCAL and FORTRAN through courses taught by faculty of the Department of Computer Science, the students were not expected to exit the course as expert SAS programmers.

The students were introduced to SAS programming features through instruction provided in the NCSU Statistics Instructional Computing Laboratory (SICL). The SICL includes seven workstations and 14 terminals networked in a single system to allow access to three laser printers. Students met once a week for hands-on demonstration of SAS features. This laboratory
which was open 12 hours per day during the week and several hours on the weekend was available for assignment use by students when it was not being used for instructional purposes. Each student is assigned disk space within the system which allows them to start, interrupt, or modify data sets/programs without fear of losing their work.

The availability of the networked SAS system allowed for storage of data to allow students to practice programming skills. While these data sets were generally small, they presented features that are likely to be found in real life situations which have, in many cases, non-standard data format situations. In addition, students have access to data sets which have been gathered because of their interesting features by J. Timothy Arnold, Director of the Statistics Instructional Computer Laboratory.

**Instructional Focus**

The SAS Language and Procedures Manuals served as course texts and were supplemented by handouts prepared by the instructor. The list of SAS topics presented in the data management part of the course are listed below:

I. The Data Step:

1. Reading of Raw and Stored Data.
2. SAS Formats and Informats.
3. Transformation of Data using Functions.
4. SAS Arrays.
5. Iterative Processing using SAS Programming Language.
6. Identification of Errors and Missing Values in Data Sets.

II. Data Set Management:

1. Listing and Sorting Data.
2. Merging and Updating SAS Data Sets.
3. Reading and Saving Data Sets to/from SAS Libraries.

III. Use of SAS Procedures for Report Construction:

1. Formatting SAS Output through use of File Writing.

2. Outputting of Data Sets from SAS Procedures:
   a. PROC CORR.
   b. PROC FREQ.
   c. PROC MEANS

   a. PROC PLOT
   b. PROC CHART
   c. PROC TIMEPLOT

It will be noted that there is no coverage listed above for the general statistical PROCs generally encountered in statistical analysis situation. The reason for this is that students will have already been exposed to various PROCs such as ANOVA, GLM, MEANS, REG, and UNIVARIATE as well as several PROCs in the SAS/QC® program in their previous statistics methods courses.

Individual course modules have been developed to introduce other aspects of the SAS system that would be of potential use in Data Management. One module introduces the student to the use of SAS/GRAPH® as a high quality graphics output alternative to the use of PROC CHART. This module was developed by Paul Marsh of the Department of Statistics and provides a walk-through of the use of the features of SAS/GRAPH. By asking students to download a series of programs, the students are led through a simulation of the iterative process of adding SAS/GRAPH features to enhance the communication of information presented within the graph.

A second module, developed by Ms. Sandy Donaghy, introduces the students to the use of SAS/IML® through the use of matrix algebra as a vehicle to carry out a regression analysis. Students are exposed to the movement of data from Base SAS to SAS/IML and back again and of the use of the matrix manipulation features to write programs for statistical tests not yet available in the SAS system.

A final instructional module is designed to provide students with experiences of using SAS to interface with other statistical computer packages such as BMDP® and SPSSX® which are only available to students on the NCSU mainframe. Students are given instructions for setting up the necessary JCL required for the interchange of information between SAS and the other statistical packages. This is accomplished by communicating via terminals/modems with the North Carolina State University IBM® mainframe.
located at the university computing center. In addition, students are given the opportunity to utilize a PC to carry out course assignments using the SAS language.

Course evaluation consists of three different components. Homework is regularly given that counts for 10 percent of the final grade in the course. Each student is also assigned to present to the class in a written and oral manner the use of a particular SAS PROC not covered in ST445 and provides another 10 percent of the course grade. A midsemester examination is given roughly half way through the semester to evaluate student grasp of their knowledge of SAS and their ability to use SAS programming principles to carry out basic database operations and counts for 40 percent of the course grade. A course project serves as an opportunity for students to demonstrate their ability to integrate the learning of SAS database management and programming principles in a "real life problem" and is used in lieu of a formal final examination. The class project grade provides 40 percent of the final course grade.

A Recent ST445 Offering

The most recent offering of ST445 utilized the information provided on 386/33 microcomputers in a recent review conducted by PC Magazine (December 25, 1990 and January 15, 1991) to develop a database on microcomputers. Students were introduced to the concept of variable definition and the coding of variables focusing on such issues as to whether a character or numeric codes should be used for to represent different types of variables. Teams of students were then formed to work on individual sections of the review to provide variable definitions, variable codes, and proposed formats.

Each student was then given a section of the review to code for data entry purposes. Unbeknownst to the students, two students were assigned to independently code and enter each of the pieces of information concerning the microcomputer into a unique database. This feature was used to demonstrate rather forcefully the potential for intra-coder variability in individual coding even in relatively straightforward variable coding and data entry situations. At this particular point, each student's data was personally entered into a database. The individual data sets were then inspected for obvious errors and the best one of each pair of student generated database inputs was combined to provide a total class database. This database was then used by the students to complete the required end-of-course project by performing a statistical analysis and to write up a report of the finding utilizing writing skills developed in a required technical writing course.

A second project is also performed by students as the last homework assignment in the course. The purpose of this project was to provide the student with an opportunity to gain experience in the use of SAS to perform computer simulations. The project used as the last homework assignment asked students to replicate a well known experiment of the violation of assumptions underlying the use of the t test (Bonneau, 1960). Students were assigned one of situation where one or more of three factors related to the use of the comparison of two groups could be varied. These manipulations allowed for variation of the following data distribution characteristics: (1) sample size (5 versus 15); (2) equality of variances (1*variance versus 5*variance); and/or (3) type of underlying distribution (normal, rectangular, and exponential). Students then wrote up their findings in a report format with a requirement that they include graphical presentation of the results of their simulation experiment.

Course Acceptance

The acceptance of the course is probably best reflected in the uniformly positive endorsement given by students who have left the campus and then came back later to either visit or to begin work on a graduate degree. The often heard statement from many students that "ST445 was the best course I had at NCSU" may be reflecting the fact that they likely obtained a job due to the knowledge gained in this particular course as well the teaching job performed in the course.

Course Changes

The most likely change in the teaching of this course in the near future will be the movement towards the adoption of a formal text for use in the course. One of the major criticisms of the course by the students was that it was difficult "to find anything in the manuals." This led to a search for alternative sources of primary information about using SAS.

The widespread use of SAS as the statistical computing tool of choice in statistical
methods courses has resulted in the writing of textbooks designed to teach students how to use the SAS programming language to write programs to perform basic data management tasks as well as to run SAS statistical procedures. For example, Freund (1989) focused on the use of SAS for programming to perform simulations as a tool for understanding statistical concepts in addition to teaching the use of SAS for statistical computing. In another orientation, Cody and Smith (1991) used the teaching of statistics as a vehicle for teaching database management principles.

Surprisingly, the writing of textbooks focusing on the use of SAS as a general database management tool has been a relatively recent phenomenon. Some of these textbooks were written for practicing programmers. For example, Aronson and Aronson (1990) wrote for experienced programmers without SAS experience while Aster and Seidman (1991) wrote a text designed to assist programmers already familiar with the SAS programming syntax of SAS (e.g. Aster & Seidman, 1991). Other textbooks have been written for the purpose of introducing the SAS programming principles and use of PROCs to interested potential users. These more general texts include Mastering the SAS System by Jay Jaffe (1989) and SAS Applications Programming by Frank Dilorio (1991).

It is likely that one of these books will be adopted as the primary instruction tool with one or more manuals purchased to serve in a reference role in the course.

Conclusions

The initiation of a course in Statistical Computing and Data Management at the undergraduate level has been a successful experience. A major part of the success has to be attributed to the use of the SAS language as the primary programming tool within the course. The use of SAS in ST445 to accomplish tasks in a data management/data processing project has provided students with very marketable scientific programming skills. At the same time, it is felt that the course has provided students with an opportunity to apply their inquiry and problem solving skills developed through the educational process provided by the experiences associated with attending a research oriented university such as North Carolina State University.

References


London, Wendy, "Teaching the SAS Programming Language to Programmers and Non-programmers." Proceedings of the Fifteenth

SAS, SAS/GRAPH, SAS/IML, and SAS/QC are registered trademarks of SAS Institute, Inc., Cary, NC, USA.

BMDP is the registered trademark of BMDP Statistical Software, Inc., Los Angeles, CA, USA.

IBM is the registered trademark of the International Business Machines Corporation.

SPSSX is the registered trademark of SPSS, Inc.

The author can be contacted at:

Department of Statistics
North Carolina State University
Box 8203
Raleigh, NC 27695-8203