

## Developing Statistical Application Workshops with SAS Software on a PC

Dr. Priscilla Hancock

Working as a statistical consultant for a large university, I was assigned the task of developing a series of statistical workshops using PC SAS. The university needed training in how to program SAS to perform the required data manipulations and statistical analyses and how to interpret the obtained results. It was not uncommon for our users to know statistics but still have difficulty writing the SAS code and interpreting the output. Over the next 18 months, I developed a series of workshops to meet this need.

The first workshop in the series is PC SAS Basics. This workshop was designed for the true beginner. No knowledge of the operating system or PC SAS is assumed. The focus of the workshop is on learning SAS programming in general. It is not a task specific workshop. The PC SAS Basics workshop presents the most commonly used SAS statements and informs the user what these statements do and when they are used.

The outline for the PC SAS Basics course is:

- 1) Operating system interface
- 2) How to execute SAS
- 3) Basic Data Step processing
  - a. How to read external data file and create a SAS dataset
  - b. How to modify an existing SAS dataset via assignment, KEEP, DROP, and RETAIN statements
  - c. Basic programming: IF and WHERE statements, Do Loops, functions, and merging and concatenating datasets
  - d. Making results more informative: PROC FORMAT, LABEL and FORMAT statements
- 4) Proc Step Programming
  - a. Basic Procedures: PROC PRINT, PROC CONTENTS, PROC SORT
  - b. Summarizing Procedures: PROC FREQ, PROC CHART, PROC MEANS, PROC PLOT
  - c. Beginner-Level Statistical Procedures: PROC CORR, PROC TTEST

The next workshop in the series is Introduction to Data Analysis. The workshop is designed for users who have never analyzed data before but who are familiar with PC SAS. The users are assumed to know how to write and execute a program that creates a SAS dataset from an external data file.

The whole approach of the Introduction to Data Analysis workshop is task oriented. The focus on the course is to step users through the process of coding data, entering data, massaging the data into shape, choosing the appropriate statistics, and interpreting the results from descriptive procedures. All of the output produced by the descriptive procedures is explained.

The outline for the Data Analysis workshop is:

- 1) How to prepare the data for data analysis
  - a. Coding Data
  - b. Writing a Codebook
- 2) Ways of getting data into the computer
  - a. PC SAS editor
  - b. Wordprocessors
  - c. Spreadsheets
  - d. Databases
- 3) Ways of changing data into a SAS dataset
  - a. From an flat Ascii file to a SAS dataset: SAS Data Step
  - b. From a spreadsheet to a SAS dataset: PROC DIF
  - c. From a database to SAS dataset: PROC DBF
  - d. SAS's data entry system: PROC FSEDIT
- 4) Manipulating data in a SAS dataset
  - a. Creating new variables
  - b. Changing only certain cases (subsetting IF statement)
  - c. Selecting only certain cases (conditional IF and WHERE statements)
  - d. Recoding variables (ARRAY and DO statements)
  - e. Preparing for by-group processing (PROC SORT)
  - f. Labelling the output (PROC FORMAT, FORMAT and LABEL statements)

- 5) Descriptive Statistics
  - a. Key concepts: levels of measurement and continuous vs discrete variables
  - b. Summary statistics for discrete variables (PROC FREQ)
  - c. Summary statistics for continuous variables (PROC MEANS and PROC UNIVARIATE)
  - d. Looking at the shape of the distribution (PROC CHART)
  - e. Looking for linear relationships (PPROC PLOT AND PROC CHART)
- 6) Choosing the appropriate inferential statistic
  - a. Independent vs dependent variables
  - b. Appropriate statistical procedures for various designs

Two advanced statistics workshops are included in the series. The Analysis of Variance and Regression workshops are for SAS users who need to use these specific statistical applications. Both workshops assume that the user knows how to create a SAS dataset and has some knowledge of the statistical application being discussed. The workshops do not teach the statistics underlying the analysis. Instead the workshops focus on how to set up the data for the analysis, how to program the analysis (including appropriate follow-up techniques and assumption testing), and how to interpret the output. Each piece of the output is explained.

The Analysis of Variance workshop teaches users how to write PROC ANOVA programs that analyze one-way and factorial designs including how to obtain post hoc comparisons. Users also learn how to write PROC GLM programs that analyze one-way and factorial designs including how to obtain a priori contrasts and least square means. The use of PROC GLM to perform analysis of covariance is also presented, including how to test the homogeneity of slopes assumption and obtain adjusted means.

The outline for the Analysis of Variance workshop is:

- 1) Uses and Definitions of analysis of variance
- 2) PROC ANOVA vs PROC GLM: When to use one or the other
- 3) One-way ANOVA with Balanced Data
  - a. Data File Organization
  - b. PROC ANOVA Syntax
  - c. How to obtain post hoc comparisons
- 4) One-way ANOVA with Unbalanced Data
  - a. Data File Organization
  - b. PROC GLM Syntax
  - c. How to obtain a priori contrasts
- 5) Factorial Designs with Balanced Data
  - a. Data File Organization
  - b. PROC ANOVA Syntax
  - c. Obtaining the cell means
  - d. How to perform simple effects tests
- 6) Factorial Designs with Unbalanced Data
  - a. Data File Organization
  - b. PROC GLM Syntax
  - c. Obtaining the least square means
- 7) Assumptions for analysis of variance designs
- 8) Analysis of Covariance
  - a. Data File Organization
  - b. PROC GLM Syntax
  - c. Obtaining adjusted means
  - d. Testing homogeneity of slope assumption

The Regression workshop teaches users how to program PROC REG to perform simple and multiple regression. The focus of the workshop is on the various PROC REG options and statements that produce regression statistics, request model selection techniques, test the model, test the assumptions of regression, and detect multicollinearity. Each piece of output that is produced is interpreted.

The outline for the Regression workshop is:

- 1) Basic Concepts of Regression
- 2) Simple Regression
  - a. How to specify the model
  - b. How to obtain predicted values
  - c. How to obtain residuals
  - d. How to obtain standard beta weights

- 3) Multiple Regression
  - a. Model selection methods
  - b. Model testing methods
- 4) Testing Regression Assumptions
  - a. Outlier detection methods
  - b. Influence diagnostics
  - c. Specification error detection
- 5) Detecting Multicollinearity

Future workshops have been requested. Next spring, the workshop series will include two days of course work on the statistical theory of analysis of variance and regression. The purpose of these theory courses is to provide users with the necessary statistical background to take the current workshops. A SAS/Graph course will also be added to the series this spring. Even further into the future, there are plans to add a time series course and a contingency table analysis course. Both have been requested by users. These workshops will also be expanded to the mainframe platforms.

The current series is conducted in a hands-on facility with 25 PS/2 Model 50's from IBM. The teaching facility comes equipped with a SONY terminal projection system and overhead projector. Both of which allow the instructor to provide excellent audio-visual aids to the lecture.

The workshops are a full day from 8:30 to 4:30 with two 15 minute breaks and an hour lunch.

The basic format of each workshop is that the instructor discusses and demonstrates a unit of information. Users are encouraged to do the examples along with the instructor. After one or two units have been completed, there is an exercise for users to complete. After the users have completed the exercise, the instructor goes over the exercise program and output. Questions are welcome throughout the workshop.

Users are provided an excellent resource manual at the beginning of the workshop so they can follow along with the instructor. These manuals range from 60 to 100 pages in length. They include an explanation of each command, command syntax, sample programs, and examples of the output. The examples of the output are thoroughly explained in the text. There is not a single piece of output that is not documented.

In the past two years that these workshops have been given, they have been well received by the university community and the state agencies. Although there are still workshops to be developed in this series, the existing ones have already met the goal of helping users write statistical programs and understand the obtained results.