# MIS Applications of the Statistical Analysis System

Dr. Douglas G. Bonett, Business Administration, University of Wyoming

## Abstract

*The Statistical Analysis System (SAS) is a powerful and flexible computer software package for management information system applications. The database management, data manipulation, data presentation, and report generation capabilities of SAS are described.*

## Introduction

Mangers are decision makers. In many instances, managers need timely and accurate information to make good decisions. The information that managers need to make decisions will be referred to as management information. Some management information must be extracted from a variety of different sources and some data sets may be large and complicated. Furthermore, the data may require extensive manipulation before meaningful management information can be obtained. Finally, it may be necessary to summarize the management information in a tabular or graphical display which can then be quickly and easily comprehended. A general system of data acquisition and database management, data manipulation, and data presentation methods that yield management information will be referred to here as a management information system (MIS).

The Statistical Analysis System (SAS) is a multipurpose computer software package possessing highly developed database management, data manipulation, data presentation, and report generation capabilities. SAS is available for both mainframe and microcomputers. As such, SAS is an important MIS tool. Surprisingly, MIS textbooks give only passing mention of SAS. Some MIS applications of SAS are described in the sections that follow.

## Database Management

In essence, a database management system provides a simple way to 1) create single or multiple files, 2) enter data into the files, 3) edit existing data, 4) manipulate data to quickly obtain summaries and reports. Although SAS is not a true database management system, it can perform the data entry/edit, select, retrieve, index, sort, summary, and report generation functions of most database management systems.

The SAS/FSP package provides general procedures for browsing, searching, and editing data sets in a full screen mode. SAS/FSP also contains very powerful spreadsheet capabilities. Using the basic SAS language, multiple data files can be merged horizontally or vertically, data can be subsetted, and extracted into other files, and data can be sorted by one or more variables. The powerful merging capabilities of SAS provides a means of performing the functions of a true relational database management system. In many applications, however, SAS will not be an adequate substitute for a true database management system since a true database management system may be more conserving of computer resources (Jaffe, 1989). In this section, some applications of SAS for statistical database management are described.

One important component of statistical data base management involves data screening, a process by which the accuracy of the data is determined (Afifi and Azen, 1979). If the amount of data to be analyzed is not large, then each data point can be examined and checked for accuracy. In many MIS applications, very large amounts of data must be analyzed and data screening becomes more difficult. A variety of methods can be used to assist in the data screening process. These methods include frequency distributions, scattergrams, and descriptive statistics.

A frequency distribution is a listing of the frequency counts associated with each category of a specific variable. If the variable is Type of Product where five different products are assigned the values 1 through 5, a frequency distribution of this variable would give the number of occurrences of each product type. This kind of analysis is useful in detecting certain kinds of transcription errors. An example of a frequency distribution for Type of Product is given below.

| Type | Frequency |
|---|---|
| 1 | 196 |
| 2 | 889 |
| 3 | 42 |
| 4 | 761 |
| 5 | 324 |
| 6 | 1 |
| $ | 2 |

This frequency distribution indicates that three entries (one 6 and two dollar signs) require correction. The FREQ procedure in SAS can produce a frequency distribution as shown above. Using select capabilities of SAS, it is possible to list the observation numbers associated with Type of Product equal to 6 or $ in the data editing process.

For numeric variables that may have many valid values, such as dollar amounts, the frequency distribution can be performed on class intervals of the numeric variable. Dollar amounts can be categorized into a small number of meaningful class intervals as shown below. In this example, no product sells for less than $5.00 or more than $100.00 so the frequency distribution has revealed four errors.

| Sales in Dollars | Frequency |
|---|---|
| 1.50 | 1 |
| 5.00 - 10.99 | 199 |
| 11.00 - 25.99 | 865 |
| 26.00 - 50.99 | 991 |
| 51.00 - 99.00 | 223 |
| 999.00 | 3 |

The use of a frequency distribution can only help identify out-of-bound values. The frequency distribution does not reveal errors such as a type 3 product incorrectly coded as a type 4 product. In some applications, additional information can be used to reveal errors of this kind. For example, if products 1, 2, and 3 are sold only in region A and products 4 and 5 are sold only in region B, then a two-dimensional frequency distribution as shown below may be informative.

| Region | Type | Frequency |
|---|---|---|
| A | 1 | 196 |
| | 2 | 889 |
| | 3 | 41 |
| | 6 | 1 |
| B | 3 | 1 |
| | 4 | 761 |
| | 5 | 324 |
| | $ | 2 |

From this table we detect an additional error. A type 4 or type 5 product has been incorrectly coded as a type 3 product, or a type 3 product has incorrectly been associated with region B. Multidimensional frequency distributions can be generated using the FREQ procedure in SAS. In some MIS applications, two or more dimensions may be needed to discover erroneous data.

The bivariate scattergram is very useful in screening the values of two numeric variables that are logically related (Seber, 1984). For example, if the price of a product is determined by its weight, then price should be related to weight. The bivariate scattergram is a two-dimensional plot of numeric pairs. For example, the price may determine the vertical position of the point and the weight may determine the horizontal position of the point in two dimensional space (Figure 1). Errors may not show up in a single dimension frequency table of class intervals. For example, in Figure 1, there appears to be one error, a unit weighing 9 ounces has a price of only 30 dollars. Note that this error would not have shown up in a single dimension frequency table of price alone or weight alone. The PLOT procedure in SAS can be used to generate bivariate scatterplots.

Descriptive statistics can provide information regarding the integrity of quantitative data. The UNIVARIATE procedure in SAS reports the number of observations and missing values, mean, sum, mode, standard deviation, variance, skewness, kurtosis, minimum and maximum values, range, 1st, 5th, 10th, 25th, 50th, 75th, 90th, 95th, and 99th percentiles. The number of observation and missing values, as well and the minimum, maximum, and various percentiles also may reveal errors in the data (Dixon, 1985).

## Data Manipulation

The most impressive features of SAS are its data manipulation capabilities. SAS contains many intrinsic mathematical, statistical, string, and date functions. SAS also contains numerous statistical data analysis capabilities which have important MIS applications in the area of statistical decision support. The statistical data analysis procedures in SAS are the finest available. Some of the data analysis capabilities include very general time series and forecasting methods such as ARIMA models, multiple regression models with autoregressive errors, and simultaneous equation econometric models. In addition, very general mathematical programming and quality control procedures are available.

Regarding SAS functions (routines that return a value computer from arguments), arguments can simply be variable names or constants, or they can be expressions, including expressions involving other functions. Arithmetic functions include ABS, MAX, MIN, MOD, SIGN, SQRT. Truncation functions include CEIL, FLOOR, FUZZ, INT, and ROUND. Mathematical functions include EXP, GAMMA, LOG, TAN, COS, and SIN. Probability functions include PROVBF, PROBCHI, PROBT, PROBNORM, PROBGAM, POISSON, and PROBIT. Statistical functions include MIN, MAX, MEAN, VAR, SUM, RANGE, KURTOSIS, and SKEWNESS. Random number functions generate realizations from various distributions including the normal, uniform, binomial, Cauchy, Poisson, exponential, gamma, and triangular. A variety of string functions are available. These functions may be used to scan for words, remove trailing blanks, convert to upper case, left or right align a character string, search for pattern of characters, or extract of substring. SAS contains several date and time functions. Some examples are current day, month, or year, current time in hours, minutes, and seconds, Julian date, and day of month.

## Data Presentation

The information in large data sets can be presented graphically to show the value each variable or the relationships between two or more variables (Mendenhall and Reinmuth, 19-82). The CHART procedure in SAS can produce vertical or horizontal bar charts, three dimensional block charts, pie charts, star charts. The charts can be constructed from frequency counts, percentages, cumulative frequencies, cumulative percentages, totals, or averages. Bar charts may standard, stacked, or side-by-side.

The PLOT procedure graphs one variable against another to produce a bivariate plot as in Figure 1. Scaling of the horizontal and vertical axes is performed automatically or can be user specified. The symbols used in plots also can be user specified. Multiple plots can be overlayed to form a single graph with different symbols used in each graph. For example, a plot of observed sales over time can be overlayed with a plot of the predicted sales over time with the symbol "O" denoting observed and the symbol "P" denoting predicted.

It is possible to represent the values of three variables in a two-dimensional plot by declaring one of three variables to be a contour variable in the PLOT procedure. When the value of the contour variable is high, it is represented by a dark point on the plot; when the value is low, is represented by a light point. Presentation quality graphics can be obtained using the SAS/GRAPH package (which requires access to a graphics printer or plotter) and contains GCHART and GPLOT procedures to replace the CHART and PLOT procedures described above.

## Report Generation

Tables of variable values can be generated with the PRINT procedure to produce transaction or information reports (Parker, 1989). The variable values may be frequencies, percentages, totals, means, or any other variable computed in a SAS analysis. The variable can be displayed at each level of one or more other variables such as department, quarter, or product type. One or more variables can be displayed in a table and column sums of each variable can be requested. Title, line spacing, rounding, and format options also are available.

The PRINT procedure is useful for generating "quick and dirty" reports. SAS also has a full compliment of programming functions, such as looping, if/then, if/then/else, go to, locate, and print so that elegant customized reports can be programmed as in any other high level language.

## Summary

Detailed descriptions of the PRINT, PLOT, and CHART procedures, as well as functions and the programming language are given in the SAS User's Guide: Basics (1985). A full range of univariate and multivariate statistical procedures is presented in the SAS/STAT User's Guide (1988). A complete description of numerous time series and forecasting procedures is given in the SAS/ETS User's Guide (1984). Quality control procedures are described in the SAS/QC User's Guide (1988) and management science tools are documented in the SAS/OR User's Guide (1985). Finally, interactive facilities for full screen data entry and editing are described in the SAS/FSP User's Guide (1988).

## References

1. Afifi, A.A. and S.P. Azen. *Statistical Analysis: A Computer Oriented Approach.* Academic Press, New York, 1979.
2. Dixon, W.J. *BMDP Statistical Software Manual.* University of California Press, Los Angeles, 1985.
3. Jaffe, J.A. *Mastering the SAS System.* Van Nostrand Reinholf: New York, 1989.
4. Mendenhall, M. and J.E. Reinmuth. *Statistics for Management and Economics, 4th Ed.* Duxbury Press, Boston, 1982.
5. Seber, G.A.F. *Multivariate Observations.* Wiley, New York, 1984.
6. *SAS User's Guide: Basics*, Release 5.0, 1985
7. *SAS/STAT User's Guide*, Release 6.03, 1988
8. *SAS/ETS User's Guide*, Release 5.0, 1984
9. *SAS/OR User's Guide*, Release 5.0, 1985
10. *SAS/QC User's Guide*, Release 5.0, 1985
11. *SAS/FSP User's Guide*, Release 6.03, 1988
12. Parker, C.S. *Management Information Systems.* McGraw-Hill, New York, 1889.

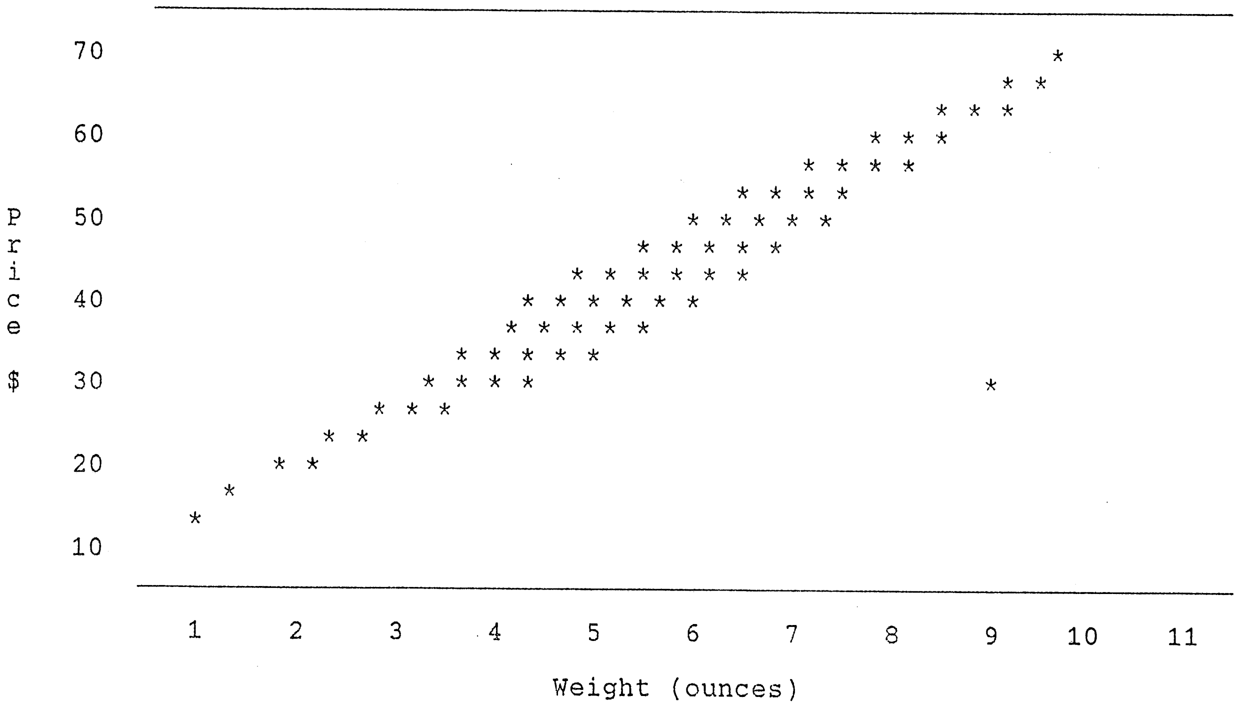

Figure 1. Scattergram of Price versus Weight