DATA-ANALYSIS WITH SAS/GRAPH® SOFTWARE

An alternative view.

Henk van der Knaap, Unilever Research Laboratorium Vlaardingen

Abstract

SAS/GRAPH® software supplies a reasonable number of procedures to present the results of statistical data-analyses. For that purpose handy use can be made of the possibility to output the results of statistical procedures into datasets. Besides simple one- or two-way data structures these procedures can also present multi-dimensional structures fairly well. If SAS® software is used for doing this and one also wants indicate the plausibility or generalizability of the conclusions then the use of the ANNOTATE facility is indispensable. Moreover it appears in practice that the use of this facility makes the interpretation of 3D-plots easier. Examples of this use of SAS® software will be presented. A number of the ANNOTATE-applications are so trivial that one wonders why in relevant procedures these options are not included.

Introduction

SAS/GRAPH® software supplies a reasonable number of procedures to present the results of a statistical data-analysis, performed with one or more procedures out of the other modules of SAS software. In most cases these statistical analyses give a lot of information about the plausibility or generalizability of the conclusions. To incorporate this extra information in the output of the graphical procedures the use of a graphical editor is necessary. In future releases of SAS software such an editor is available and the benefit of it will be proven in practice. However, if the output from analyses on a lot of variables or datasets has to be treated in the same way, the use of this editor will be boring and time-consuming because of its 'data-independent' nature. Moreover this way of post-processing of graphical output may induce an extra risk of making mistakes.

Therefore it is better to do this work 'data-dependent' and by means of the ANNOTATE facility we are able to do so. In practice however (probably due to its high threshold) it is treated as the Cinderella.

The base for this facility is the so-called ANNOTATE= dataset. This is a special SAS dataset with predetermined variables to define the functions available within the ANNOTATE facility and can be created directly or by reshaping ordinary SAS datasets. Because of the fact that most statistical procedures can output the results directly into a dataset, the combination of these procedures and the ANNOTATE
facility is ideal for data-dependent editing of output of graphical procedures. Furthermore it offers the possibility to upgrade graphics in general. Examples of 'annotations' of the output of the procedures G3D, GCHART and GPLOT will be discussed. A number of the ANNOTATE-applications (and certainly those in the procedure G3D) are so trivial that one wonders why in relevant procedures these options are not included.

G3D

Figure 1 presents the results of a regression analysis in which the influence of two independent variables (A and C) on a third variable (R13) is investigated. The practical value of this plot is to deliver a three dimensional impression of the shape of the response surface. However, for location of the optimum of the response variable one can better use the procedure GCONTOUR. Figure 2 gives its output using the same data. The dataset used to produce figure 1 can be transformed into an ANNOTATE= dataset by subsetting it in an extra data step to do the necessary manipulations and calculations. By using the appropriate ANNOTATE functions figure 3 can now be produced by proc G3D annotated with this dataset. The shape of the response surface and its projection on the base can be inspected at a glance. Furthermore the legend of the colours used is projected on the response axis. Another use of ANNOTATE= datasets is given in the figures 4, 5 and 6 based on different data. Besides the information about the response surface these figures show also the datapoints from which the response equation has been calculated in the regression analysis. By means of colours or symbols it is shown whether these points lie above or below the surface. By projecting these datapoints on the base (see figures 5 and 6) one can also indicate the actual experimental area.

GPLOT

Figures 7, 8 and 9 show graphs produced with the procedure GPLOT. They are based on the outcomes of statistical analyses on data out of a factorial screening test (a so-called Taguchi experiment). Figures 7 and 8 present the main effects and two interactions including their 95 % confidence intervals. The text strings in these plots and the confidence interval are produced using an ANNOTATE= dataset created from an output dataset out of the procedure REG. However, this experiment dealt with more than one response variables and a statistical analysis has been performed on each of them. To compare the results of the statistical analyses at a glance figure 9 is produced. To achieve this the effects presented in figure 7 have been standardised. After this step it is possible to combine the 'effect profiles' of more than one variable in one plot. By means of the vertical decision lines one can draw conclusions about the significance of the effects. Or in other words: Figure 9 presents a Multivariate look on Univariate analyses.

As already said: The base for these figures was output produced by the procedures REG and GPLOT. But, also for a lot of other cases, a combination of some statistical and graphical procedures and the ANNOTATE facility could produce similar types of output.
Figure 10 shows bar charts produced by the procedure GCHART and presents an overview of the staffing of a research division in our laboratory in 1990. These charts give the number of employees per category (midpoint variable and coded as characters) for each of the sections (group variable and coded as numbers) of this research division.

In fact this figure is the result of combining two pages of output of the procedure GCHART using the GREPLAY procedure and its TEMPLATE facility. From this figure it is not easy to get a quick overview of the change in staff during 1990. One can obtain a better impression of that change from figure 11. In this figure the same data are presented in one bar chart, but now the category type is taken as the subgroup variable and the moment (S for the start of the year and E for the end) as the midpoint variable. Although the change in staff seems to be presented more clearly the conclusions, however, can be misleading. If one looks e.g. at the bars for section 3092 one can conclude that there is not a change at all. In practice, however, there was a change, but the net result was zero. Therefore figure 12 has been produced. In this figure the subbars on top of the bars give the number of employees left and joined during 1990. Furthermore the bar for the ARU employees in section 3091 is broken down into pieces and a new response axis value is written besides the bar. Presenting the data in this way one can see the staffing at the beginning of 1990 (the empty bar plus the solid black subbar), at the end (the empty bar plus the gray subbar) and the actual change.

To draw figure 12 one only needs to have a dataset with variables in which the section codes, category type and the numbers for the staffing (at the begin, at the end and the change) are given. After some calculations and manipulations in a data step one can create from this dataset two new ones. One prior dataset contains the information about the number of persons that were present, both at the beginning and at the end, and can be passed through to the procedure GCHART. If one uses a number of the normally available options this will then result in the figure with only the empty bars. All the other information in figure 12 is the result of the second dataset that has been used as a ANNOTATE= dataset.

Conclusions and remarks

Figure 13 gives a brief outline of a way of acting to get from raw data to a clear presentation using the SAS software system and only showing the information we need in the way we want.

In figure 14 a few conclusions are given. But also a few remarks, because a number of the editing actions using the ANNOTATE facility are very laborious but also very trivial. Therefor it is advised to include them as options in relevant procedures in future releases of the SAS software system.

Where possible, the figures in this presentation are prepared using a two-colour device. This solves in daily practice a lot of problems if the graph has to be copied using an ordinary photocopy machine. Furthermore it is easier for the receiver of your graphs
(human or instrument) in case it, he or she cannot handle colours at all.

**Hardware and Software**

This presentation was prepared using the modules BASE, STAT and GRAPH of the SAS® software system. As hardware an IBM/PS2 model 80 under PC-DOS 3.31, equipped with a HP7550 plotter and a QMS-PS-810 as hard copy devices, was used. Copies of SAS programs to produce one or more of the presented figures can be obtained from the author.

Henk van der Knaap
Unilever Research Laboratorium Vlaardingen
Statistics Section
Olivier van Noortlaan 120
3133 AT Vlaardingen
The Netherlands
Figure 7
Example of an analysis of a Taguchi Experiment.
Variable = VARI

Figure 8
Example of an analysis of a Taguchi Experiment.
Variable = VAR

Figure 9
Example of an analysis of a Taguchi Experiment.
Confidence (%)
Figure 10
Staffing PAS Division in 1990
1 January

31 December

Figure 11
Staffing PAS Division in 1990

Figure 12
Staffing PAS Division in 1990

Left

Joined
Figure 13

Data
- Calculations, Transformations

Analysis
- FREQ, TABULATE, SUMMARY
- ANOVA, CLM, REG
- CLUSTER, DISCRIM

Graphics → Edit
- G3D → GREPLAY
- GCONTOUR → ANNOTATE
- GPLOT → Reports

Presentation

Figure 14

Conclusions:
- G3D needs revision.
  (Contourplots, datapoints)
- Extra options in GCHART.
  (Breaking of bars, Start and End values)
- 'Multivariate' inspection of
  'univariate' analyses with
  SAS/GRAPH® possible.
- Graphics Editor necessary.