Introduction

Anyone who analyzes data needs certain readily available tools for the task. Statisticians, researchers and managers of today require data analysis tools from calculators to data management, statistics and graphics. While good data analysis can involve all of these tools, well-done graphics is surely the one which stimulates our inquisitive mind the most and, commonly, the results of that thinking process. Statistical graphics embodies the concepts of good graph design and is extending our repertoire of graphs from what is ordinary today to more powerful data displays that will become ordinary tomorrow. The SAS® family of products is well-suited for participating in this evolution. By combining the SAS tools of calculations, data management, statistics, basic graphics, and now ANNOTATE, SAS users can develop programs that will add statistical graphics to their list of data analytic capabilities.

What's in a Name

The ANNOTATE facility is much more than its name might imply. Surely, it does provide the much-needed flexibility for annotating graphs [1], yet it doesn't stop there. The secrets to its hidden richness are the ability for accurate placement on the graph, the library of graphics primitives, and its integration with a powerful statistics package. All are basic requirements for producing statistical graphics. Accurate placement is required because one of the basic tenets of statistical graphics is to portray information accurately, thereby avoiding any tendency to mislead the viewer. Having a library of graphics primitives is a great help because many of the statistical graphics of today rely, by and large, on simple geometric shapes which are assembled together to form the graph design. Being part of an integrated statistics package makes for efficient use of well-developed statistical routines for preprocessing the data.

Alternatives

An inspection of the alternatives for statistical graphics programs does not reveal any superior options [2]. Programming in a higher level language, such as FORTRAN, and using callable graphics subroutines gives great flexibility, yet is costly in development, maintenance and modification time and effort. A few software packages might have more built-in statistical graphics but might not be compatible with the existing computer environment.

SAS/GRAPH® with ANNOTATE has the advantage from the developers viewpoint, that SAS/GRAPH can be used as the workhorse for creating the basic graph layout, for example, performing axis scaling and titling, as a direct consequence of SAS/GRAPH, many graphics devices are already supported. Thus the development effort can get to work immediately on constructing the elements that make up the statistical graph. As for maintainability and modification effort for graphics programs, my experience says that well-documented SAS code is easier to support than its FORTRAN counterpart which relies on subroutine calls to a graphics library. Another important advantage for using the SAS products is that the SAS Users community which provides a means of sharing the effort and the potential for new ideas.

Background

At the Chemical Research Division of American Cyanamid Company, computing capability is driven primarily by the needs of scientific investigations as opposed to business applications. Graphics needs characteristic of an R&D environment include graphs for problem solving and graphs for exploratory investigations. Statistical graphics are well-suited to meeting both types of problems.

With the current trend toward end-user computing and with user-friendly interfaces, the population of graphics software users at Cyanamid, as elsewhere, is evolving from the statistician, data analyst, programmer to the researchers themselves. To make this transition effective requires, among other things, educating the masses on the concepts of good graph design, improving visual thinking skills and providing good graphics software with room for customization and expansion. ANNOTATE
enhances the SAS products position as a part of that growth plan. An issue yet to be resolved for the portable systems is the user-friendly graphics interface, something that is expected in today's software market. ANNOTATE is not a product for the novice or occasional user, yet it is quite learnable by the experienced SAS programmer. So, an appropriate approach is for the experienced user to develop general statistical graphics programs which would appear as options on the user interface of the basic SAS/GRAPH options.

Until version 5.0 of SAS, ANNOTATE was available only on the portable version of SAS/GRAPH. This gave the portable SAS users an unusual advantage over their IBM counterparts. Yet, perhaps to even the score, a few things are missing from portable SAS in the commercially available version, 4.07, which would have simplified the task of developing the statistical graphics programs. Portable SAS for VAX's® does not have the programming power of MACRO's nor INCLUDE. Without these higher programming steps, one must accept longer and more tedious codes and some restrictions in program generalization. This has resulted in some limitations in attempting to make statistical graphics tools available to my user community. As a stop-gap measure, BY group processing can be used to overcome some of the limitations. MACRO's and INCLUDE are promised in subsequent releases of the portable systems. This would be a great programming benefit.

Examples

Some examples will serve to illustrate the discussion to this point. The intent is two-fold, first to teach by example some concepts of programming with ANNOTATE, and second to identify some quirks or limitations that surfaced in the course of my using the language. Stop-gap measures are given which overcame these problems, and hopefully, long-term solutions will be implemented as the product matures. It should be remembered that ANNOTATE is a new product, and as with any new product, minor problems can arise. What is important to keep in mind is the capability that is intended and now feasible.

Box-and-Whisker plots

The box-and-whisker plot, and its many variations, is surely a favorite among statistical graphics. Two items on the SASware ballot were related to this type of plot this year. It is elegant in its ability to summarize univariate data in a simple and easily comprehended form. One might use this type of statistical graphic instead of a histogram. It summarizes distributional characteristics in one dimension, such as the range, center and bulk of the data. Outliers and distributional shape information are also presented by this clever graph. In this mode, the graph is a tool for data exploration. Because it is familiar, it is an appropriate starting point to introduce ANNOTATE and to illustrate several points that have been made so far.

Developing the box-and-whisker plot provides an opportunity to evaluate the capabilities of SAS/GRAPH and ANNOTATE as a statistical graphics programming package. Specifically, before any drawing is done, PROC UNIVARIATE is used to preprocess the data. This step determines the summary statistics which define the box and other statistics useful for completing the graph.

SAS/GRAPH is used as the workhorse for the graph layout including one of the newer SYMBOL options, I=HILOTJ, used to create the whisker portion of the graphic. The box portion is created using one of the graphics primitives of ANNOTATE, namely the bar. Here, the accurate placement of the bar on the graph is crucial; otherwise the graph would be meaningless. In this graph, all drawing is done within the axes boundaries using (x,y)-coordinates.

Since the box-and-whisker plots were intended to become one of several generally available graphing options, generalizing the code to fit most situations was important. The height of the box was automatically generalized because UNIVARIATE was used to furnish the quartiles for the top and the bottom. However, the width of the box had to be determined in such a way that it too would be automatically adjusted depending on the data at hand. An algorithm was developed which determines the box width based on the overall range in the x-dimension. The halfwidth of the box was defined as 1 percent of the range of x. Because it is fashioned in the same manner that one would scale the x-axis, this approach has given boxes of acceptable widths for plots covering many different scales of x. Listing 1 is an excerpt of the details [3]. See also Illustration 1.

Putting more than one box-and-whisker plot on a page is where the power of the graph begins to prove itself. BY group processing can be used to create boxes for different
sets of conditions. Figure 1 is an example where box-and-whisker plots were drawn for each of several conditions of coating and exposure time. With the data displayed in this manner, the graph becomes a decision-making tool because the viewer can simultaneously assess several aspects of the data that are important for choosing the best coating. Shown here are the failure of the uncoated samples, the improvement due to each of the coatings, the longevity of the effects, and the inherent variability in the samples. A couple bonuses were obtained as a direct result of making the plot, namely observing the dip in the uncoated samples at four days of exposure and the increase in variability with time for these same samples. Both were of particular interest to the researcher in understanding the physical phenomenon.

Two minor problems were encountered and are worth noting because they were not what one might expect. In both cases, fixes were found which overcame the problems. First, it was a surprise to find that the bars have no bottom line. When the bars are filled solid, this omission is irrelevant. But, when solid bars are inappropriate, such as in non-color printings, the missing bottom makes for incomplete boxes. Bars are the preferred choice of graphics primitives because they are completely specified, and filling them is accomplished very quickly by simply selecting a fill pattern. Short of a change in ANNOTATE to add bottoms to the bars, couple stop-gap fixes are possible. First, whenever possible use only solid bars. A second option is to add the code to draw the line segment along the bottom of the bar, remembering to draw it in the same color as the bar. Thirdly, one could replace the bar with a rectangle drawn as a polygon by specifying the four corners, and remembering to fill the rectangle with the pattern desired [4]. All of these fixes will work, but a bar with a bottom line would be preferred.

The second item relates to the new symbol HILOTJ [5]. When any line style other than solid is specified, the broken line is applied to the vertical member in addition to the joining segments. This often gives the effect of chopped off crossbars and/or crossbars that don't touch their vertical line, making a graphic that is unnecessarily difficult to read. If a simpler plot statement was possible, the data could be plotted twice, using OVERLAY and a second symbol statement with I=HILOT and a solid line. In this case, however, the plot statement is of the form Y*X=%, so OVERLAY is not allowed. Here, one has to resort to filling in the gaps by hand. It would be preferred if only the joining lines were affected by the line style, and the vertical lines and crossbars were always drawn as solid lines. A cleaner and therefore better graphic would result.

Quality Control Charts

A family of graphs which surely has grown in popularity these days is the set of quality control charts. These graphs belong to the domain of statistical graphics, having existed long before there was a name for them. They too appeared on the SASware ballot this year. Many different SAS programs exist for producing the plots and the important control lines which give these graphs their decision-making power. For an example, see Reference 6. As an enhancement to existing QC programs, ANNOTATE can be used to label the control limit lines to clarify their use. This use of ANNOTATE is probably more in line with one's initial concept of the use of this programming capability.

In the QC chart shown in Figure 2, both warning and control lines at 2- and 3-sigma, respectively, are drawn on an xbar chart. The locations of the lines are determined by preprocessing the data through DATA steps and statistical procedures. ANNOTATE is used to draw the lines by connecting (x,y)-coordinates, although other approaches would be equally satisfactory. The primary advantage that ANNOTATE offered in this example was being able to accurately place the labels on the plot adjacent to these lines. Without ANNOTATE, one must use NOTES and moves, and change HPOS and VPOS which, for all the effort involved, does not have the same level of accuracy of placement and almost never results in exactly placed labels. Using ANNOTATE is much quicker because the label location is sure to come out right the first time. The locations of the labels were defined in terms of the y-coordinate of the limit line, but the x-coordinate was defined in terms of data %. Centering of the title in the y-dimension was also specified. The code in Listing 2 gives more details.

As foolproof as this might sound, a quirk did occur due to writing at greater than 100% of the data range. Usually the labels work as desired. However, when the x-axis reaches some length, yet unknown, the labels apparently extend too far to the right
and get clipped as if there were a solid boundary. This does not always occur, but seems to depend on the x-axis scaling, which in turn depends on the data. Adding blank footnotes along the right side did not help. A fix was found for the problem which was to adjust HPOS and VPOS. A more complete solution would be to have SAS/GRAPH include the annotation space in its plan of the plot, as is done with footnotes and titles, and to thereby ensure that all will fit within the plotting window.

Sunflower Plots

When data are two-dimensional, it is often helpful to try to summarize graphically the bi-variate density. A statistical graph that does this very well is the sunflower plot. It brings out patterns in the data that can be lost in scatter plots. To construct the sunflower plot, the data space is divided into a square grid, and the number of data points in each cell is determined. Although some newer methods use interlocking hexagons instead of square cells, the square cell approach was used here for its simplicity. For cells with only one point, the center of the cell is marked with a small circle. For cells with more than one point, a geometric shape is constructed where the number of spokes emanating from the center of the cell is equal to the number of points in the cell. Figure 3 is an example.

DATA steps and BY group processing were used to put the points into cells and to determine how many occupied each cell. Using geometry and polar coordinates, the endpoints of the spokes can be calculated quite nicely. By defining each spoke as a line segment, ANNOTATE is required only to draw the segments and thus construct the sunflowers. Listing 3 gives the details. See also Illustration 2.

The plot in Figure 3 is an example of a sunflower plot that revealed the possibility of two overlapping populations. The scatterplot of these data is included for comparison, Figure 4, because it did not indicate any such possibility. This use of the sunflower plot is another example of a data screening graphics tool that served to sharpen our two-dimensional image of these data. As is turns out, the data were from two different types of materials and the review of this plot led to an improved statistical analysis of the data.

It is not likely that the first grid pattern that is tried will give the sharpest image. So, one criterion for this program was to keep to a minimum, the number of things that would have to be changed in order to try different grids. To change the grid, only three parameters have to be changed, namely XLOW, YLOW and CELLSIZE. One proceeds in this manner toward a grid which gives the sharpest image.

Using ANNOTATE is probably the simplest part of developing this plot. It only has to move to the center of a cell and draw straight line segments using (x,y)-coordinates. What was and still is difficult to achieve, is the proper dimension on the axes so that square cells come out reasonably square, especially on a CRT where grid iteration is best performed. Without strong control of the axes lengths and since they are usually not of the same length, 10 units along the x-axis will not measure the same as 10 units along the y-axis. So, the sunflowers can come out either too short and squatty or too tall and skinny. Best results were obtained when using a plotter, where HSIZE and VSIZE have an effect. They provide some indirect control over the axes lengths. To ensure the proportions of the cells and of the sunflowers would require more control over the axes on both plotters and CRT's.

Conclusions

Using ANNOTATE opens up a new spectrum of graphics capability to SAS users. Because of some well-done features in ANNOTATE, the programmer has the capability to program graphs that might not be available by other means. The statistical graphics capability demonstrated here is but one example. Knowledge of geometry and trigonometry is very helpful for developing the algorithms that translate the mental images into the computer program equivalent. For further reading on statistical graphics, the reader is referred to Reference 7.

Having developed graphics systems for end-users, a few comments are worth making which would make the SAS graphics packages a stronger competitor in the graphics market. Slowness is the major complaint of users; it is hoped that the dynamic loading of version 5.0 will help here. For the portable systems, the higher level programming steps, such as MACRO's and user-written procedures, would be a benefit to any SAS programming, whether for graphics or not. Also for the portable systems, some mechanism for a user-interface to SAS graphics would be
well-used and would help SAS to compete in a market that expects easy and quick access to graphs. One comprehensive graphics manual would be welcomed. Increased access to axis operations would help to meet some of the requirements of statistical graphics. Including all ANNOTATE drawings in the scan of the plot layout prior to scaling is recommended. Lastly, it is hoped that more statistical graphics are added to the SAS library of readily available graphs.

References:


For further information, contact the author at:

Patricia Benoit
American Cyanamid Company
Chemical Research Division
1937 West Main Street, P.O. Box 60
Stamford, CT 06904
Phone: (203) 348-7331

Trademark Notice:

SAS, SAS/GRAPH and SAS Communications are registered trademarks of SAS Institute Inc., Cary, NC, USA. VAX and VMS are trademarks of Digital Equipment Corporation, Maynard, MA.
Illustration 1.
Constructing the Box

LISTING 1.
* Define the lower left-hand corner of the box in plot units at the lower quartile
  FUNCTION='MOVE'
  XSYS='2';
  YSYS='2';
  X=X-DAY-BOXHALF;
  Y=Q3Y;
  OUTPUT;
* Solid fill to the upper right-hand corner of the box at the upper quartile
  FUNCTION='BAR';
  STYLE='SOLID';
  IF COATING='U' THEN COLOR='RED';
  IF COATING='B' THEN COLOR='YELLOW';
  IF COATING='A' THEN COLOR='BLUE';
  XSYS='2';
  YSYS='2';
  X=X+DAY+BOXHALF;
  Y=Q3Y;
  OUTPUT;

Figure 2.
AMERICAN CYANAMID COMPANY
Laboratory Quality Control Chart

LISTING 2.
* Add line label for the upper control line of XBAR chart
  FUNCTION='LABEL';
  XSYS='2';
  YSYS='2';
  X=X+161;
  Y=Y+ULCBAR;
  POSITION='6';
  COLOR='BLACK';
  STYLE='DUPLEX';
  TEXT='Control';
  OUTPUT;
Figure 3.
A Sunflower Plot that shows two overlapping populations

Figure 4.
... and the corresponding Scatter Plot

Illustration 2.
Constructing a Sunflower

\[ \begin{align*}
& (x_1, y_1) \quad (x_2, y_2) \quad (x_3, y_3) \\
\end{align*} \]

A cell having \( n=3 \) points and centered at \((x_{\text{center}}, y_{\text{center}})\)

\[ \begin{align*}
& k=1, 2, \ldots, m \\
& x_{\text{center}} + r \cos \theta_k \\
& y_{\text{center}} + r \sin \theta_k \\
\end{align*} \]

\[ \begin{align*}
& k=0, 1, 2, \ldots, n \\
& r = 0.3 \times \text{cellsize} \\
& \theta_k = 90 + k \times \frac{360}{n} \\
\end{align*} \]

LISTING 3.

* Construct the arms of the symbols using polar coordinates and make the ANNOTATE dataset, SYMBOLS, of short line segments

DATA SYMBOLS; SET GRID;
KEEP FUNCTION COLOR XSYS YSYS X Y LINE SIZE;
LENGTH FUNCTION COLOR 50;
IF FREQ GT 1;
ANGLE=360.*FREQ/
LINESIZE=CELLSIZE*0.30;
DO COUNT=1 TO FREQ;
ANGLEP=ANGLE*COUNT; 
ANGLEP=ANGLEP+3.1415926/180.;
XEND=XCENTER+(LINESIZE*COS(ANGLEP));
YEND=YCENTER+(LINESIZE*SIN(ANGLEP));
X=XCENTER;
Y=YCENTER;
FUNCTION='DRAW';
OUTPUT;
END;