Paper 70-26

Data Visualization of Outliers from a Health Research Perspective Using SAS/GRAPH® and the Annotate Facility

Nadia Redmond
Kaiser Permanente Center for Health Research, Portland, Oregon

ABSTRACT

SAS/GRAPH® is a powerful tool for customizing the box plot to detect and identify outliers. This paper shows how to use the ANNOTATE facility and annotate data set to customize box plots and profile plots of outliers using data from a dietary-health study.

This paper assumes:

- A working knowledge of basic SAS/GRAPH procedures.
- The ability to display or print graphics on your operating system.
- A thorough understanding of the DATA step, since the ANNOTATE facility requires SAS data set.
- An understanding of SAS/BASE® global statements (e.g. TITLE, FOOTNOTE, OPTIONS, etc.).

SAS skill level of this paper is intended for beginners to intermediate programmers.

KEYWORDS
Outliers, Global, ANNOTATE, PROC GPLOT

INTRODUCTION

During a clinical trial study, many monitoring reports are created to assure the quality of outcome measurements. This paper uses data from a dietary study that aimed to reduce blood pressure by varying sodium levels in the meals. Subjects who participated in a controlled-intervention feeding for 14 weeks were weighed weekly to determine if they maintained their baseline weight (Svetkey, LP. et. al., 1999).

Box plots of the within-subject standard deviation of subjects' weekly weights were used to determine whether a subject experienced a major weight loss or gain. For the original box plot, the GPLOT procedure used an annotated data set to label the total number of subjects at each clinic site (see Figure 1). The boxes represent 25th percentile, median, and 75th percentile. The whiskers represent 5th and 95th percentiles. The points represent outliers with major weight gain or loss.

This paper demonstrates how to use the ANNOTATE facility to identify the outliers on the box plot and then plot a profile of each subject's weight over the study period for clarification.

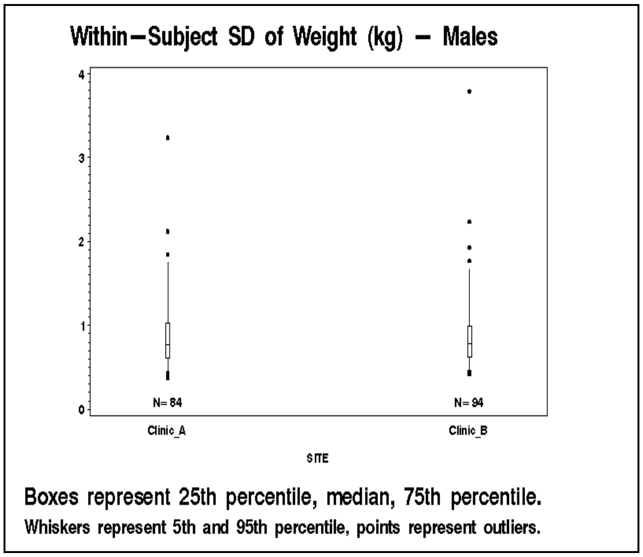

Figure 1 – Box Plot using PROC GPLOT and ANNOTATE data set to label the number of subjects at each clinic site.

GLOBAL STATEMENTS

SAS/GRAPH offers a number of graphic-specific global statements, plus options on the TITLE and FOOTNOTE statements, to customize graphics procedures. Global statements are executed immediately upon reaching those statements in the code. In this paper, examples of GOPTIONS and SYMBOL statements are described.

The GOPTIONS statements are used to temporary set the graphics options, overriding the defaults in the SAS/GRAPH software. Graphics options control many aspects of the graphics environment, for example, output device (TARGETDEVICE), fonts (FTEXT), and colors (CTEXT). The plots in this paper use the following code after resetting some of the options. (See Chapter 12, "The GOPTIONS Statement", *SAS/GRAPH Software: Reference, Volume 1, Version 6, First Edition,* pp. 291-294 for details.) Example code:

```
goptions reset=(axis, legend, pattern,
          symbol, title, footnote)
          rotate=landscape hpos=0 vpos=0;
 goptions targetdevice=winprtc
            ctext=black ftext=swissb
            interpol=join noprompt;
```

The SYMBOL statement is the key to controlling the appearance of the plot. The option INTERPOL=BOX or I=BOX produces box and whisker plots ("SYMBOL

Statement Options" in Chapter 16, *SAS/GRAPH Software: Reference, Volume 1,* pp. 413-418). The BOX option specifies the percentile to control the length of the whiskers within the range 00 through 25. BOX05 was used to represent the 5th and 95th percentile. Example code:

```
   symbol1 interpol=box05 value=dot
           height=.5 colorvalue=black;
```

ANNOTATE FACILITY

The ANNOTATE facility is an integral portion of SAS/GRAPH. It is a set of step by step instructions for labeling or drawing on a graph. The annotate data set stores the instructions for the SAS/GRAPH procedure with an option on the procedure line that specifies the name of the annotate data set: ANNOTATE=< data set name >.

THE ANNOTATE DATA SET

The ANNOTATE data set contains annotate variables that can be manipulated to customize the plots. A partial list of the annotate variables are summarized in Table 1.
The annotate data set is manipulated in a DATA step .
Example code:

```
*create annotate data set for the box plot;
    data annowtm;
     length function color style $ 8;
     retain function 'label' color 'black'
             when 'a' style 'swissb'
             xsys '2' ysys '2' position ' '
               size 1.2 hsys '3';
        set wtout;
         by site sex sdwt;
        x=site;  y=5;
        if sex=1 and (sdwt<pctlwt5 or
                       sdwt>pctlwt95)
           then do;
                text=subjectid;
                 y=sdwt; /*standard dev. */

              retain out 0;
              if first.site then out=0;
              out+1;

/* use of mod() function helped to
alternate the subjectID position labeling
the outlier points on the graph, see below
for further details and figure 3*/

        if mod(out,3)=0 then do;
          position='3'; end;
       else if mod(out,3)=1 then do;
          position='1'; end;
       else do;
          position='7'; end;
          output annowtm;
       end; run;
```

The annotate variable POSITION tells where to draw the TEXT variable in relation to the point (x, y) that is plotted on the graph. Values are typically '1' to '9' or 'A' to 'F' in relation to a grid around the point. For example, position '9' positions the text just below the point and left aligns it. See chart on page 522 of the *SAS/GRAPH Software: Reference, Volume 1* for further details.

Example of annotate data set from PROC PRINT output:

```
data=annowtm

Obs function color style when xsys ysys position
 1   label   black swissb   a   2     2        1
 2   label   black swissb   a   2     2        7

Obs size  hsys    site     sex   numobs   pctlwt5
 1   1.2   3    Clinic_A  MALE    24     0.46548
 2   1.2   3    Clinic_A  MALE    24     0.46548

Obs  pctlwt95   sdwt    x    y       text     out
1    1.84939   0.46089  1  0.46089  IDN030     1
2    1.84939   3.23946  1  3.23946  IDN093     2
```

TABLE 1 - Partial List of Annotate Variables

Variable	Description
Variable that defines an action	
Function (with task 'Label')	Draws text
Positioning variables that determine coordinate values	
X	Specifies a numeric horizontal coordinate
Y	Specifies a numeric vertical coordinate
Positioning variables that specify coordinate systems	
HSYS	Specifies type of units for the SIZE variable
XSYS	Specifies coordinate system for X or XC coordinates
YSYS	Specifies coordinate system for Y or YC coordinates
Attribute variables	
ANGLE	angle of text label or starting angle of a pie slice
CBORDER	colored border around text or symbol
COLOR	color of a graphics primitive
POSITION	Placement and alignment for text strings
ROTATE	angle at which to place individual characters in a text string or the delta angle (sweep) of a pie slice
SIZE	size of an aspect of a graphics primitive; depends on FUNCTION variable (for TEXT, height of characters; for PIE, pie slice radius; for DRAW, line thickness; and so on)
STYLE	font or pattern for a graphics element, depends on the FUNCTION variable
TEXT	text to use in a label, symbol, or comment
WHEN	whether a graphics element is drawn before or after procedure graphics output
Web variable- *New in SAS® Version 8	
HTML	Specifies link information for a drill-down graph

In Figure 2, the annotate variable 'TEXT' displays the subject identification of each outlier point. In the creation of the annotate data set, the use of MOD(argument-1,argument-2) function allows the 'POSITION' of the 'TEXT' to be alternated for readability, especially if several outlier points are next to or on top of each other. Depending on the number of outliers 'clustering' together, the integer quotient of argument-1 divided by argument-2 may need to be changed in the MOD ( ) function (see *SAS Language, Version 6, First Edition,* page 571).

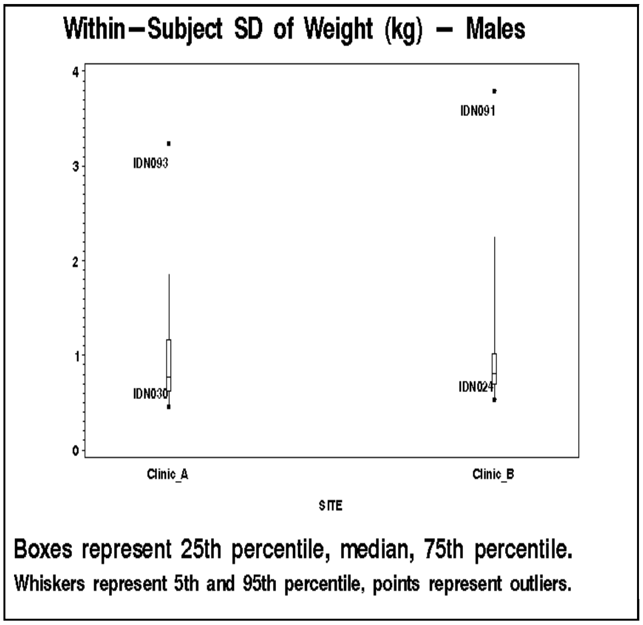

Figure 2 – Box Plot customized with the annotate data set to label the individual outliers.

The following code produced the graph in Figure 2 with ANNOTATE option in PROC GPLOT.

```
*******************************************;
PRODUCE GRAPH for the box plot
*******************************************;
   proc gplot data=sumall2m ;
      plot sdwt * site / haxis=axis1
      vaxis=axis2
     /annotate=annowtm; *annotate data set;
        format site siteft.;
    title h=2 "Within-Subject SD of Weight
              (kg) - Males";
    footnote h=1 j=l "Boxes represent 25th
       percentile, median,75thpercentile.";
    footnote2 h=1 j=l "Whiskers represent
       5th and 95th percentile, points
       represent outliers.";
   run; quit;
```

AXIS< n> is a statement that contains AXIS definitions. It is used to control the location, values, and appearance of the axes on plots and charts. The example code below defines the COLOR of the box plot axes (AXIS1 and AXIS2). WIDTH is the thickness of the axes line. STYLE is the line type, which has values 0 through 48. MAJOR and MINOR specify the tick marks on the axes. ORDER specifies the data values in the order they are to appear on the axes, and OFFSET specifies the amount of space to offset the first major tick mark. Chapter 9 of the *SAS/GRAPH Software: Reference, Volume 1* is dedicated to the AXIS statement.

Example code used for the box plot axes:

```
/* The AXIS statement defines the axis
characteristics and can be placed prior or
after the SAS/GRAPH procedure */

   axis1 color=black width=1.0 style=1
         major=none minor=none
         order=(1 to 2 by 1) offset=(4 cm);
   axis2 color=black width=2.0 style=1
         label=none;
```

The AXIS definitions are then used when assigned by an option within SAS/GRAPH procedure: GCHART, GCONTOUR and GPLOT. Example code:

```
   proc gplot data=sumall2m;
      plot sdwt *site / haxis=axis1
      vaxis=axis2;
```

The two outliers from Clinic A in Figure 2 are then displayed on a profile plot (Figure 3).

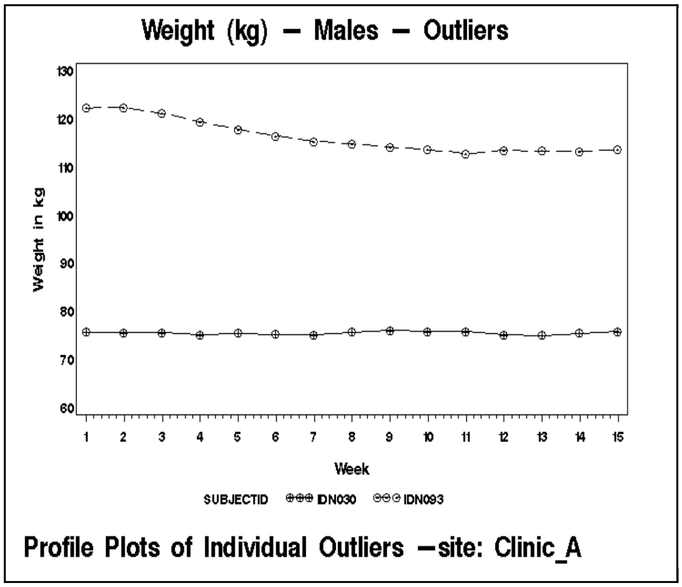

Figure 3 – Profile Plot of the subjects' weekly weights.

Michael Friendly (1991) describes a *profile plot* as a set of variables that simultaneously sets line segments with connecting points, one for each variable. The PROC GPLOT is used to build the profile plot in Figure 3 (See Chapter 31, "The GPLOT Procedure", *SAS/GRAPH Software: Reference, Volume 2, Version 6, First Edition,* pp. 1073-1130). The SYMBOL< n > statements are used to distinguish each subject on the graph.

The first symbol definition, assigned with the SYMBOL1 statement, is used for the first set of points plotted. Subsequent symbol definitions, SYMBOL2, SYMBOL3, etc., are assigned to subsequent plots when more than one plot is displayed in a set of axes.

Example code:

```
  symbol1 interpol=join value=+ height=1
          colorvalue=BLACK line=1;
  symbol2 interpol=join value=- height=1
          colorvalue=BLACK line=21;
```

As shown in Figure 3, the first subject is graphed with + ‘s. The second one is graphed with –‘s in the VALUE= options. Example code:

```
*****************************************;
*Entire code of the profile plot - figure 3
*****************************************;
   goptions reset=(axis, legend, pattern,
             symbol, title, footnote)
             rotate=landscape;
   goptions targetdevice=winprtc
            ctext=black ftext=swissb
            interpol=join noprompt;

   /*note short abbreviations used for the
    symbol options */
   symbol1 i=join v=+ h=1 cv=black l=1;
   symbol2 i=join v=- h=1 cv=black l=21;
   axis1 color=black width=2.0 style=1
         major=none minor=none
         order=(60 to 130 by 10)
         label=(a=90 h=1.2 'Weight in kg');
   axis2 color=black width=2.0 style=1
         offset=(2) label=(h=1.2 'Week');

  %let site1=Clinic_A;
  %let site2=Clinic_B;

  %macro gplot_m;
  %do I=1 %to 2;
  proc gplot data=idplot_m ;
     where site=&I and ((sdwt<pctlwt5 or
                       sdwt>pctlwt95);
    plot weight * week=subjectid /
         haxis=axis2
         vaxis=axis1 skipmiss;
  title h=2 "Weight (kg) - Males -
             Outliers";
  footnote h=1 j=l "Profile Plots of
            Individual Outliers -site:
            &&site&I";
  run;
  %end; quit;
  %mend gplot_m;
  %gplot_m;
```

CONCLUSION

SAS/GRAPH software produces presentation quality custom graphics. The production of graphics to address specific health-measurement quality checks can be done very efficiently, especially where many sets of these graphics are regularly needed.

The ANNOTATE data set is very useful when these graphics need to be updated later. For one-time only customizations, use the interactive graphics editor by typing EDIT at the command line of the GRAPH window.

NEW TOOLS

Version 8 SAS/STAT® software now includes the BOXPLOT procedure for generating box plots. Detail information is found at the following SAS web site: http://www.sas.com/rnd/app/da/new/daunivariate.html

REFERENCES

Carpenter, Arthur L. and Shipp, Charles E. (1995), *Quick Results with SAS/GRAPH Software,* Cary, N.C.: SAS Institute Inc.

Friendly, Michael. (1991), *SAS® System for Statistical Graphics, First Edition*, Cary, N.C.: SAS Institute Inc.

Gilbert, Jeffery D. (1999), "Customizing SAS Graphs Using the Annotate Facility and Global Statements", in *Proceedings of the Twenty-fourth Annual SUGI Conference*, Cary, N.C.: SAS Institute Inc. pp. 1002-1005.

SAS Institute Inc. (1990), *SAS Language Reference, Version 6, First Edition*, Cary, N.C.: SAS Institute, Inc.

SAS Institute Inc. (1999), *SAS OnlineDoc®,Version 8, SAS/GRAPH Software: Reference*, Cary, N.C.: SAS Institute, Inc.

SAS Institute Inc. (1990), *SAS Procedures Reference*, Cary, N.C.: SAS Institute, Inc.

SAS Institute Inc. (1990), *SAS/GRAPH Software: Reference, Volume 1, Version 6, First Edition*, Cary, N.C.: SAS Institute, Inc.

SAS Institute Inc. (1990), *SAS/GRAPH Software: Reference, Volume 2, Version 6, First Edition*, Cary, N.C.: SAS Institute, Inc.

SAS Institute Inc. (1990), *SAS/STAT Software, Version 6, Fourth Edition*, Cary, N.C.: SAS Institute, Inc.

Svetkey, LP, Sacks, FM, Obarzanek, E, Vollmer, WM, Appel, LJ, Lin, P, Karanja, NM, Harsha, DW, Bray, GA, Aickin, M, Proschan, MA, Windhauser, MM, Swain, J, McCarron, PB, Rhodes, DG, Laws, RL, for the DASH-Sodium Collaborative Research Group (1999). "The DASH Diet, Sodium Intake and Blood Pressure Trial (DASH-Sodium): Rationale and design"*. J. Am Diet Assoc*. 99:8 (suppl): pp. S96-S104.

ACKNOWLEDGEMENTS

I would like to thank Lauren Haworth for her original code of the box plot. It challenged me to learn and use SAS GRAPH procedures and ANNOTATE data set for the first time.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Nadia Redmond
7470 SW Alpine Drive
Beaverton, OR 97008
Work Phone: 503-528-2474
Email: nredmond@jps.net