JITTERing and Other Graphics Macros for Exploratory Data Analysis

Lawrence H. Muhlbaier, Duke University Medical Center

ABSTRACT

There are many techniques for exploring the structure of data that are awkward to use with existing SAS* procedures, as pointed out so graphically by Paul Tukey at SUGI 11. This paper discusses implementation of a macro to make some of these methods of exploratory data analysis (EDA) more accessable to the SAS user. In particular, a SAS macro for JITTERing in one dimension, with or without grouping levels is provided. EDA macros and procedures available to users in the SAS procedures, contributed procedures, and SUGI proceedings are reviewed.

INTRODUCTION

Techniques for graphical display of data are taking on increasing importance in understanding data. Modern computing equipment allows us to collect data faster than we can analyze it. We once had the time to plot the data by hand and could become intimately familiar with it. Most of us no longer have that luxury. John Tukey, in EXPLORATORY DATA ANALYSIS (Tukey, 1977) started people looking at new methods of displaying data. Some of these, like boxplots, have been available for some time in the SAS* system. Others, such as draftsman's plots, are new and somewhat awkward to use in Version 5 of SAS/GRAPH*. Multivariate symbol plots (trees, stars, or faces) are not readily accessable.

Paul Tukey, in his presentation at SUGI 11, opened many eyes to the variety of graphical analyses that SAS users could use, but find hard to do. The SAS macros presented here are based on graphical displays in GRAPHICAL METHODS FOR DATA ANALYSIS by Chambers, et al (Chambers, 1983).

There are several of these procedures available in the SAS system, in the Supplemental Procedures, and in the SUGI proceedings. The SAS system provided a rudimentary boxplot in PROC UNIVARIATE (SAS Basics Manual). PROC SPLOT (Gerig, 1983) provides printer boxplots. The SAS/GRAPH ANNOTATE capability has been described by Benoit (Benoit, 1985) for creating hoxplots on a plotter, followed by Stock's (Stock, 1985) and Olmstead's (Olmstead, 1985) macros to produce box-plots. Benoit (Benoit, 1986) extends her use of the ANNOTATE facility to enhance a GPLOT scatterplot to display sunflower plots. Gugel (Gugel, 1985) describes procedures that he has written for providing multivariate profiles via STARs and FACEs, but gives no information on how other SAS users can obtain these procedures for use. Gugel's procedures require additional software and an IBM mainframe to run. The sources of these tools are diverse and the tools do not have the consistency of implementation that are common to procedures provided by SAS Institute. As SAS/GRAPH reaches a more interactive design, perhaps users will really be able to do exploratory data analysis as it was designed to be used.

JITTERING IN ONE DIMENSION

The histogram (PROC CHART; VBAR ...) is well known but is limited in the detail it can present. Some distortion occurs with the abrupt shifts in the bars, and the individual data points are inaccessable. Plotting the points in one dimension is fine if there are few points:

```
---X-X-XX-XXX--X--X---X---X-XX-XXXX-X-X-X-----
```

but the points are obscured for a moderate sample size or if there is clumping of the data. Jittering creates an accessory variable of random noise that can separate out the clumping. This "two" dimensional plot can show the distribution of the variable more clearly:

```
      X     X          XX          X   X
        X   X  X       X X   X    X   X
    X   X  X       X X  X  X  X X X X      X
-----------------------------------------------
```

Note that the vertical axis is not needed. Figures 1 and 2 show similar plots using SAS/GRAPH.

```
JITTERP X;
```

a statement form macro, creates a plot such as the above for variable X. X is required. Multiple variables can be requested using one call of JITTER and a printer plot or graphics plot can be requested. BY-variables, whose levels define separate pages of plots, are also allowed. Jittering can be performed for multiple levels of a GROUPing variable, in much the same fashion that a GROUP option is provided in PROC CHART. The user can specify the degree of jittering, which is useful to control the amount of "white space" between columns when GROUPing is used.

```
JITTERP x GROUP=y PLOT=2 HAXIS="0 TO 4";
```

produced the plots shown in Figure 3. PROC GPLOT; PLOT x*y/HAXIS='0 TO 4'; was used to generate the comparison plot in Figure 4. The complete text of the JITTERP macro is included in Figure 5.

SUMMARY

Jittering is a very useful tool for the graphical display of univariate and some bivariate data. There are many other techniques described in GRAPHICAL METHODS FOR DATA ANALYSIS that would be very useful in the SAS environment for understanding data. I encourage SAS

Institute to provide us with a graphics workstation to make these analyses easier. In the meantime, I invite SAS users to publish additional macros to make these display's easier to create.

Future Plans: JITTERING IN TWO DIMENSIONS

A scatter plot in two variables (PROC PLOT; PLOT X*Y) can be very useful to examine the relationship between the two variables, but what one or both of the variables have limited resolution or precision? The values clump and you see the familiar B or C or D to indicate 2 or 3 or 4 values at that spot. The problem is worst when the variables are ordered or binary, and not really continuous at all.

```
JITTERXY "Y*X=Z";
```

would provide a two dimensional plot of Y and X, displaying tbe value of Z at each x-y pair, and jittering the values of X and Y.

These macros are currently in a stage of evolution of features, and are expected to change in the future. For copies of these macros and subsequent updates, contact:

Lawrence H. Muhlbaier, PhD
Assistant Professor of Biometry & Medical Informatics
Assistant Professor of Experimental Surgery
Box 3865
Duke University Medical Center
Durham, North Carolina 27710

Electronic Mail:
BITNET: DGTDOC@TUCC
USENET: decvax!duke!mcnc!ecsvax!doc

References:

Benoit P (1985). Box-and-whisker plots using the ANNOTATE facility, SAS Communications, Winter 1985.

Benoit PMD (1986). Statistical graphics made possible by the SAS/GRAPH ANNOTATE facility under VMS*, Proceedings of the Eleventh Annual SAS Users Group International Conference. SAS Institute, Inc. Cary, North Carolina, pp 228-234.

Chambers JM, Cleveland WS, Kleiner B, and Tukey PA (1983). Graphical Methods for Data Analysis. Duxbury Press, Boston.

Gerig TM (1986). in SUGI Supplemental Library User's Guide, Version 5 Edition. SAS Institute, Inc, Cary, NC, pp 531-534.

Gugel HW (1985). STARS and FACES -- procedures for constructing graphical profiles of multivariate data, Proceedings of the Tenth Annual SAS Users Group International Conference. SAS Institute, Inc. Cary, North Carolina, pp 253-258.

Olmstead A (1985). Box plots using SAS/GRAPH software, Proceedings of the Tenth Annual SAS Users Group International Conference. SAS Institute, Inc. Cary, North Carolina, pp 888-984.

Stock DL (1985). Boxplots and more boxplots -- a generalized SAS macro, Proceedings of the Tenth Annual SAS Users Group International Conference. SAS Institute, Inc. Cary, North Carolina, pp 259-264.

Tukey JW (1977). Exploratory Data Analysis. Addison-Wesley, Reading, Massachusetts.

Figure 1

Frequency without Jitter

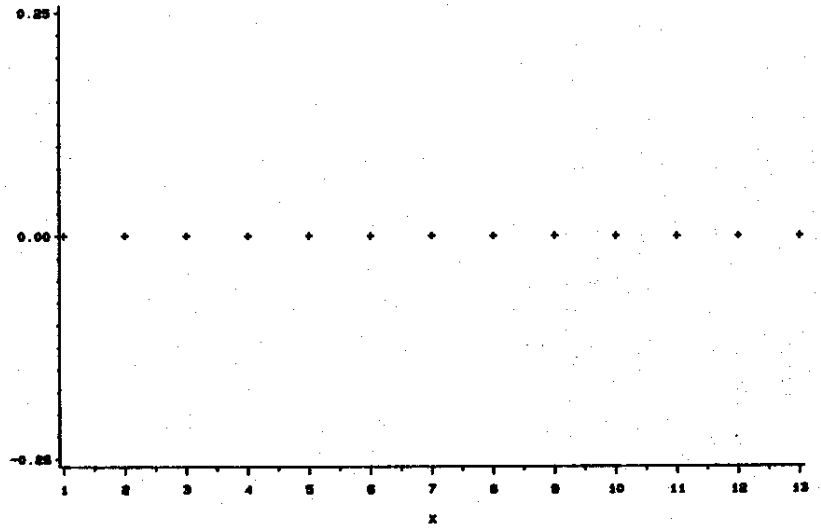
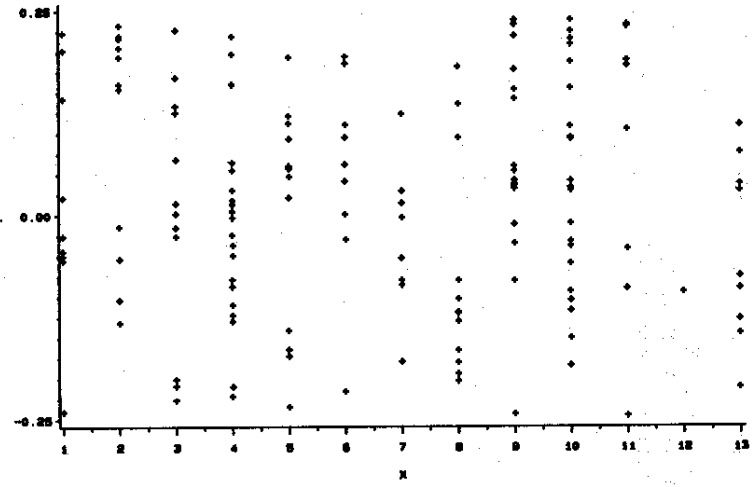


Figure 2

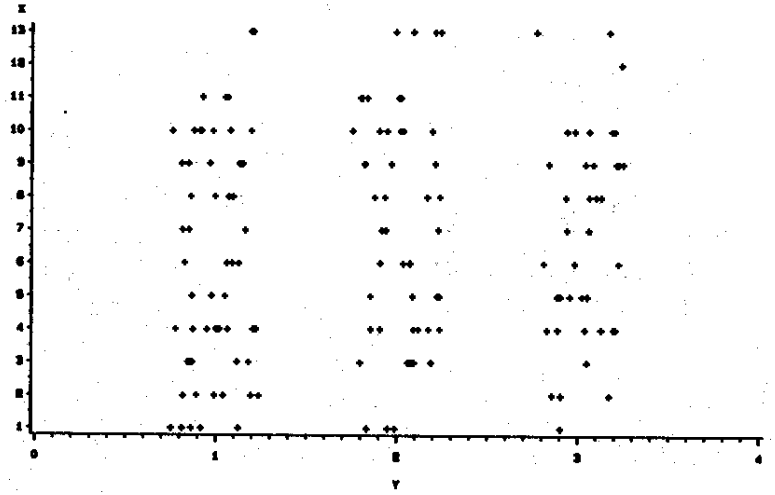
Frequency with Jitter



JITTERP x PLOT=2;

Figure 3

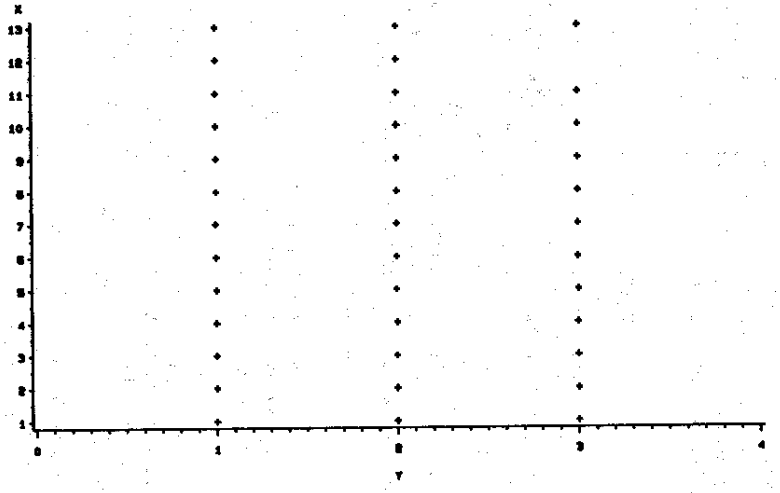
Frequency within Groups defined by Y, with Jitter



JITTERP x GROUP=y PLOT=2 HAXIS="0 to 4";

Figure 4

Frequency within Groups Defined by Y, without Jitter



*PROC GPLOT; PLOT x*y/HAXIS=0 to 4;*

```
 /*
JITTERP

JITTERP SAS MACRO PROCEDURE:  Jittering Plots in One Dimension

Reference:  Chambers JM, Cleveland WS, Kleiner B, and Tukey PA (1983),
            Graphical Methods for Data Analysis.  Duxbury Press, Boston.

JITTERP is a statement form macro for the creation of jittered plots of
one or more continuous variables.  JITTERP can create multiple plots on
one page classified by another GROUPing variable, and can create
separate multiple pages based on BY variable(s).

Usage:

JITTERP "list of variables to plot"
        DATA=input dataset (default=last created)
        BY="list of classification variables".  These are the
             variables whose values create different plots.
             "BY-variables".
        PLOT=plotting option
             1=line printer (default)
             2=graphics device using SAS/GRAPH PLOC GPLOT
             3=both line printer and graphics device.
        GROUP=optional numeric variable that groups the data to be
             jittered on a plot.  If present, the plots are rotated and
             the group variable defines the X-axis and the original
             variable's values are on the Y-axis.
        HAXIS=optional range of values for the grouping variable.
             May have one of the following forms:
                HAXIS="0 to 5" (default increment is 1)
                HAXIS="0 to 5 by 2"
                HAXIS="1,3,5"
             HAXIS is not used for PLOT=1, but is recommended
             otherwise.
        UNIT=(Optional) units of the GROUPing variable (Default=1).
             UNIT>0.  UNIT is the minimum distance between values of
             the GROUPing variable.
        JDEGREE=(Optional) degree of jittering (Default=100).  JDEGREE>0
             JDEGREE is a % multiplier of the spread of the jittering.
             Smaller values compress the jitter, larger spread it.
             JDEGREE=200 provides jitter that is 1 UNIT wide in the
             GROUPing variable, and thus provides no white-space for
             separation.

Author   :  Doc Muhlbaier, Box 3865 DUMC, Durham, NC 27710
            BITNET:  dgtdoc@tucc
            USENET:  decvax!mcnc!ecsvax!doc
Date     :  1986
Modified:   11 Dec 86 - formalized specifications.
            11 Jan 1987 - Coded.
            2 Feb 1987 - Revised Specifications & Code.
            7 Feb 1987 - Added %QUOTE to HAXIS to handle negatives.
 */

%MACRO JITTERP(PLIST,DATA=_LAST_,BY=,PLOT=1,GROUP=,HAXIS=,
        UNIT=1,JDEGREE=100)/STMT;
%LOCAL  lastds ;
%LET plist=%SCAN(&plist,1,'"');
%LET by=%SCAN(&by,1,'"');
%LET group=%SCAN(&group,1,'"');
%LET haxis=%SCAN(&haxis,1,'"');
RUN;
%LET  lastds =&sysdsn;
%IF &_lastds_^= NULL  %THEN
   %LET  lastds_=%SCAN(&sysdsn,1).%SCAN(&sysdsn,2);
%IF &unit<= %THEN %DO;
   %PUT ERROR:  UNIT <= 0 not allowed.;
   %GOTO thatsall; %END;
%IF &jdegree< %THEN %DO;
   %PUT ERROR:  JDEGREE < 0 not allowed.;
   %GOTO thatsall; %END;
%IF &plist=   %THEN %DO;
   %PUT ERROR:  Nothing to plot.;
   %GOTO thatsall; %END;
DATA  _dat_ ; SET &data(KEEP=&plist &by &group);
%* _j_ is distributed U(-.25,.25)*(JDEGREE/100);
 _j_=((UNIFORM()-.5)/2)*(&jdegree/100);
LABEL _j_='.';
%* to center on the value of group;
%IF &group^=  %THEN
   &group=&group+_J_*&unit;;
%IF &by^=  %THEN %DO;
   PROC SORT; BY &by; %END;
%IF &plot^=2 %THEN %DO;
   PROC PLOT;
   %IF &by^=  %THEN BY &by;;
   %IF &group=  %THEN %DO;
      PLOT _j_*(&plist)/VAXIS=-.25 TO .25 BY .25 VPOS=20;
      %END;
   %ELSE %DO;
      PLOT (&plist)*&group
      %IF %QUOTE(&haxis)^=  %THEN /HAXIS=&haxis ;
      ;
      %END;
   %END;
%IF &plot^=1 %THEN %DO;
   PROC GPLOT;
   %IF &by^=  %THEN BY &by;;
   %IF &group=   %THEN %DO;
      PLOT _j_*(&plist)/VAXIS=-.25 TO .25 BY .25;
      %END;
   %ELSE %DO;
      PLOT (&plist)*&group
      %IF %QUOTE(&haxis)^=  %THEN /HAXIS=&haxis ;
      ;
      %END;
   %END;
%thatsall: RUN; OPTIONS _LAST_=&_lastds_;
%MEND jitterp;
```