ABSTRACT

Boxplots are a useful tool for exploratory data analysis. This paper describes a macro written with the SAS® (SAS Institute, Inc., 1982) macro language which produces multiple pages of boxplots appropriately labeled for inclusion into reports or memos. The macro has extended capabilities over the previously documented boxplot procedure PROC SPLUS (Gerig, 1983) such as automatic fence plotting and a capacity for 22 boxplots per page. Labels for each plotted page as well as for each box may be easily changed to allow for production of report quality plots, and plots are routed to a user-specified device such as graphics terminal or graphics plotter.

GENERAL OVERVIEW

The rapid development of computer graphics in recent years may be seen as support to the old adage: "A picture is worth a thousand words", or as one might add, a thousand numbers. Graphical methods are being used increasingly as data exploration and summarization tools and, as such, are enabling the viewer to better discover the characteristics of the data by which he hopes to increase his knowledge and from which he may, perhaps, draw some conclusions.

Boxplots

Boxplots are among the more recently developed techniques which give the researcher a visual impression of the data at hand (Emerson and Streneo, 1983). Figure 1 shows the anatomy of a boxplot. The 'box' of the boxplot represents 50% of the data and is drawn to a Y scale; that is, the top aligns with the 75th percentile (upper fourth) and the bottom lies at the 25th percentile (lower fourth) of the distribution of values. Inside the box a line is drawn at the point corresponding to the median (50th percentile). The distance between the upper and lower fourths is called the fourth spread. When 1.5 times this distance is added to the upper fourth and subtracted from the lower fourth the locations of the fences have been found. Any points lying beyond these fences are, in some sense, outliers and subject to close scrutiny. The adjacent values are defined as the points nearest to and inside of the fences. Lines are drawn from the box to these points and are sometimes termed "whiskers". These boxplots, sometimes called box-and-whisker plots or schematic plots, can be useful in data examination and analysis, because they summarize the distribution of the data, giving a measure of central tendency (median) and spread (fourths and/or whiskers). Also, changes in the distribution are easily perceived when two or more boxplots are compared on one page.

Motivation for MACRO BOXPLOT

The motivation for MACRO BOXPLOT arose with the analyses of several small data sets where summarization in plots by the distribution means alone was not felt to be appropriate. It was desired to use nonparametric methods of analysis and corresponding methods for plotting.

FIGURE 1

Anatomy of a Boxplot

Outlier (Maximum)

Approximate Distribution 75th percentile (Upper Fourth)

Median

Approximate Distribution 25th percentile (Lower Fourth)

Minimum

Upper fence to indicate Outliers (Upper Fourth + 1.5 * (Fourth Spread))

Fourth Spread

Lower fence to indicate Outliers (Lower Fourth + 1.5 * (Fourth Spread))

••••••

Adjacent Value (greatest non-outlying value)

that would reveal both the central tendency and spread of the data. MACRO BOXPLOT was written to produce high quality graphs on a user-specified plotting device or graphics terminal. It provides several features not available with PROC SPLIT (Gerig), e.g., automatic fence plotting, a capacity for 22 boxplots per page, and the ability to use plotting devices. Labels for each plotted page as well as each box may be easily changed to allow for production of report quality plots. A more detailed comparison of MACRO BOXPLOT and PROC SPLIT will be given later. Single boxplots may also be obtained by using the PLOT option in PROC UNIVARIATE (SAS Institute, Inc.).

MACRO BOXPLOT

General Description

MACRO BOXPLOT is a macro written in SAS language and utilizing the SAS 1982 macro facility to plot multiple pages of report quality boxplots. The macro is designed for general use and can reside in a user-defined macro library.

There are four sections in the basic structure of the macro text: 1) macros, 2) data input, 3) variable initialization and 4) macro call and variable reinitialization. Data on the population of the ten largest cities of each of sixteen different countries (Emerson and Sirene) will be used to explain and illustrate the use of the four sections which make up MACRO BOXPLOT. By creating a boxplot from the ten city populations for each county (within a continent), general trends or oddities in city size between the countries may become apparent.

The MACRO BOXPLOT text for plotting this data is given in Figure 2 (Section 1 is only summarized), and the resulting plots are shown in Figures 4 and 5. Section 1, containing the macros text, should never need to be modified. However, appropriate changes to Sections 2, 3 and 4 can adapt MACRO BOXPLOT to other data sets. References in this paper to the Zeta 887 should be modified to correspond to one's own graphics device.

FIGURE 2
MACRO BOXPLOT

********************************************************************** ; ***********
BOXPLOT MACRO FOR PLOTS ON THE ZETA PLOTTER ********************;
**********************************************************************; ******************
PROGRAM BY: D. STOCK ****************************; ******************
ADVISORY COMMITTEE: J. BOLOGNESE ****************; ******************
PROGRAMMING ASSISTANCE: D. SHAPIRO **************; ******************
GRAPHICS ADVISOR: J. PERRY ****************************; ********************
WRITTEN FEBRUARY, 1984 ****************************;
REVISION 1 JULY 11, 1984: BOXES,DATA(BOX1) ************;
REVISION 2 DEC. 24, 1984: BOXES,DATA(BOX2) ************;
REVISION 3 MARCH 1, 1985: BOXES,DATA(BOX3) ************;
MACROS FOR BOXPLOTS IMMEDIATELY FOLLOW. *********************;
**********************************************************************; ***********
SECTION 1

**********************************************************************; ***********
*MACRO SETUP SETS UP DATA FOR PLOTTING.
*MACRO PRINTV INTERNAL TO *MACRO SETUP; PRINTS "BY" VARIABLES AND PAGE NUMBERS.
*MACRO SYMBOL ASSIGNED PLOTTING SYMBOLS.
*MACRO FORHAT ASSIGNING LABELS TO BOXES.
*MACRO TOPLOT PLOTS THE BOXES.
**********************************************************************; ***********
INSTRUCTIONS
**********************************************************************;
**********************************************************************;
**********************************************************************;
OPTIONS NODCENT; TITLE; PROC PRINTTO NEW UNIT=82;
*** READ DATA:ZZZZZZ.BOXES.DATA;DATA;***;
DATA A; INFILE DATA;
INPUT CI VI-V10;
DATA A1 SET A;
IF CI=1 OR CI=2 OR CI=4 OR CI=7 OR CI=9 OR CI=10 THEN CONTIN=1;
IF CI=3 OR CI=5 OR CI=6 OR CI=11 OR CI=12 OR CI=14 THEN CONTIN=2;
IF CI=13 OR CI=15 OR CI=16 THEN CONTIN=3;
CITY1=V1 VAL=W1 JOUTPUT;
CITY2=V2 VAL=W2 JOUTPUT;
CITY3=V3 VAL=W3 JOUTPUT;
CITY4=V4 VAL=W4 JOUTPUT;
CITY5=V5 VAL=W5 JOUTPUT;
CITY6=V6 VAL=W6 JOUTPUT;
CITY7=V7 VAL=W7 JOUTPUT;
CITY8=V8 VAL=W8 JOUTPUT;
CITY9=V9 VAL=W9 JOUTPUT;
CITY10=V10 VAL=W10 JOUTPUT;
PROC SORT; BY CONTIN CI;

**********************************************************************;

SECTION 2

19960
SECTION 3

%LET ONZETA = YES;
%LET PRINTBY = YES;
%LET PLOTVAR = VAL;
%LET CI VAR = CI;
%LET PAGEVAR = CONT;
%LET HIGHERLV = EU;
%LET DATASET = A;
%LET XLABEL = COUNTRY;
%LET YLABEL = POPULATION;
%LET TTEXT = FIGURE 3;
%LET TITLE =EUROPEAN COUNTRIES;
%LET FOOT = EACH BOX REPRESENTS THE POPULATIONS X 10E-5;
%LET FOOT1 = ;
%LET F1 = SWEDEN;
%LET F2 = NETHERLANDS;
%LET F3 = FRANCE;
%LET F4 = UNITED STATES;
%LET F5 = ENGLAND;
%LET F6 = ITALY;
%LET F7 = GERMANY;
%LET F8 =;
%LET F9 =
%LET F10 =
%LET F11 =;
%LET F12 =;
%LET F13 =
%LET F14 =
%LET F15 =;
%LET F16 =;
%LET F17 =;
%LET F18 =;
%LET F19 =
%LET F20 =
%LET F21 =;
%LET F22 =;
%LET F23 =;

%SETUP
%FORMAT;
%STOPLOT;
%LET PAGEVAR =;
%LET TTEXT = FIGURE 4;
%LET TITLE = AMERICAN COUNTRIES;
%LET F1 = CANADA;
%LET F2 = MEXICO;
%LET F3 = ARGENTINA;
%LET F4 = BRAZIL;
%LET F5 = UNITED STATES;
%LET F6 =;
%LET F7 =;
%FORMAT;
%STOPLOT;

SECTION 4

The user can insert the proper SAS code to read and format his data in Section 2 (see Figure 2, Section 2). The macro portion of MACRO BOXPLOT requires a single line for each of the values to be boxed on a single page. In other words, to create a boxplot for all the values of Sweden (CI=1) along side the corresponding box for the Netherlands (CI=2), the data input in Figure 2 required reformating as is schematically presented below.
FIGURE 3 (Continued)

<table>
<thead>
<tr>
<th>OBS</th>
<th>CI</th>
<th>VAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>val(1,1)</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>val(2,1)</td>
</tr>
<tr>
<td>10</td>
<td>.</td>
<td>val(1,10)</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>val(2,1)</td>
</tr>
<tr>
<td>20</td>
<td>2</td>
<td>val(2,10)</td>
</tr>
<tr>
<td>160</td>
<td>16</td>
<td>val(16,10)</td>
</tr>
</tbody>
</table>

Variable Initialization (Section 3)

The user must initialize the values of the macro variables in Section 3. These variables are defined in Table 1.

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>DEFINITION</th>
</tr>
</thead>
<tbody>
<tr>
<td>ONZETA</td>
<td>Set this variable equal to &quot;yes&quot; if the graphs will be plotted on the graphics device, and to &quot;no&quot; otherwise.</td>
</tr>
<tr>
<td>PRINTBY</td>
<td>If a printout of the BY variables is desired then this variable should be set to &quot;yes&quot;, otherwise it should be set to &quot;no&quot;. A printout of the BY variables can be useful to identify which subsets are graphed on which page.</td>
</tr>
<tr>
<td>PLOTVAR</td>
<td>This is the variable whose distribution is plotted as the box.</td>
</tr>
<tr>
<td>BOXVAR</td>
<td>This is the variable each value of which gets a box made. All of these boxes will be on one plot page.</td>
</tr>
<tr>
<td>PAGEVAR</td>
<td>This variable indicates pages of box-plots. It is like the last variable in a 'by' statement.</td>
</tr>
<tr>
<td>HIGHERLV</td>
<td>These variables are the higher level variables in the 'by' statement.</td>
</tr>
<tr>
<td>DATASET</td>
<td>This variable indicates which SAS dataset contains the data to be plotted.</td>
</tr>
<tr>
<td>XLABEL</td>
<td>Text for the X-axis label.</td>
</tr>
<tr>
<td>YLABEL</td>
<td>Text for the Y-axis label.</td>
</tr>
<tr>
<td>TTEXT1</td>
<td>Text for the first line of titles.</td>
</tr>
<tr>
<td>TTEXT2</td>
<td>Text for the second line of titles.</td>
</tr>
<tr>
<td>FOOT1</td>
<td>Text for the first footnote.</td>
</tr>
<tr>
<td>FOOT2</td>
<td>Text for the second footnote.</td>
</tr>
<tr>
<td>FOOT3</td>
<td>Text for the third footnote.</td>
</tr>
<tr>
<td>F0</td>
<td>This variable defines the space before the first box format. It should usually be set to ' '.</td>
</tr>
<tr>
<td>F1</td>
<td>X-axis label of the first box.</td>
</tr>
<tr>
<td>F2</td>
<td>X-axis label of the 2nd box.</td>
</tr>
</tbody>
</table>

The proper format is `%LET VARIABLE-YOURVARIABLE`. Extra blanks are not needed, but blanks are required between a series of variables or words. All variables must be specified. If the variable is not needed, blank spaces may be put before the semicolon. Some variables must have associated text. These are: ONZETA, PRINTBY, PLOTVAR, BOXVAR, PAGEVAR, and DATASET. If there is only one page to be plotted, then a variable to be used as the page variable should be defined and all values set to 1 before the macro is called. F1-F22 define formatted X variable labels and need text only if the Zeta plotter is used. F1-F22 are set so that the proper number or label will be printed under each box (This option is not available unless the program is run in batch). If F1-F22 contain = - 0 (or) then the values should be enclosed in single quotes.

Macro Call and Reinitialization (Section 4)

This section is actually used in two subsections; the first part sets up the data and the second part does the actual plotting. To call the first part, all the variables must be initialized as indicated for Sections 2 and 3 above. When all of the `%LET` statements in Section 3 are initialized, `%SETUP` is called to set up the data, after which `%FORMATX` is called to define the X-label formats. `%SETUP` is only called once unless the data set used in plotting is changed. `%FORMATX` need only be called once also unless F1-F22 are changed partway through the run.

F21 The plotting part of the macro is designed to enable changes to labels, titles and footnotes for every page. The plotting part of Section 4 is called with the following statements:

```lessen`%
%LET PAGENUM=1;%TOPLOT
```

These two statements must be used for each page (change the page number accordingly). When titles, labels or footnotes are changed, the corresponding `%LET` statement is retyped and put between the above two statements.
If the titles etc. will be the same for all pages or a series of pages the following shortcut statements might be useful:

MACRO SHORTCUT;
%DO P=1 %TO 10;
%LET PAGENUM=&P;
%TOPLOT %END;
%MEND;
%MSHORTCUT;

The beginning and ending numbers in the %DO statement correspond to the page series.

In summary, the whole Section 4 of MACRO BOXPLOT may be called by the following statements:

%SETUP
%FORMATX
%LET PAGENUM=1;
%TOPLOT
%LET PAGENUM=2;
%TOPLOT

In the example, the data have been subdivided by the geographical regions Europe, America and Asia. Only the first two pages (European and American countries) are called to be plotted and those plots are shown in Figures 4 and 5.

COMPARISON OF MACRO BOXPLOT AND PROC SPLIT

A comparison of MACRO BOXPLOT and PROC SPLIT is given in Table 2.

<table>
<thead>
<tr>
<th>Features</th>
<th>MACRO BOXPLOT</th>
<th>PROC SPLIT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graphs</td>
<td>Graphics Device, Printer</td>
<td>Graphics Terminal, Terminal</td>
</tr>
<tr>
<td>Output to</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiple boxes per page</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Multiple pages of multiple boxes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>SAS dataset easily specified</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Boxed variable easily specified</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Upper and Lower limit for Y-axis</td>
<td>Set by GLOT</td>
<td>user-specified</td>
</tr>
<tr>
<td>Limit for Y-axis</td>
<td>automatically</td>
<td></td>
</tr>
<tr>
<td>Titles and footnotes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Axes labels easily specified</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Formatted box labels</td>
<td>yes-easily</td>
<td>yes-easily</td>
</tr>
<tr>
<td>Graph printing selectability</td>
<td>yes-easily</td>
<td>yes-easily</td>
</tr>
<tr>
<td>Number of possible boxes per page</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>Mean plotted</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Far out values</td>
<td>differentiates from outliers</td>
<td>yes</td>
</tr>
<tr>
<td>Fence plotted</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Listing of page subgroup</td>
<td>available</td>
<td>yes</td>
</tr>
</tbody>
</table>

ADDITIONAL NOTES ON MACRO BOXPLOT

The quantiles in MACRO BOXPLOT are computed as indicated by Emerson and Streneo by using DEF=4 in PROC UNIVARIATE (SAS Institute, Inc.). However, it has been recently noted that the four options of quantile definition available with PROC UNIVARIATE are different from the quantile definition used in PROC SPLIT as well as from one another (Miller, 1982). Thus, slightly different boxes may result from the same data set depending upon the plotting procedure and quantile definition employed.

SUMMARY

MACRO BOXPLOT provides additional boxplot software which enables the user to quickly and easily obtain high quality plots from graphics devices. These plots would be suitable for data screening or for inclusion into reports and summaries.

ACKNOWLEDGMENTS

The author wishes to express gratitude to Mr. J. A. Bolognese for instruction in the use of SAS language, and the helpful suggestions and encouragement received during the writing of this paper. An additional thank you goes to Mrs. Mar-yonna Kersten for her expert typing services.

REFERENCES


For copies of the complete text of MACRO BOXPLOT or for further information, please contact: Donna L. Stack (WB D-216), Merck, Sharp & Dohme Research Laboratories, P.O. Box 2000, Rahway, NJ 07065-0914 - (201) 750-8686

263
FIGURE 4
EUROPEAN COUNTRIES

FIGURE 5
AMERICAN COUNTRIES

COUNTRY
EACH BOX REPRESENTS THE POPULATIONS
OF THE TEN LARGEST CITIES ( \times 10^8)

COUNTRY
EACH BOX REPRESENTS THE POPULATIONS
OF THE TEN LARGEST CITIES ( \times 10^8)