

Abstract

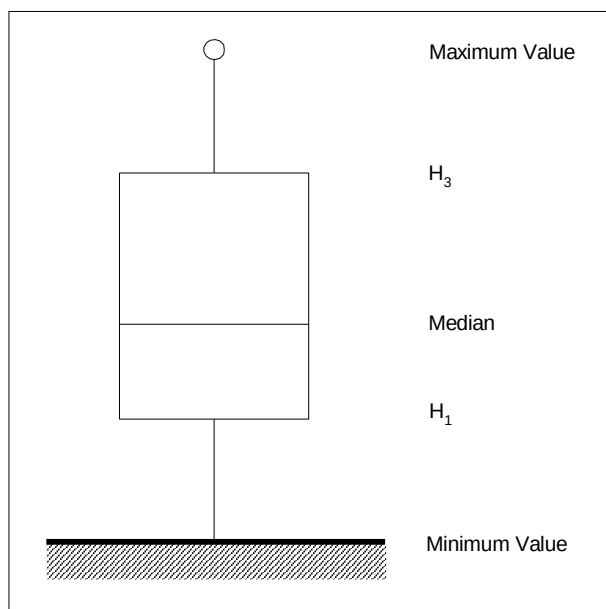
One regularly used graphical method of presenting data is the box-and-whisker plot. Whilst the vast majority of pharmaceutical statisticians produce their statistical output with The SAS System, until V8 producing box-and-whisker diagrams to a high standard was only achievable through custom macros and annotating graphics.

With the introduction of Proc BOXPLOT to the SAS/STAT module statisticians now have the power to produce several styles of box-and-whisker plots, thus enabling comparative displays of data groups to be easily presented. This paper examines the statistical capabilities of Proc BOXPLOT, the styles of box-and-whisker plot it can produce and points out the pitfalls those programming the procedure should consider.

Introduction

The box-and-whisker plot, referred to as a box plot, was first proposed by Tukey in 1977.

Figure 1: Box & Whisker Diagram, Tukey, 1977.



Tukey's original box-and-whisker plot used the less familiar *hinge* instead of upper and lower quantile measurements.

The whiskers were drawn all the way to the upper and lower observations, were a dot and hatched line represented these values, respectively.

Since then several variations and enhancements from the original definition have been published, notably McGill, Tukey and Larsen in 1978 which introduced the notched box plot for locating confidence limits of the median within the box-and-whisker plot.

As a graphical means of exploring distributions of qualitative data, many researchers, not just in the pharmaceutical industry, use this technique for examining data and presenting findings.

Commonly box-and-whisker plots are used to show trends of a distribution through time, or for side-by-side comparisons of groups of data.

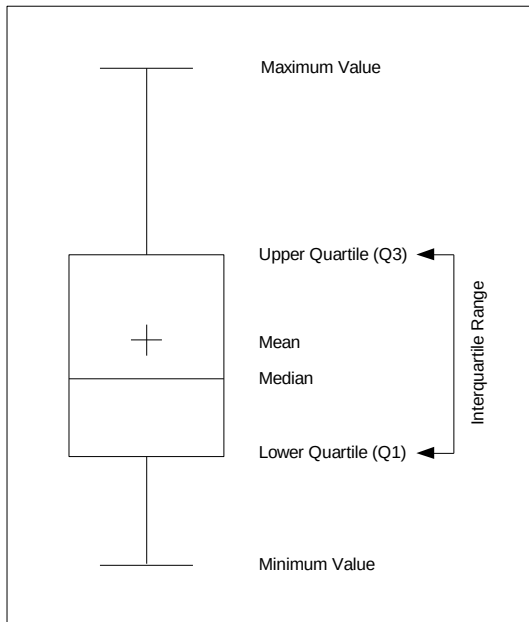
This paper examines how the pharmaceutical statisticians most commonly available tool, The SAS System, can produce box-and-whisker plots. It should be noted that in addition to Base SAS, the SAS/STAT and SAS/GRAPH modules are required to use the BOXPLOT procedure.

How are Box-and-Whisker Plot Constructed?

Box-and-whisker plots are constructed from the data groups mean, median, quartiles and outlying observations.

The two common styles of box-and-whisker plot used in statistics today are the skeletal and schematic plots.

Figure 2: Skeletal Box & Whisker Style



The skeletal plot (shown in Figure 2), draws its whiskers to the maximum and minimum values in the group of data plotted.

The upper and lower quartiles make up the boundaries of the box, the height of the box therefore represents the interquartile range (IQR).

The arithmetic mean of the group of data is plotted with a symbol, in this example a plus sign is used.

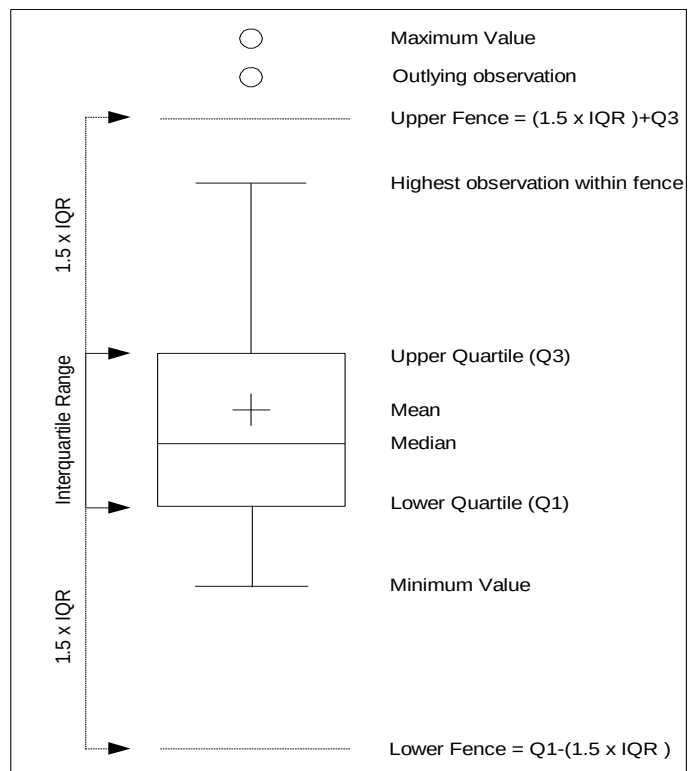
The schematic style of box-and-whisker plot is shown in Figure 3. This also draws the box height as the interquartile range, and plots a symbol at the arithmetic mean.

However, a skeletal plot determines where the lower and upper fences are. These are located at $1\frac{1}{2} \times \text{IQR}$ either side of Q3 and Q1. The fences are not actually plotted on the graph.

The whiskers are drawn to the value nearest to, but within, each fence. Any observations beyond a fence is plotted with a symbol.

In Figure 3 the lower whisker is drawn to the minimum value because it falls within the lower fence.

Figure 3: Schematic Box & Whisker Style



The SAS System's Proc Boxplot

Prior to version eight of SAS, box-and-whisker plots could be produced with Proc UNIVARIATE. Proc GPLOT could produce high resolution box-and-whisker plots as an interpolation option, although their is little flexibility over methodology and appearance.

Many variations of macros to draw plots with the annotate facility can be found, however such approaches require potential maintenance of code and are not necessarily supported by SAS Institute.

With the release of V8 the BOXPLOT procedure became production, allowing side-by-side box-and-whisker plots to be produced on groups of data.

The minimum syntax of the procedure is as follows:

```
PROC BOXPLOT      ;
      PLOT analysis-variable * group-variable / options;
RUN;
```

Where:

analysis-variable is the measurement of interest. This will be plotted on the y-axis.

group-variable is plotted on the x-axis.

options can be any of 89 options which control the box-and-whisker plot drawing methodology and appearance.

Both analysis and group variables are required.

Interestingly, the input data set must be sorted on the group variable, otherwise an error is returned for numeric variables and character variables, whilst appearing to take on a life of their own, are actually treated similarly to a class variable that has not been consolidated into unique groups.

Statistical Capabilities

The BOXPLOT procedure draws the box-and-whisker plots directly from the raw data. There is no method of saving the quantile and mean statistics generated by the procedure, used to in the plot.

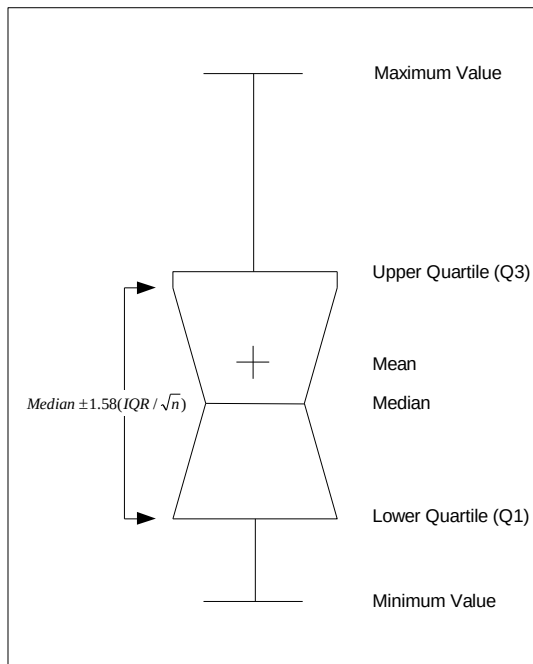
There is one option on the plot statement (NLEGEND) which allows the group sample sizes to be included in a legend on the plot.

Control over the methodology used to calculate the quantile statistics can be exercised, as in Base SAS procedures which calculate quantile statistics. The PCTLDEF=*index* option allows any of the following five definitions to be used. The default is 5:

Index	Definition
1	Weighted average x_{np}
2	Observation numbered closed to np
3	Empirical distribution function
4	Weighted average $x_{p(n+1)}$
5	Empirical distribution function with averaging

The NOTCHES option can be used to apply the McGill, Tukey and Larson variation on a box-and-whisker plot which draws approximate 95% confidence intervals of the medians.

Figure 4: Box & Whisker Plot with Notches



The notches in Figure 4 are shown between the arrows. The confidence interval of the median (identified by the notch) is located between:

$$\text{Median} \pm 1.58 \left(\frac{\text{IQR}}{\sqrt{n}} \right)$$

When like groups of data are plotted side-by-side and the notch boundaries do not overlap, it can be concluded there is a difference between the medians at around the 5% level of statistical significance.

In small sample sizes a notch may appear to fold back on its self, e.g. the lower quartile is higher than the lower confidence limit of the median.

Types of Box-and-Whisker Plots Produced by SAS

Skeletal

By default Proc BOXPLOT produces a skeletal plot.

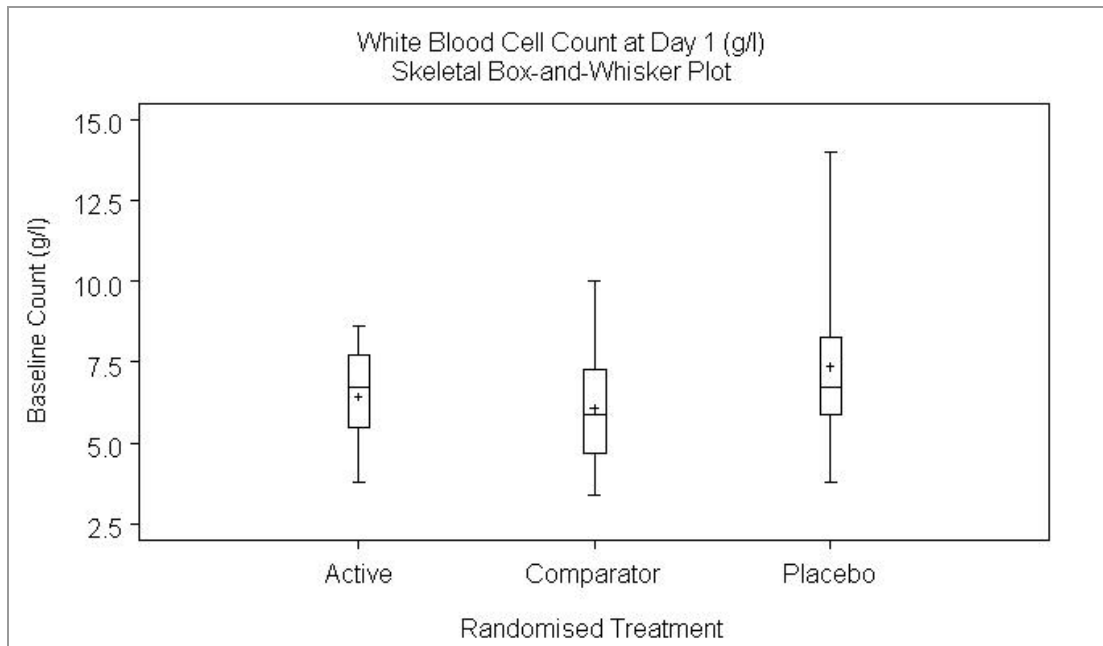
The following example takes a data set of white blood cell counts and generates box-and-whisker plots for each treatment group at Day 1.

```
* Data must be sorted by the grouping variable;
proc sort data=boxplot.wbc out=work.skeletal;
  by trtcd;
run;

* Generate skeletal plot for the BASELINE analysis variable;
proc boxplot data=work.skeletal(where=(day=1));
  title1 "White Blood Cell Count at Day 1 (g/l)";
  title2 "Skeletal Box-and-Whisker Plot";

  plot baseline*trtcd;
run;
```

Figure 5: Skeletal Box & Whisker Plot from Proc Boxplot



Schematic

A schematic plot can be requested by altering setting the BOXSTYLE option on the plot statement as follows:

PLOT analysis-variable * group-variable / BOXSTYLE= SCHEMATIC ;

Schematic ID

An extension of the schematic plot is to include a label to each of the outlying observations. For example, this may be a subject number or observed value.

The value from the first variable given in the ID statement is drawn against the outlying value.

The following code demonstrates requesting a schematic box-and-whisker plot with outlying values labelled with the subject identifier. The BOXWIDTHSCALE option is included to draw boxes with widths proportional to $\sqrt{n}$ in each group.

```
proc boxplot data=work.schematic(where=(trtcd=1));
  plot supdia*day / boxstyle=schematicid
                    boxwidthscale=0.5;
  id subjid;
run;
```

Figure 6: Schematic Box & Whisker Plot with Labelled Outliers

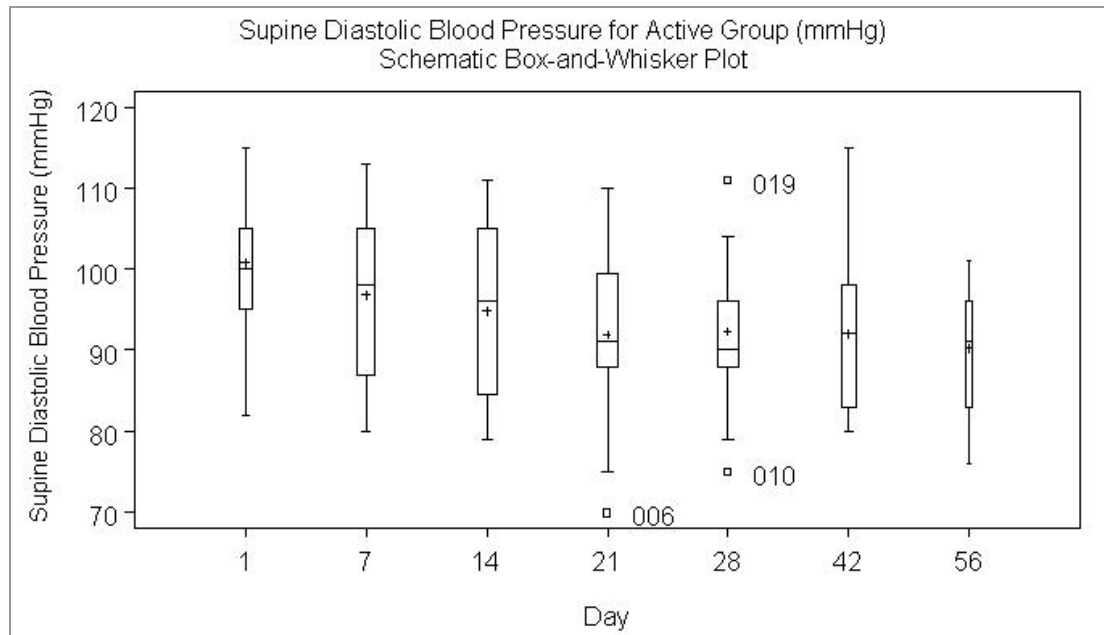


Figure 6 shows subject 006, 010 and 019 identified as observations falling outside either the upper or lower fences. Notice also the widths of bars varies through time as the number of subjects with observed data fluctuates.

Schematic ID Far

When the schematic `BOXSTYLE=SCHEMATICIDFAR` option is used in combination with an ID variable on the ID statement, only observations falling outside the upper and lower far fences are labelled.

Upper and lower far fences are determined with similar methodology as described in Figure 3 above, except the far fences are located at $3 \times \text{IQR}$ either side of the upper and lower quartiles.

Box Widths

All styles of box-and-whisker plots allow the `BOXWIDTHSCALE` option to draw each box width in proportion to the sample size in that boxes group.

The usage of `BOXWIDTHSCALE=` is as follows:

PLOT *analysis-variable * group-variable* / `BOXWIDTHSCALE= value`;

Where value determines the function used to calculate the proportion as follows:

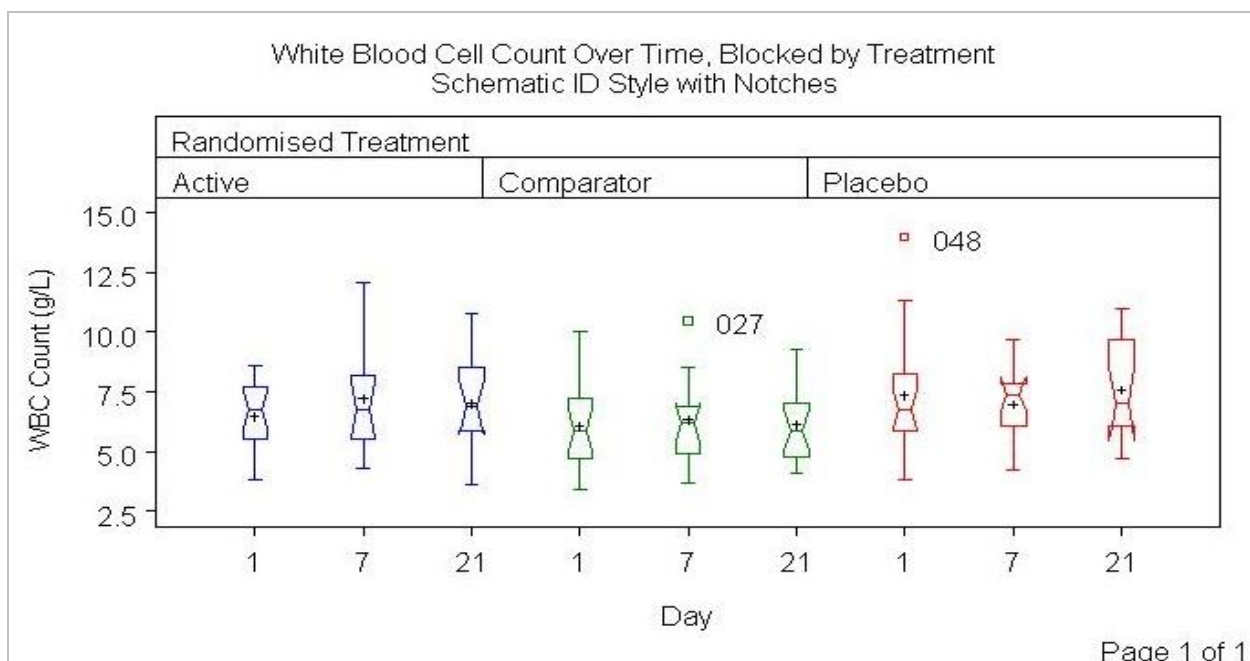
n^{value}

Therefore when value is 0.5 widths are proportional to $\sqrt{n}$ as shown in Figure 6. If desired, the width of all boxes can set to a percentage of the screen width or absolute number of pixels with the `BOXWIDTH=` option.

Grouping Box-and-Whisker Plots within Blocks

In addition to grouping data along the x-axis an additional blocking variable can be specified. This allows data to be grouped within groups. For example, supposing the white blood cell data are to be presented through time, for each treatment group plot can be achieved as demonstrated in Figure 7.

Figure 7: Box & Whisker Plot with Blocking and Notches



The syntax for including a blocking variable is as follows:

```
PLOT analysis-variable * group-variable(block-variable) / options ;
```

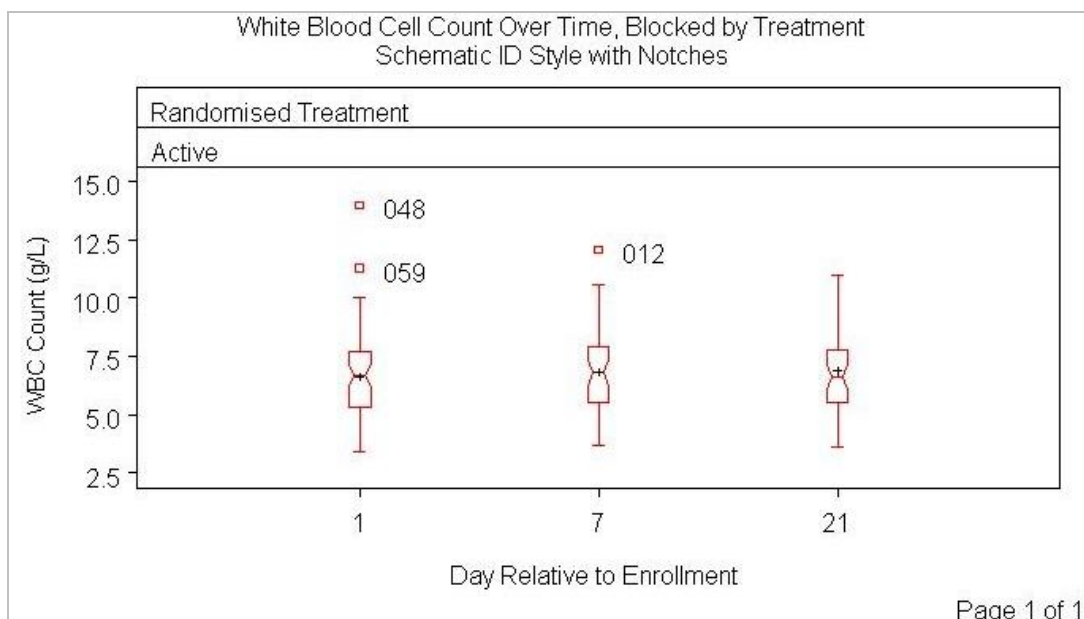
Therefore it would be expected that the following syntax would produce the figure above:

```
PROC BOXPLOT data=...;
  PLOT result * day (trtcd) / options ;
  ID subjid;
RUN;
```

However, because the group variable is numeric the input data set must be sorted on that variable, but to allow the blocks the data must also be sorted on the block variable, in this example TRTCD which represents the treatment code.

In actual fact, the code above (when all variables in the plots statement are numeric) produces Figure 8. Notice that only the first level of the block variable has been displayed, and within that block the whole data set has been used to draw the three boxes for days 1, 14 and 21.

Figure 8: Box & Whisker Plot with Incorrect Blocking



To avoid such a potentially dangerous mistake it is important to understand how the input data set must be constructed for Proc BOXPLOT when blocking is required. Firstly the group variable must be converted to a character field, this avoids the need for the input data set to be sorted on this variable first. The block variable is therefore the first sort variable with the character version of the day variable sorted within the block variable. The input data set now looks like the following sample:

VIEWTABLE: Work.Wbc						
	result	trtcd	day	dayc	colorb	subjid
117	6.4	Placebo	1	1	red	057
118	5.8	Placebo	1	1	red	058
119	11.3	Placebo	1	1	red	059
120	7.9	Placebo	1	1	red	060
121	7.7	Placebo	7	7	red	041
122	4.2	Placebo	7	7	red	042
123	6.6	Placebo	7	7	red	043
124	7.9	Placebo	7	7	red	044
125	7.5	Placebo	7	7	red	045
126	7.8	Placebo	7	7	red	047
127	9.7	Placebo	7	7	red	048
128	7.6	Placebo	7	7	red	049
129	7.3	Placebo	7	7	red	050
130	6.4	Placebo	7	7	red	051
131	5.7	Placebo	7	7	red	052

Notice that in Figures 7 & 8 the NOTCHES option is used. Also used is the CBOXES = (variable) option. This allows a variable in the input data set to hold the text of a valid SAS colour to which is used to draw each blocks' box-and-whisker plots in that colour.

The code used to produce Figure 7 is shown below, where DAYC is the character representation of the numeric DAY variable.

Some additional options are also used to give an indication of the flexible appearance options, of which there are far too many to discuss.

```
proc sort data=boxplot.wbc out=work.wbc;
  by trtcd dayc ;
run;

proc boxplot data=work.wbc;
  plot result * dayc(trtcd) / boxstyle=schematicid
                             notches
                             blockpos=2
                             vaxis=axis1
                             cboxes=(colorb)
                             noframe
                             pagenum='Page # of #';

  id subjid;
  label dayc ='Day';
run;
```

Conclusions

Box-and-whisker plots are a useful means of exploring and presenting data. There are several variations of presentation style any of which may be valid presentations of a data set.

Producing high resolution box-and-whisker plots in SAS has only previously been available as an interpolation option through Proc GPLOT with SAS/GRAPH. The level of control is limited through this route, requiring potentially complex macros to be required used to achieve the desired result.

Now Proc BOXPLOT allows box-and-whisker plots to be drawn in several styles of plot, whilst allowing extensive control over appearance itself. As Proc BOXPLOT adheres to the rules of SAS/GRAPH any graphic options are also honoured in the usual way.

Users of the procedure should be aware that the input data set must be sorted appropriately for grouping data. Particular attention must be given to the set-up of input data when blocking is also applied to the box-and-whisker plots, as potentially inaccurate figures could be misleading.

Contact

I very much welcome your feedback on this paper, or your thoughts on the subject of statistical graphics with The SAS System. Contact the author:

David Shannon
Amadeus Software Limited
Orchard Farm
Witney Lane
Leafield
Oxfordshire
OX29 9PG
United Kingdom

Phone: +44 (0)1993 878287
Fax: +44 (0)1993 878042
Email: david.shannon@amadeus.co.uk
Web: www.amadeus.co.uk

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.

© Amadeus Software Limited, 2003.