PharmaSUG 2012 - Paper DG10

Taming the Box Plot

Sanjiv Ramalingam, Vertex Pharmaceuticals, Inc.

ABSTRACT

Box plots are used to portray the range, quartiles and outliers if any in the data. PROC BOXPLOT can be used to obtain the necessary graphic but certain situations may arise where direct application of the procedure and use of the plethora of options that come along with it may not be sufficient to get the desired graph. One such application is discussed in detail. This should empower the reader upon understanding to create any box plot that the reader may wish. The methodology discussed assumes that the reader has at least a modicum of understanding of the annotate feature.

INTRODUCTION

In performing a clinical trial the primary objective is to test the safety and efficacy of a drug. This process involves comparing the drug of interest with a placebo and/or already existing drugs in the market. For a number of clinical tests that are performed it is desired to have a graphical summary of both the central tendency and variation in that clinical test. The boxplot is the desired graphic to portray the range and the quartiles of the data and possibly some outliers.

There is no direct approach to present box plots when all of the following tasks have to be implemented:

(A) Multiple treatments have to be presented for multiple visits with a box plot for each treatment per visit.
(B) The medians or any statistical parameter for each of the treatments have to be connected by separate lines.
(C) The counts (number of subjects) for each of the treatments per visit are to be displayed. Assuming that there are a number of visits to be displayed on a single page, a large number of visits necessitate the counts to be displayed on a slant with counts for each treatment on a separate line.

A step-by-step procedure is described below to create the boxplot described above. It is assumed that there are two treatments involved, namely ‘Drug A’ and ‘Placebo’.

METHODOLOGY

Step 1

The visits involved in the box plot are re-assigned if necessary so that they are in a sequential order represented by smaller integers such as 0, 2, 4, 6, 8, 10, 12, etc., if originally the visit numbers were represented as 100, 200, 700, and so forth. The visits for one of the treatments and in this example for DrugA are shifted by a small numerical value (0.5). Let this dataset be ‘dataX’ with parameter results under the variable name ‘aval’.

Step 2

The counts for each treatment per visit are obtained using PROC FREQ between the treatment and visits. The counts are then converted to macro variables. There are N number of ways of reaching a solution; the following approach was taken so as to create an array of macro variables that will be used in a PROC FORMAT.
Two datasets (DrugA and Placebo), containing records only for that treatment, are created using the output statements from the output of the PROC FREQ procedure.

Macro variables (nobs_A, nobs_pl) are created to represent the total number of observations in each of the datasets, DrugA and Placebo.

The following code accomplishes this task.

```
data _null_;
set DrugA end=eof;
if eof then
      call symput('nobs_A',trim(left(input(_n_,best12.))));
run;

data _null_;
set Placebo end=eof;
if eof then
      call symput('nobs_pl',trim(left(input(_n_,best12.))));
run;
```

Using PROC SQL and into: statements, an array of macro variables is created as shown below.

```
proc sql noprint;
select drugAcnt
into :drugAcnt1 -:drugAcnt&&nobs_A
from DrugA
;
select plcount
into :plcnt1 -:plcnt&&nobs_pl
from Placebo
;
quit;
```

Step 3

Formats are now assigned so that the counts can be displayed.

```
picture trtoth
0.5=" "
1.5=" "
2.5=" "
...
;
value cnts
0="Baseline,(N2=&plcnt1/N1=&drugAcnt1)"
1="Month 3,(N2=&plcnt2/N1=&drugAcnt2)"
2="Month 6,(N2=&plcnt3/N1=&drugAcnt3)"
...
other=[trtoth.]
;
```

The concept for the counts to be displayed on a slant with counts of each treatment on a separate line can be implemented using the annotate feature [1]. In this annotate dataset a different co-ordinate system, i.e. the procedure output area is used, unlike the data area that is used in the next step.

Step 4

The medians associated with each of the treatments are obtained using PROC MEANS. The structure of the 'Median' dataset will be as follows:

TRT	TRTN	VISITN	MEDIAN
DrugA	1	0	XX
DrugA	1	2	XX
DrugA	1	4	XX
…	…	…	…
Placebo	2	0	XX
Placebo	2	2	XX
…	…	…	…

The next step involves connecting the median from one visit to the next visit which is achieved using the annotate feature. The resulting annotate dataset is named 'anno1'. Part of the implementation involves automatically assigning the 'move' and 'draw' functions.

A. In the first step the parameters for the annotate data set are set for all records.
Line=20 assigns a dashed line. The dashed line is preferred for placebo and a solid line(line=1) is assigned to DrugA, which however is assigned later.
Only the data area will be used and hence XSYS and YSYS are set to 2.
The 'when' option is set to 'a' so that lines are drawn 'after' the figure is plotted. Function='move' is assigned to all records.

B. All records are output and the output statement is used again when the visit is not equal to the initial visit (visitn=0). The output dataset will be similar to the structure below.

TRTN	VISITN	MEDIAN	FUNCTION	COLOR	XSYS	YSYS	WHEN	LINE
1	0	2.1	move	black	2	2	a	20
1	1	4.2	move	black	2	2	a	20
1	1	4.2	move	black	2	2	a	20
1	2	3.1	move	black	2	2	a	20
1	2	3.1	move	black	2	2	a	20
…	…	…	…	…	…	…	…	…

C. The above dataset is sorted by treatment and visit. Using the FIRST. and LAST. option the last unique visit but with visit not equal to the initial visit(visitn=0) is assigned a 'move' function and the first unique visit but with visit not equal to the initial visit is assigned a 'draw' function. The last record with the function 'move' is deleted as it is not required. The contents of the dataset will then be similar to the dataset below.

TRTN	VISITN	MEDIAN	FUNCTION	COLOR	XSYS	YSYS	WHEN	LINE
1	0	2.1	move	black	2	2	a	20
1	1	4.2	***draw***	black	2	2	a	20
1	1	4.2	move	black	2	2	a	20
1	2	3.1	***draw***	black	2	2	a	20
1	2	3.1	move	black	2	2	a	20
...	...	...	...	...	...	...	...	...

Step 5

As part of the annotate dataset the legend is also constructed using the text, move and draw statements.

```
length function $8 text $100;
retain  xsys ysys '2'  style '' when 'a'  size 1;
function='label';
text="_ _ _ _  Median,N2= Placebo";
x=xx;
y=yy;
color='black';
output;
function='label';
text="______  Median,N1= DrugA";
x=xx;
y=yy;
color='black';
output;
function='move'; x=x1; y=y1; output;
function='draw'; x=x2; y=y1; line=1; output;
function='draw'; x=x2; y=y2; line=1; output;
function='draw'; x=x1; y=y2; line=1; output;
function='draw'; x=x1; y=y1; line=1; output;
```

Step 6

Since the legend has been created using the annotate feature the default legend that appears should be prevented from appearing. The following options are used in the legend statement to ensure that the default legend is invisible. One of the key options is the shape option. Shape can either be LINE or SYMBOL. The LINE option with a minimal value of 1 is chosen as this option omits the symbols[2].

```
legend1 down=4
label=(justify=l ' ')position=(bottom center outside )
value=(color=white)
shape=line(1);
```

Step 7

Assign formats so that for each visit, the treatments, DrugA and Placebo are represented as N1 and N2 respectively.

```
value vst
0="N2"
0.5="N1"
1="N2"
1.5="N1"
...
;
```

Step 8

The desired graphic can now be created using PROC BOXPLOT. The following options should be used.

`boxstyle=schematic` <To draw whiskers from the outer edge of the box to the highest value within the upper fence and lower edge to smallest value within lower fence>

`idsymbol=dot` <Symbol to denote outliers>

`cboxfill=white` <Specifies the interior fill colors for the box and whisker plot>

`cboxes=black` <Color for outline of box and whisker plot>

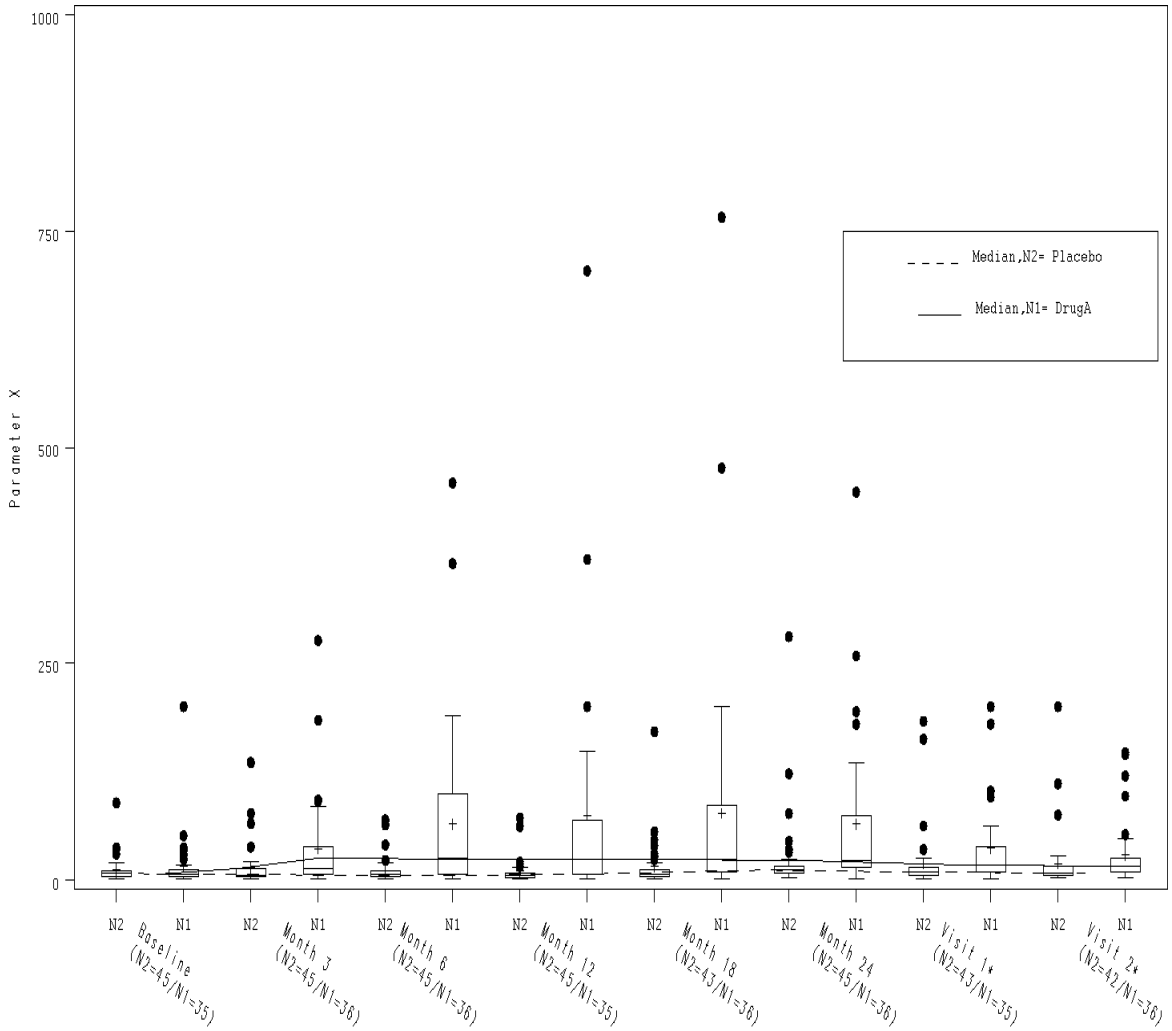
`nohlabel` <To suppress the label for the horizontal axis>

Symbols (symbol1 and symbol2) are used to denote the symbol to be used to represent the means for each of the treatments.

```
symbol1 value=plus color=black;
symbol2 value=plus color=black;
axis1 minor=none color=black label=(angle=90"ParameterX")
order=(0 to 1000 by 250);

proc boxplot data=dataX;
plot aval*visitn=trtn/anno=anno1 boxstyle=schematic
idsymbol=dot  cboxfill=white cboxes=black
nohlabel symbollegend=legend1 vaxis=axis1;
format visitn vst. ;
label visitn=" "
trtn=" ";
run;
```

The final figure will then be similar to the figure below.



CONCLUSION

A methodology has been shown that maximizes the use of the PROC BOXPLOT procedure. An understanding of this methodology should enable the reader to manipulate the box plot in many other ways.

REFERENCES

1. Rick Edwards , It's Not All Relative:SAS/Graph® Annotate Coordinate Systems. PHARMASUG 2007.
2. Wendi L. Wright, A Legend is Not Just a Legend, SAS Global Forum 2007.

CONTACT INFORMATION

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

Sanjiv Ramalingam
Vertex Pharmaceuticals, Inc.
130 Waverly Street
Cambridge, MA 02139
Email : sanjiv_ramalingam@vrtx.com