PharmaSUG China 2016 - Paper 72

Clip Extreme Values for a More Readable Box Plot

Mary Rose Sibayan, PPD, Manila, Philippines
Thea Arianna Valerio, PPD, Manila, Philippines

ABSTRACT

The BOXPLOT procedure creates box-and-whiskers plots of measurement that displays the mean, quartiles, minimum, maximum and outlier values for one or several groups. When handling interim or "dirty" data, there can be numerous outliers that may affect the readability of the plots. This issue can result into compressed plots and wide y-axis ranges, from which no clear conclusions can be drawn unless summary statistic values are presented.

The CLIPFACTOR option has the ability of clipping extreme values to produce readable and useful plots without affecting the summary statistic values. This paper aims to explore the CLIPFACTOR option of PROC BOXPLOT and present a macro that will compute for the optimal *factor* value based on the data. In addition, the macro also produces a data set which contains the extreme observations that were clipped from the plots.

INTRODUCTION

In clinical trial analysis, box plots are commonly used to visualize and compare variabilities and summary statistics within and between one or more treatment groups. These plots offer a quick and simple way to assess the shape and distribution of the data. The degree of dispersion, skewness and even outliers can easily be estimated just by looking at the plots.

However, when dealing with interim or “dirty data”, extreme values are often observed. Due to these extreme values (or otherwise known as outliers), the outputs often have wide axis ranges leading to overly compressed box plots. This defeats the practical use of box plots as summary statistics may not be easily estimated and the distribution of the data may be hard to distinguish in a compressed plot. This may prohibit comparison of results across different groups and may limit drawing sound conclusions from the plot.

Using CLIPFACTOR=*factor* option in PROC BOXPLOT is an excellent way to clip extreme values in the plot without actually affecting the summary statistics derived from the data. Outliers falling outside the clipping range derived from the specified factor are removed from the box plot. In effect, axis ranges can be adjusted to give more emphasis on the main body of the boxplot where most of the summary statistics can be found. Question is, how to derive the appropriate factor value for a given data? What are the customizations or clipping options that can be used to further improve the plot? How can we make use of the clipped values?

COMPRESSED PLOTS

As an illustration, below is one example of a compressed boxplot due to extreme values.

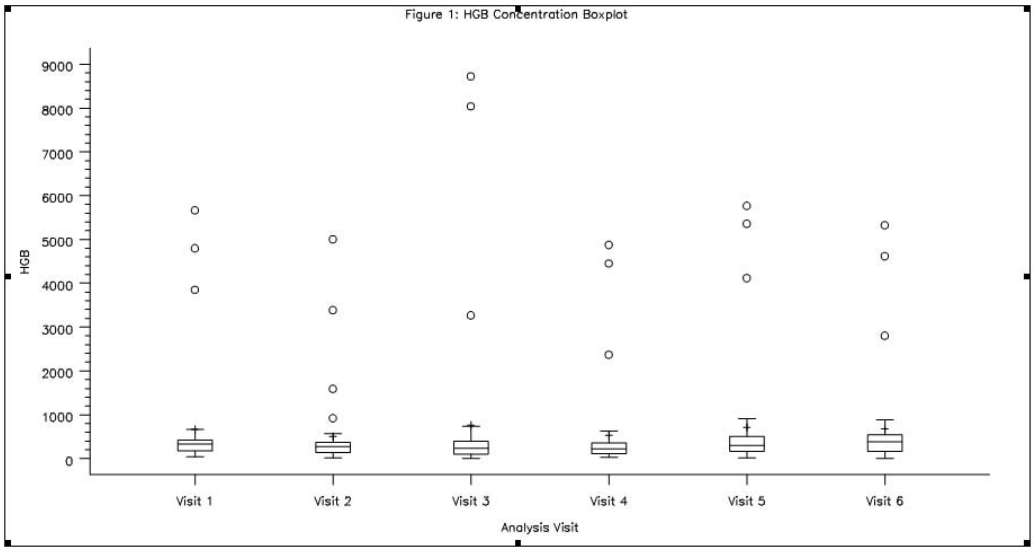

Figure 1: HGB Concentration Boxplot

As observed from figure 1, the plots are too compressed and the ranges are too wide leading to a difficulty in interpreting the results. Since the plots look similar with each other, comparing between groups would not be preferable. Even the summary statistics cannot be estimated just by looking at the plots. These are the problems that the CLIPFACTOR option resolves.

CLIPFACTOR AND ITS RELATED OPTIONS

These extreme outliers can be clipped from the plot to produce a more readable and useful output without changing the actual summary statistics. The factor value specified in the CLIPFACTOR option will determine the extent to which extreme values are clipped without affecting the main body of boxplot. Note that the valid factor is any value greater than 1. The minimum and maximum range for clipping values is defined below:

$$y_{max} = \overline{Q1} + (\overline{Q3} - \overline{Q1}) * factor$$

$$y_{min} = \overline{Q3} - (\overline{Q3} - \overline{Q1}) * factor$$

Observations outside of derived y_{max} and y_{min} values will be removed from the plot. The algorithm of deriving the first quartile $\overline{Q1}$ may vary, depending on whether summary statistics are taken per different groups or across all records. If summary statistics are to be taken per different groups, then $Q1$ will be derived by variable level first. The minimum of $Q1$ across all of the by variable levels will then be set as $\overline{Q1}$. If summary statistics are to be taken across all records, $\overline{Q1}$ is derived as the first quartile across all observations. The same algorithm was also followed in deriving the third quartile $\overline{Q3}$, only that the maximum of all $Q3$ among all variable levels was set as $\overline{Q3}$ when summary statistics are to be derived per by variable level. The formula for $y_{\max}$ was mainly used in deriving the formula for the optimal CLIPFACTOR and this is discussed further in the next sections.

For summary statistics with by variables specified (say TRT01A and AVISITN are the grouping variables), the codes below will be used in deriving $\overline{Q1}$ and $\overline{Q3}$:

```
proc means data=dummydata noprint;
      by trt01a avisitn;
      var aval;
      output out=summ_ q1(aval)=q1 q3(aval)=q3 mean(aval)=mean;
run;
proc means data=summ_ noprint;
      var q1 q3;
      output out=fsumm max(q3)=q3 min(q1)=q1 ;
run;
```

Aside from the CLIPFACTOR option, there are other clipping options available in PROC BOXPLOT to help further improve the plot. Here is the list of commonly used clipping options and their use:

- The CLIPLEGEND= option allows the user to indicate the label for the legend that specifies the number of boxes clipped. The label must not be more than 16 characters and must be enclosed in quotes.
- The CLIPLEGPOS= option specifies the position for the legend. Possible values may be TOP or BOTTOM.
- The CLIPSUBCHAR= option specifies the character which will be replaced with the number of boxes clipped in the label of legend.
- The CLIPSYMBOL= option allows the user to specify the symbol to mark plots with clipped values.
- The CLIPSYMBOLHT= option specifies the height of the symbol marker for outliers. If not specified, then the default height specified with the H= option in the symbol statement would be used.

SAMPLE OF CLIPPED OUTPUTS

Below are samples of codes applying the CLIPFACTOR option in the BOXPLOT procedure, followed by the resulting plot:

Figure 2 is a clipped version of Figure 1, applying the different clipping options in PROC BOXPLOT.

```
title1 'Figure 2: Clipped HGB Concentration Boxplot (Clipped version of Figure 1)';

proc boxplot data=dummy(where=(trt01a="Treatment A")) ;
       format avisitn visit.;
       plot aval*avisitn/ clipfactor=1.7 cliplegend='Boxes clipped=#'
       clipsubchar='#' cliplegpos=bottom clipsymbol=square noframe idsymbol=circle
       boxstyle=schematic cboxfill=white vaxis=axis1 haxis=axis2 outbox=boxval1 ;
run;
```

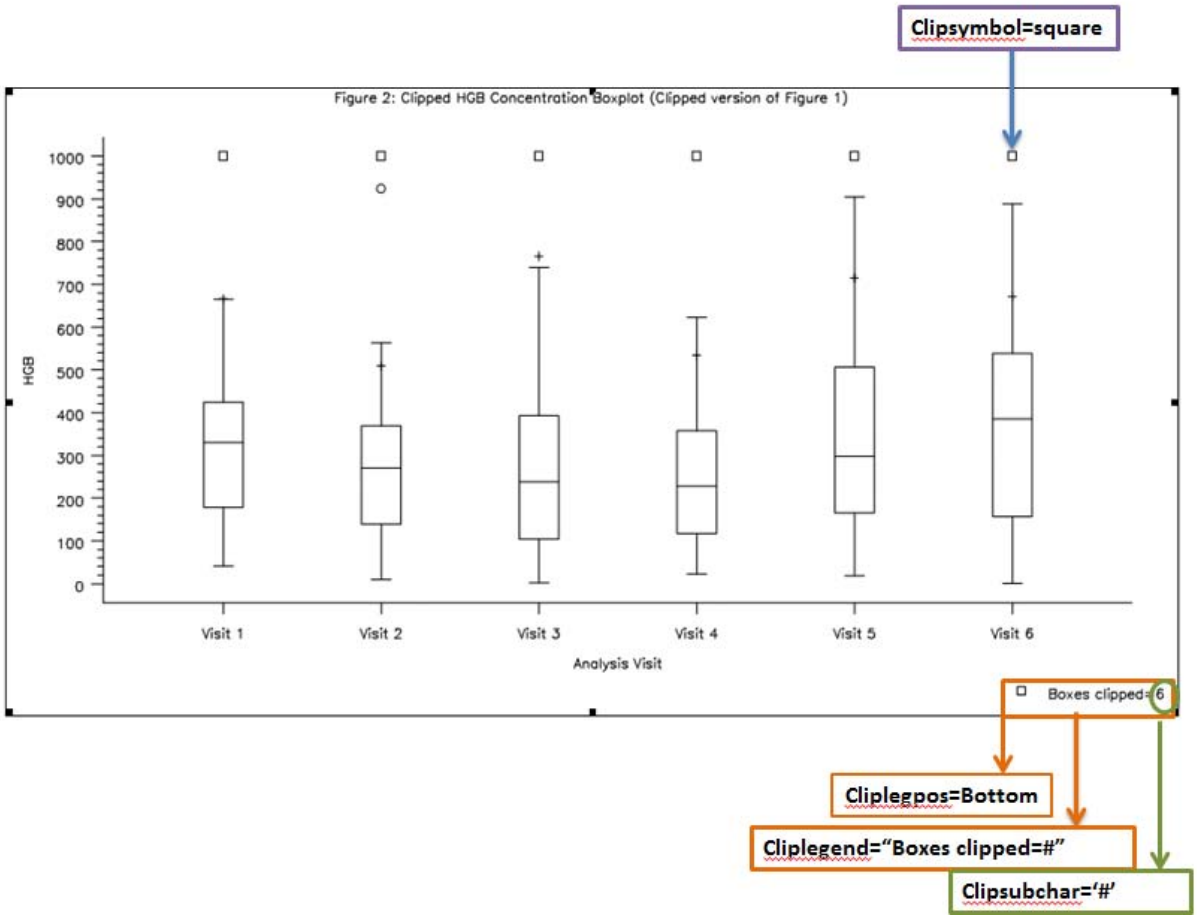

Figure 2: Clipped HGB Concentration Boxplot (Clipped Version of Figure 1)

In contrast to Figure 1, the boxplots are now more readable after using the CLIPFACTOR option. Since the extreme values were removed, the axis range has been adjusted to focus on the main body of the boxplots (which includes the median, quartiles, etc). This time, comparison among the different groups can be easily made because the difference among the summary statistics is more evident. The squares indicate which among the boxplots were clipped and also denote whether clipped values are extremely high or low value (depending on where the squares are placed in the axis). In this figure, since the squares are at the top of the plot, it means that extremely high values were clipped. The legend gives information on how many boxes were clipped. For this plot, outliers were removed from 6 boxes or the 6 visits.

It can be seen how the figure improved with the use of the CLIPFACTOR option but how should the appropriate *factor* value be determined?

%CLIPFACTOR MACRO

The CLIPFACTOR=*factor* option is an easy and effective way of removing extreme values from the boxplots. However, estimating the optimal factor value is not always easy. This issue led to the development of the %CLIPFACTOR macro which utilizes the same data and determines the optimal factor value. This macro uses the same formula in finding the maximum (y_{max}) and minimum (y_{min}) points as described earlier.

The first step is to compute for $\overline{Q1}$ and $\overline{Q3}$ using the same algorithm mentioned earlier. Then the formula commonly used in finding the upper whiskers for boxplots will be used.

$$upwhisk = \overline{Q3} + (\overline{Q3} - \overline{Q1}) * 1.5$$

where $(\overline{Q3} - \overline{Q1})$ is known as the interquartile range (IQR) (Stapel, 2010) and the value 1.5 determines how wide the clipping range will be.

However for this macro, the user is given the option to specify the extent of the clipping range which is stored in the *cutoff* macro parameter. If the *cutoff* value is not specified, the macro will set the value to either 1.5 or 3, depending on which *cutoff* value produces more outliers. Then the $upwhisk$ value will be substituted in y_{max} giving us the initial formula used to find the *factor* is:

$$factor = (upwhisk - \overline{Q1})/(\overline{Q3} - \overline{Q1})$$

The codes below will be used to derive the *factor* value:

```
**Calculate the interquartile range, lower and upper whiskers/fences**;

data whisk;
       set fsumm;
       iqr = (q3-q1);
       lowwhisk=q1-(&cutoff.*IQR);
       upwhisk=q3+(&cutoff.*IQR);
run;

**compute for the clipfactor using formula with Q1**;
proc sql noprint;
       select max(upwhisk) into :maxwhis
       from whisk;

       select min(lowwhisk) into :minwhis
       from whisk;

**get minimum value that is above the lower whiskers**;
       select min(aval) into :minbs
       from dummydata
       where aval > &minwhis.;

**get maximum value that is lower than upper whiskers**;
       select max(aval) into :maxbs
       from dummydata
       where aval < &maxwhis.;

       select round((&maxbs.-Q1)/(Q3-   Q1), .1) into :factor
       from fsumm;
quit;
```

Since the $upwhisk$ value is determined from the $\overline{Q3}$ value, numerous extremely low values will not be considered and this will still result to a wide axis range. Therefore, a second formula is introduced to account for the extremely low values by using $\overline{Q1}$. The *factor* will then be re-derived using the formula for y_{min}. The second formula used is:

$$factor = (\overline{Q3} - y_{min})/(\overline{Q3} - \overline{Q1})$$

The codes for deriving the *factor* value using the second formula is almost the same as the codes above except for the final section:

```
proc sql noprint;
       select round((Q3-&minbs.)/(Q3-Q1), .1) into :factor
       from fsumm;
quit;
```

The %CLIPFACTOR macro also considers reference ranges when supplied. The user can set the *refl* macro parameter to Y and specify the low and high reference ranges. These ranges will be considered in the computation of *factor* value so that the reference values will still be displayed in the axis. The specified reference ranges will be checked if it is within the computed maximum and minimum whiskers (*maxbs* and *minbs*). If the low reference range is lower than the computed *minbs*, then *minbs.* is set as the low reference range. If the high reference range is higher than the computed *maxbs*, then *maxbs* is set as the high reference range.

The optimal *factor* value derived is then stored in the macro variable *factor*, which will then be plugged in the CLIPFACTOR= option of the PROC BOXPLOT.

ADDITIONAL FEATURES OF THE %CLIPFACTOR MACRO

In addition to its main function of obtaining the *factor* value, the %CLIPFACTOR macro also outputs the details pertaining to the clipped values. A data set (*clipped_values.sas7bdat*) will be produced which contains information on the extreme values removed from the plot.

Image 1 is an example of *clipped_values* data set where the list of clipped values and its details are contained.

VIEWTABLE: Here.Clipped_values

	Subject Identifier for the Study	Actual Treatment for Period 1	Analysis Visit (N)	Analysis Value
1	041	Treatment A	1	4801
2	044	Treatment A	1	3853
3	045	Treatment A	1	5667
4	041	Treatment A	2	3393
5	044	Treatment A	2	5007
6	045	Treatment A	2	1594

Image 1: Clipped Values Data Set

Another data set containing the remaining values after the removal of outliers (input data set's name concatenated with '_') will also be produced.

To be able to make use of the clipped values, below are some suggestions on how the *clipped_values* data set can be used in the report:

- Annotate the range of clipped values in the plot

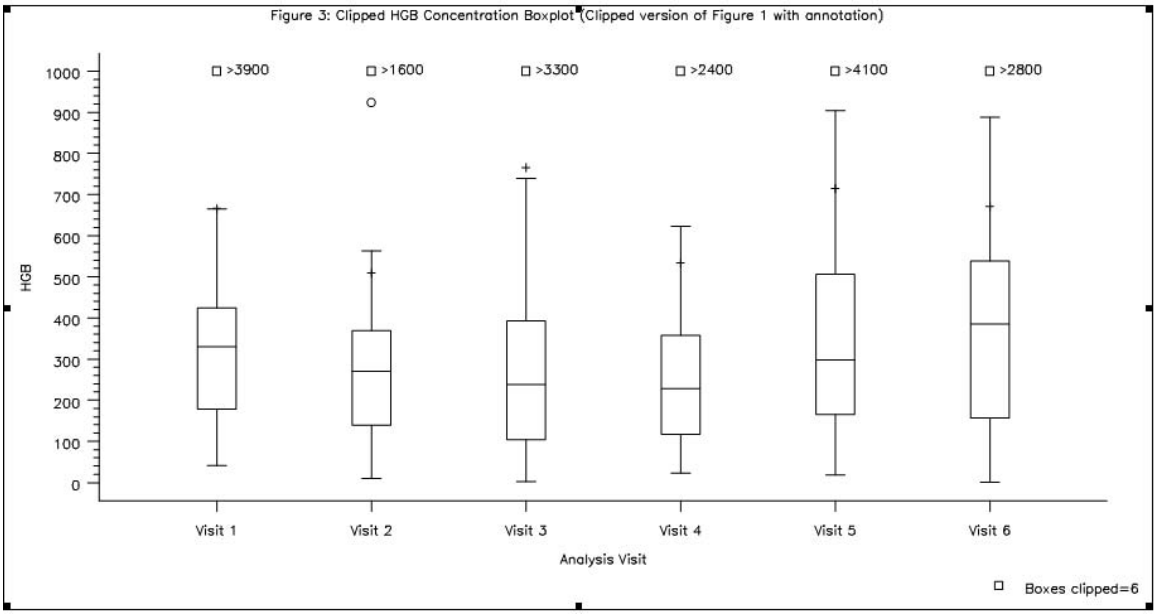

Figure 3: Clipped HGB Concentration Boxplot (Clipped Version of Figure 1 with Annotation)

Although the extreme values were removed from the plot, general information on these values can still be displayed. In this example, the ranges of clipped values are annotated to give an idea on how far the removed points are from maximum axis value.

- Add details pertaining to the outlier in the footnote

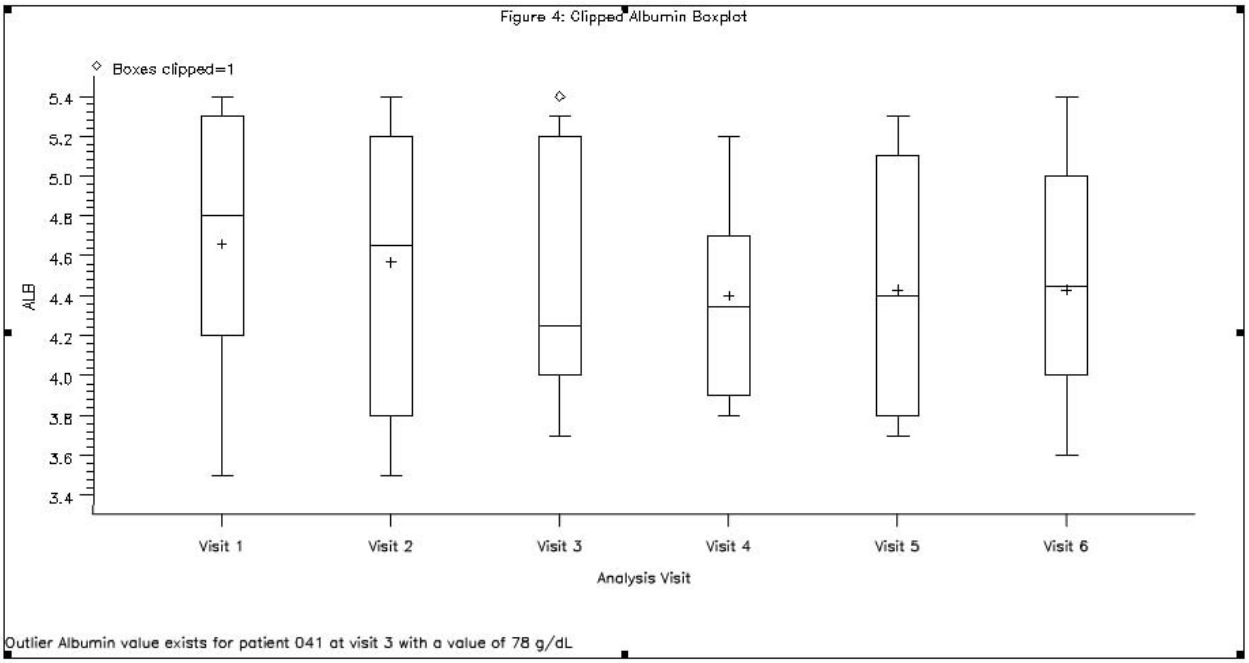

Figure 4: Clipped Albumin Boxplot

If there are very few extreme outliers (say less than 5 outliers), the *clipped_values* data set can be used to extract pertinent information and present them in the footnotes. Adding these details in the footnote will somehow give emphasis on these values.

- Use for data cleaning process

The list of extreme values contained in the *clipped_values* data set can be used to query the database for confirmation and resolution when needed.

CONCLUSION

Producing boxplots with severe outliers in the data usually results to compressed plots which are unreadable and almost not useful. Interpretation and comparison of these plots can be improved by clipping these extreme values using the CLIPFACTOR=*factor* option in PROC BOXPLOT. The %CLIPFACTOR macro proves to be helpful in determining the optimal *factor* value appropriate for the given data. This macro also provides information on the clipped values for other relevant use. This way, even if the outliers are removed from the plot, the values are not completely ignored.

REFERENCES

- Gemperli, Armin. 2009. “Zooming in on graphics”. Proceedings of the Pharmaceutical Users Software Exchange 2009. Lex Jansen’s Homepage. Available at http://www.lexjansen.com/phuse/2009/cs/CS04.pdf. Accessed 28 July 2016.

- SAS 9.22 Online product documentation. Available at https://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#titlepage.htm. Accessed 28 July 2016

- Stapel, Elizabeth. "Box-and-Whisker Plots: Interquartile Ranges and Outliers." Purplemath. Available from http://www.purplemath.com/modules/boxwhisk3.htm. Accessed 28 July 2016

ACKNOWLEDGEMENT

The authors would like to thank Rick Edwards, Jason Ralph Isturis and PPD Biostatistics and Programming Manila team who provided valuable inputs to improve this paper.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. You may contact the authors at:

Mary Rose Alpha Sibayan
PPD
Taguig City, Philippines
E-mail: MaryRoseAlpha.Sibayan@ppdi.com

Thea Arianna Valerio
PPD
Taguig City, Philippines
E-mail: Thea.Valerio@ppdi.com