Paper 190-2010

A Picture is Worth a Thousand Words: Integrated Visualization of Clinical Data

Vikash Jain, eClinical Solutions, A Division of Eliassen Group

ABSTRACT

Researchers in the Clinical domain are more interested in presenting their data and compiled text using visual graphics. This not only increases the clarity of the presentation, but also makes it easier for their target audience to understand the distribution of data and to draw inferences right away when they eyeball it at first glance. With this objective in mind, this paper will concentrate on how to produce various kinds of box plots with an integrated mean distribution and a reference line. We will be incorporating an array of BOXPLOT procedure options available to the user. We will also show how to annotate descriptive statistics and data outliers dynamically in a more efficient manner using SAS®9 PROC BOXPLOT options. Also covered is a discussion on powerful but simple techniques to incorporate the tabulated summary of statistics or a list report of outliers based on user business needs. These will be demonstrated in this paper on a case by case scenario with a visual presentation.

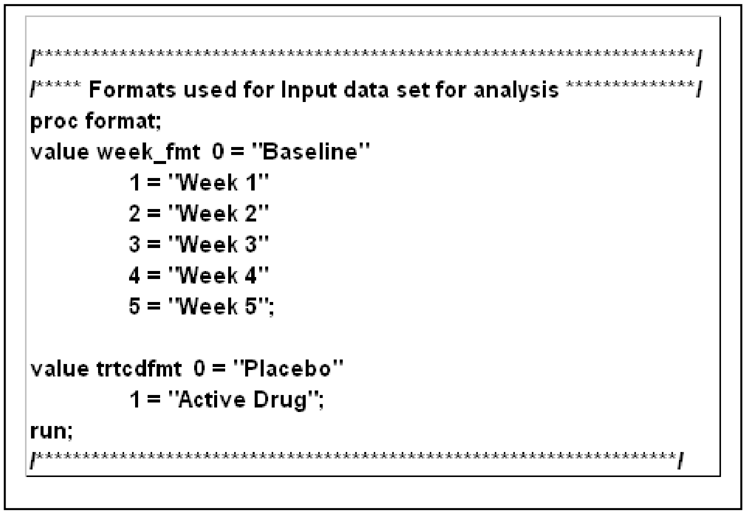
INTRODUCTION

In a clinical trials setting, laboratory data is one of the key elements of the safety profile for the drug agent being studied. Often clinicians want to see what happens to the patient's response by collecting lab samples before and after therapeutic intervention. Did certain lab parameter values drop below or above expected values that may cause any clinical concerns for the safety of the patients. This can be initially answered with visual presentation of particular labs of interest.

Having said that lab results are good candidates for graphical summarization, a graph can give us a quick visualization of increases and decrease in lab values over time as well as a comparison between treatment groups. They will also reveal data points which are outliers that bear further investigation. With respect to the scope of this paper our discussion will cover various graphical options which a user can incorporate in PROC BOXPLOT on a case by case basis.

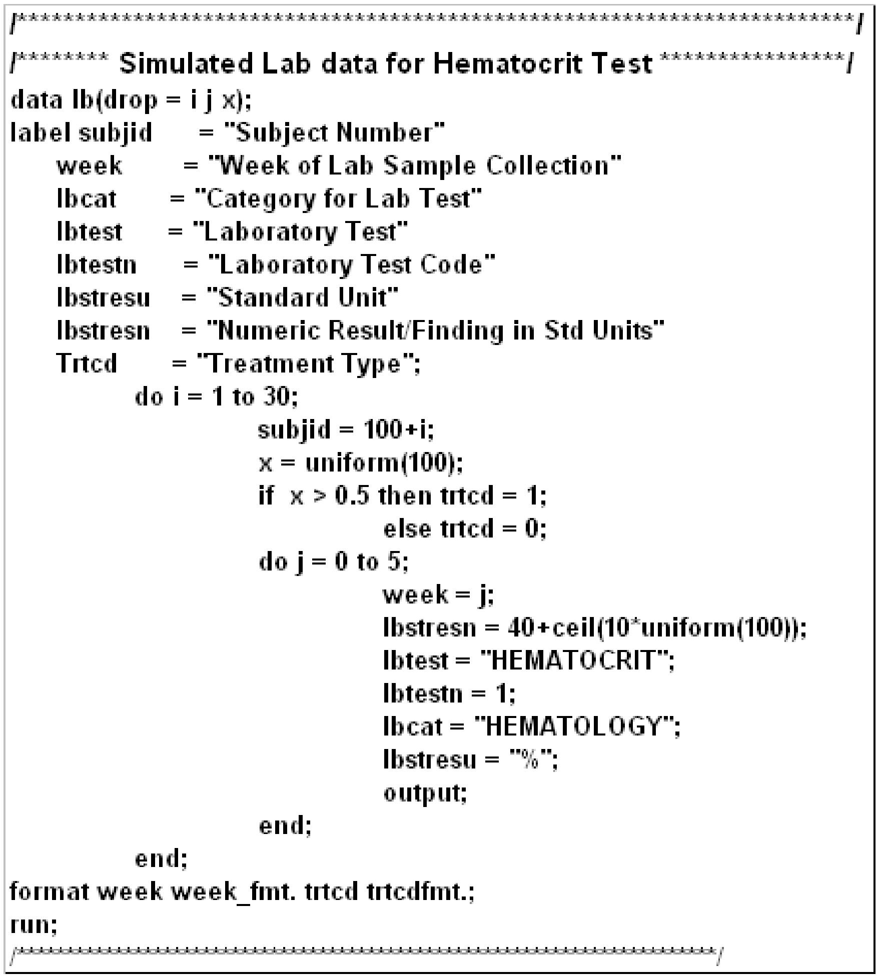
OVERVIEW OF DATA USED FOR ANALYSIS

The following data shown in Figure 1 will be used for illustrative purposes throughout this paper. The data contains a sample of 30 subjects which have been randomized to represent either an active or a placebo treatment. The data was collected on Hematocrit lab tests on baseline or Pre-Treatment (Week 0) and follow-up visits during study treatment (Week1 to Week 5) phase respectively from each of the patients which is been simulated with a code presentation of dataset below. Figure 1 shows the partial data set which is been used for analysis.



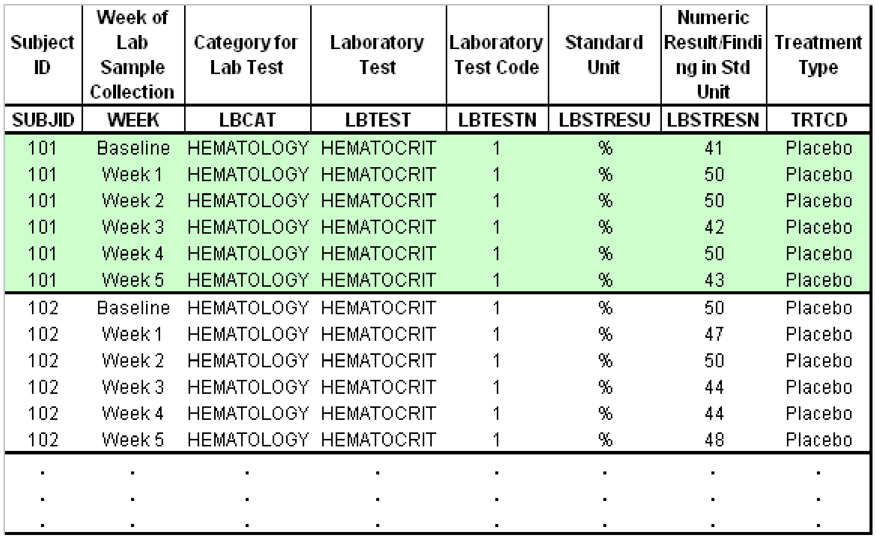
```
/**********************************************************************/
/***** Formats used for Input data set for analysis **************/
proc format;
value week_fmt  0 = "Baseline"
            1 = "Week 1"
            2 = "Week 2"
            3 = "Week 3"
            4 = "Week 4"
            5 = "Week 5";

value trtcdfmt  0 = "Placebo"
            1 = "Active Drug";
run;
/*********************************************************************/
```



```
/*****************************************************************************/
/******** Simulated Lab data for Hematocrit Test *****************/
data lb(drop = i j x);
label subjid     = "Subject Number"
      week       = "Week of Lab Sample Collection"
      lbcat      = "Category for Lab Test"
      lbtest     = "Laboratory Test"
      lbtestn    = "Laboratory Test Code"
      lbstresu   = "Standard Unit"
      lbstresn   = "Numeric Result/Finding in Std Units"
      Trtcd      = "Treatment Type";
            do i = 1 to 30;
                        subjid = 100+i;
                        x = uniform(100);
                        if  x > 0.5 then trtcd = 1;
                                    else trtcd = 0;
                        do j = 0 to 5;
                                    week = j;
                                    lbstresn = 40+ceil(10*uniform(100));
                                    lbtest = "HEMATOCRIT";
                                    lbtestn = 1;
                                    lbcat = "HEMATOLOGY";
                                    lbstresu = "%";
                                    output;
                        end;
            end;
format week week_fmt. trtcd trtcdfmt.;
run;
/***********************************************************************/
```

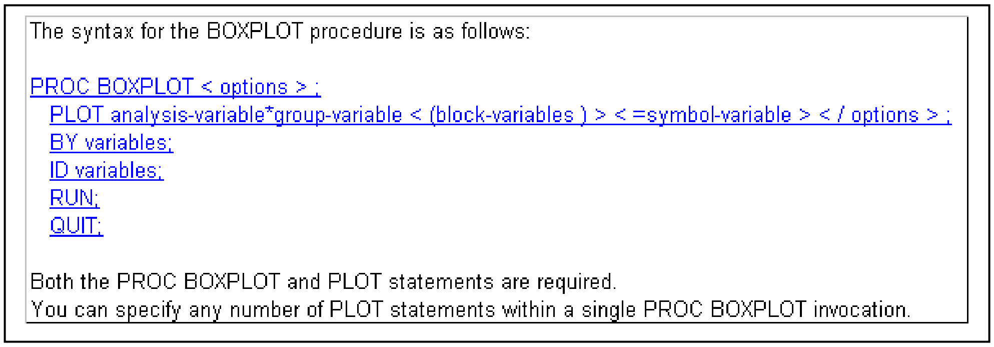
Figure 1: Simulated Lab Dataset (Partial data)



Subject ID	Week of Lab Sample Collection	Category for Lab Test	Laboratory Test	Laboratory Test Code	Standard Unit	Numeric Result/Finding in Std Unit	Treatment Type
SUBJID	WEEK	LBCAT	LBTEST	LBTESTN	LBSTRESU	LBSTRESN	TRTCD
101	Baseline	HEMATOLOGY	HEMATOCRIT	1	%	41	Placebo
101	Week 1	HEMATOLOGY	HEMATOCRIT	1	%	50	Placebo
101	Week 2	HEMATOLOGY	HEMATOCRIT	1	%	50	Placebo
101	Week 3	HEMATOLOGY	HEMATOCRIT	1	%	42	Placebo
101	Week 4	HEMATOLOGY	HEMATOCRIT	1	%	50	Placebo
101	Week 5	HEMATOLOGY	HEMATOCRIT	1	%	43	Placebo
102	Baseline	HEMATOLOGY	HEMATOCRIT	1	%	50	Placebo
102	Week 1	HEMATOLOGY	HEMATOCRIT	1	%	47	Placebo
102	Week 2	HEMATOLOGY	HEMATOCRIT	1	%	50	Placebo
102	Week 3	HEMATOLOGY	HEMATOCRIT	1	%	44	Placebo
102	Week 4	HEMATOLOGY	HEMATOCRIT	1	%	44	Placebo
102	Week 5	HEMATOLOGY	HEMATOCRIT	1	%	48	Placebo
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.

CASE BY CASE APPROACH ON PICTORIAL PRESENTATION OF BOX PLOTS

With respect to the objective of this paper we will first review the basic building block of PROC BOXPLOT. Then we will proceed to incorporate more enhancements to this by adding in the array of options available based on the case objective which will follow the subsequent sections below.



The syntax for the BOXPLOT procedure is as follows:

```
PROC BOXPLOT < options > ;
   PLOT analysis-variable*group-variable < (block-variables ) > < =symbol-variable > < / options > ;
   BY variables;
   ID variables;
   RUN;
   QUIT;
```

Both the PROC BOXPLOT and PLOT statements are required.
You can specify any number of PLOT statements within a single PROC BOXPLOT invocation.

For our discussion of this paper we will have “Numeric Result/Finding in Std Unit” as our analysis variable and “Week of Lab Sample Collection” as our group variable in the plot statements. The BY variable will be “treatment type” and the ID variable will be “Subject ID”.

ENVIRONMENT SETUP

The code below will help supplement the environmental setup for the generation of plots by defining the system and graphic options.

```
/****************************************************************/
/********* System and plot options defined for plots *************/

options Orientation=portrait nodate nonumber;

goptions reset=all colors=(black brown) device=sasemf gunit=pt
         htext=14 htitle=18  ftext=duplex rotate=landscape
         display hby = 16
         gsfmode=replace goutmode=append xmax=23cm ymax=16cm;

title;

/*** Global Box plot title ***/
Title1 j=c  "Box plot of Hemotocrit Lab results for each week by treatment
group";

/** Sorting data by BY variables **/
proc sort data= lb out= lb_rep;
   by trtcd week;
run;

/** Axis statement for the vertical axis (Lab Value) **/
axis1 order = (35 to 55 by 5);
/****************************************************************/
```

CASE 1: BOXPLOTS WITH BASIC PROCEDURE OPTIONS

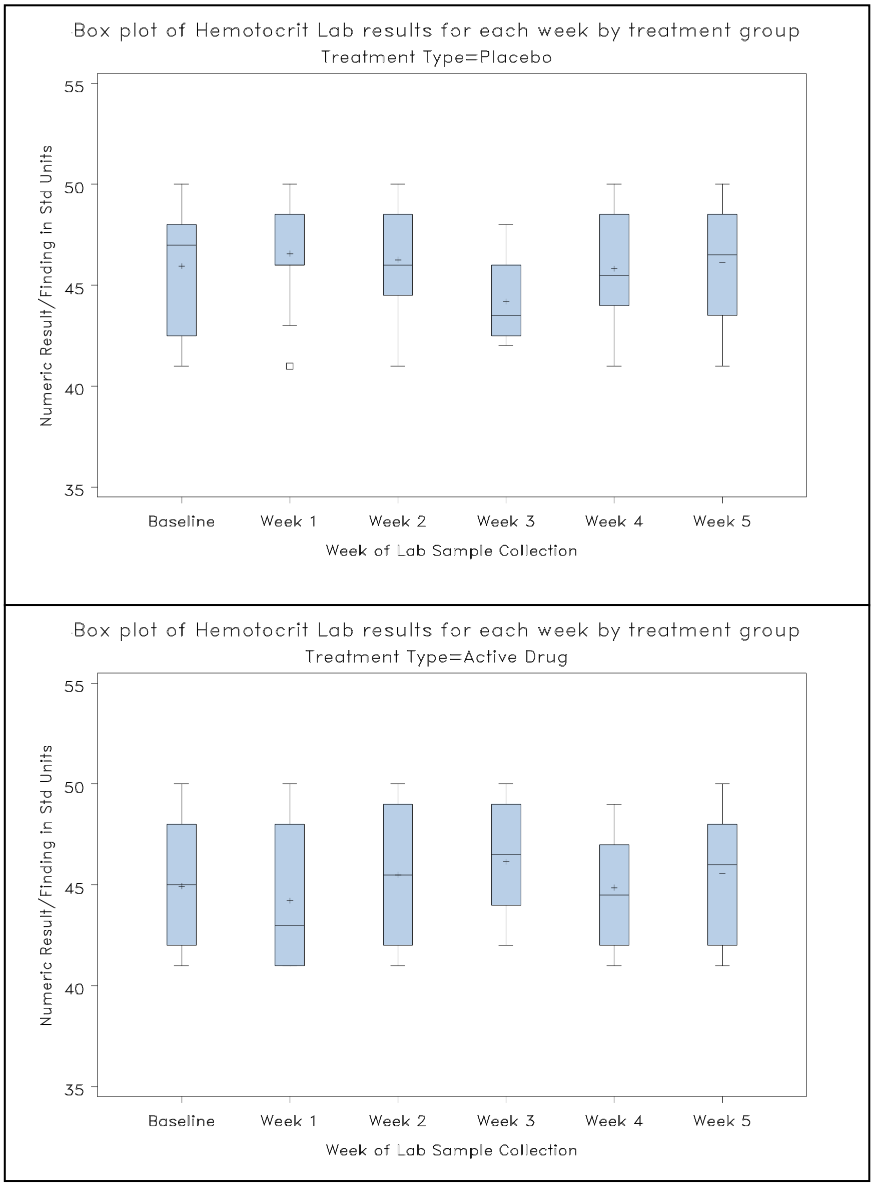


Figure 1: PROC BOXPLOT with basic procedure options

The code to produce the above plots in Figure 1 is:

```
proc boxplot data=lb_rep;
   plot lbstresn*week/VAXIS = axis1
                     BOXSTYLE=SCHEMATICID;
   by trtcd;
run;
quit;
```

In this case the plots generated in Figure 1 are based on very basic procedure options like the VAXIS option which defined the axis statement for the Numeric Results column with the BOX STYLE defined as SCHEMATICID.

CASE 2: BOXPLOTS WITH INTEGRATED MEAN PLOTS

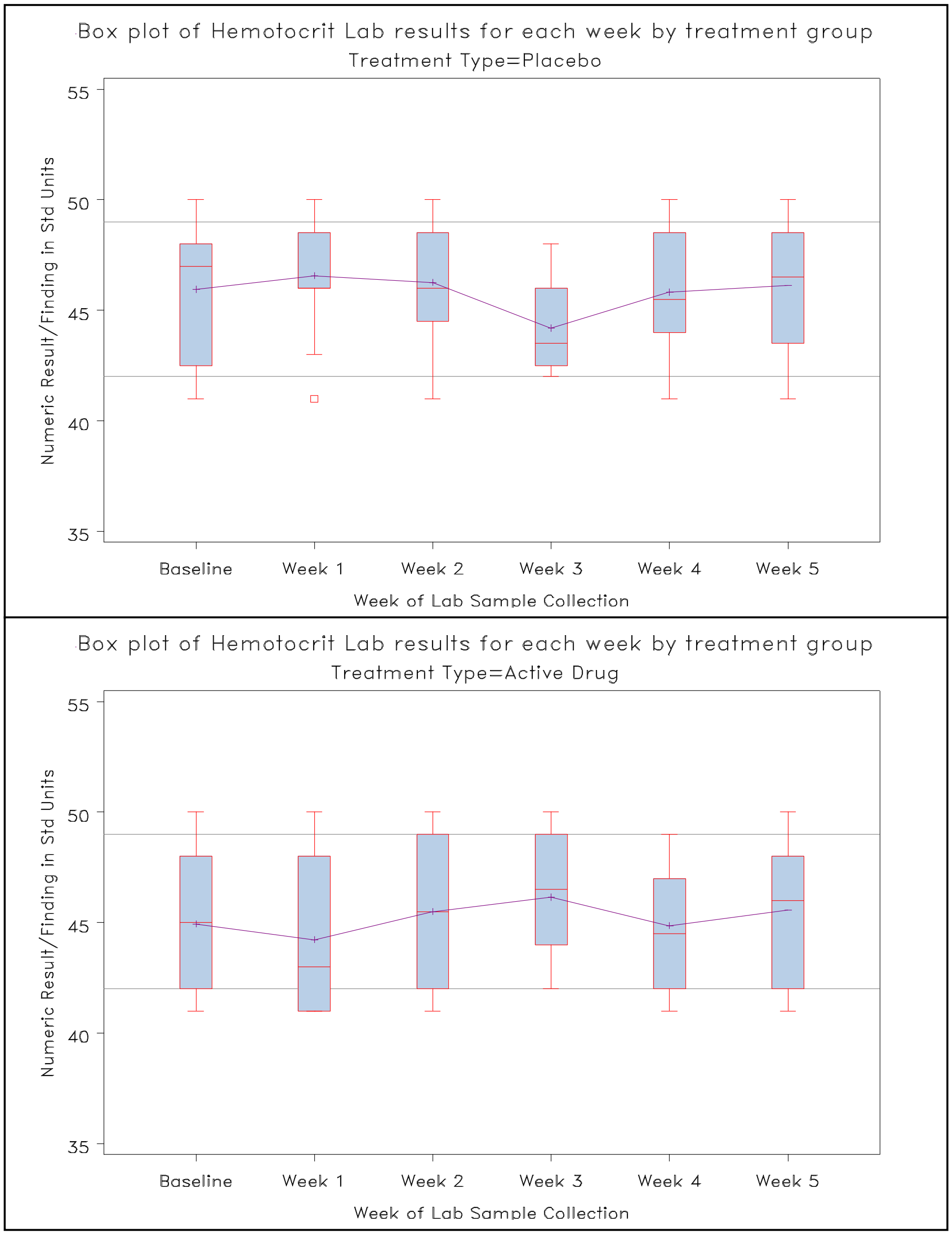


Figure 2: PROC BOXPLOT with vertical reference and means connect options

The code to produce the above plots in Figure 2 is:

```
symbol1 v=plus h=10 c=purple;
proc boxplot data=lb_rep;
   plot lbstresn*week
          / VAXIS = axis1 VREF= 42 49
            BOXCONNECT=mean BOXSTYLE=SCHEMATICID
            CBOXES= red;
   by trtcd;
run;
quit;
```

In this case the plots generated in Figure 2 display the box plots with mean plot integrated into one plot which shows the distribution of mean lab values at respective weeks with a reference on to distribution of normal range. This can be accomplished by using the BOXCONNECT and VREF options for means and showing reference lines on the plots. Also the symbol statement defined use of the "+" sign for the mean with a purple color and to connect the mean points within the boxes. The CBOXES option will display the color of the box plots outline with red (see above code).

CASE 3: MEANS/LINE PLOT GENERATED FROM PROC BOXPLOT

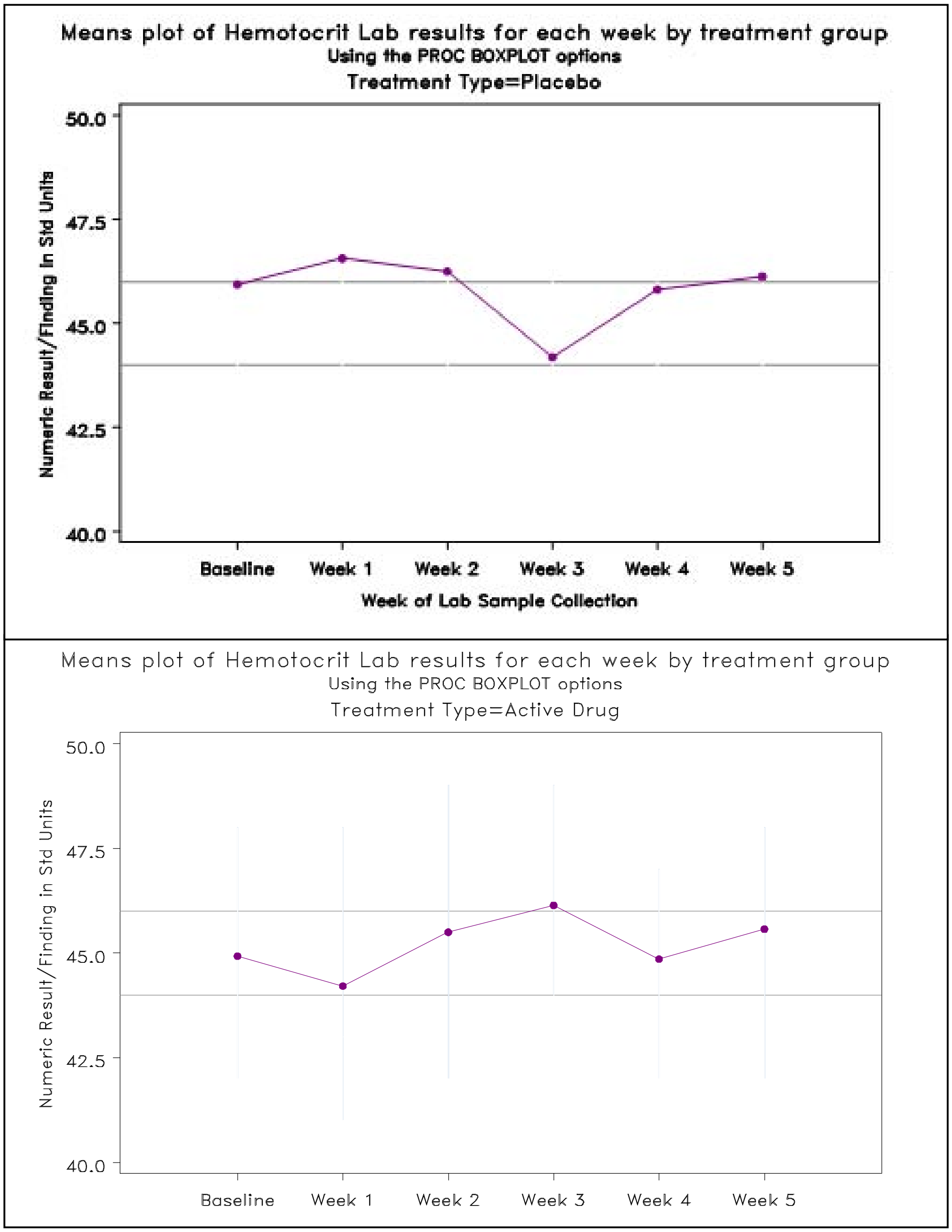


Figure 3: Means plot generated from PROC BOXPLOT

The code to produce the above plots in Figure 3 is:

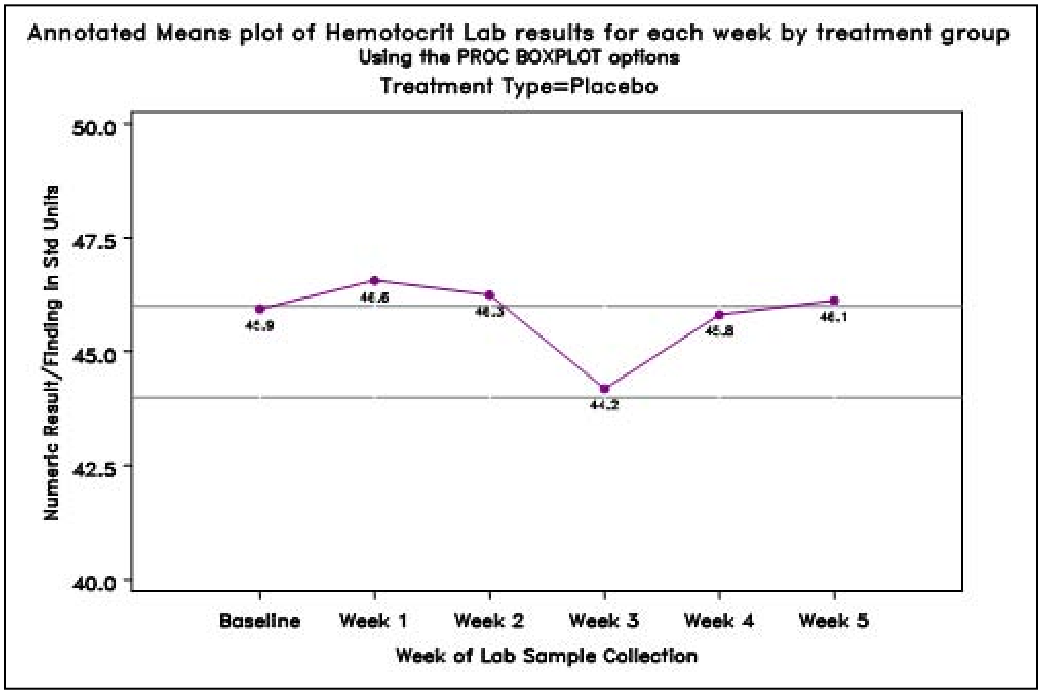
```
Title1 j=c  "Means plot of Hemotocrit Lab results for each week  by
treatment group";
Title2 j=c  "Using the PROC BOXPLOT options";

symbol1 v=dot  h=10 c=purple ;

proc boxplot data=lb_rep;
   plot lbstresn*week
          / VREF= 44 46
            BOXWIDTH=0
            BOXCONNECT=mean
            BOXSTYLE=SCHEMATICID
            CBOXES= white;
   by trtcd;
run;
quit;
```

In this case Figure 3, displays the mean plots which have been generated using the above PROC BOXPLOT code. The plots incorporate very simple but powerful options which help the user to generate the Means plots directly without calculating the means for each week and then plotting the points using PROC GPLOT. PROC BOXPLOT does this by itself and shows the results all in a one step procedure with various options used in it. The above BOX PLOT options used were BOXCONNECT = mean which will help connect all the mean points of the boxes at each week. The BOXWIDTH = 0 option helps to shrink each of the boxes to a simple vertical line and at last we attach the color of the box lines with the color of boxes CBOXES = white, which make the line invisible since the background color is also white. By this mean we can produce the Means plots by specifying certain BOX PLOT options.

CASE 4: ANNOTATED MEANS PLOT GENERATED FROM PROC BOXPLOT



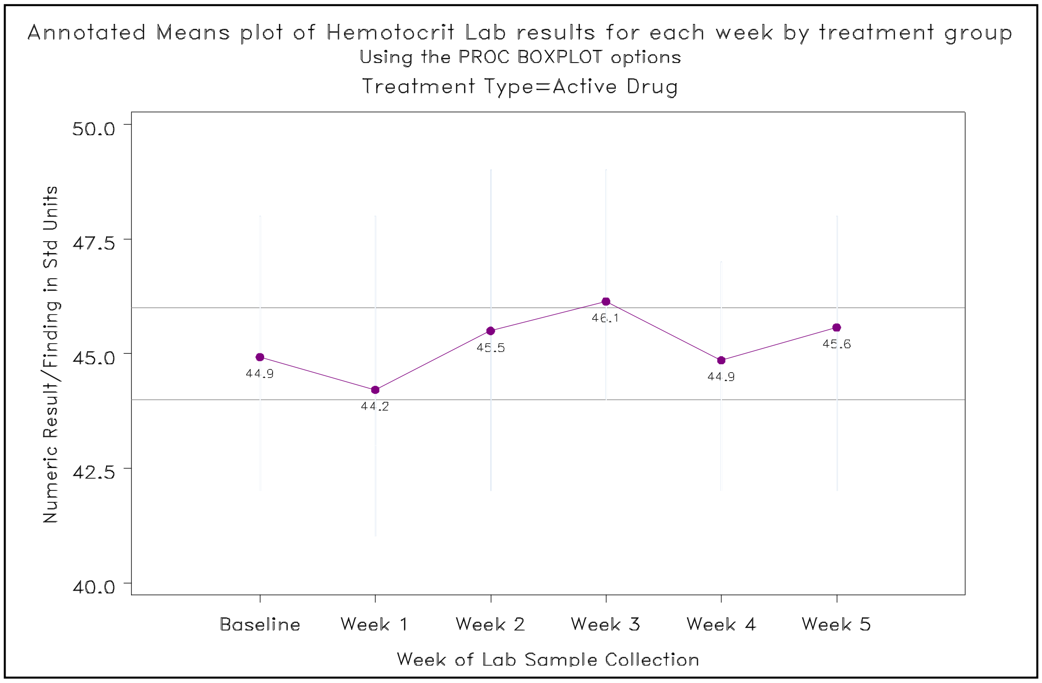


Figure 4: Annotated Means plot generated from PROC BOXPLOT

The code to produce the above plots in Figure 4 is:

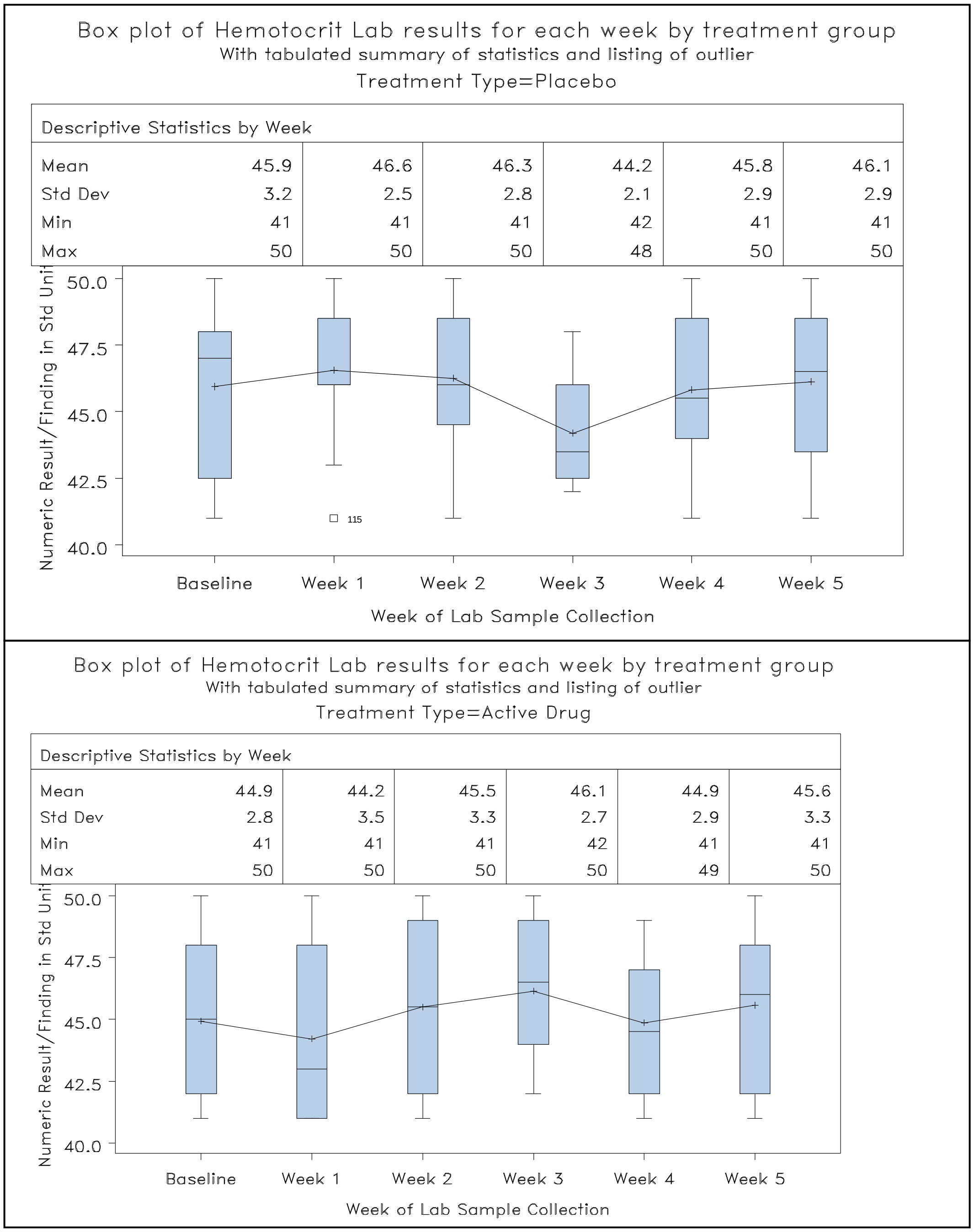
```
/* Calculate the mean points for each week */
    proc univariate data=lb_rep noprint;
          by trtcd week; var lbstresn;
          output out=desc_stats mean=mean n=n max = max min=min;
    run;
/* Create the annotate dataset which will be supplemented to PROC BOXPLOT for
annotation of mean values on to the plot */
    data final_anno(keep=xsys ysys function size text angle
                         style position y  x trtcd);
    set desc_stats;
         length function $8;
         retain xsys ysys '2' position "8" angle 0
                function 'label' size 1 style 'simplex'
                hsys '3';
         x = week; y = mean;
         text = compress(left(put(mean,4.2)));
    run;

Title1 j=c  "Annotated Means plot of Hemotocrit Lab results for each week by
treatment group";
Title2 j=c  "Using the PROC BOXPLOT options";

symbol1 v=dot  h=10 c=purple ;
proc boxplot data=lb_rep;
   plot lbstresn*week / ANNOTATE = final_anno
            BOXWIDTH=0 VREF= 44 46 BOXCONNECT=mean
            BOXSTYLE=SCHEMATICID CBOXES= white;
   by trtcd;
run;
quit;
```

Figure 4 displays the mean plots which are the same as generated in Figure 3 except that these plots have been annotated to display the mean values for each week of labs collected. This was done using the ANNOTATE options in the PROC BOX PLOT using the traditional method to first create an annotate dataset which will house all the pre-requisite values for the PROC BOXPLOT to display when it has been plotted.

CASE 5: ANNOATED BOX PLOTS USING SAS®9 PROC BOXPLOT OPTIONS



Box plot of Hemotocrit Lab results for each week by treatment group
With tabulated summary of statistics and listing of outlier
Treatment Type=Placebo

Descriptive Statistics by Week

Mean	45.9	46.6	46.3	44.2	45.8	46.1
Std Dev	3.2	2.5	2.8	2.1	2.9	2.9
Min	41	41	41	42	41	41
Max	50	50	50	48	50	50

Box plot of Hemotocrit Lab results for each week by treatment group
With tabulated summary of statistics and listing of outlier
Treatment Type=Active Drug

Descriptive Statistics by Week

Mean	44.9	44.2	45.5	46.1	44.9	45.6
Std Dev	2.8	3.5	3.3	2.7	2.9	3.3
Min	41	41	41	42	41	41
Max	50	50	50	50	49	50

Numeric Result/Finding in Std Uni
50.0
47.5
45.0
42.5
40.0
Baseline
Week 1
Week 2
Week 3
Week 4
Week 5
Week of Lab Sample Collection

Figure 5: Tabulate Summary of Statistics generated from PROC BOXPLOT

The code to produce the above plots in Figure 5 is:

```
Title;
Title1 j=c  "Box plot of Hemotocrit Lab results for each week by
treatment group";
Title2 j=c  "With tabulated summary of statistics and listing of
outlier";

symbol1 v=plus h=10 c=black ;

proc boxplot data=lb_rep;
   plot lbstresn*week
          / BOXSTYLE=SCHEMATICID
            CBOXES= black
            BOXCONNECT=mean;

   INSETGROUP mean (5.1) stddev (5.1) min max /
                    header = "Descriptive Statistics by Week"
                    position = top
                    cfill = white;
   by trtcd;
   ID subjid;
run;
quit;
```

Figure 5 displays an integrated visualization of Box Plots, Means Plot, Tabulated Summary of Statistics and Listing of Outliers all into one single plot. This array of reporting options which were illustrated before in cases 2 to 4 are now compiled into one code. This code includes a new enhancement to the above plot by using the INSETGROUP option into the PROC BOXPLOT procedure. This helps the user tabulate the requested statistics in the BOX PLOT with a user requested label as “Descriptive Statistics by Week” in the HEADER option. The ID option allows the user to report the outliers which in our case was selected to outline the Subject ID’s.

CONCLUSIONS

The various cases demonstrated during this paper helps the user incorporate the array of options at his disposal within PROC BOXPLOT. This can be one of the most powerful procedures in SAS graphics which helps the target audience to have a better overview on the distribution of data under consideration across a span of time in Clinical studies. As a next step to further enhancement to these graphics the user can incorporate these set of options as building blocks into their respective macro programs. The macro programs can report the data in a more dynamic and tailor made fashion based on the clinical reporting needs in a more sophisticated manner.

REFERENCES

SAS Online Documentation http://support.sas.com/onlinedoc/913/docMainpage.jsp

ACKNOWLEDGMENTS

I would like to acknowledge the following from whom I have received continuous mentorship, support & help:

Lance K. Heilburn & Judith Abrams, Biostatistics Unit, Karmanos Cancer Institute
Parag Shiralkar, Statistical Programming and Analysis, eClinical Solutions

CONTACT INFORMATION

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

Vikash Jain email: **vjain@egistar.com** or **jainvikash77@yahoo.com**