PharmaSUG 2013 - Paper DG02

Dances with Box Plot

Xuefeng Yu, Celgene Corporation, Summit, NJ, USA

ABSTRACT

Box plot is one of the basic, but commonly used graphic tools to display the distribution of data. Making a box plot with SAS® can be as easy as using a single SAS® procedure, such as PROC UNIVARIATE, with specific options. It can also be complicated in some situations where the data distribution needs to be displayed in details. This paper illustrated various SAS® procedures for box plot, including the new procedures in SAS 9.2® and SAS 9.3®. An advanced programming approach is discussed in detail to show how to combine a box plot with scatter plot.

KEYWORDS

Box Plot, Scatter Plot, PROC UNIVARIATE, PROC BOXPLOT, PROC SGPLOT

INTRODUCTION

A box plot is a commonly used graphic tool to display statistical distribution of the data, such as mean, median, 25th percentile, 75th percentile, minimum, maximum and outliers. SAS® provides several procedures to create a box plot. Some can quickly and easily generate a graphic scratch of the data while some procedures need more comprehensive knowledge and skills to create more sophisticated graphs. In this paper, several ways to create a box plot using PROC UNIVARIATE, PROC BOXPLOT and PROC SGPLOT are introduced, and the advantages and disadvantages of each way are compared. A unique program will be demonstrated in detail to show how to combine a box plot and scatter plot.

EXAMPLE 1: PROC UNIVARIATE

PROC UNIVERIATE is the easiest, and probably the most commonly used way to create a box plot. The SAS® code below shows a simple way to generate multiple side by side box plots for the variable CHOLESTEROL, one for each SMOKING group. Option PLOT in proc univariate statement specifies that a box plot will be generated along with a stem-and-leaf plot and a normal probability plot in line printer output. Figure 1 shows the box plot created.

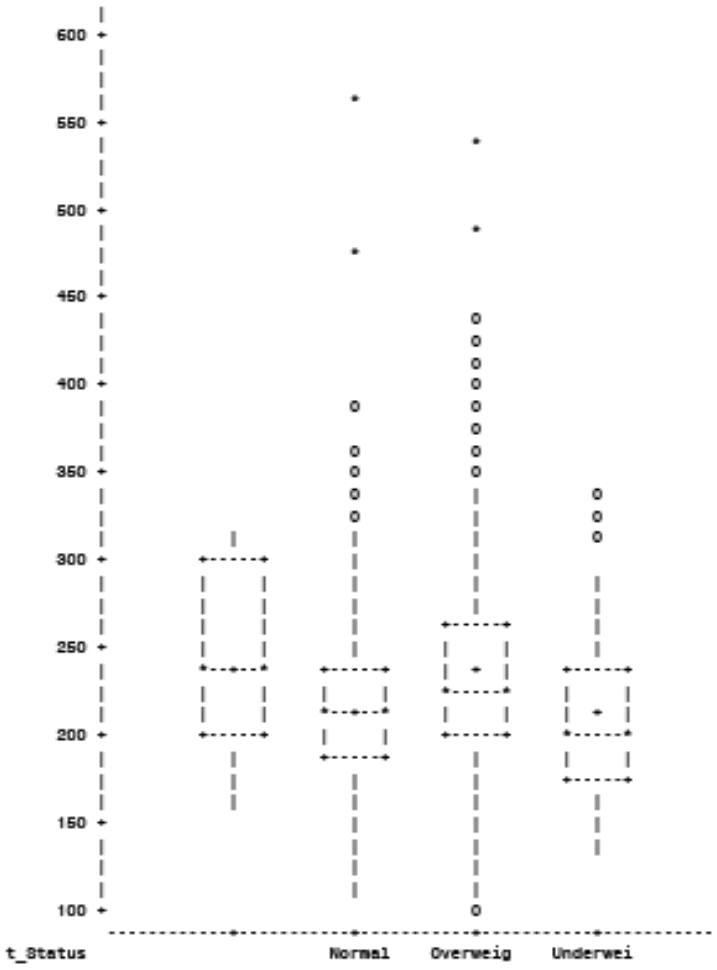
```
proc sort data=sashelp.heart out=heart;
  by weight_status;
run;
proc univariate data=heart plot;
  var cholesterol;
  by weight_status;
run;
```

Although it's fairly easy to create a box plot in this way, the graphs are low-resolution and the appearances are not satisfactory in most practical cases. It's impossible to add legends, text, or color to the figure.

Figure 1: Cholesterol Distribution by Weight Class by PROC UNIVARIATE

The UNIVARIATE Procedure
Variable: Cholesterol

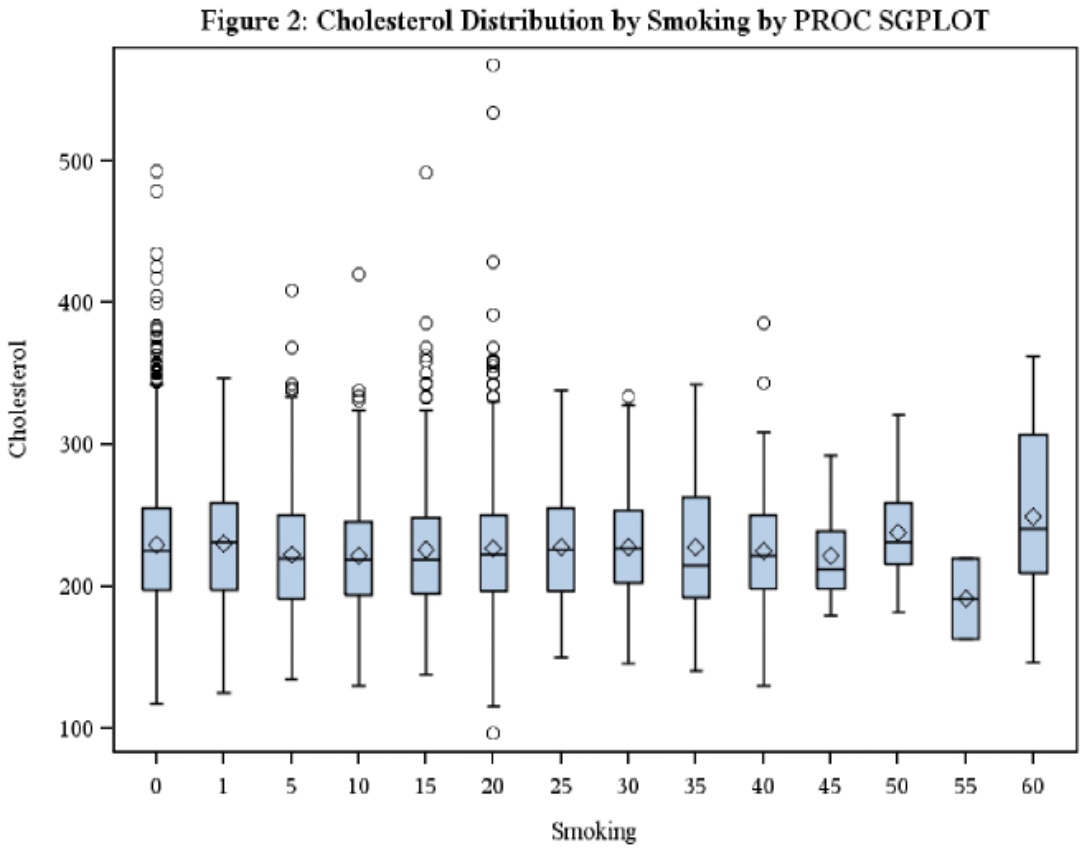
Schematic Plots



EXAMPLE 2: PROC SGPLOT

PROC SGPLOT is a powerful procedure available in SAS 9.2® or higher version. It can be used to produce many kinds of plots, such as a scatter plot, histogram, bar chart or vertical/horizontal box plot. Figure 2 is created with the simple statement below. The resolution is good enough for publication purposes. The color, text, format of number, tick marker and legends can be customized with appropriate options and statements of PROC SGPLOT. PROC SGPLOT is easy and flexible, which makes it a good alternative to PROC UNIVARIATE in terms of creating a box plot. For the syntax details, readers can refer to the SAS® Document, SAS/GRAPH(R) 9.2: Statistical Graphics Procedures Guide, Second Edition.

```
title "Figure 2: Cholesterol Distribution by Smoking by PROC SGPLOT";
proc sgplot data=sashelp.heart;
  vbox cholesterol / category=smoking;
run;
```



EXAMPLE 3: PROC BOXPLOT

Another SAS® procedure that is dedicated to create a box plot is PROC BOXPLOT. The advantage of using PROC BOXPLOT is that by combining the statements of AXIS and SYMBOL, and GOPTIONS, one can make a very sophisticated box plot. In this example, tick mark values and labels were specified by the AXIS statement, and assigned to PROC BOXPLOT with the option of VAXIS and HAXIS. SYMBOL statement specified the symbol style and size. This specification is a universal change to all graphics following SYMBOL statement. INSETGROUP statement presents the numeric values of mean, minimum and maximum as a label on the top of the plot.

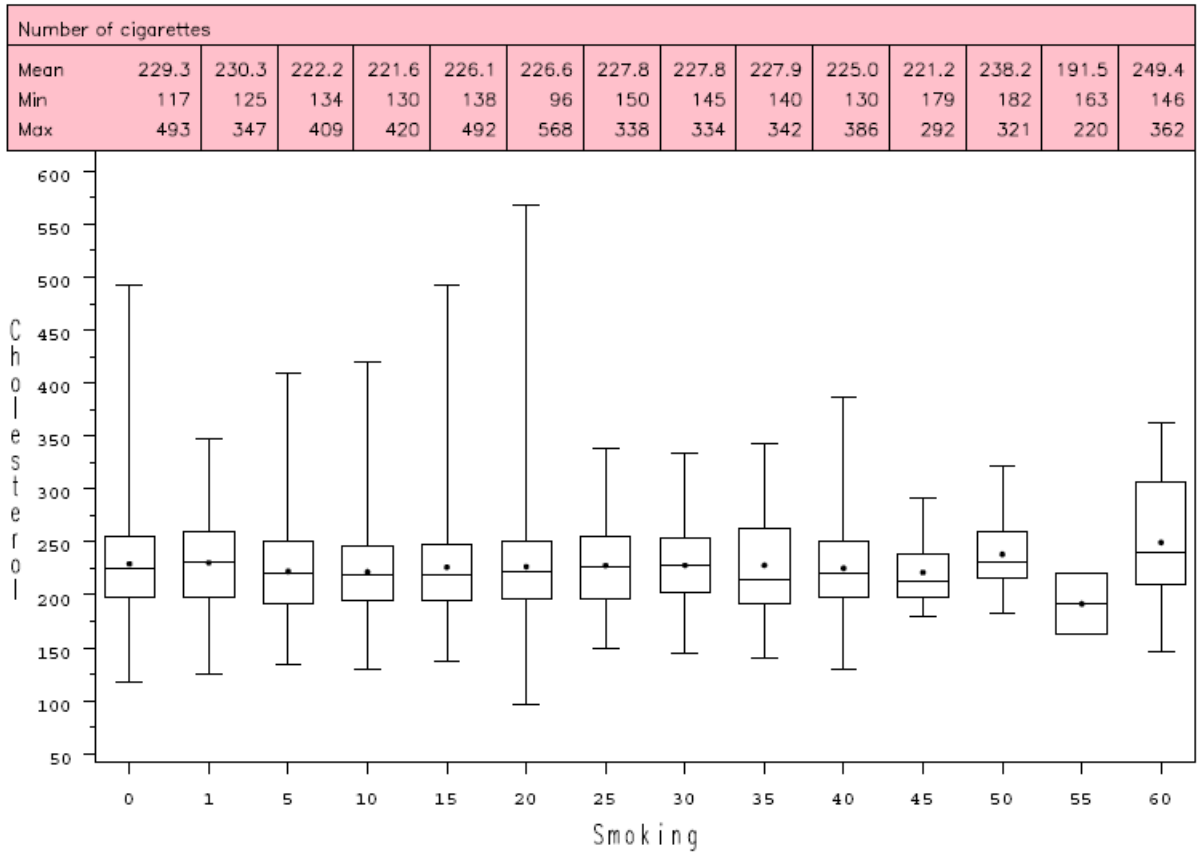
An potential issue was found when preparing Figure 3 with PROC BOXPLOT. The SAS® codes below only work properly on SAS 9.1.3®. If using SAS 9.2®, the table created by INSETGROUP statement will be displayed at the bottom of the figure, mean values will be missing, and the pink color is not presented. Readers may need to pay more attention when using PROC BOXPLOT procedure on SAS 9.2®.

```
axis1 order = (50 to 600 by 50)
```

```
label = (height = 1.25 'Cholesterol')
minor = (number = 1) ;
axis2 order = (0 to 20 by 5)
label = (height = 1.25 'Smoking') ;
symbol value = dot
height = 0.5 ;

proc boxplot data = heart ;
  plot (cholesterol) * smoking / vaxis = axis1
  haxis = axis2
  cboxes = BL ;
  insetgroup mean (6.1) min max / header = 'Number of cigarettes'
  pos = top
  cfill = pink;
run;
```

Figure 3: Cholesterol Distribution by Smoking by PROC BOXPLOT SAS 9.1.3



EXAMPLE 4: USING PROC SGPLOT TO COMBINE SCATTER PLOT WITH BOX PLOT

In Example 2, the outliers were presented as circle symbols. However, as all the points lay along the same vertical line in each group, it's hard to identify each data point. Thus it is difficult to appreciate the total number of outliers. It would be more desirable to have the data points scattered in each group and more easily distinguishable. In some environments, all data points, not only outliers need to be presented within the box plot, so that researchers can better understand the data distribution. It is a combination of a box plot and scatter plot, rather than a simple box plot. Unfortunately, we can't make such a plot in a few simple statements with any current SAS® procedures.

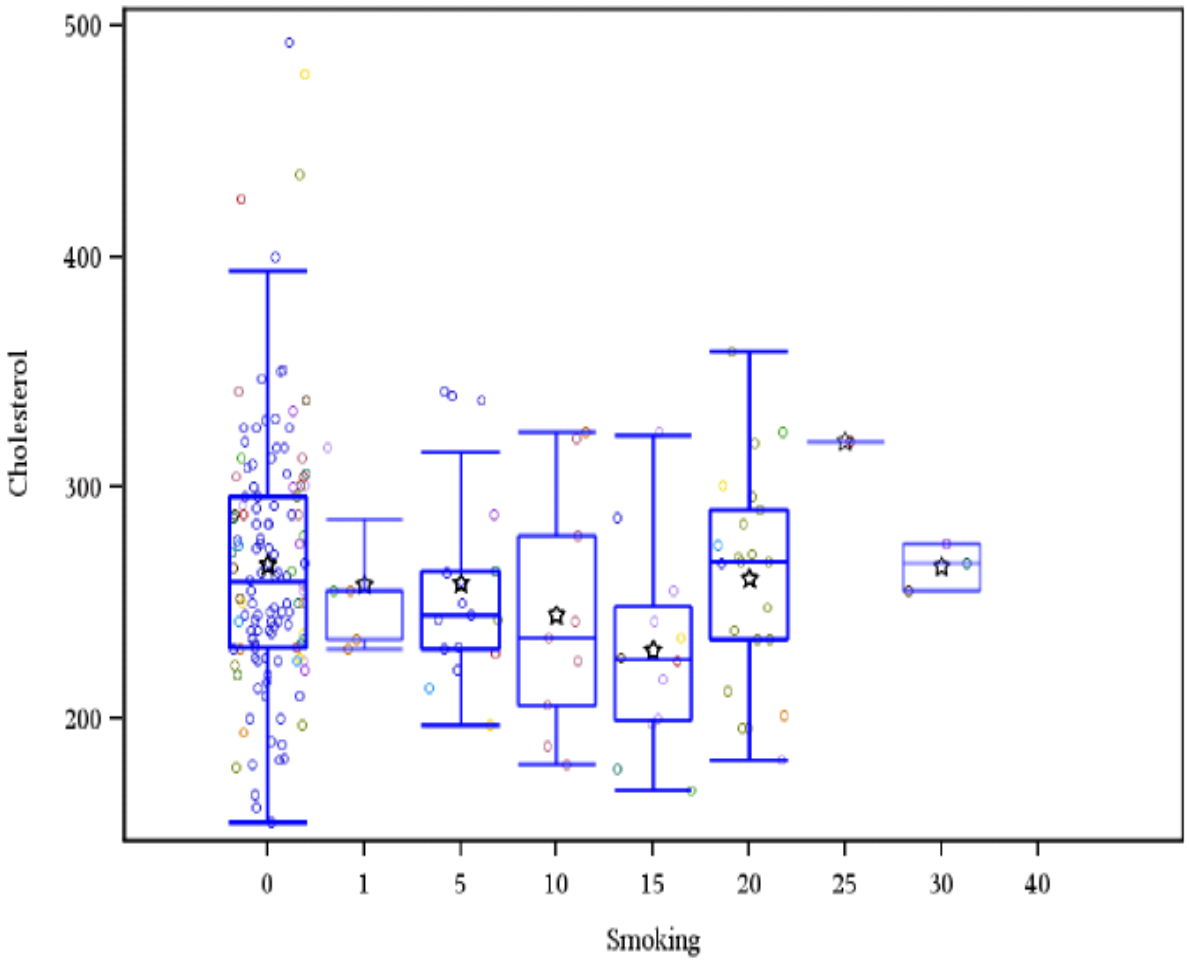
To make such a plot, the following steps have to be performed:

I. Calculate the descriptive statistics of the data by each group.
II. Manually draw the box plot with the values of mean, median, 25th percentile, 75th percentile and the range by using VECTOR statement in PROC SGPLOT. That is to draw a box with the 25th percentile as the bottom line, with the 75th percentile as the top line, with median as the middle line, etc.
III. Draw all the data points with the value of group variable as the x-axis value, and the key variable value as the y-axis value.
IV. Add a random number to the x-axis value of each point. The random number should be small enough to keep the point within the box, and also big enough to distinguish from each other point. This random number can be compared to white noise, making the data point "vibrate" around the vertical line of each group. Thus we call this random number as vibration number in this paper.

The above four steps were organized in order to present the main ideas of the method. In programming practices, Step I and Step IV need to be implemented first to prepare all of the values needed for the plot. Both Step II and Step III will be accomplished in PROC SGPLOT with the statement of VECTOR and SCATTER respectively. The SAS® code of this example is provided in the appendix section.

The vibration number is defined as: (uniform random number – 0.5)*box_width. For the details of how to define box_width and x-axis coordinate values of the group variable, readers can refer to the SAS® code in the appendix section.

Figure 4: Boxplots of Cholesteral by Smoking
Data Source: sashelp.heart
Population: Female Death Due to Coronary Heart Disease



With the limitation of the box size, we can't display thousands of subjects from the sample data of sashelp.heart. We selected a subset of the subjects (females who died from coronary heart disease) to illustrate this method in Figure 4.

As shown in Figure 4, the mean was specified as a star and the data points were presented in different colors for each smoking group. The data points were scattered in each group. The data points in the same group with the same cholesterol value can still be distinguished.

CONCLUSION

This paper introduced three ways to create a box plot with PROC UNIVARIATE, PROC BOXPLOT and PROC SGPLOT. Many features of annotated facility can be borrowed to enhance the box plot created with PROC BOXPLOT. A unique method to combine a scatter plot and box plot with PROC SGPLOT was illustrated. The combined plot presents the data more efficiently and is an ideal graphic tool for clinical data with multiple treatments or visits.

RECOMMENDED READING

SAS/GRAPH(R) 9.2: Statistical Graphics Procedures Guide, Second Edition

CONTACT INFORMATION

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

Xuefeng Yu
Celgene Corporation
86 Morris Ave
Summit, NJ 07901
Work Phone: (908) 860-4343
E-mail: xuyu@celgene.com

TRADEMARKS

APPENDIX

```
option  orientation=landscape;

%macro boxscatter(
                    data = boxplots,
                    byvar= test,
                    group_code = group,
                    group_label = treatment,
                    response = results,
                    xaxis_label = "Treatment",
                    bxwidth = 5
                    );

/* Sorting the data by the group to be vibrated  */
proc sort data = &data;
  by &group_code;
run;

/* Assessing the number of distinct groups */
proc sql;
  create table distinct_groups as
  select distinct &group_code., count (distinct &group_code.) as count_groups
  from &data;
quit;

/* the number of distinct groups */
%let dsid=%sysfunc(open(distinct_groups,i));
%let num_TN=%sysfunc(varnum(&dsid,count_groups));
%let rc=%sysfunc(fetch(&dsid,1));
%let TN=%sysfunc(getvarn(&dsid,&num_TN));
%let rc=%sysfunc(close(&dsid));

/* Multiply the fetched number of observations by 10 */
%let TN10 = %eval(&TN*10);

/* Recoding the distinct groups to be 10,20,... etc */
data new_codes;
  set distinct_groups;
  new_group = _n_ * 10;
run;

/* Create the new vibrated variable using the ranuni function
Setting up the value ranges for the format statement, i.e. 9.5 to 10.5 */
proc sql;
create table merged as
  select a.*, b.new_group,
  (ranuni(44)-0.5)*&bxwidth as vibrate,
  b.new_group + calculated vibrate as new_group_vibrated label= &xaxis_label.,
  b.new_group - &bxwidth/2 as startfmt,
  b.new_group + &bxwidth/2 as endfmt
  from &data as a inner join new_codes as b
  on a.&group_code. = b.&group_code.
  order by b.new_group;
quit;

/* Select the distinct groups to enable the formatted values to be saved */
proc sql;
  create table merged_distinct as
  select distinct &group_label., new_group, startfmt, endfmt
  from Merged;
quit;


/* Calculate the Mean and Median */
```

```
proc means data = merged noprint;
  by new_group;
  var &response.;
  output out = means_medians mean = mean median = median q1 = q1 q3 = q3
                qrange = qrange min = min max = max;
run;

proc sql;
  create table merged_final as
  select a.*, b._FREQ_, b.mean, b.median, b.q1, b.q3, b.qrange,
         (1.5 * qrange) + b.q3 as top_whiskers,
         b.q1 - (1.5 * qrange) as bottom_whiskers, b.min, b.max,
  case when max > calculated top_whiskers then calculated top_whiskers
       else max end as top_whiskers_final,
  case when min < calculated bottom_whiskers then calculated bottom_whiskers
       else min end as bottom_whiskers_final
  from merged as a inner join Means_medians as b
  on a.new_group = b.new_group;
quit;

data merged_final;
  set merged_final;
run;

/* Create the format ranges in the datastep */
data _fmt(keep=fmtname start end label);
  length start end 8 label $50;
  retain fmtname 'treat' type 'n';
  set merged_distinct end=eof;
  if _n_=1 then do;
    start=0;
    end=0;
    label=' ';
    output;
  end;
  start=startfmt;
  end=endfmt;
  label=&group_label.;
  output;
  if eof then do;
    start+3;
    end+3;
    label=' ';
    output;
  end;
run;

/* Creating the format using the cntlin argument */
proc format cntlin=_fmt;
run;

/* Plotting the vibrated values */
goption device=sasprtc reset=all;
ods graphics / reset noborder imagefmt=bmp noscale;
title1 j=c h=1pct "&title1";
title2 j=c h=1pct "&title2";
title3 j=c h=1pct "&title3";

proc sgplot data = merged_final noautolegend DESCRIPTION="";
scatter x = new_group_vibrated y = &response. / group = new_group_vibrated
       transparency = 0 markerattrs = (size=1 mm symbol = circle );
scatter x = new_group y = mean / markerattrs = (symbol = star color = black );
format new_group_vibrated treat.;
/* Leaving some room between the left hand and right */
```

```
xaxis values=(10 to &TN10 by 10) valueshint offsetmax = 0.1 offsetmin = 0.1 ;

vector x = endfmt y = median / noarrowheads lineattrs = (thickness = 1
       color = blue pattern = 1 ) transparency = 0.8
       xorigin = startfmt yorigin = median;
vector x = endfmt y = q1 / noarrowheads lineattrs = (thickness = 1
       color = blue pattern = 1) transparency = 0.8
       xorigin = startfmt yorigin = q1;
vector x = endfmt y = q3 / noarrowheads lineattrs = (thickness = 1
       color = blue pattern = 1) transparency = 0.8
       xorigin = startfmt yorigin = q3;
vector x = startfmt y = q3 / noarrowheads lineattrs = (thickness = 1
       color = blue pattern = 1) transparency = 0.8
       xorigin = startfmt yorigin = q1;
vector x = endfmt y = q3 / noarrowheads lineattrs = (thickness = 1
       color = blue pattern = 1) transparency = 0.8
       xorigin = endfmt yorigin = q1;

vector x = startfmt y = top_whiskers_final / noarrowheads lineattrs =
       (thickness = 1 color = blue pattern = 1) transparency = 0.8
       xorigin = endfmt yorigin = top_whiskers_final;
vector x = startfmt y = bottom_whiskers_final / noarrowheads lineattrs =
       (thickness = 1 color = blue pattern = 1) transparency = 0.8
       xorigin = endfmt yorigin = bottom_whiskers_final;
vector x = new_group y = top_whiskers_final / noarrowheads lineattrs =
       (thickness = 1 color = blue pattern = 1) transparency = 0.8
       xorigin = new_group yorigin = q3;
vector x = new_group y = bottom_whiskers_final / noarrowheads lineattrs =
       (thickness = 1 color = blue pattern = 1) transparency = 0.8
       xorigin = new_group yorigin = q1;
run;
%mend;

%macro mallfig(     outnm=,
                    title1=,
                    title2=,
                    title3=);
data heart;
  length test testtoc $47;
  set sashelp.heart;
  where deathcause='Coronary Heart Disease' & sex='Female' & smoking ne .;
  run;

ods pdf  file="S:\PHARMASUG2013\scatter_box.pdf";

%boxscatter(
       data = heart,
       byvar = smoking,
       group_code = smoking,
       group_label = smoking,
       response = cholesterol,
       xaxis_label = "Smoking",
       bxwidth = 8
       );
ods pdf close;
%mend mallfig;

%mallfig(outnm=scatter_box1,
   title1=%str(Figure 4: Boxplots of Cholesteral by Smoking),
   title2=%str(Data Source: sashelp.heart),
   title3=%str(Population: Female Death Due to Coronary Heart Disease ));
```