Introduction and Application of a New Type of Jittered Scatter Plot - The Line-up Jittered Scatter Plot or Bee Swarm Plot

Charlie C. Liu, PhD and Todd M. Gross PhD
Kythera Biopharmaceuticals, Inc.
30930 Russell Ranch Road, Third Floor, Westlake Village, CA 91362

Abstract

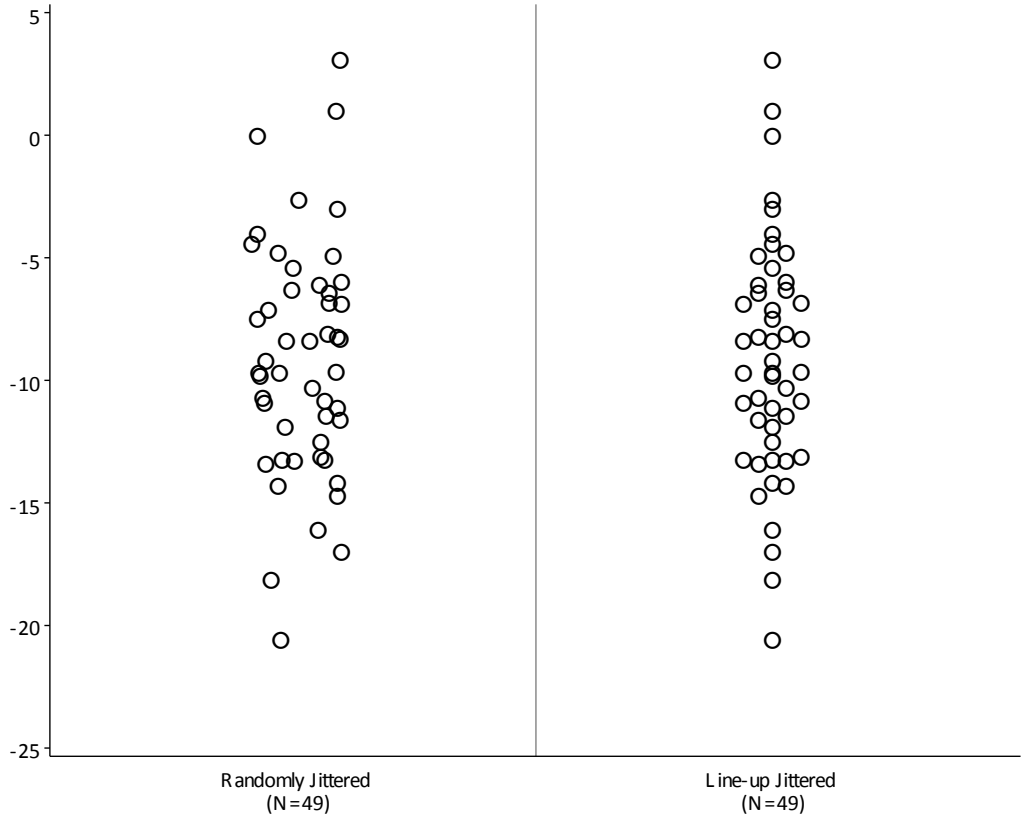
Jittered scatter plots are often used when there are overlapping data values in one of the axes, and use random or a pre-set jittering noise to separate the overlapping data. This paper introduces a new type of jittered scatter plot, known as line-up jittered where the data points are jittered systematically within a group instead of randomly. When the data points with differences within a certain limit on one of the axes are jittered and positioned by group, the data points are "lined-up", hence the name. A line-up jittered scatter plot is also known as a bee swarm plot because it resembles bees swarming around a hive. A line-up jittered scatter plot provides a more organized and better data visualization than its randomly jittered counterpart. The algorithms and a SAS macro performing the line-up jittering are provided. The applications of the new scatter plot are illustrated in examples, including in a customized box-plot with outliers presented in line-up scatters.

Key Words: Scatter Plot, Jittered, Line-up, Bee swarm plot, Box-plot, Data Visualization

Introduction

The creation of data visualization tools that allow readers to see the overall data pattern and explore inherent relationships has been an interesting yet challenging job (Tufte 1983, 1987, 2006). Jittered scatter plots have been widely used as they display all data points of a sample, including outliers that allow the data to speak for themselves. A jittered scatter plot normally uses a random or a pre-set jittering noise level to separate the overlapping data, and the resulting jittered plot is sometimes not very well organized or beautifully presented. This paper introduces a new type of jittered scatter plot, known as line-up jittered where the data points are jittered systematically within a group, rather than randomly. When the data points with differences within a certain limit on one of the axes are jittered and positioned by group, the data points are "lined-up", hence the name. A line-up jittered scatter plot is also known as a bee swarm plot because it resembles bees swarming around a hive. Figure 1 displays a fictitious data set using a randomly jittered and line-up jittered scatter plot side-by-side. The figure demonstrates the value of the line-up littered version to provide a more organized and better data visualization than its randomly jittered counterpart.

Figure 1. Randomly Jittered vs. Line-up Jittered Scatter Plots



Algorithms and a SAS Macro for Performing Line-up Jittering

The detailed algorithms for line-up jittering are described in Liu's book "*Producing High-quality Figures Using SAS/GRAPH and ODS Graphics Procedures*" (Liu, 2015). The algorithms begin by sorting the data from the minimum to the maximum and assigning them into small groups by a specified difference level. The data within the same group are then jittered based on the number of data points within that group by a specified jitter level - if the data point is only one then no jitter is needed; if the data point is an odd number, do not jitter the first number, then jitter one data point to the left and another to the right by the specified jittering level, and continue this way until all data points in the group are jittered; if the group data count is an even number, jitter the first number ½ of the jittering level to the left, the second number ½ of the jittering level to the right, then jitter one data point to the left and another to the right by the jittering level, and continue this way until all data points in the group are jittered. Plot the data points by the line-up jittered value then a line-up jittered scatter plot is produced.

A SAS Macro for line-up jittering based on the algorithms described above has been developed (Liu, 2015).

```
*************************************************************************;
* Macro name: LineUp_Jitter(idn=, xvar=, yvar=, lvl=, jitter=, odn=)    ;
* Input variables:                                                       ;
*      idn    = input dataset name                                      ;
*      xvar   = variable for x-axis                                     ;
*      yvar   = variable for y-axis                                     ;
*      lvl    = interval level/chosen difference within the group       ;
*      jitter = jitter level                                            ;
*      odn    =output dataset name                                      ;
*************************************************************************;
%macro LineUp_Jitter (idn=, xvar=, yvar=, lvl=, jitter=, odn=);
** Get the Min and Max values and save them into macro variables;
PROC MEANS DATA=&idn. MIN MAX noprint;
 VAR &yvar. ;
 OUTPUT OUT=tmp_STAT min=min max=max;
RUN;
data _null_;
  set tmp_stat;
  call symput('Min', min);
  call symput('Max', max);
run;

proc sort data=&idn. out=d_tmp;
  by &yvar.;
run;

data add_tmp;
  set d_tmp;
  do i=floor(&min.) to ceil(&max.) by &lvl.;
    if  i <= &yvar. < i+1 then yvar_int = i;
  end;
run;

** get the number of data points at each interval by x-axis value;
proc freq data=add_tmp noprint;
  tables &xvar.*yvar_int / out=freqdata;
run;

data jitter_sum;
  retain  pos_neg 1;
  set freqdata;
  pos_neg=1;

  if count=1 then do;
     xvar_j=&xvar.;
     output;
  end;

  if count>1 then do;
    if mod(count,2) > 0 then do; ** odd count number;
      xvar_j=&xvar.; pos=&xvar.; neg=&xvar.;  output;
      do i=1 to count-1 by 1;
        if pos_neg=1 then do;
          xvar_j=neg-&jitter;
         neg = xvar_j;
        end;
```

```
        else do;
          xvar_j=pos+&jitter;
         pos = xvar_j;
        end;
        output;
        pos_neg=-1*pos_neg;
      end;
    end;

    if mod(count,2) = 0 then do; ** even count number;
      pos=&xvar.; neg=&xvar.;
      do i=1 to count by 1;
        if pos_neg=1 then do;
          xvar_j=neg-&jitter/2;
           neg = xvar_j-&jitter/2;
        end;
        else do;
          xvar_j=pos+&jitter/2;
         pos = xvar_j+&jitter/2;
        end;
        output;
        pos_neg=-1*pos_neg;
      end;
    end;
  end;
run;

proc sort data = add_tmp;
  by yvar_int &xvar.;
run;

proc sort data=jitter_sum;
  by yvar_int &xvar. ;
run;

data jitter_sum;
  set jitter_sum;
  ord=_n_;
run;
data add_tmp;
  set add_tmp;
  ord=_n_;
run;
** Output dataset: Jittered dataset for ploting;
data &odn.;
  merge jitter_sum (in=a) add_tmp;
  by ord;
 * if a;
  keep  &yvar. &xvar.  yvar_int xvar_j;
run;
%mend LineUp_Jitter;
```

The macro can be modified and used easily to produce your own customized line-up jittered scatter plots.

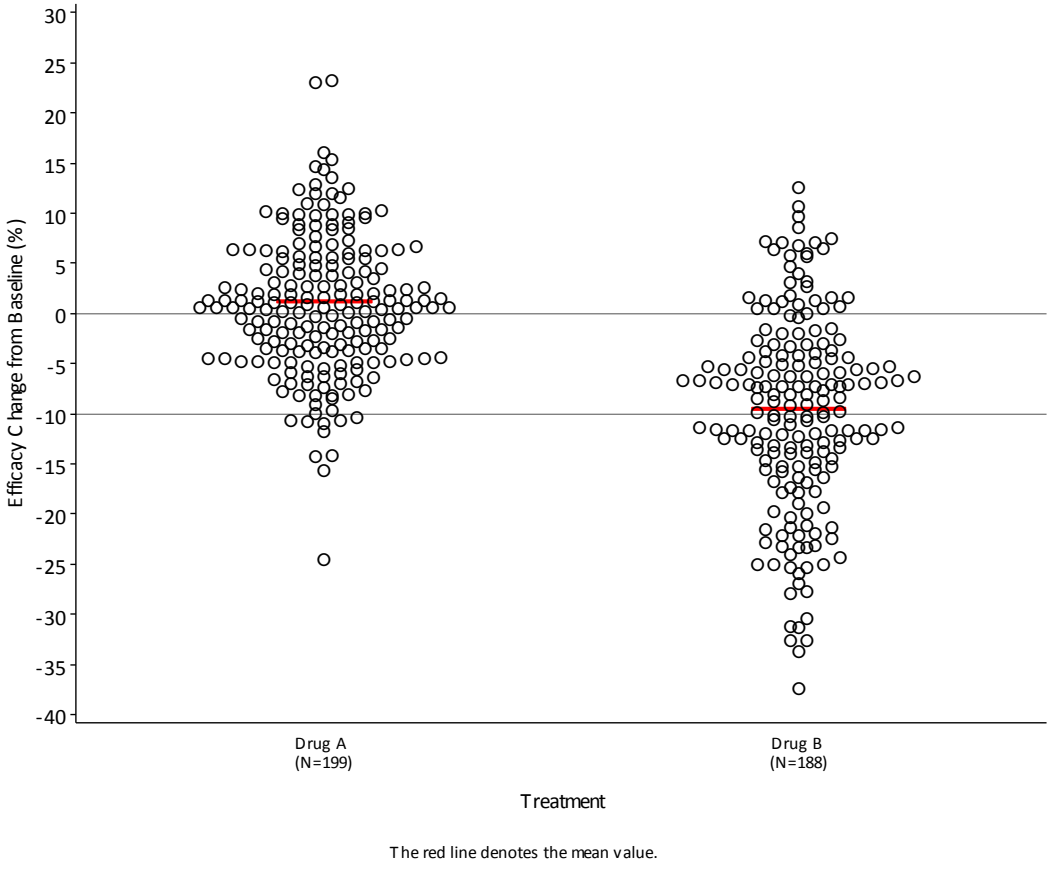
Application of Line-up Jittered Scatter Plots

A Simple Line-up Jittered Scatter Plot

In clinical studies, we usually measure the efficacy endpoints at baseline and post-

baseline visits and compare the difference in change from baseline between the drugs studied. Let's assume a clinical study evaluating the effects of two drugs (Drug A and Drug B) where the primary efficacy measurements are made at the baseline visit before the treatment and at the primary time point after the last treatment. A line-up jittered scatter plot displaying the change from baseline in efficacy by the two drugs is shown in Figure 2. The line-up scatter plot or bee-swarm plot is very organized and provides a powerful comparison of outcomes for the two treatment groups. The plot provides good data visualization for all data including the minimum, the maximum values and the outliers if any. The plot, together with the two reference lines (0 and -10%) on the y-axis easily allow readers to find out the data points below 0 (with reduction) and below 10% reduction which might be a clinically meaningful endpoint. We can easily see that more subjects in the Drug B treatment group had efficacy reduction and had the bigger reduction than those in the Drug A group.

Figure 2. Line-up Jittered Scatter Plot for Efficacy Percent Change by Treatment

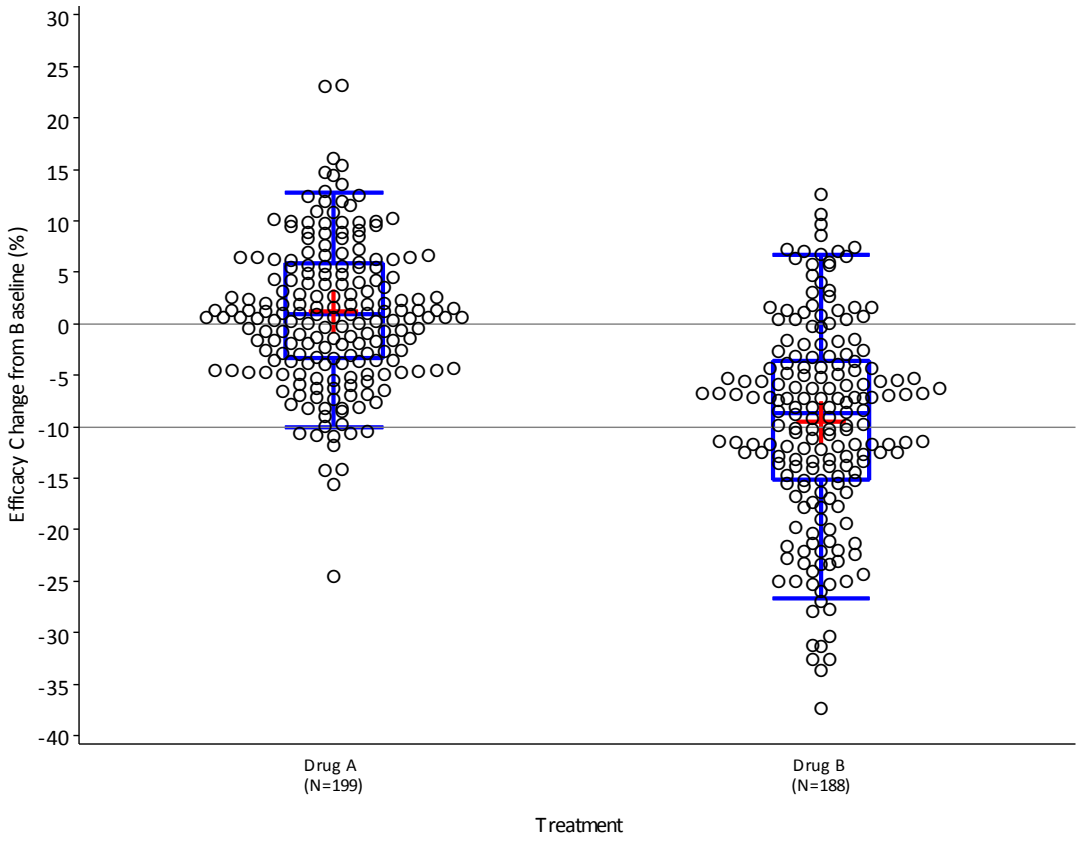


A Line-up Jittered Scatter Plot Overlaid with a Box Plot

Besides displaying all the data points in a more organized way in a line-up jittered scatter plot (Figure 2), it is also a good idea to present some data summaries (mean, median,

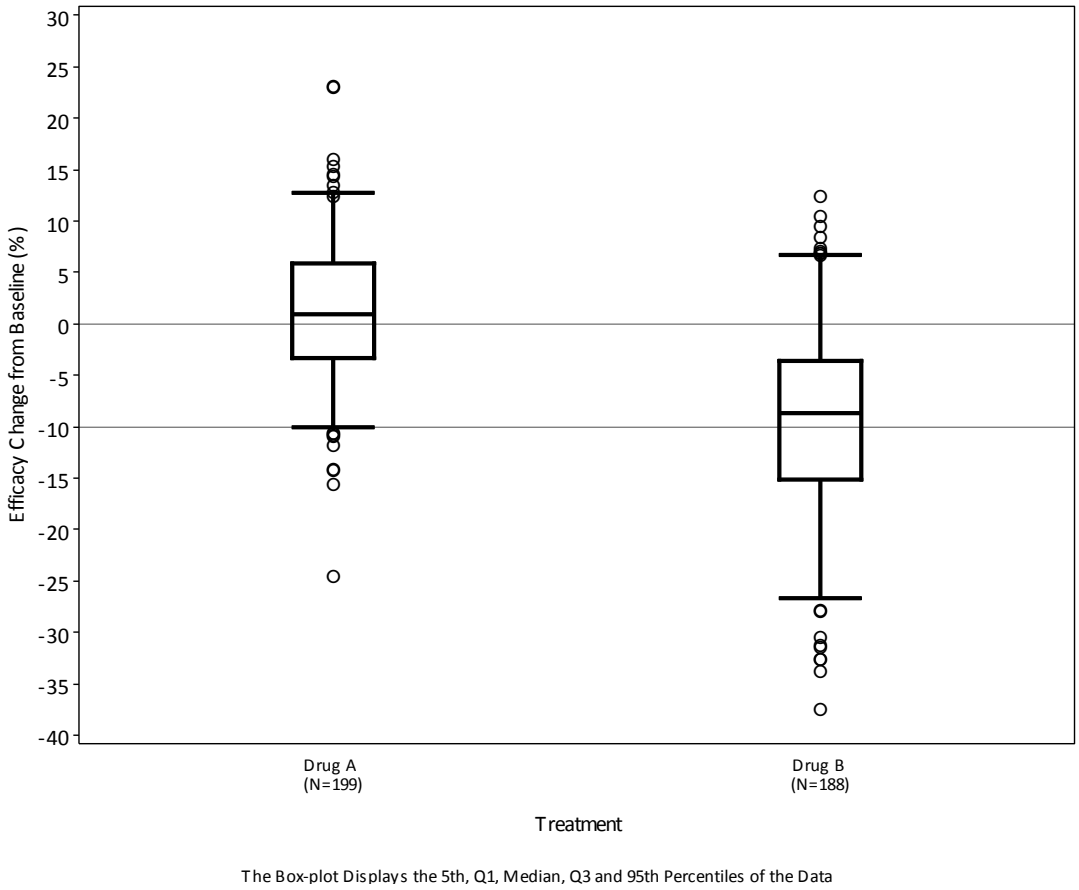
percentiles etc.) on the same plot. Figure 3 is the same as Figure 2 but overlaid with a box-plot that displays the 5th, 25th (Q1), Median, 75th (Q3) and 95th percentiles which are normally produced in a box-plot, and the mean value for each treatment group (the plus sign) which is not included in a typical SAS box-plot. A typical box-plot produced in SAS `PROC GPLOT` (Using `INTERPOL=BOXT5` in the `SYMBOL` statement) is included in Figure 4. Figure 3 is preferred to Figure 4 as it allows us to visualize all the actual data within the "black-box" and also distinguish the outliers. Such a combination of the bee swarm and box-plot can also be a useful educational aid in teaching the meaning of each of the markings on a traditional box-plot.

Figure 3. A Line-up Jittered Scatter Plot Overlaid with a Box-plot with Mean Values Indicated



The Box-plot Displays the 5th, Q1, Median, Q3 and 95th Percentiles of the Data. The plus (+) sign denotes the mean value.

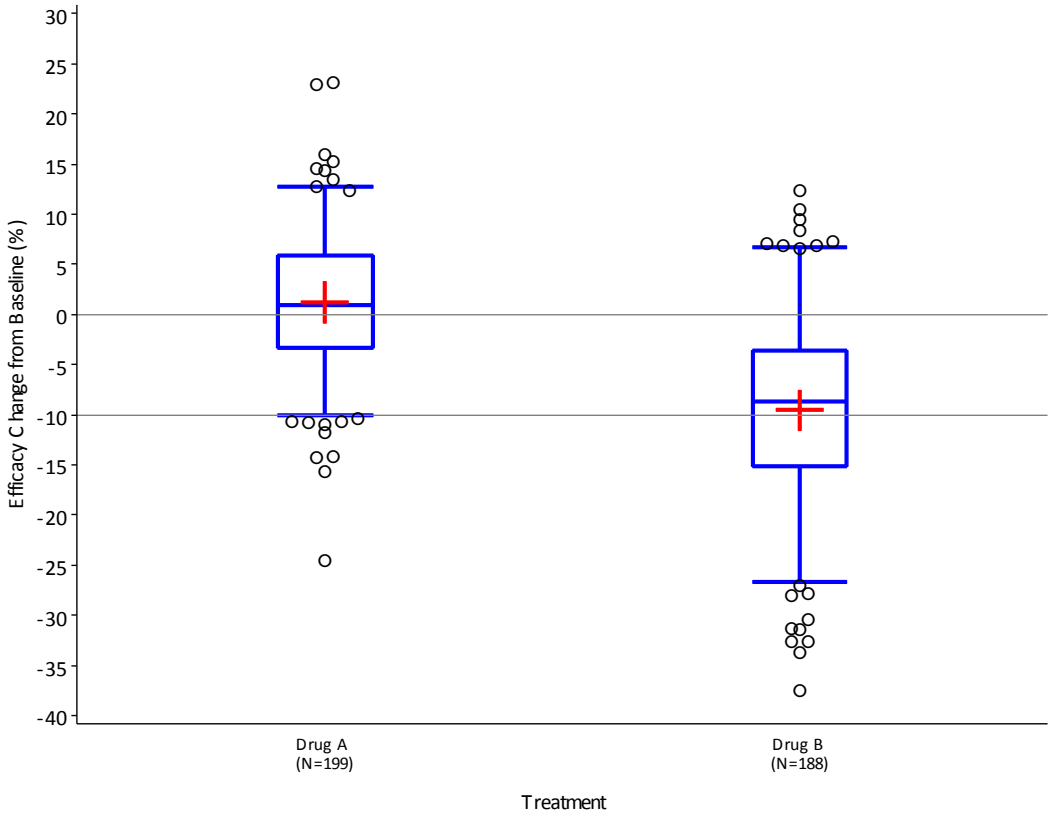
Figure 4. A Typical Box-plot Produced in SAS



A Customized Box-plot with Outliers Presented in Line-up Scatters

A typical box-plot produced in SAS/GRAPH using a symbol statement with PROC GPLOT normally only displays median values together with various percentiles (5^{th}, 25^{th}, 75^{th}, 95^{th} etc.) (SAS Institute, 2012). Using the algorithms of line-up jittering we can produce a customized box-plot with median and mean values displayed together with various percentiles and have the outliers (those outside the lowest and highest percentiles) presented as line-up scatters (Figure 5).

Figure 5. A Customized Box-plot with Outliers Presented as Line-up Scatters



The custom box plots with mean/median values and various percentile data are produced using the SAS Annotate facility. The lines that make up a box-plot and the cross sign denoting the mean value are drawn using SAS *%line()* function. The annotated dataset is produced using the codes below.

```
** Generate an annotation dataset that draws the custom box plot with
median/mean and
   various percentiles;
DATA ANNO;
  SET tmp_STAT;
  %DCLANNO;
  RETAIN SHIFT .1;
  SIZE=2; XSYS='2'; YSYS='2';
  ** horizontal lines for percentiles;
  %LINE(index-SHIFT, Q1, index+SHIFT, Q1, Blue, 1, SIZE);
  %LINE(index-SHIFT, MD, index+SHIFT, MD, Blue, 1, SIZE);
  %LINE(index-SHIFT, Q3, index+SHIFT, Q3, Blue, 1, SIZE);
  %LINE(index-SHIFT, P5, index+SHIFT, P5, Blue, 1, SIZE);
  %LINE(index-SHIFT, P95, index+SHIFT, P95, Blue, 1, SIZE);

  ** vertical lines ;
  %LINE(index-SHIFT, Q1, index-SHIFT, Q3, Blue, 1, SIZE);
  %LINE(index+SHIFT, Q1, index+SHIFT, Q3, Blue, 1, SIZE);
  %LINE(index, P5, index, Q1, Blue, 1, SIZE);
```

```
  %LINE(index, P95, index, Q3, Blue, 1, SIZE);

  ** draw a cross sign for the mean value;
  %LINE(index-SHIFT2, MN, index+SHIFT2, MN, Red, 1, SIZE);
  %LINE(index, MN-2, index, MN+2, Red, 1, SIZE);
  KEEP X Y FUNCTION COLOR SIZE HSYS XSYS YSYS text;
RUN;
```

Conclusion

A line-up jittered scatter plot provides a more organized and better data visualization than its randomly jittered counterpart. Using the algorithms and the SAS Macros for line-up jittering, one can easily produce his/her own customized line-up jittered scatter plots or bee swarm plots. Using SAS annotation and the line-up jittering technique, a custom box-plot can be produced to display mean, median and various percentiles with outliers presented in line-up scatters.

References

Liu, C. 2015. Producing High-quality Figures Using SAS/GRAPH® and ODS GRAPHICS Procedures. CRC Press, Taylor & Francis Group.

SAS Institute Inc. 2012. *SAS/GRAPH® 9.3: Reference, Third Edition*. Cary, NC: SAS Institute Inc.

Tufte, Edward. 1983. *The Visual Display of Quantitative Information.* Graphics Press.

Tufte, Edward. 1997. *Visual Explanations*. Graphics Press.Tufte, Edward. 2006. *Beautiful Evidence.* Graphics Press.