

Jittering: A SAS® Macro to Shift Overlapping Plots

Nathaniel L Baker, Medical University of South Carolina, Charleston, SC

Kevin M Gray, Medical University of South Carolina, Charleston, SC

ABSTRACT

In many clinical trials, interim and final analysis data presentations require complex plots to describe the data distributions by subgroups at several points within the study. Preparing graphics that are adequate for comparing the distributions of multiple groups at multiple time points can be very tedious and time consuming. With each new data run, additional subject information becomes available and the SAS® code written to create the plot must be adjusted and reformatted to the new constraints. This paper presents a unique and easy to implement SAS® macro designed to jitter bar charts, box and whisker plots, and longitudinal mean \pm standard error plots (repeated measure plots) in order to present to a general clinical audience with minimal SAS/GRAPH experience.

THE PROBLEM

Over plotting of data poses a problem in graphical representation of clinical trial and experimental results in many formats. With simple point plots, different units of observation may have the same coordinates. Box plots representing the distribution of different groups at discrete time points can have overlap and become tough to decipher (specifically when the number of groups becomes large). Adding a small amount of random noise to the X or Y values will generally work, but when the group size becomes larger, the plots tend to shift away from the true X (or Y) value and possibly begin to overlap once again. Several SAS® macros have been created to handle the over plotting of individual points on a simple scatter plot with the addition of random noise, but random noise addition is not designed to systematically jitter bar, box and whisker, and repeated measure plots. The proceeding macro adjusts, internally, the amount of systematic noise added based on the total number of groups/individuals within the data set to assure ease of visualization.

THE SOLUTION

MACRO USAGE

This SAS® macro is designed to perform a jitter function on box and whisker, longitudinal, and various other SAS/GRAPH plots with possible overlap due to grouping characteristics of the data. The data steps within the macro identify characteristics of the data in order to determine, systematically, what constant (noise) needs to be added to create the new (jittered) X values that will no longer overlap. Also, imbedded within the jitter macro is a SYMBOL (%*symb*) macro designed to assign a unique color to each grouping assignment in the data set; this provides consistent and recognizable group assignment over the entirety of the X axis range. The %*Jitter* macro also contains the necessary PROC GPLOT statement for the *by group* plots (with minimal changes to the default parameter values). The macro was written with very general and easy to understand SAS data steps; SAS programmers with no experience, as well as advanced SAS programmers, could easily make adjustments to suit their needs.

INPUT PARAMETERS

This macro has eight required parameters and 2 default parameter:

<i>Data</i>	= SAS data set name
<i>Group</i>	= Grouping variable, id, or treatment allocation; must be numeric
<i>X</i>	= Independent variable name (usually time or visit #); this is the variable to be jittered
<i>Y</i>	= Dependent variable name
<i>Num_groups</i>	= Total number of groups, ids, or treatment allocations in the data set
<i>Htitle</i>	= "X variable label" for plot statement; must be in Quotations
<i>Vtitle</i>	= "Y variable label" for plot statement; must be in Quotations

Mtitle = "Main title" of plot; must be in Quotations
Interpol = Interpolation type requested for plot; Default = *box*
Width = line width for plot; Default = 2 but with more than 5 groups
 (*num_groups*), it is recommend using a value of 1.

The general data structure for the macro is as follows and is in general "long" form:

Group/Treatment/ID	X Variable	Y Variable
1	0	2.7
1	1	4.1
1	2	3.3
...
3	0	2.1
3	1	3.6
3	2	5.0

A good tutorial on "Reshaping Wide to Long" can be found in the following link
http://www.ats.ucla.edu/stat/sas/seminars/SAS_arrays/default_new.htm#wide_to_long

THE MACRO

In order to call the macro, the dataset name, the total number of groups (or ID's) within the data set, as well as the X and Y variable names must be known. The title statements (*htitle*, *vtitle*, and *mtitle*) are used as axis and plot labels within the SAS/GRAPH PROC Gplot statement.

```
%macro jitter( data,
               group,
               x,
               y,
               num_groups,
               htitle,
               vtitle,
               mtitle,
               interpol=none,
               width=2);
```

During the following data step, each group is counted and assigned a number from 1 to n (number of unique groups or id's). This number will be used later to determine the amount of systematic noise to add to the plot in an attempt to plot groups in the same order over time.

```
data temp; set &data; j=0;
    if first.&group then j = 1; by &group; run;
data temp; set temp;
    where j=1; count + 1;
    if first.&group then count = 1;run;
```

The variable *new_x* is created with the systematic noise added based on the total number of groups in the data set and the previously formulated count variable. This calculation puts group number 1 closer to the true X value than, say, group 3 or group 5. The macro also uses the modulus function to determine which side of the true X value that the *new_x* variable will be plotted.

```

data &data; set &data;
  adjust=5;
  if &num_groups > 9 then adjust=3;
  if &num_groups < 4 then adjust=10;
  i=1;
  if mod(&group,2)=0 then i=-1;
  new_x = &x + (count* i*(1/(adjust*&num_groups)));
run;

```

An imbedded SYMBOL macro is used to assign unique colors and patterns to each group within the plotting data. This statement allows adjustments to be made to non parameter values to create plots that suit individual needs.

```

%macro symb(num, color=);
  %do n=1 %to &num;
    symbol&n i=&interpol value=dot font=marker
      h=0.5 line=1 width=&width r=1 c=&color;
  %end;
%mend;

%symb(&num_groups);

```

The AXIS option statements use the labels from the macro call to label the X and Y axis. The Legend option statement places a group assignment legend in the top left corner of the plot; this can be changed easily to suit the needs of the individual SAS® User. The included PROC GPLOT statement provides the framework needed to create the plots that may be necessary for presentation, publication, or personal understanding. The FORMAT call requires that the necessary PROC FORMAT statement be run prior to the macro call, however, this can be omitted from the macro if no format is desired.

```

axis1 label=(f="arial/bo" h=1.5 &htitle) style=1;
axis2 label=(a=90 f="arial/bo" h=1.5 &vttitle)
  style=1;
legend1 across=1 down=&num_groups position=(top left
  inside) mode=protect;

proc gplot data=&data;
  title &mtitle;
  plot &y*new_x=&group / haxis=axis1 vaxis=axis2
    legend=legend1;
  format new_x x. &y y. &group group.; run;

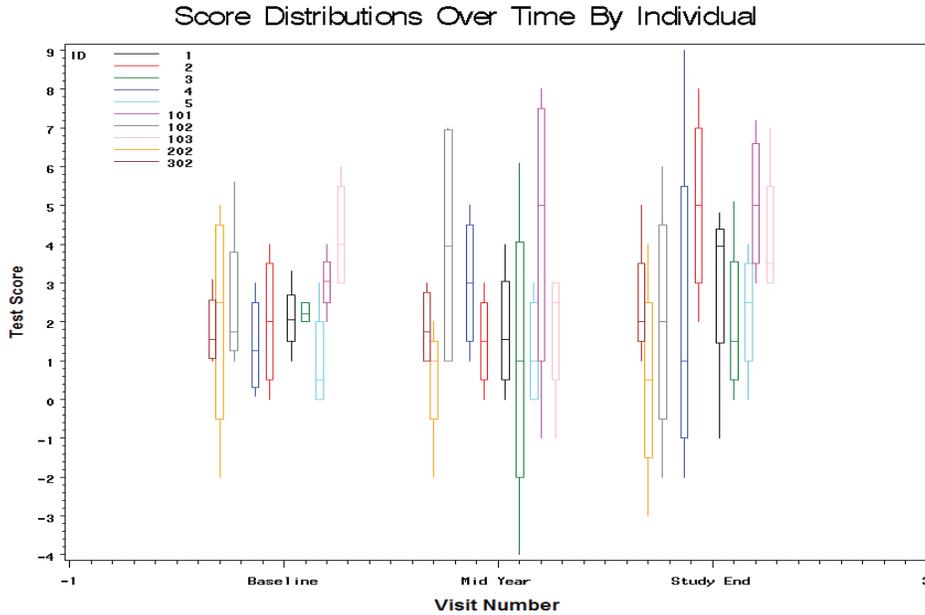
```

The full macro is provided in the appendix.

EXAMPLES

EXAMPLE 1

Box and whisker plots to track test score ranges by visit number for a moderate amount of individuals. At various stages of the investigation, more subjects and time points would be added to the analysis. The investigator needs consistent plots time over time such that records could be compared at several instances.

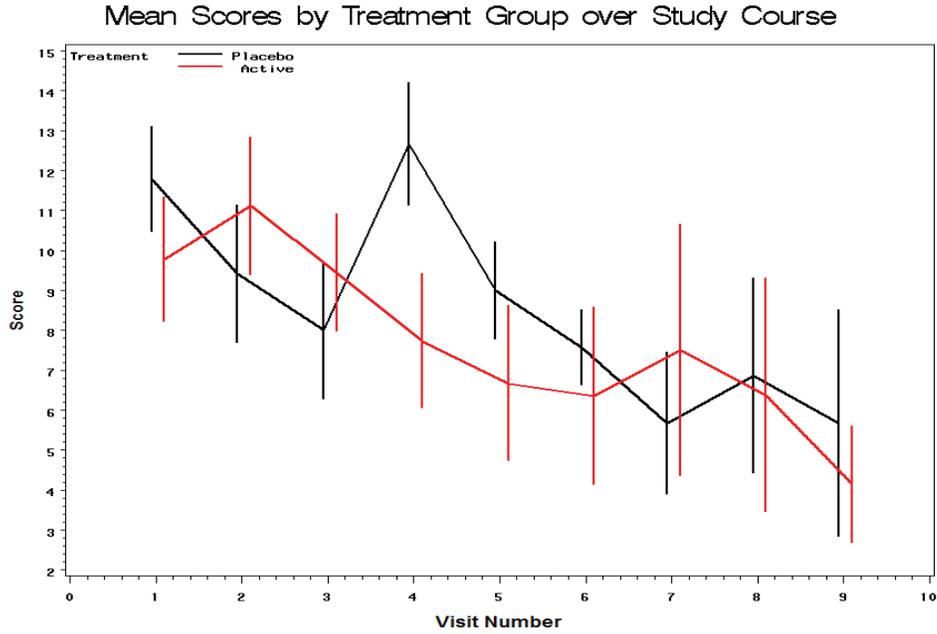


This plot can be accomplished using the general data structure mention earlier and with no adjustment in default settings.

EXAMPLE 2

Another useful situation presents itself with longitudinal analysis graphs that include standard error bars. Overlap of the error bars can make graphical determination of differences and overall results difficult and challenging to interpret. Longitudinal plots can be created by outputting means and standard errors using PROC MEANS (Y=Mean +/- SE) by treatment group for each time point (X Value) and changing the INTERPOL option such that INTERPOL = HILOJ.

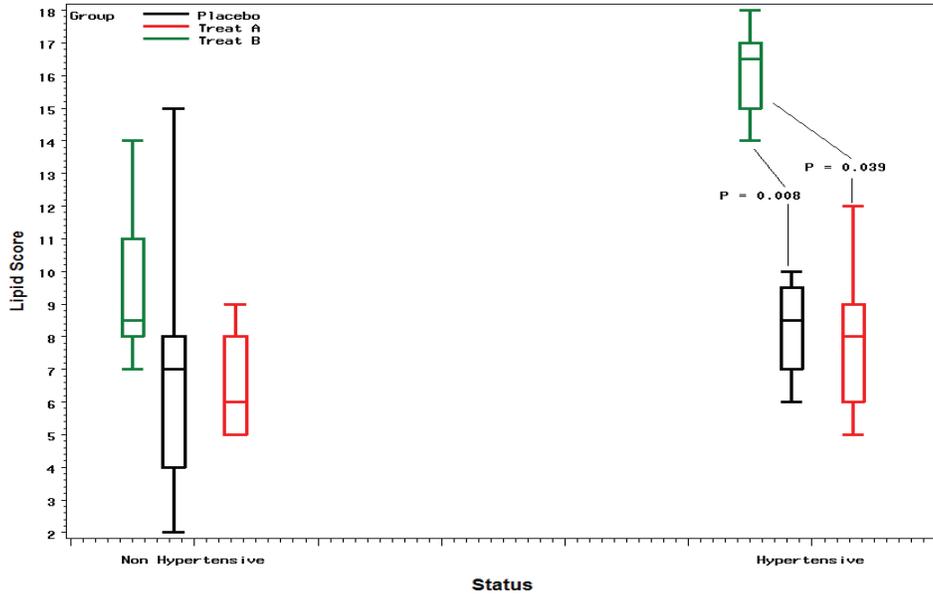
In a recent Adolescent smoking cessation study (R01 DA 17460 Gray / Upadhyaya), Smoking urges were examined on a weekly basis over the course of the 9 week study for differences based on treatment allocation. In an attempt to visualize the data for the Principal Investigator, Means plots were also used to examine treatment group differences in measures such as ADHD Rating Scales and Nicotine Withdrawal Symptoms.



EXAMPLE 3

By simply adding formats for the Grouping variable to the PROC GPLOT data step, *by group* plotting can be done with ease and consistency. The following plot shows differences between lipid scores for randomized treatments within hypertensive and non-hypertensive subgroups at the baseline visit.

Lipid Scores by Treatment Group and Hypertension Status at Baseline



CONCLUSION

As collaborations between clinical investigators and statisticians have gained prevalence, the graphical representation of data has become a critical area of clinical analysis and data understanding. This

macro provides a means of creating jittered plots using SAS/GRAPH without reassigning treatment group variable information with the addition of new subject data. With the inclusion of *systematic* noise, rather than *random* noise, group measurements are jittered identically over multiple time points and measurements assuring that plots containing several X axis points are consistent.

REFERENCES

Introduction to SAS. UCLA: Academic Technology Services, Statistical Consulting Group. From http://www.ats.ucla.edu/stat/sas/seminars/SAS_arrays/default_new.htm#wide_to_long (accessed March 16, 2009).

ACKNOWLEDGMENTS

The author would like to thank Sharon Yeatts and Amy Wahlquist for providing the necessity and Jennifer Waller for providing the ability. Portions of the development of this SAS® macro were accomplished during research supported by NIDA R01 DA 17460 Kevin M Gray / Himanshu P Upadhyaya and M01 RR 01070 MUSC GCRC.

CONTACT INFORMATION

Nathaniel L Baker
Department of Biostatistics, Bioinformatics, and Epidemiology
Medical University of South Carolina
135 Cannon Street, Suite 303
Charleston, SC, 29425
843-792-5028
bakern@musc.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies

APPENDIX

```
%macro jitter(      data,
                    group,
                    x,
                    y,
                    num_groups,
                    htitle,
                    vtitle,
                    mtitle,
                    interpol=none,
                    width=2);

proc sort data=&data; by &group; run;
data temp; set &data; j=0;
  if first.&group then j = 1; by &group; run;
data temp; set temp;
  where j=1; count + 1;
  if first.&group then count = 1;run;
  /* Counts # of unique groups for later adjustment */

proc sort data=temp;
  by &group; run;
```

Paper SD-002

```
data &data; merge temp &data;
    by &group; run;
    /* Merges count data with SAS data set */

proc datasets library=work; delete temp; run;

data &data; set &data;
    adjust=5;
    if &num_groups > 9 then adjust=3;
    if &num_groups < 4 then adjust=10;
    i=1;
    if mod(&group,2)=0 then i=-1;
    new_x = &x + (count* i*(1/(adjust*&num_groups))); run;
    /* Creates Systematic noise based on group order */
    /* and total number of group */

%macro symb(Num, color=);
    %do n=1 %to &num;
        symbol&n i=&interpol value=dot font=marker h=0.5
        line=1 width=&width r=1 c=&color;
    %end;
%mend;
%symb(&num_groups);
    /* Macro to create unique symbol statement for */
    /* each treatment group, needs no call */

axis1 label=(f="arial/bo" h=1.5 &htitle) style=1;
axis2 label=(a=90 f="arial/bo" h=1.5 &vtitle) style=1 ;
legend1 across=1 down=&num_groups position=(top left inside)
    mode=protect;

proc gplot data=&data;
    title &mtitle;
    plot &y*new_x=&group / haxis=axis1 vaxis=axis2 legend=legend1;
    format new_x x. &y y. &group group.; run;
    /* PROC GPLOT statement for plots, can adjust */
    /* as needed. PROC FORMAT must be run prior to */
    /* Macro call or format stmt must be removed */

QUIT;RUN;
%mend;
```

