Paper 140-28

Innovative Graph for Comparing Central Tendencies and Spread at a Glance

Varsha C. Shah, CSCC, Dept. of Biostatistics, UNC-CH, Chapel Hill, NC
Ravi M. Mathew, CSCC,Dept. of Biostatistics, UNC-CH, Chapel Hill, NC

ABSTRACT

The methodology described in this paper includes overlay and other features of PROC GPLOT, a little bit of the ANNOTATE facility, and PROC TEMPLATE for construction of this comparative plot. The graph enables quick visual assessment of effect magnitude and direction, and the code can be easily changed to accommodate new and different numbers of subgroups. Though the mechanism of plotting the graph is more general, it is easy to illustrate with the common concepts of central tendencies and spread. Therefore, the data for this paper is based on randomly generated numbers. The appendix includes the sample figure, data used in the example, and the SAS code needed to produce the figure.

INTRODUCTION

The program develops a side-by-side pictorial comparison of various categories with important statistical information listed next to the figure. Hypothetical data ALLSTAT, listed in Table 1, is processed by the code to produce Figure 1. The algorithm was originally developed since PROC GPLOT does NOT overlay graphs that are of the form: PLOT Y*X=Z**,** even when Z is always an integer. So it is necessary to generate several variables Y1, Y2, Y3, etc., plot Y1*X=1, Y2*X=2, ..., and overlay all such plots.

Let us consider hypothetical data with 6 subcategories and an overall category as shown in table 1. The data has seven observations, one for each subcategory and one for overall. It consists of six variables: CAT=category number, LCI=lower CI, UCI=upper CI, AVG=mean, MEDIAN=median value, and N=sample size.

BASIC CONCEPT

The figure is developed in three major parts which are then linked together by superimposing. The first part is the vertical reference line with a horizontal line of spread for each category, along with the respective labels and dots representing the category means. The reference line can be arbitrarily placed on the x-axis. For this illustration, the vertical reference line is set at x=1.05, the 'overall' sample mean. The second part is the two columns of the information chosen for display. The third and last part contains the title for the graph, titles for the columns of information, the labels below the x-axis, and any footnotes if desired.

The development of the steps is easiest and most straightforward for the third part. The second part is a little more complex, and the first part even more. The logic for each step is explained here, starting with the last one first.

LOGICAL DEVELOPMENT

CREATING THE INFORMATIVE TITLES

The title of the figure, the column titles, the labels for the horizontal axis, and footnotes are developed in the annotate dataset. The text strings are developed and placed by choosing appropriate "x *by* y" coordinates in reference to the paper size. The style, size, fonts, etc., are chosen in this data step. This step is the most enjoyable, like using a paintbrush, creating the picture till it reaches perfection.

CREATING THE COLUMNS OF INFORMATION

Assume that the coordinates where the columns will be located are x1 and x2. These values are arbitrarily chosen after some experimenting with respect to the coordinates of the frames. Next, a data set is created by adding seven variables (=the total number of categories). The values for each observation are assigned in a given pattern. Suppose x1=0.6, x2=1.5, and the names of the added variables are cat1– cat7. The work dataset created from 'ALLSTAT' in the SAS code is ALLSTAT1' and will look something like Table 2.

Note that the values of medians and the sample sizes are converted to macro variables at a later stage in the program, so they are available at any subsequent stage of the process. In Ithis example, the symbols used for plotting at location x1 are the macro values of medians and at location x2 are the macro values of sample sizes.

CREATING THE FIGURE

This is the trickiest part. Data set 'ALLSTAT2' is created from 'ALLSTAT', not only by adding variables, but by adding additional observations as well. Seven variables cat1-cat7 are added just as explained in the previous step. In addition, variables are added for representing LCL, mean, and UCL. Let us call these variables SPREADVAL and CENTERVAL. For each original observation, two observations are generated. The two additional variables are assigned values in this way: the first observation has values of SPREADVAL=LCL and CENTERVAL=mean, and the second observation has values of SPREADVAL=UCL and CENTERVAL=mean. The data set looks like Table 3.

When the graph routine moves from the first to the second observation, the values of SPREADVAL are plotted and connected. On the other hand, when the values of CENTERVAL are plotted, the program stays at the same point since these are identical values. CENTERVAL is plotted using the symbol DOT for a significantly outstanding impression. The resulting graph can be asymmetric.

The Z values 1, 2, 3,…7 in this example are given labels using PROC FORMAT, so they appear in the plot as Overall, Single, etc. and not as 1,2, etc.

CONNECTING THE LOGIC

Using PROC GREPLAY as described in the source (Table 4) combines the three parts developed above. Obviously the three steps are intricately connected. So the three major parts are linked as parts of one big macro.

The resulting graph can be created in any form – graphics stream file, CGM, HTML, PDF, RTF, etc. depending on the GOPTIONS chosen or use of ODS. The background and foreground colors can be changed, and some of them render very pleasant results. However, if the program does not specify colors for connecting points, then different colors are automatically used for each different category.

OTHER APPLICATIONS

This idea can be applied to many different fields. For example, the performance of various production lines could be compared with respect to the established guidelines while the sample sizes are listed in an adjoining column. Similarly it could be applied to the performance of students in a school district with respect to state or national averages while the percentages of the categories are listed in an adjoining column. An epidemiological application could plot odds ratios and their 95% confidence intervals, listing the p-values in the column.

CONCLUSION

This algorithm can be applied in various fields and in variable forms. For example, management could choose to display the figure only, or the figure with only one column of information, depending on needs and desires.

REFERENCES

SAS Institute Inc.:*SAS/GRAPH® Software: Reference, Version 8*, Cary, NC.
SAS Institute Inc.:*SAS Macro Language: Reference, Version 8*, Cary, NC.

ACKNOWLEDGMENTS

The work could not have been put together without support and important input from our supervisor, Kathy Roggenkamp, at the Collaborative Studies Coordinating Center.

CONTACT INFORMATION

Varsha C. Shah
CSCC, Department of Biostatistics, CB #8030
University of North Carolina
Chapel Hill, NC 27514
Phone: 919-966-5160
Fax: 919-966-6335
Email: uccvcs@mail.cscc.unc.edu

Ravi M. Mathew
CSCC, Department of Biostatistics, CB #8030
University of North Carolina
Chapel Hill, NC 27514
Phone: 919-966-3008
Fax: 919-966-6335
Email: uccrmm@mail.cscc.unc.edu

APPENDIX

Table 1: Hypothetical Sample data

Category		N	Mean	LCL	UCL	Median
Overall	1	1000	1.0513	0.7341	1.3685	1.0701
Single	2	425	1.1773	0.8033	1.5513	1.1163
Married	3	575	1.0521	0.6528	1.4513	0.9027
Females	4	713	1.2148	0.9034	1.5262	1.2613
Males	5	287	0.9533	0.6511	1.2555	0.8518
Minority	6	479	1.4087	0.9929	1.8244	1.2882
Non-minority	7	521	1.0132	0.8568	1.1695	1.0008

Figure 1: Illustration of the Hypothetical data for the Innovative Graphic Representation

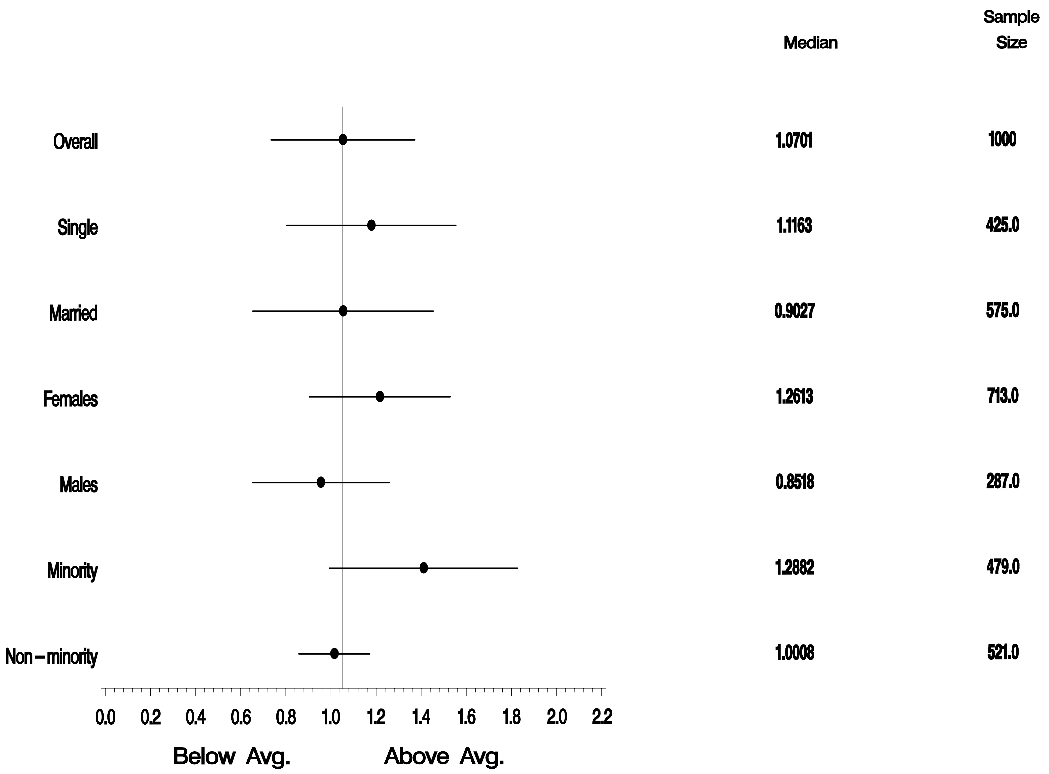
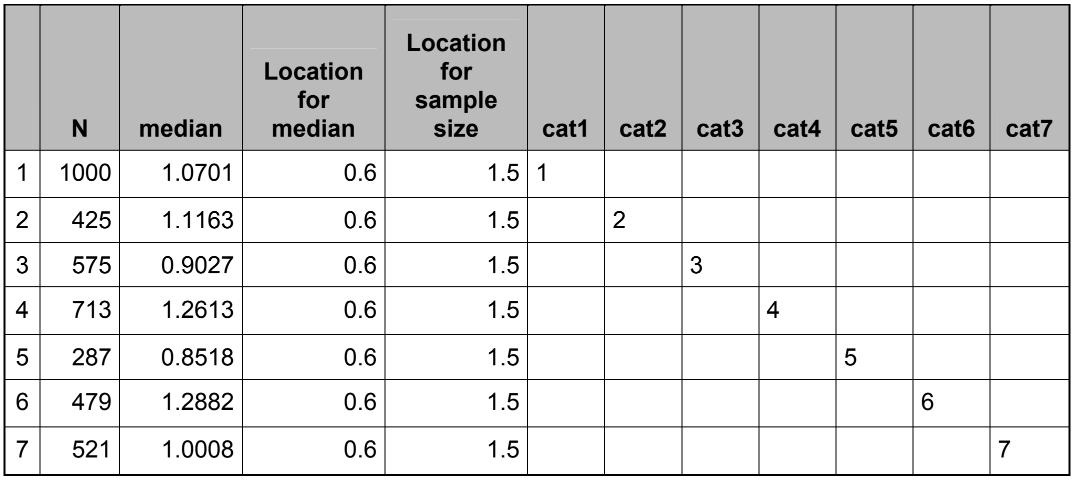


Table 2: Data for column values and their locations



	N	median	Location for median	Location for sample size	cat1	cat2	cat3	cat4	cat5	cat6	cat7
1	1000	1.0701	0.6	1.5	1						
2	425	1.1163	0.6	1.5		2					
3	575	0.9027	0.6	1.5			3				
4	713	1.2613	0.6	1.5				4			
5	287	0.8518	0.6	1.5					5		
6	479	1.2882	0.6	1.5						6	
7	521	1.0008	0.6	1.5							7

Table 3: Data for the graph for plotting the central and extreme values.

	Mean	LCL	UCL	spreadval	centerval	cat1	cat2	cat3	cat4	cat5	cat6	cat7
1	1.0513	0.7341	1.3685	0.7341	1.0513	1						
1	1.0513	0.7341	1.3685	1.3685	1.0513	1						
2	1.1773	0.8033	1.5513	0.8033	1.1773		2					
2	1.1773	0.8033	1.5513	1.5513	1.1773		2					
3	1.0521	0.6528	1.4513	0.6528	1.0521			3				
3	1.0521	0.6528	1.4513	1.4513	1.0521			3				
4	1.2148	0.9034	1.5262	0.9034	1.2148				4			
4	1.2148	0.9034	1.5262	1.5262	1.2148				4			
5	0.9533	0.6511	1.2555	0.6511	0.9533					5		
5	0.9533	0.6511	1.2555	1.2555	0.9533					5		
6	1.4087	0.9929	1.8244	0.9929	1.4087						6	
6	1.4087	0.9929	1.8244	1.8244	1.4087						6	
7	1.0132	0.8568	1.1695	0.8568	1.0132							7
7	1.0132	0.8568	1.1695	1.1695	1.0132							7

Table 4: Source Code with Detailed Internal Documentation.

```
*****************************************************
* Source code for presentation of                  *
* "Innovative Graph for Comparing Central  *
*  Tendencies and spread at a Glance"       *
*****************************************************;

options nocenter MPRINT;
%let job= SUG_v1;
*libname my 'c:\sug';
*filename  outg  "c:\sug\&job..cgm" ;

%annomac;
  goptions device=cgmmppa rotate=landscape gsfmode=replace  gsfname=outg
         gunit=pct htext=3 ftext=swiss goutmode=append nodisplay;

proc format;
VALUE cat_fmt
                 1 = 'Overall'
                 2 = 'Single'
                 3 = 'Married'
                 4 = 'Females'
                 5 = 'Males'
                 6 = 'Minority'
                 7 = 'Non-minority' ;
run ;
data allstat;
     set my.allstat;
     length ccat_no $15;
       ccat_no=put(cat_no,cat_fmt.);
run;
```

```
%let stats=allstat;
%let max_cat=7; ** 'Max_cat' is Maximum Number of Categories. **;

   **----------------------------------------------------------------------------------------**;
   ** CREATE DATASETS to be used as INPUT TO PROC GPLOT.  **;
   **----------------------------------------------------------------------------------------**;

DATA allstat2(keep=cat_no avg lcl ucl spreadval centerval cat1-cat7);
    SET &stats  ;
  ** PROC GPLOT does NOT overlay graphs that are of the form: **;
  **    plot y*x=z,  or  cat_no*avg=cat_no, even when 'z'is               **;
  **    always an integer. So, it is necessary to generate                 **;
  **    several variables: cat1,cat2, .., and plot cat1*avg=1,             **;
  **    cat2*avg=2, ...., and overlay all such plots.                        **;

    array cat(&max_cat) cat1-cat&max_cat ;
    do i=1 to &max_cat ;
      cat(i) = . ; ** Default. **;
      if (cat_no = i) then cat(i) = i ;
    end ;

    ** "spreadval" is for plotting the UCL and LCL."centerval"          **;
    ** is for plotting the mean, with a symbol 'dot'. Create              **;
    ** these variables for each line of the input dataset.                 **;
      spreadval = lcl ; centerval = avg ; output ;
      spreadval = ucl  ; centerval = avg ; output;
run ;


DATA allstat1(keep=cat_no totn median x1 x2 cat1-cat7) ;
    SET &stats  ;

    ** "allstat1" is for plotting Median and Sample Size       **;

    x1 = 0.6 ;** Median will be plotted at this coordinate     **;
    x2 = 1.50;** Sample Size will be plotted at this location**;

    ** From each observation, create variables:             **;
    **    cat1=1 on line with cat_no=1,                         **;
    **    cat2=2 on line with cat_no=2, ...., etc.              **;

    array cat(&max_cat) cat1-cat&max_cat ;
    do i=1 to &max_cat ;
      cat(i) = . ; ** Default. **;
      if (cat_no = i) then cat(i) = i ;
    end ;

    ** NOTE: The values of Median and Sample Size will be     **;
    ** converted into a Macro variables in MACRO GRAPHIT.   **;
run ;

** Creation of Input datasets for Graphs is complete at this point. **;

   **-------------------------**;
   **  PLOT GRAPHS  **;
   **------------------------**;

%MACRO GRAPHIT (x_max= ) ;
    ** x_max = must be at least Max. of UCL expected or  **;
    **        larger, to accommodate all the values to           **;
    **        be included in the graph.                              **;

    ** FIRST PLOT:                                                    **;
    **Subgroups of data on the Y-axis, denoted by cat_no,**;
    **vs. UCL and LCL. Join each pair of extreme values   **;
    **by a line. Also, plot cat_no. vs. AVG where the          **;
    **symbol is a black dot. Overlay these plots.              **;

       ** Nullify all symbol assignments:                         **;
     %do i= 1 %to %eval(&max_cat + &max_cat) ;
        symbol&i ;
     %end ;

     axis1 order=(0.0 to &x_max by 0.2) label=none ;
```

```
 axis2 order=(&max_cat to 1 by -1) major=none minor=none label=none
    style=0 offset=(5,5) value=(justify=left) ;

 proc gplot data=allstat2 ;
    format cat1-cat&max_cat cat_fmt. ;
 plot
  %do i= 1 %to &max_cat ;
    cat&i*spreadval=&i
    cat&i * centerval = %eval(&max_cat + &i)
  %end ;
     /overlay vaxis=axis2
          haxis=axis1 href=1.05 nolegend ;

  ** e.g., plot cat1*spreadval=1 cat1*centerval=8    **;
  **          cat2*spreadval=2 cat2*centerval=9       **;
  **            ...........etc.................     **;
  **          cat7*spreadval=7 cat7*centerval=14 /    **;
  **      overaly vaxis= ..., where max_cat=7.         **;

 %do i= 1 %to &max_cat ;
   symbol&i v=none i=join l=1 w=2 color=black;
     %let ii = %eval(&max_cat + &i) ;
   symbol&ii  v='dot' h=3 w=3 color=black;
 %end ;

 run ;

** SECOND PLOT:                                       **;
** Subgroups of data on the Y-axis, denoted by      **;
** cat_no, vs. Median, where the symbols are         **;
** the macro variables of Median.  Same for the      **;
** Sample Size values. Overlay these plots.          **;


** Convert Median and AVG into Macro Variables. **;
DATA _null_ ;
   SET allstat1 ;
    %do i= 1 %to &max_cat ;
     if (cat&i = &i) then do ;
      call symput("median&i", trim(left(put(median, 6.4))));
      call symput("totn&i", trim(left(put(totn, 6.1))));
     end ;
    %end ;
run;

  ** Nullify all symbol assignments: **;
 %do i= 1 %to %eval(&max_cat + &max_cat) ;
   symbol&i ;
 %end ;

 axis1 order=(0.0 to &x_max by 0.2) label=none
     value=none  major=none minor=none style=0 ;
 axis2 order=(&max_cat to 1 by -1) major=none minor=none
     label=none style=0 offset=(10,4) value=none ;

 proc gplot data=allstat1 ;
    format cat_no cat_fmt. ;
 plot
  %do i= 1 %to &max_cat ;
    cat&i * x1 = &i
    cat&i * x2 = %eval(&max_cat + &i)
  %end ;
    /overlay vaxis=axis2 haxis=axis1 nolegend ;

 %do i= 1 %to &max_cat ;
   symbol&i  font=swissb v="&&median&i" w=3 h=3 color=black;
     %let ii = %eval(&max_cat + &i) ;
   symbol&ii  font=swissb v="&&totn&i" w=3 h=3 color=black;
 %end ;

 * e.g., for max_cat = 7:
 *      plot cat1*x1=1 cat2*x1=2 .... cat7*x1=7
 *           cat1*x2=8 cat2*x2=9 .... cat7*x2=14
 *                / overlay ... ;
```

```
       *
       *  symbol1  font=swissb v="&median1" w=3 h=3 ;
       *  symbol2  font=swissb v="&median2" w=3 h=3 ;
       *     ..............etc.................
       *  symbol7  font=swissb v="&median7" w=3 h=3 ;
       *
       *  symbol8   font=swissb v="&totn1" w=3 h=3 ;
       *  symbol9   font=swissb v="&totn2" w=3 h=3 ;
       *     ..............etc.................
       *  symbol14  font=swissb v="&totn7" w=3 h=3 ;

       run ;
quit ;

goptions vsize=0.0 cm hsize=0.0 cm;

data annodata;
     length function style color $ 8 text $115 ;
     retain when 'a' function 'label'  xsys '3'  ysys '3'  hsys '3'
      x 50  position '5'  style 'swissb'  color 'black'  y  97  size 1.8;

       y=100 ;
        text="Figure 1: Illustration of the Hypothetical data for the Innovative Graphic Representation";
              style='swiss' ;size=2.5;
        output;

      y=5 ;
      x=20; size=2.5; style='swiss' ; text="Below Avg." ;
        output ;
      x=38; size=2.5; style='swiss' ; text="Above Avg." ;
        output ;

      y=90 ;
      x=86; size=1.8; style='swiss' ; text="Sample" ;
        output ;
      y=87 ;
      x=69; size=1.8; style='swiss' ; text="Median" ;
        output ;
      x=86; size=1.8; style='swiss' ; text="Size" ;
        output ;
      stop;
run;

proc ganno anno=annodata;
run;

goptions display;

proc greplay igout=work.gseg tc=look nofs ;

     ** Defining the 2 panels of the Template named TWO **;
     **NOTE:                                              **;
     **     1=Left, 2=Right.                              **;
     **     The 3rd panel is for the annotate dataset.    **;
     tdef two
        1/ llx=0     lly=7
           ulx=0     uly=85
           lrx=53    lry=7
           urx=53    ury=85

        2/ llx=55    lly=7
           ulx=55    uly=85
           lrx=100   lry=7
           urx=100   ury=85

        3/ llx=0      ulx=0
           lrx=100  urx=100
           lly=0      uly=100
           lry=0      ury=100;

   template two ;
   treplay  1:1 2:2 3:3;
run;
```

```
%MEND GRAPHIT ;

%GRAPHIT (x_max=2.2) ;
quit;
run;
```