Plots With Generalized Confidence Intervals Using SAS/GRAPH™

Linda C. Whitehand & Bruce E. Mackey
USDA, Agricultural Research Service, PWA
Albany, CA

ABSTRACT

Plotting of group means and confidence intervals is a popular method with the statistical community for summarizing data. The technique provides more information than multiple comparison procedures and is easier to comprehend than tables of the same statistics. Although SAS™ PROCs GPLOT and GCHART provide ways of plotting means and confidence intervals, they are quite restrictive for many types of complex analyses. For example, they do not easily plot asymmetric confidence intervals, which occur for binomially distributed data. Asymmetry also arises with data requiring statistical analysis in a transformed scale, but for which presentation of means and confidence intervals is desired in the original scale. Even with symmetric cases, derivations may be complex and/or confidence intervals may be unequal among groups, such as with those produced using PROC MIXED. We present techniques and examples to "trick" GPLOT and GCHART into correctly plotting user defined means and confidence intervals, generalizations which should increase the usefulness of these procedures.

INTRODUCTION

As a convenient method of summarizing data, the statistical community often uses plots of group means with appropriate confidence intervals. One reason for the popularity of this technique is that it provides more information than presenting results of a multiple comparison procedure. Also, due to its visual nature, the information is more readily understandable than if tables of the same statistics are used. Although SAS™ PROCs GPLOT and GCHART provide ways of plotting means and confidence intervals, they do not allow for the complexities that arise from different types of analyses.

In particular, correct confidence intervals are often asymmetric, as is the case for binomially distributed data. The need for asymmetry also arises for data that requires transformation prior to statistical analysis. For example, data often must be analyzed in a transformed scale to correct for heterogeneity of variance, but means and confidence intervals are transformed back to the original scale for ease of interpretation in their presentation. In other cases although symmetric, confidence intervals may involve complex derivations and/or they may be unequal among groups. One example would be confidence intervals produced using PROC MIXED. We present techniques of generalization that allow PROCs GPLOT and GCHART to correctly plot user specified means and confidence intervals. The former addresses the problem of asymmetry, and the latter simply allows the user to specify the proper confidence intervals.

PROC GPLOT

In the case of symmetric confidence intervals PROC GPLOT can directly plot means and their confidence intervals connected by lines if the data presented to PROC GPLOT consists of just three points for each group: the mean and the upper and lower confidence limits. The desired plot would be created using the I=HILLOTJ option of the SYMBOL statement, which plots each group mean connected to its maximum and minimum data points, the upper and lower confidence limits in the case described. Such data can be obtained through creation of an output data set from many SAS procedures. Means (or predictions) along with 95% confidence intervals can be output directly to a data set, or means and standard errors
can be output followed by a data step to create the appropriate limits.

The problem arising in the asymmetric case is that the mean will not be the average of the upper and lower limits. We hypothesized that a fourth "observation" exists which in combination with the mean and the confidence limits would give the correct, asymmetric mean. That is, a value \( y_4 \) must be found which satisfies the following formula:

\[
y = \left( \frac{\bar{y} + LL + UL + y_4}{4} \right)
\]

The required "observation" will fall between the upper and lower limits, thus affecting only the mean and not the limits connected with the HILOTJ option. A little computation shows the solution for the hypothetical fourth data point to be:

\[
y_4 = 3\bar{y} - (LL + UL)
\]

**Example 1:** You would like to compute and plot 95% confidence intervals on a binary response with 10 observations per group. Naturally, the confidence intervals will be symmetrical only at \( p = 0.5 \). The following are calculated and presented to SAS as a data set:

<table>
<thead>
<tr>
<th>Group</th>
<th>( \bar{y} )</th>
<th>LL</th>
<th>UL</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.2</td>
<td>0.025</td>
<td>0.556</td>
</tr>
<tr>
<td>B</td>
<td>0.5</td>
<td>0.187</td>
<td>0.813</td>
</tr>
<tr>
<td>C</td>
<td>0.8</td>
<td>0.444</td>
<td>0.975</td>
</tr>
</tbody>
</table>

The following SAS code produced Figure 1:

```sas
title 'Confidence Intervals Computed from Binomial';
data ex1;
  input group$ ybar 11 ul;
  /* ybar is the mean, and ll and ul are 95% confidence limits */
  /* the value \( y_4 \) is calculated from formula (2) */
  y4=3*ybar-(ll+ul);
  /* statements to create 4 needed "observations" */
array y{4} ybar ll ul y4;
do i=1 to 4;
z=y{i};
output; end;
cards;
A .2 .025 .556
B .5 .187 .813
C .8 .444 .975
/* statements to produce plot shown in Figure 1 */
options dev=os2 targetdevice=os2prtm
  htext=1.3 ftext=centb htitle=1.5;
proc gplot;
plot z*group/vaxis=axis1 haxis=axis2;
symbol i=hilog v=none c=black w=3;
axis1 w=3 label=(a=90 'ESTIMATED PROPORTION');
axis2 w=3;
footnote 'Fig. 1: GPLOT Output from Ex. 1';
run;
```

**Example 2:** To correct for heterogeneity of variance among data taken at different weeks, you need to transform by logs prior to running a one-way analysis of variance. However, you desire to transform computed means and 95% confidence intervals back to the original scale of measurement for graphical presentation. A data set with 6 observations taken at weeks 3, 4, 6, and 7 and 4 observations at week 9 was transformed to log scale and analyzed using PROC GLM. An output data set was created which contained predictions (ln \( p \)) and upper (ln \( ul \)) and lower (ln \( ll \)) confidence intervals on the transformed means. These three resulting "observations" were transformed back to the original scale by exponentiation, causing the resultant confidence limits to be asymmetrical about the prediction.

Formula (2) was then employed, substituting the prediction value for the mean, to compute the necessary 4th "observation" to allow HILOTJ to plot the correct means and 95% confidence limits shown in Figure 2. The calculated predictions and confidence intervals are shown below in log, and original scale along with the arithmetic means of the original data:
The following SAS code produced the above table and Figure 2:

```
data orig;
  infile 'd:\sasrun\wuss96m8\ex2.dat';
/* input group and response variable */
  input week c s;
y=c+s;
/* take log of response variable for analysis */
  ln=log(y);
/* run GLM analysis on log variable, */
  proc glm data=orig;
    class week;
    model ln=week;
/* output data set containing predictions and 
   confidence limits */
    output out=outglm p=ln_p 195m=ln_jl_u95m=ln_ul;
/* DATA step transform values to original units */
data backtran;
  set outglm;
  ll=exp(ln_l); p=exp(ln_p);
  ul=exp(ln_ul);
/* PROC MEANS to eliminate duplicate data */
  proc means noprint data=backtran;
    by week;
    var ln_l p ll ul y p ll ul;
    output out=backmean n(ln)=n mean= ;
data graph;
  set backmean;
/* compute "trick" data value */
  y4=3*p-(ll+ul);
/* and prepare data for plotting */
  array x{4} p ll ul y4;
  do i=1 to 4;
    z=x{i};
    output; end;
```

### PROC GCHART

PROC GCHART has the capability of plotting confidence intervals based on variances within groups. However, these are often not the confidence intervals appropriate for a specific statistical analysis. Our objective is to be able to specify desired confidence intervals, i.e., trick GCHART into using previously calculated ones. Using a pre-calculated mean and upper limit, a constant can be computed such that, if it is added to and subtracted from the mean, the resultant three observations will force GCHART to give the desired confidence interval. The formula is as follows:

$$c = \sqrt{3} (UL - y)/t,$$

where $t = \text{Student's} t$ for two degrees of freedom and the desired $\alpha$ level.

#### Example 3:
We use the ERRORBAR=TOP option in order to draw only the upper confidence limit. Part of the data from table 17.5 in Milliken & Johnson is used as follows:

<table>
<thead>
<tr>
<th>Group</th>
<th>n</th>
<th>y+5</th>
<th>se</th>
<th>UL+5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1W</td>
<td>3</td>
<td>2.667</td>
<td>3.085</td>
<td>8.79</td>
</tr>
<tr>
<td>2B</td>
<td>7</td>
<td>4.286</td>
<td>2.020</td>
<td>8.30</td>
</tr>
<tr>
<td>2W</td>
<td>8</td>
<td>4.375</td>
<td>1.889</td>
<td>8.13</td>
</tr>
<tr>
<td>3B</td>
<td>6</td>
<td>1.333</td>
<td>2.182</td>
<td>5.67</td>
</tr>
</tbody>
</table>
The following SAS code produced figure 3:

data ex3;
/* input mean, std err, and df */
  input mean std err df;
/* set desired alpha level, calculate corresponding prob */
  alpha=0.05; prob=1-alpha/2;
/* calculate appropriate upper confidence limit */
  ul=mean+inv(prob,df)*std;
/* compute needed constant */
c=sqrt(3)*(ul-mean)/inv(prob,2);
/* output 3 needed 'observations' */
/* NOTE: 5 added to make all means positive */
z=mean+5; output;
z=mean+c+5; output;
z=mean+c+5; output;
cards;
1W 3  3.085
2B 7  -7.14  2.02
2W 8  -6.625  1.889
3B 6  -3.667  2.182
3H 2  2.5  3.779
3W 20  -2  1.195
4B 1  0  5.344

title 'Confidence Intervals Computed from Means & Standard Errors';
/* statements to produce plot shown in Figure 3 */
goptions dev=os2 targetdevice=os2prtm
  htext=1.3 ftext=centb htitle=1.5;
proc gchart;
  vbar group / discrete sumvar=z
    type=mean
    errorbar=top /* the default a is 0.05 */
    raxis=axis1
    maxis=axis2;
  axis1 label=(a=90 'POPULATION CELL MEAN') order=0 to 20 by 5 w=3;
  axis2 w=3;
footnote 'Fig. 3: GCHART Output from Ex 3';
  run; quit;

CONCLUSION

Although useful for presentation of results from statistical analysis, GPLOT and GCHART appear to be quite restrictive for the frequently desired plotting of means and associated confidence intervals. However, on closer examination it becomes possible to devise some fairly simple tricks that force these PROCs to plot exactly what you want. Two generalizations are provided that make these PROC's even more useful for summarizing data. The principles presented can be easily extended to other examples.

References:


SAS and SAS/GRAPH 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.
Fig. 1: GPHOT Output from Example 1

Fig. 2: GCART Output from Example 3

Fig. 2: GPHOT Output from Example 2