Using Annotate Datasets to Enhance Charts of Data with Confidence Intervals: Data-Driven Graphical Presentation

Gwen D. Babcock, New York State Department of Health, Troy, NY

ABSTRACT

Data and accompanying confidence intervals (e.g. 1.2 95% CI 1.0-1.5) are often presented in tabular formats that complicate the ability to compare data sets. A graphical format, as opposed to tabular, can be superior when it is desirable to quickly compare data sets. However, standard SAS graphs are not well suited to such presentation. This paper explores the use of the Annotate facility with PROC GLPLOT to produce graphical presentations of data with confidence intervals. Annotate datasets are powerful tools for customizing charts or graphs. These datasets can be used to produce data-driven changes of the color and size of symbols, lines, text, and other chart features. Using a graphical display instead of tables eases comprehension of data and comparison between categories.

INTRODUCTION

Part of my work for the New York State Department of Health includes preparing reports with numerous standardized mortality ratios (SMRs) and corresponding confidence intervals. A SMR is defined as the number of observed deaths in the study population divided by the number of expected deaths. The death rate in a reference population by age group is applied to the number of persons in each age group in the study population to obtain the expected number of deaths in the study population. My initial attempts to display data using PROC REPORT resulted in output consisting of innumerable mind-numbing tables. The data were not easily grasped; patterns were obscured; unusual numbers were buried in text and did not stand out. Therefore, I decided to present the data using range bar charts with horizontal bars. Range bar charts show the upper and lower boundaries of data groups. In this paper, the boundaries are upper and lower 95% confidence intervals, but could be maximums and minimums, multiples of the standard error, upper and lower specification values, or 10th and 90th percentile values, etc., as required by the project in question. In addition, range bar charts can show central values, such as the SMRs used here, or the mean, median, first value, last value, etc (Harris, 1999). Castellanos and Spanos (2004), Elkin et al (1997), Roohan and Zdeb (1989), and Mitchell (2003) describe how to produce output similar to that which I was seeking. I adapted these methods to produce the desired charts. I used Michell’s (2003) method to display the confidence intervals (boundaries), but used an Annotate dataset to display the ratios (central values) and to customize the graphs (Castellanos and Spanos, 2004; Elkin et al, 1997). Annotate datasets can easily be used to produce multiple graphs with BY variables. This paper assumes the reader has knowledge of the SAS data step and the SAS GPLOT procedure and related commands. I used SAS 9.1 running under Windows XP.

BACKGROUND

I calculated SMRs and 95% confidence intervals comparing mortality in selected areas of the state to statewide mortality rates. The SMRs for each area were calculated for males, females, and both sexes for several causes of death (Fig. 1). I wanted to produce horizontal range charts of the ratios and their confidence intervals for each cause of death by location and sex. Symbols and colors should be different depending on whether the data are unusually low or high. The exact data values are given to the right of the chart.

READING IN THE DATA

A sample of the data is shown in Fig. 1. The first step was to read in the data and assign each cause of death a

<table>
<thead>
<tr>
<th>cause</th>
<th>location</th>
<th>sex</th>
<th>observed</th>
<th>expected</th>
<th>SMR</th>
<th>lower CL</th>
<th>upper CL</th>
</tr>
</thead>
<tbody>
<tr>
<td>all cancers¹</td>
<td>Metropolis</td>
<td>both sexes</td>
<td>5705</td>
<td>5043.12</td>
<td>1.13</td>
<td>1.10</td>
<td>1.16</td>
</tr>
<tr>
<td>all cancers¹</td>
<td>Metropolis</td>
<td>females</td>
<td>2873</td>
<td>596.12</td>
<td>1.11</td>
<td>1.07</td>
<td>1.15</td>
</tr>
<tr>
<td>all cancers¹</td>
<td>Metropolis</td>
<td>males</td>
<td>2832</td>
<td>2288.89</td>
<td>1.24</td>
<td>1.19</td>
<td>1.28</td>
</tr>
<tr>
<td>COPD*</td>
<td>Metropolis</td>
<td>both sexes</td>
<td>99</td>
<td>81.29</td>
<td>1.22</td>
<td>0.99</td>
<td>1.48</td>
</tr>
<tr>
<td>COPD*</td>
<td>Metropolis</td>
<td>females</td>
<td>52</td>
<td>37.01</td>
<td>1.40</td>
<td>1.05</td>
<td>1.84</td>
</tr>
</tbody>
</table>

*chronic obstructive pulmonary disease
number for plotting purposes, using the variable “smrno” (Program 1). These numbers form the y-axis. Although I could have plotted cause of death as a qualitative variable, assigning a number allowed me to plot vertical bars at the ends of the confidence intervals and to have more flexibility when creating the annotate dataset later.

/*SAS® Program 1: read in data*/
data SMR;
  infile datalines dsd delimiter="," missover;
  length cause $49 location $12 sex $10;
  input cause $ location $ sex $ expected observed smr smr_LL smr_UL;
  /*create the labels to annotate the chart later*/
  smrlabel=trim(left(put(SMR,8.2)))||" ("||trim(left(put(smr_LL,8.2)))||","||trim(left(put(smr_UL,8.2)))||")";
  /*Assign numbers to causes for graphs. Graphing by number rather than categori-
  cal variable allows more flexibility in plotting and annotating*/
  if cause= "cancer of digestive organs" then smrno=6;
  else if cause="cancer of lung and bronchus" then smrno=5;
  else if cause="cancer without specification of site" then smrno=4;
  else if cause="Diseases of pulmonary circulation" then smrno=3;
  else if cause="COPD" then smrno=2;
  else if cause="all cancers¹" then smrno=1;
  datalines;
  all cancer¹, Metropolis, both sexes, 5043.12, 5705, 1.13, 1.10, 1.16
  ...
  Run;

PREPARING THE DATA

I could have used the “INTERPOL=HILO” option in a SYMBOL statement to create the plot, but the ratios and confidence intervals would be displayed vertically, not horizontally. To display the data horizontally, I assigned six coordinates to plot the confidence intervals as described by Mitchell (2003). The six coordinates form an I-beam shape and are connected using the “Interpol=JOIN” option in the SYMBOL statement.

/*SAS® program 2: assign coordinates to plot confidence intervals*/
proc sort data=smr;
  by location sex;
run;

/*create six coordinates for each location and sex which will be joined to form the horizontal range lines*/
data smrsix (keep=location cause sex smrno xx yy);
  set smr;
  /*the order of the following statements is important. If the order is changed, the confidence intervals may not appear correctly*/
  if smr_LL NE 0 then do;
    xx=smr_LL;
    yy=smrno+0.2; output;
    yy=smrno-0.2; output;
    yy=smrno; output;
    xx=smr_UL;
    yy=smrno; output;
    yy=smrno+0.2; output;
    yy=smrno-0.2; output;
  end;
else
  xx=0;
  yy=smrno;
output; /*need to plot zeros as well*/
run;
Once the coordinates were assigned, the data was plotted. The "yy*xx=smrno" syntax creates separate plots for each "smrno" on a single graph. Symbol statements are used to join the 6 points in each plot with thin black lines. The NOBYLINE option combined with the "#byval1","byval2"..."byvaln" syntax used in the TITLE statements allows me to use the by-values in the titles so they clearly explain the data used in the plot.

/*SAS® Program 3: plot the data*/
goptions reset=all ftext="CENTB" htext=24pt;

options nobyline;
symbol1 ci=black font=marker v=none line=1 width=2 interpol=join;
symbol2 ci=black font=marker v=none line=1 width=2 interpol=join;
symbol3 ci=black font=marker v=none line=1 width=2 interpol=join;
symbol4 ci=black font=marker v=none line=1 width=2 interpol=join;
symbol5 ci=black font=marker v=none line=1 width=2 interpol=join;
symbol6 ci=black font=marker v=none line=1 width=2 interpol=join;

proc gplot data=smrsix;
  title1 ls=0 height=24pt "Standardized mortality ratios (SMR's)";
  title2 ls=0 height=24pt "and 95% confidence intervals";
  title3 ls=0 height=20pt "comparing mortality within #byval1 to national";
  title4 ls=0 height=20pt "rates for #byval2 by cause of death, 1986-1995";
  by location sex;
  plot yy*xx=smrno /href=1 lhref=2 nolegend;
run; quit;

Figure 2. One of the plots of confidence intervals generated by program 3.

Standardized mortality ratios (SMR's) and 95% confidence intervals comparing Smallville rates to national rates for males by cause of death, 1986—1995

The resulting plot (Fig. 2) shows the ranges (confidence intervals) horizontally as desired, but it is very bare bones. The next step is to label the axes and change the numbers on the y-axis to descriptive labels using AXIS statements. As a further enhancement, symbols will be added and the data values will be displayed to the right of the graph. This can be accomplished using an Annotate dataset. It is also possible to use the Annotate dataset to plot the confidence intervals (Castellanos and Spanos, 2004; Okerson, 2002; Elkin et al, 1997), but then SAS® will be unable to automatically determine the length of the x-axis. Therefore, it must be specified in the AXIS statement.

ANNOTATING: MOVING TO ROCKY ROAD

An Annotate data set was used to add the symbols marking the SMR values to the plots and to place the values of the ratios to the right of the plot. In an annotate dataset, each observation is an instruction to SAS® to do something on a graph or chart. An annotate dataset can be used to add symbols, text, lines, rectangles, borders, pictures, pie slices, points, and polygons to a graph or to create a custom graph. The remainder of this paper will focus on adding symbols, text, and lines to the plot.
ADDITION SYMBOLS TO THE PLOT

The addition of symbols to the chart to mark the SMRs, required the creation of an annotate dataset containing
the SYMBOL function as shown in SAS® program 4. This program creates the dataset shown in Figure 3.

/*SAS® Program 4: create annotate dataset*/
data ptanno (keep=xsys ysys hsys when color text function size color x y location sex style);
  retain hsys '3' when 'a';
  /*when='a' indicates the annotate command is executed after the graph is drawn*/
  /*HYS='3' indicates that 1 size unit is 3% of the graphics output area*/
  format color $8. text $32. style $16. size 8.;
  set smr;
  by location sex;
  xsys='2'; ysys='2';
  /*create points for each SMR*/
  /*the type and color is based on statistical significance*/
  function='symbol';
  x=smr; y=smrno;
  style='marker'; size=2;
  if smr=0 then do;
    color='black';
    text='X';
  end;
  else if smr_UL<1 then do;
    color='green';
    text='D';
  end;
  else if smr_LL>1 then do;
    color='magenta';
    text='C';
  end;
  else do;
    color='black';
    text='P';
  end;
run;

The SYMBOL command requires several arguments, which are stored in annotate dataset variables. Some ar-guments are different, depending on the data. The X and Y variables indicate the location of the symbol. I chose
to place the symbols on the graph using the ratio ("smr") as the X value and the number of each cause of death ("smrno") as the Y value. IF-THEN statements allow me to change the color and shape of the symbol depending on whether the confidence intervals include one (black diamonds), the confidence limits are both greater than one (magenta triangles), or the confidence limits are both less than one (green triangles). The COLOR variable indi-cates the color of the symbol, and the TEXT variable indicates the symbol shape. I chose to keep the remainder of the arguments the same for all symbol functions. STYLE="marker" tells SAS® the font to use. WHEN="a" indi-

Figure 3. An annotate dataset.

<table>
<thead>
<tr>
<th>location</th>
<th>sex</th>
<th>function</th>
<th>hsys</th>
<th>xsys</th>
<th>ysys</th>
<th>when</th>
<th>color</th>
<th>text</th>
<th>style</th>
<th>size</th>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gotham City</td>
<td>both sexes</td>
<td>symbol</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>a</td>
<td>green</td>
<td>D</td>
<td>marker</td>
<td>2</td>
<td>0.86</td>
<td>1</td>
</tr>
<tr>
<td>Gotham City</td>
<td>both sexes</td>
<td>symbol</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>a</td>
<td>green</td>
<td>D</td>
<td>marker</td>
<td>2</td>
<td>0.72</td>
<td>2</td>
</tr>
<tr>
<td>Gotham City</td>
<td>both sexes</td>
<td>symbol</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>a</td>
<td>black</td>
<td>P</td>
<td>marker</td>
<td>2</td>
<td>0.80</td>
<td>3</td>
</tr>
<tr>
<td>Gotham City</td>
<td>both sexes</td>
<td>symbol</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>a</td>
<td>black</td>
<td>P</td>
<td>marker</td>
<td>2</td>
<td>1.01</td>
<td>6</td>
</tr>
<tr>
<td>Gotham City</td>
<td>both sexes</td>
<td>symbol</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>a</td>
<td>green</td>
<td>D</td>
<td>marker</td>
<td>2</td>
<td>0.76</td>
<td>5</td>
</tr>
<tr>
<td>Gotham City</td>
<td>both sexes</td>
<td>symbol</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>a</td>
<td>black</td>
<td>P</td>
<td>marker</td>
<td>2</td>
<td>0.84</td>
<td>4</td>
</tr>
<tr>
<td>Gotham City</td>
<td>females</td>
<td>symbol</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>a</td>
<td>green</td>
<td>D</td>
<td>marker</td>
<td>2</td>
<td>0.92</td>
<td>1</td>
</tr>
<tr>
<td>Gotham City</td>
<td>females</td>
<td>symbol</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>a</td>
<td>black</td>
<td>P</td>
<td>marker</td>
<td>2</td>
<td>0.69</td>
<td>2</td>
</tr>
<tr>
<td>Gotham City</td>
<td>females</td>
<td>symbol</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>a</td>
<td>black</td>
<td>P</td>
<td>marker</td>
<td>2</td>
<td>0.85</td>
<td>3</td>
</tr>
<tr>
<td>Gotham City</td>
<td>females</td>
<td>symbol</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>a</td>
<td>black</td>
<td>P</td>
<td>marker</td>
<td>2</td>
<td>1.16</td>
<td>6</td>
</tr>
</tbody>
</table>
cates that the command is to be executed after the graph is drawn. The SIZE variable tells SAS® to use a text size of “2”, and the HSYS variable indicates text size units—in this case 3% of the graphics output area. The XSYS, and YSYS variables indicate the units and origin of the X and Y coordinates; in this case, the data values are used, and the origin of the plot is the 0,0 point. Another possibility is to use a percentage of the graphics output area, with the origin being the lower left corner of the area. When plotting the graphs using PROC GPLOT, the ANNO= option of the PLOT statement tells SAS® to use the commands in the annotate dataset. If by-variables are used, the annotate dataset should contain the same by-variables as the plot dataset.

/*SAS® Program 5: plot the data with the ANNOTATE dataset*/

proc gplot data=smrsix;
   title1 ls=0 height=24pt "Standardized mortality ratios (SMR's)";
   title2 ls=0 height=24pt "and 95% confidence intervals";
   title3 ls=0 height=20pt "comparing mortality within #byval1 to national";
   title4 ls=0 height=20pt "rates for #byval2 by cause of death, 1986-1995";
   by location sex;
   plot yy*xx=smrno /href=1 lhref=2 nolegend annotate=ptanno;
run;quit;

Figure 4. One of the plots generated by Program 5, indicating how the Annotate dataset is used to provide custom symbols.

Standardized mortality ratios (SMR’s) and 95% confidence intervals comparing Smallville rates to national rates for males by cause of death, 1986—1995

ADDING TEXT TO THE PLOT

Next, I used the annotate command LABEL to place additional text next to the chart. To locate the text outside of the chart area, I changed the XSYS value to 3, which tells SAS® that the value of X will represent a percentage of the graphics output area. The STYLE variable dictates the font to be used for the text. The SIZE variable tells SAS® the text size. The POSITION=">" statement tells SAS® the text should be centered vertically and left aligned horizontally on the coordinate given. The COLOR variable carries over from the symbol commands, so the text will be the same color as the corresponding symbol.

I also used annotate commands to create a legend for the chart and to place a header over the additional text to the right of the chart. Since these only need to be done once for each chart, I used the SAS® automatic variable first.<byvar> in a IF-THEN statement so the headings and legend were only created once for each by-group. In addition to the TEXT and SYMBOL commands, the MOVE and DRAW commands were used to place lines under the headers. The MOVE command gives the location of the start of the line, and the DRAW command gives the location of the end of the line.
/*SAS® Program 6: create annotate dataset with additional commands*/
data ptanno2 (keep=xsys ysys hsys when color text position function size color x y location sex style);
  retain hsys '3' when 'a';
  format color $8. text $32. style $16. size 8.;
  set smr; by location sex; xsys='2'; ysys='2';
/*add symbols for each SMR*/
  function='symbol'; x=smr; y=smrno; style='marker'; size=3;
  if smr=0 then do; color='black'; text='X'; end;
  else if smr_UL<1 then do; color='green'; text='D'; end;
  else if smr_LL>1 then do; color='magenta'; text='C'; end;
  else do; color='black'; text='P'; end;
  output; /*output adds the symbol commands to the dataset*/
/*put the values of each SMR to the right of the graph*/
  function='label'; style='CENTB'; text=smrlabel; y=smrno; size=2.5;
/*table to right of graph area*/
    xsys='3'; x=90; position='>'; output;
    x=78.0; text=smrlabel; output;
/*command to print the ratio to the right of the graph*/
  if first.sex then do;
/*all of the following annotations should be done once for each graph*/
    size=3; color='black'; style='CENTB';
/*header for column of the table to the right of the graph*/
    y=7;
    x=79.2; text="SMR"; position='>'; output;
    y=6.7;
    x=79.2; text="(95%CI)"; position='>'; output;
/*draw lines under the header text*/
    y=6.5;
    function='move'; x=79.2; output;
    function='draw'; x=88; output;
/*draw line under vertical axis label "cause of death"*/
    y=6.85;
    function='move'; size=0.1; x=1; output;
    function='draw'; size=0.1; x=27; output;
/*create a legend using text and symbols*/
    xsys='3'; ysys='3'; /*change to coordinate system relative to entire graphics area*/
/*text for legend*/
    function='label'; color='black'; position='6'; style='CENTB'; size=2.5; y=4;
    x=4; text='as expected'; output;
    x=28; text='lower than expected'; output;
    x=65; text='higher than expected'; output;
/*symbols for legends*/
    function='symbol'; style='marker'; size=2; y=3.3;
    x=2.5; color='black'; text='P'; output;
    x=26.5; color='green'; text='D'; output;
    x=63.5; color='magenta'; text='C'; output;
  end; /*end of the once per graph annotation commands*/
run;

To make space for the values printed to the right of the graph, I needed to make the graph smaller. I did this by using the LENGTH option of the AXIS statement. I used the AXIS statement to label the X-axis and the tick marks of the Y-axis. I chose to label the tick marks using the VALUE clause of the AXIS statement rather than formats to permit greater control of the appearance of the labels when using ODS.
/*SAS® Program 7: plot the data*/
/*define vertical axis*/
axis2 label=none
  minor=none
  order=(0 to 7 by 1)
  value=( height=20pt J=L/*note that tick numbers are not the same as smrno val-
  ues*/
    /*label y-axis*/
    tick=1 color=blue ' '
    tick=2 color=blue 'All cancers²'
    tick=3 color=black 'COPD'
    tick=4 color=black 'heart attack'
    tick=5 color=black 'unknown cancer'
    tick=6 color=black 'lung cancer'
    tick=7 color=black 'colon cancer'
    tick=8 color=black 'Cause of Death' j=L);
/*define horizontal axis*/
axis1 label=(height=20pt 'SMRs')
  order=(0,0.1,1 to 5 by 1)
  length=47pct origin=(27 pct)
  value=(height=20pt);
proc gplot data=smrsix;
title1 ls=0 height=24pt "Standardized mortality ratios (SMR's)";
title2 ls=0 height=24pt "and 95% confidence intervals";
title3 ls=0 height=20pt "comparing mortality within #byval1 to national";
title4 ls=0 height=20pt "rates for #byval2 by cause of death, 1986-1995";
by location sex;
plot yy*xx=smrno /haxis=axis1 vaxis=axis2 href=1 lhref=2 nolegend anno-
tate=ptanno2;
footnote1 height=10pt " ";
run;quit;
Figure 5. One of the plots generated by program 7, showing how the Annotate dataset is used to provide custom symbols and text.

**Standardized mortality ratios (SMR's) and 95% confidence intervals comparing Smallville rates to national rates for males by cause of death, 1986—1995**

<table>
<thead>
<tr>
<th>Cause of Death</th>
<th>SMR ( (95%CI) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>colon cancer</td>
<td>1.82 ( (1.11, 2.81) )</td>
</tr>
<tr>
<td>lung cancer</td>
<td>0.70 ( (0.39, 1.15) )</td>
</tr>
<tr>
<td>unknown cancer</td>
<td>1.74 ( (0.70, 3.59) )</td>
</tr>
<tr>
<td>heart attack</td>
<td>0.72 ( (0.01, 4.01) )</td>
</tr>
<tr>
<td>COPD</td>
<td>1.59 ( (0.96, 2.48) )</td>
</tr>
<tr>
<td>All cancers</td>
<td>0.61 ( (0.50, 0.75) )</td>
</tr>
</tbody>
</table>

**SMRs**
- ◆ as expected
- ▼ lower than expected
- ▲ higher than expected

**CONCLUSIONS**
The Annotate data set is a powerful tool for customizing plots. It can enable precise control over the plotting process, and can allow for the addition of symbols, text, and lines to existing plots. The attributes of these features can be modified based on the data plotted. Creating customized range graphs of this data makes it easier to comprehend and compare data than if it were presented in tables.
REFERENCES:


ACKNOWLEDGMENTS
SAS® is a Registered Trademark of the SAS Institute, Inc. of Cary, North Carolina.

Windows XP® is a Registered Trademark of the Microsoft Corporation, Redmond, WA.

This work was supported in part by US Department of Health and Human Services, Centers for Disease Control and Prevention grant U50/CCU422440-03 for Environmental Public Health Tracking.

CONTACT INFORMATION
Your comments and questions are valued and encouraged. Contact the author at:
Gwen Babcock
New York State Department of Health
547 River St.
Troy NY 12180-2216
Voice: 518 302 7950
Fax: 518 402 7969
gdb02@health.state.ny.us

********************************************************************