Paper 088-31

Graphing Laboratory Data and an Introduction to the Custom Macro %GRAPHME

Adrienne Groulx, Scian Services Inc, Etobicoke, Ontario, Canada

ABSTRACT

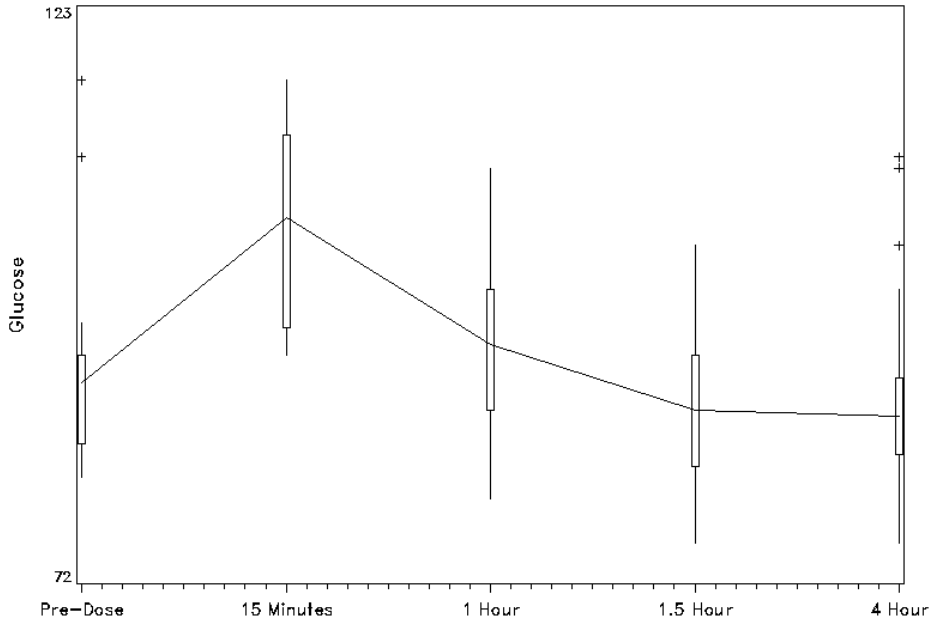
We all know that a picture speaks volumes when it comes to relaying a message. This also holds true when one is graphically displaying clinical data. Producing a clear visual representation of laboratory results is the most important method of determining the existence of abnormal values, sampling errors, or other confounding factors within a data set. This paper discusses common mistakes that are made in clinical graphs, and gives guidelines for producing plots that can be clearly interpreted by any data reviewer. The SAS macro, %GRAPHME, which was created to identify possible sources of error and display the most efficient vertical scaling of the data of interest, will also be introduced. The macro utilizes both user-defined reference parameters, as well as data-driven reference points, and uses the GPLOT procedure to plot the data. The macro is intended for beginner to intermediate users.

INTRODUCTION

Graphing laboratory data, or any data of significance, is not a trivial task. Properly displayed output is essential in order to draw the viewer's attention to important attributes of the data. In the case of laboratory data, a well-produced graph could aid in the diagnosis of a newly observed disease in a patient, or highlight a favorable (or unfavorable) trend in a laboratory parameter due to an investigational drug. A poorly produced graph, however, could lead to misinterpretations of the data such that unexpected results go unnoticed or normal results cause undue concern.

Some of the most common graphing errors include using improper scaling or layout of the x- and y-axes, and not providing enough relevant details on the graph. In Figure 1, the displayed output does not provide a clear story of the data.

Figure 1. Example of a Bad Graph



What went wrong:

- The scaling of the x-axis is inaccurate. The same distance is used between the 15-minute to 1-hour observations as is used between the 1.5-hour to 4-hour observations. This error has the potential of hiding a significantly steep increase of the parameter in question.

- The y-axis does not provide a clear reference for the values ranging between 72 to 123, nor is there a unit of measure provided for the values.

Other ways of improving the visual appearance of Figure 1 include adding a title, providing more details on the axes (was the glucose measurement taken from fasting subjects?), and providing reference lines for the expected normal range of glucose values in the population being summarized.

The use of reference lines or normal ranges can be an extremely valuable tool for highlighting suspect results. Values observed outside of a defined normal range could raise some valid questions:

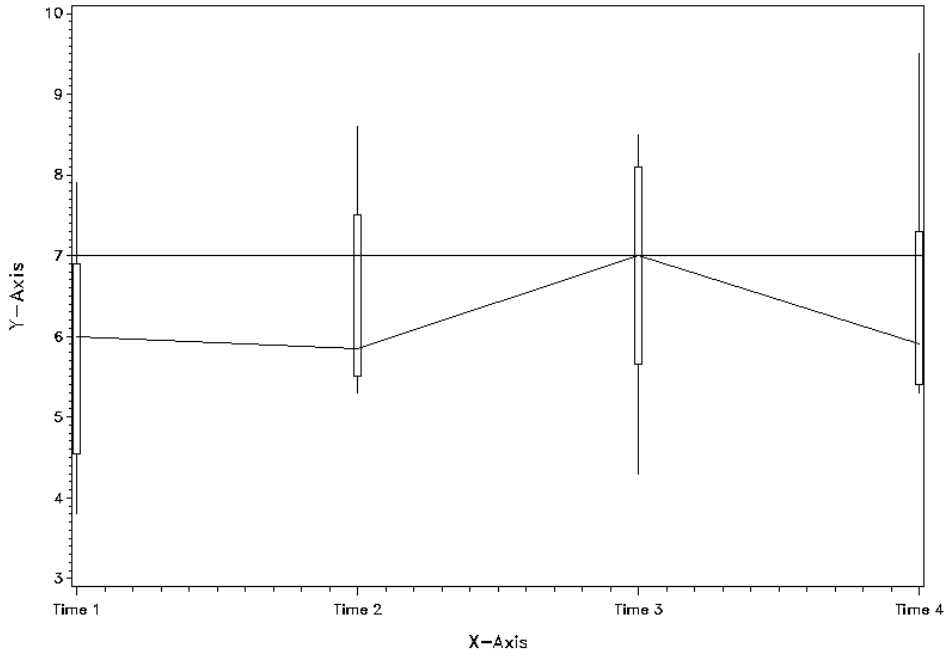
- Was the unexpected value due to poor sample handling?
- Was there any contamination in the lab equipment?
- Was the equipment calibrated properly?

If the unexpected result did not come from a sampling error, then questions could be made about the data itself:

- Was human error involved (ex. in transcribing the data)?
- Did the result in question use a different unit of measurement than the units intended for the graph?
- Was the normal range for the graph intended for a different population (ex. age-specific or gender-specific normal range values were used on an opposing age group or gender)?

SAS is already an excellent graphing tool, and will choose very suitable axes for your data. In some cases however, over-riding the SAS default axis in order to display reference lines that are outside of the observed data range may be desirable. An example of this is shown in Figure 2.

Figure 2. Example of a Cut-off Normal Range



What went wrong:

- When users look at this graph, they aren't able to determine if the reference value of '7' is an upper limit or a lower limit.
- The intended normal range for this laboratory parameter was between 2 to 7, but without printing the lower reference line onto the graph, the results are hard to interpret.

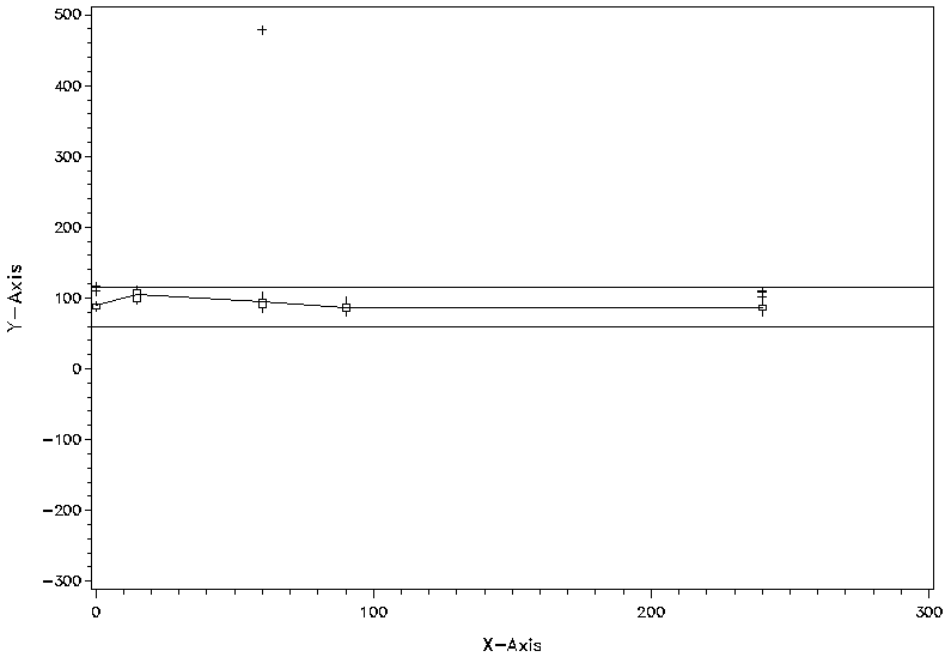
In this example, the default y-axis generated by the GPLOT procedure did not force the lower reference line to be printed, since it fell below the lowest observed value. In order to have a clear understanding of the results, both reference lines on the y-axis should be forced in the graphical display.

PLOTTING AND INTERPRETING YOUR DATA

Highlighting important properties of data requires the proper use of vertical axes, reference lines, and data-driven allowances. Remember, the purpose of visualizations is to support the reviewer in the process of diagnosis and interpretation of laboratory tests.

The choice of the y-axis range is a key element to presenting laboratory data. A nice quality to have in this axis is symmetry about the lab normal range, so that the viewer's eyes are centered at normal, and can easily view if there are any tendencies away from normal. Of course, this symmetry is not always possible in the event of extreme values in one direction (see Figure 3). In such cases, a data-driven y-axis is more appropriate.

Figure 3. Example of Bad Symmetry



What went wrong:

- One extreme value has forced an unreasonable range in the y-axis.
- The y-axis extends from negative to positive results, which is unlikely in laboratory data.

Any pros to this output?

- The extreme value is emphasized in this graph, and the user is made aware or potential errors in the transcribed data results.

Key rules to remember when plotting laboratory data are:

- show all the data, including extreme values - hide nothing!
- avoid distorting the data
- direct the reader eyes to important features of the data

When interpreting data, consider this: reference ranges of laboratory tests may differ based on age, gender, or other demographic factors. The range that is used when plotting data should be the appropriate range expected for the population that's being observed. For example, hyperglycemic patients should expect higher glucose levels than 'healthy' patients. Also be aware that scores outside of the normal range can be the result of sleep, diet, exercise, medicines, and vitamin supplements.

THE MACRO, %GRAPHME

The macro, %GRAPHME, functions on some assumptions regarding the data being submitted for graphing. The first assumption is that the data set has been reduced to include only laboratory results that are expected to fall within the normal range being provided by the user. That is, if the laboratory parameter has a gender-specific normal range, then only the appropriate gender-specified subjects are included within the submitted data set. The second assumption is that only one patient population is being summarized (i.e. only one treatment group at a time). The macro will combine all subjects in the data set together, and a boxplot will be produced to display the results. Boxplots were chosen to display the laboratory results in order to highlight extreme values of individual observations.

The x-axis variable supplied by the user is assumed to be properly ordered. If the time variable is in character format, then the order of this time variable should be properly set within the data set prior to calling upon the macro. Note that if a character time variable is entered, the macro will put equal distances between each of the elements in the time parameter. A warning will automatically be generated to warn the user of the possible hazards of using categorical time variables. The best way to avoid mistakes would be to create a time variable within the data set that reflects an appropriate numerical time. If time was captured in both hours and days throughout the lab recordings, then the best way to represent the data would be to standardize all times to one format (i.e. either hours or days).

The macro %GRAPHME will be called using the following structure:

```
%GRAPHME(
    inputdata=          /*Data set to analyze   */
    lowref=             /*Lower reference point */
    highref=            /*High reference point  */
    xaxis=              /*Time variable         */
    yaxis=              /*Laboratory variable   */
    title=              /*Title of graph        */
    xaxislabel=         /*Label of the horizontal axis                    */
    yaxislabel=         /*Label of the vertical axis                      */
    suppress=ON         /*Option to suppress notes and source code (ON/OFF)*/
    )
```

The parameter “suppress” was added in order for the user to decide how much information the log should contain. The default is set to “ON”, so that only messages written into the log from the %GRAPHME macro will be produced. In general, it is not recommended to suppress SAS notes, but the intent of this macro is to create a simple summary of the data, and to provide its highlights.

The first job of the macro is to determine whether the parameters provided are sufficient to allow a graph to be produced. A graph will not be created in certain cases. For example, if any of the crucial parameters (i.e. inputdata, lowref, highref, xaxis, or yaxis) are not provided, the following code in the macro will end the session:

```
/* verify all necessary parameters were provided*/
%if %length(&yaxis)=0 or %length(&xaxis)=0 or %length(&lowref)=0 or
%length(&inputdata)=0 or %length(&highref)=0 %then %do;
    %put; %put PROGRAM STOPPED: At least 1 key parameter was not provided;
   %goto finish;
%end;
```

Similarly, the macro %GRAPHME will be stopped in each of the following situations, and a message will be sent to the log in order to will highlight the macro parameters causing grief:

- The reference parameters, lowref and highref, were not numeric
- The data set, inputdata, does not exist
- The data set, inputdata, exists, but contains no data
- The parameter(s) xaxis or yaxis do not exist within the data set inputdata
- The parameter yaxis is not numeric
- The parameter yaxis is numeric, but contains all missing values

If all of the above validation checks are satisfied, then the next job of the macro is to find elements in the data that could lead to errors in the overall interpretation of the graphical results. The first data warning is given if the x-axis variable was provided in character format. As mentioned previously, putting the horizontal axis in numerical format is strongly advised. Another warning is generated if any negative values were observed in the y-axis variable, since this could suggest data that was transcribed incorrectly. Missing values in the y-axis variable will also produce a warning. In the case that a sample for a patient was missed at a specific time, this message will guide the user to review the data and verify “valid” missing observations. Finally, if the lower reference or upper reference values were negative, a caution message is produced, as negative values are not expected in graphs of observed laboratory data. In each of these cases, a graph will still be generated. All warning messages are printed to the log, and a total count of all warning messages created is shown in the graph’s secondary title.

Other messages are printed to the log, in order to provide the user with some general features of the data. For example, the number of non-missing observations for the x-axis and y-axis variables are printed. Also, the minimum and maximum values of the y-axis variable are printed using the macro statements below:

```
PROC SQL noprint;   /*collect the minimum and maximum lab values*/
select min(&yaxis) into :minimum from &inputdata;
select max(&yaxis) into :maximum from &inputdata;
quit;

%put The maximum value of the data set is &maximum;
%put The minimum value of the data set is &minimum;
```

To inform the user of potential outliers within their data set, results outside the normal range are printed into the log.

Ex. At time x_1, n_1 ($\%_1$) results are above the upper limit.
n_2 ($\%_2$) results are above the upper limit by more than ½ length of the normal range.
n_3 ($\%_3$) results are above the upper limit by more than 1 length of the normal range.
The number of values below the lower limit are printed as well, then the same is produced for x_2, and so on if these values exist. The percentage is based on the total number of observations at the time specified. It is up to the user to assess whether the outlier results are clinically significant.

Example - Implementing the macro %GRAPHME:

To illustrate to macro %GRAPHME in action, data collected from a clinical trial with over 200 patients was used. The laboratory data for this trial was retrieved over a period of 8 months, with blood samples drawn at Baseline (i.e. Month 0), and Months 1, 2, 3, 4, 6, and 8. The original database for this trial contained a complete laboratory data set that included results for both male and female patients. In our example, a reduced data set, "lab" is used to graph the results for aspartate aminotransferase (AST). The data set was reduced to include only male patients, who have an AST normal range between 9 U/L and 43 U/L.

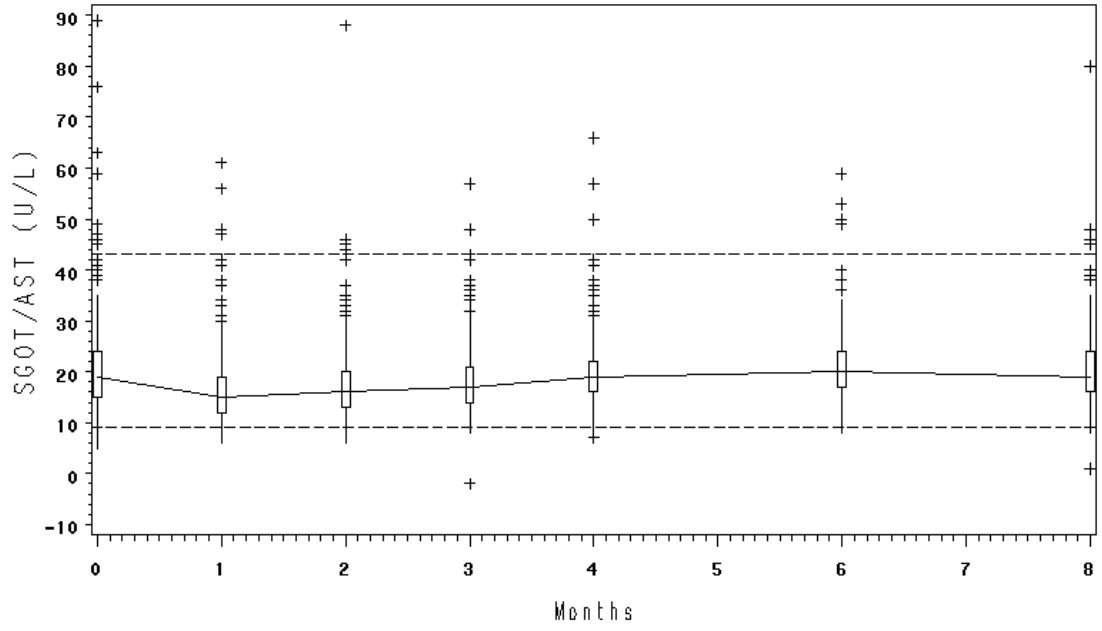
The following call to the macro was implemented:

```
%GRAPHME(inputdata=lab,
         lowref=9,
         highref=43,
         xaxis=month,
         yaxis=ast,
         heading="Figure 4. Output of AST Data Using the Macro GRAPHME",
         xaxislabel="Months",
         yaxislabel="SGOT/AST (U/L)",
         suppress=ON);
```

Figure 4 is the graph produced from the above macro call. A note is used in the secondary title to let the user know that, although the graph was produced, there were two (2) elements within the data set "lab" which should be reviewed. Another observation to make for this graph is the lack of symmetry about the normal range. In this example, results for AST can be close to zero. In such cases, the macro %GRAPHME will forego the need for symmetry, and will try its best to keep the lower limit of the axis equal to zero, unless an observation lies below this value.

Figure 4. Output of AST Data Using the Macro GRAPHME

```
Note:   2 caution messages for this table have been output into the SAS log
```



The log contents are displayed in the box below. We can see in the log that the two caution messages arose from data observed in the y-axis parameter. Namely, at least one negative value was observed, as well as at least one missing value. It is now up to the data reviewer to verify that the results are accurate. At this point, the reviewer should also consider if any of the results outside of the given normal range are of clinical significance.

```
Number of non-missing observations in the x-axis variable "month" =     1411
Number of non-missing observations in the y-axis variable "ast" =     1410

The maximum value of the dataset is       89
The minimum value of the dataset is       -2

CAUTION: negative values observed in the y-axis variable "ast"
CAUTION: missing values observed in the y-axis variable "ast"

At time 0 , 9 (  4.0%)  results are above the upper limit
At time 0 , 3 (  1.3%)  results are above the upper limit by more than ½ length of the normal range
At time 0 , 1 (  0.4%)  results are above the upper limit by more than 1 length of the normal range
At time 0 , 4 (  1.8%)  results are below the lower limit

At time 1 , 5 (  2.3%)  results are above the upper limit
At time 1 , 1 (  0.5%)  results are above the upper limit by more than ½ length of the normal range
At time 1 , 9 (  4.2%)  results are below the lower limit

At time 2 , 4 (  2.0%)  results are above the upper limit
At time 2 , 1 (  0.5%)  results are above the upper limit by more than ½ length of the normal range
At time 2 , 1 (  0.5%)  results are above the upper limit by more than 1 length of the normal range
At time 2 , 4 (  2.0%)  results are below the lower limit

At time 3 , 2 (  1.0%)  results are above the upper limit
At time 3 , 2 (  1.0%)  results are below the lower limit

At time 4 , 3 (  1.6%)  results are above the upper limit
At time 4 , 1 (  0.5%)  results are above the upper limit by more than ½ length of the normal range
At time 4 , 2 (  1.0%)  results are below the lower limit

At time 6 , 4 (  2.2%)  results are above the upper limit
At time 6 , 1 (  0.6%)  results are below the lower limit

At time 8 , 4 (  2.0%)  results are above the upper limit
At time 8 , 1 (  0.5%)  results are above the upper limit by more than ½ length of the normal range
At time 8 , 1 (  0.5%)  results are above the upper limit by more than 1 length of the normal range
At time 8 , 2 (  1.0%)  results are below the lower limit
```

CONCLUSION

When graphing laboratory data, the ability to draw a viewer's attention to important attributes of the data is a fundamental element in the making of a successful graph. This ability can lead to the early diagnosis of a newly observed disease in a patient, or suggest safety concerns of an investigational drug. A graph should neither hide data, nor allow data to become distorted and misinterpreted. The macro introduced in this article was designed to highlight features in a set of data, which could help a reviewer determine any sources of concern, and illustrate trends in laboratory results over time. Some current limits of the macro %GRAPHME include:

- Only one normal range is expected for the population being plotted (i.e. lab parameters with gender-specific or age-specific normal ranges currently need to be separated based on gender and age).
- One treatment group is plotted at a time.
- Error logs are limited to showing only messages as defined in this paper.
- X-axis limitations (character formats are not converted to numeric).
- All results observed at a given time are summarized as a whole (i.e. individual subjects are not plotted separately).

Further development of the macro is anticipated, in order to cater to a more general set of data.

REFERENCES

Tufte, E.R. (1983). *The Visual Display of Quantitative Information.* Cheshire, CT: Graphics Press.

Clay Helberg (1995). *Pitfalls of Data Analysis (or How to Avoid Lies and Damned Lies)* Retrieved August 5, 2005, from “http://my.execpc.com/~helberg/pitfalls/”.

ACKNOWLEDGEMENTS

I would like to thank Noemi Toiber Temin for helping to trouble-shoot the macro %GRAPHME, and St. Clare Chung for her encouragement of this paper.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Adrienne Groulx
Scian Services Inc
Etobicoke, Ontario Canada
E-mail: agroulx@scian.com
Web: www.scian.com