

Strategies for Data-Driven, Customized Graphical Displays

Richard W. Morris, Analytical Sciences, Inc., Durham, NC

Abstract

Some simple alternative approaches to producing multiple plots are outlined and illustrated.

Introduction

Interactive graphical software can provide deep visual insight into the structure of an individual data set. But, sometimes you want to produce a large number of graphical displays of a similar style from many data sets. If you need lots of plots, you will want to spend as little time as possible on each plot, accommodate variation in data content and structure in each plot, and employ customized graphical features when useful. In this article, I discuss strategies for using SAS® software to produce lots of plots without sacrificing flexibility of content as determined by the data. Examples are drawn from our work with the National Toxicology Program and illustrate several alternative strategies.

Rectangular Data Structure

The strategy you take in developing code to produce a number of plots depends on how your data are structured. Suppose you have a large rectangular data set in which rows are indexed by increasing values of the variable time and the many columns represent variables to be plotted against time. You begin by using PROC GPLOT to plot a single variable against time. You can produce another plot for a different variable by simply substituting its name for the name of the first variable into the PLOT statement.

You can automate this procedure by placing your code in a macro and passing the name of the variable to the PLOT statement as a macro variable. Now, you simply call the macro as many times as you wish to plot variables in the data set and you get lots of plots easily.

Linear Data Structure

Now suppose you have the same time dependent information structured as many separate data sets, each with a time variable and another variable of interest. These data sets can be stacked into a single

data set with three variables: time, the variable to be plotted, and a third variable that indexes the data set from which each section of data came. The third variable changes value each time the other two come from a new data set. Thus the third variable can be used for by-group processing.

Using this data set structure, you place a by-group statement with the name of the third variable in PROC GPLOT and execute the code. You get the same plots as before with the rectangular data set.

Strategy

A choice between these alternative approaches depends on several considerations. If your data come in one of these structures, you may be inclined to choose the approach suited to that structure. Of course, PROC TRANSPOSE makes both approaches accessible from either data structure. With either method, it is possible to include customization using the annotate facility. Annotate supports by-group processing, allowing construction of an annotate data set with by-groups.

The macro approach requires keeping track of the variables and specifying the code for each macro call. On the other hand, by-group processing does not require knowledge of the number of groups and hence is more flexible. For by-group processing, construction of a linear data set structure may require some effort since many data sets come in rectangular form. An additional consideration is that with by-group processing the variable being plotted has the same name on all plot axes with the by-group noted either in a BY= line at the top of the plot or in the plot title using a #BYVAL option. The macro approach, in contrast, has the variable name included in the macro call written on each plot axis.

An Illustration

Figure 1 shows a dense display of information about hemoglobin concentration in F344 rats. The data represent one of 44 endpoints plotted for each sex. The data consist of observations from control animals from about 40 studies with approximately 10 animals per study. The display shows the distribution of study

means as well as a frequency histogram and cumulative distribution for individual observations. Summary statistics are written on the display for convenience of access.

This display was produced using PROC GPLOT with by-group processing and the annotate facility. The bottom section of the display, depicting the cumulative distribution, was drawn with the PLOT statement while the upper two panels used the PLOT2 statement with the axis label suppressed. Vertical reference lines were used to separate the panels. Study means were plotted with + symbols and a line was drawn from the smallest to largest mean. A dot over the + symbol signals superimposed duplicate values. An annotate data set with by-groups was used to write text and numbers obtained from output of PROC MEANS.

This display illustrates that it is possible to get sophisticated custom results from PROC GPLOT with by-group processing.

Another Data-Driven Method

SAS plotting procedures have substantial flexibility, but they cannot solve all multiple-plot challenges. The most complete flexibility in graphical output is available with the Data Step Graphics Interface (DSGI). This facility allows you to draw whatever you want on a blank piece of paper, including lines, polygons, text, and numbers. You draw from within a data step which allows by-group processing and all the flexibility in output your imagination can manage. Since others have presented details of the DSGI (Whitney 1995), I do no more than note that there is a steep learning curve that is well rewarded.

An example in which we have used the DSGI is in plotting data from 13 week toxicity studies. Variation in protocol for these studies made it profitable to develop the ability to plot endpoints in several ways depending upon the data structure. Two forms of output are illustrated in figures 2 and 3.

Figure 2 shows dose-response plots for three endpoints. The three panels show body weight, heart weight, and right kidney weight. The sample mean and standard error are given along with sample size for each dose group. Placing three endpoints together conserves paper and allows endpoint comparisons in the pattern of dose response.

Figure 3 shows a dose-response plot for hemoglobin. In this study design the endpoint was observed at three separate times in the course of the study,

making it useful to display the dose-response data on a common scale. This was accomplished with DSGI by using time as one of the by-group variables to control the type of plot panel displayed. In this example vertical dotted lines are used to delimit the time that samples were taken, which is noted at the top of each panel.

These two figures were produced with DSGI and by-group processing. This approach was chosen to obtain the flexibility needed to incorporate multiple times of observation on the same axis, which would not otherwise be possible with the PROC approach.

Another Example

Figure 4 illustrates a multipanel plot depicting tumor incidence in a 2 year carcinogenicity study. The dark bars show tumor incidence in each dose group of males and females for a particular type of tumor. The light bars show tumor incidence adjusted for intercurrent mortality. The shaded field represents the range of historical control incidences for that type of tumor and the horizontal line is the median of historical control values. A legend is included at the bottom of the display.

This display incorporates by-group processing and uses the DSGI to produce the output. DSGI was used for the flexibility required to produce the legend, shading, and tumor labeling. This graphical display accesses a number of data sets to obtain the data summarized. DSGI makes it easy to incorporate all of the information into a single panel and provides flexible data step programming to modify the display for studies in which some tumor types are missing.

Conclusion

There are many ways to solve any given programming challenge. Some of the approaches to multiple plots considered here are worth reviewing when planning a data driven, custom set of graphical displays.

References

Whitney, C. W. (1995). DSGI: The OTHER Way To Do SAS Graphics. *Proceedings of the Twentieth Annual SAS Users Group International Conference*. SAS Institute Inc., Cary, NC. pp. 733-738.

SAS is a registered trademark or trademark of SAS Institute, Inc. in the USA and other countries.
© indicates USA registration.

Figure 1

Hemoglobin Concentration g/dl

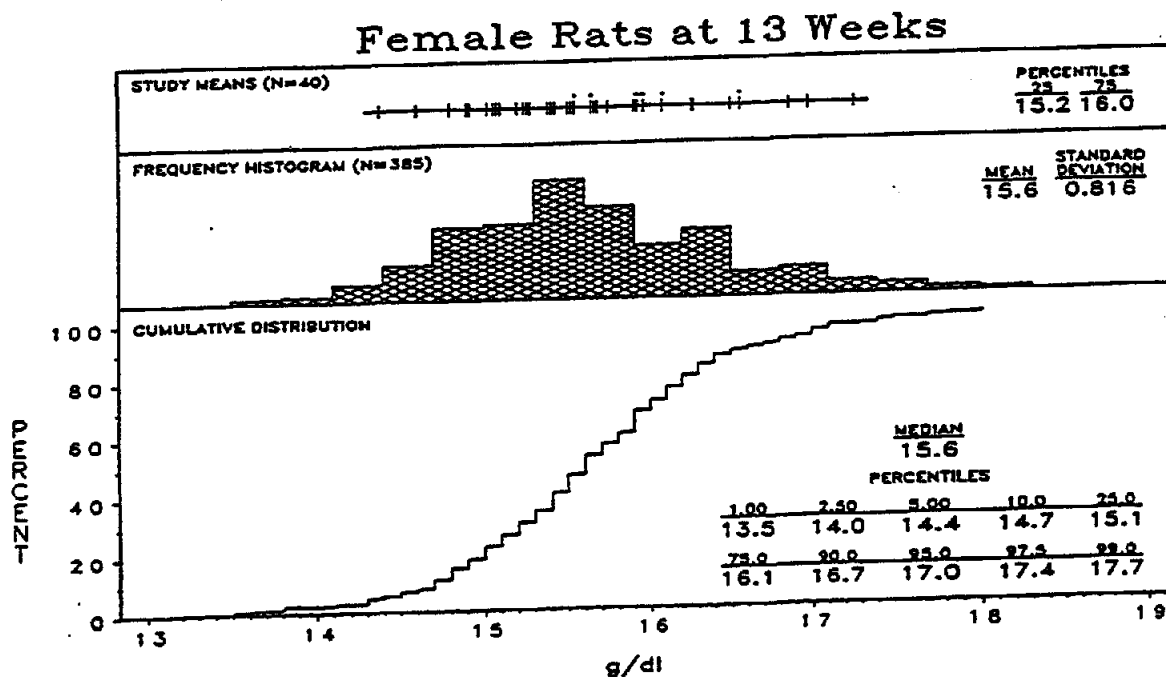
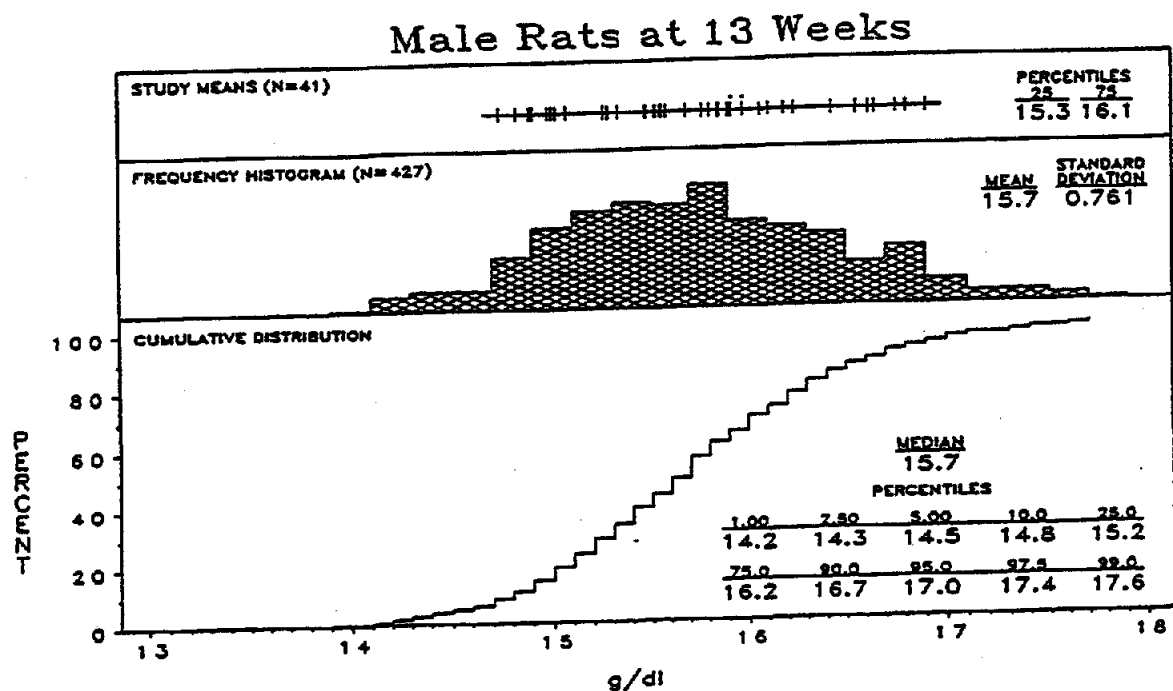


Figure 2

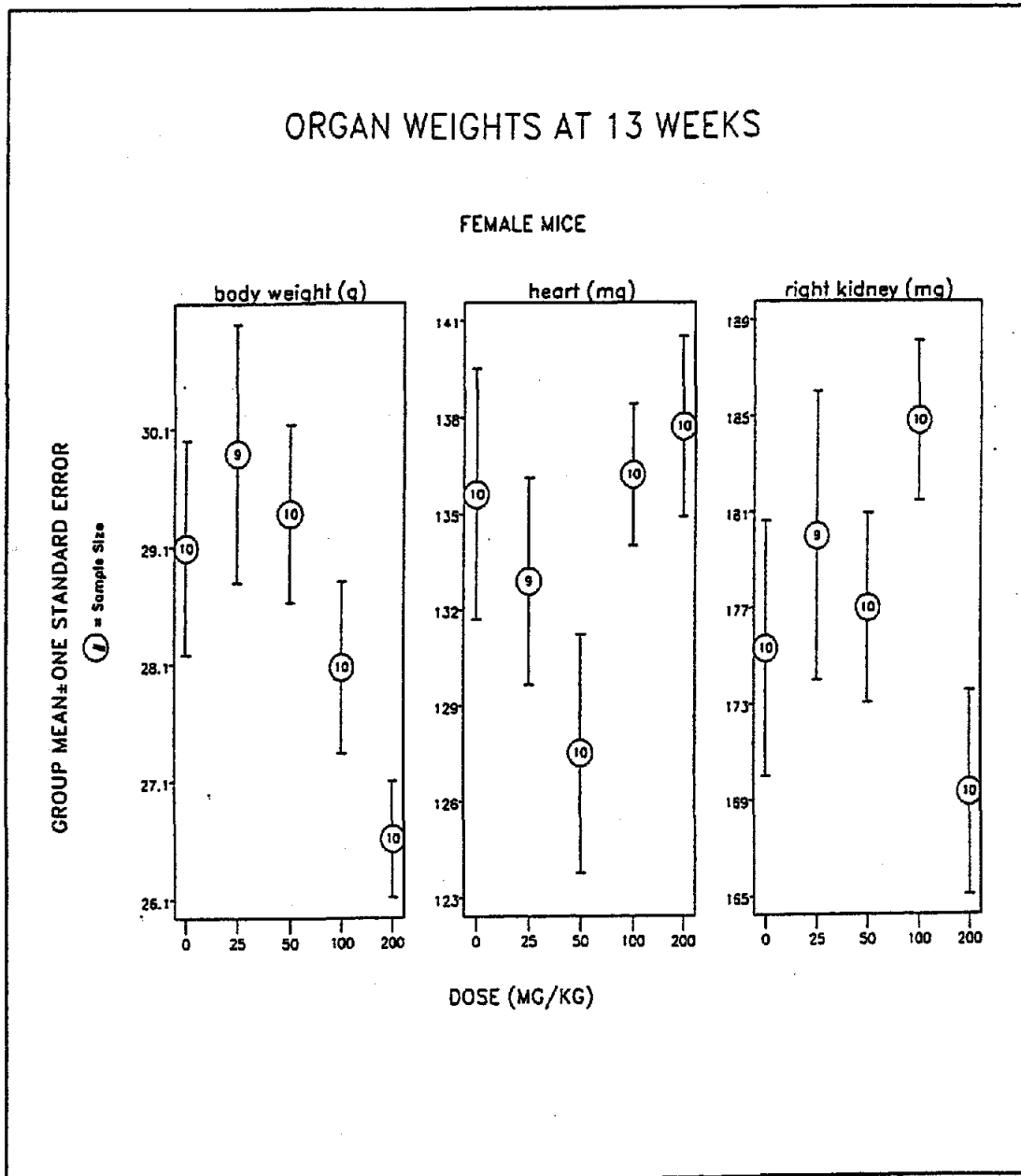


Figure 3

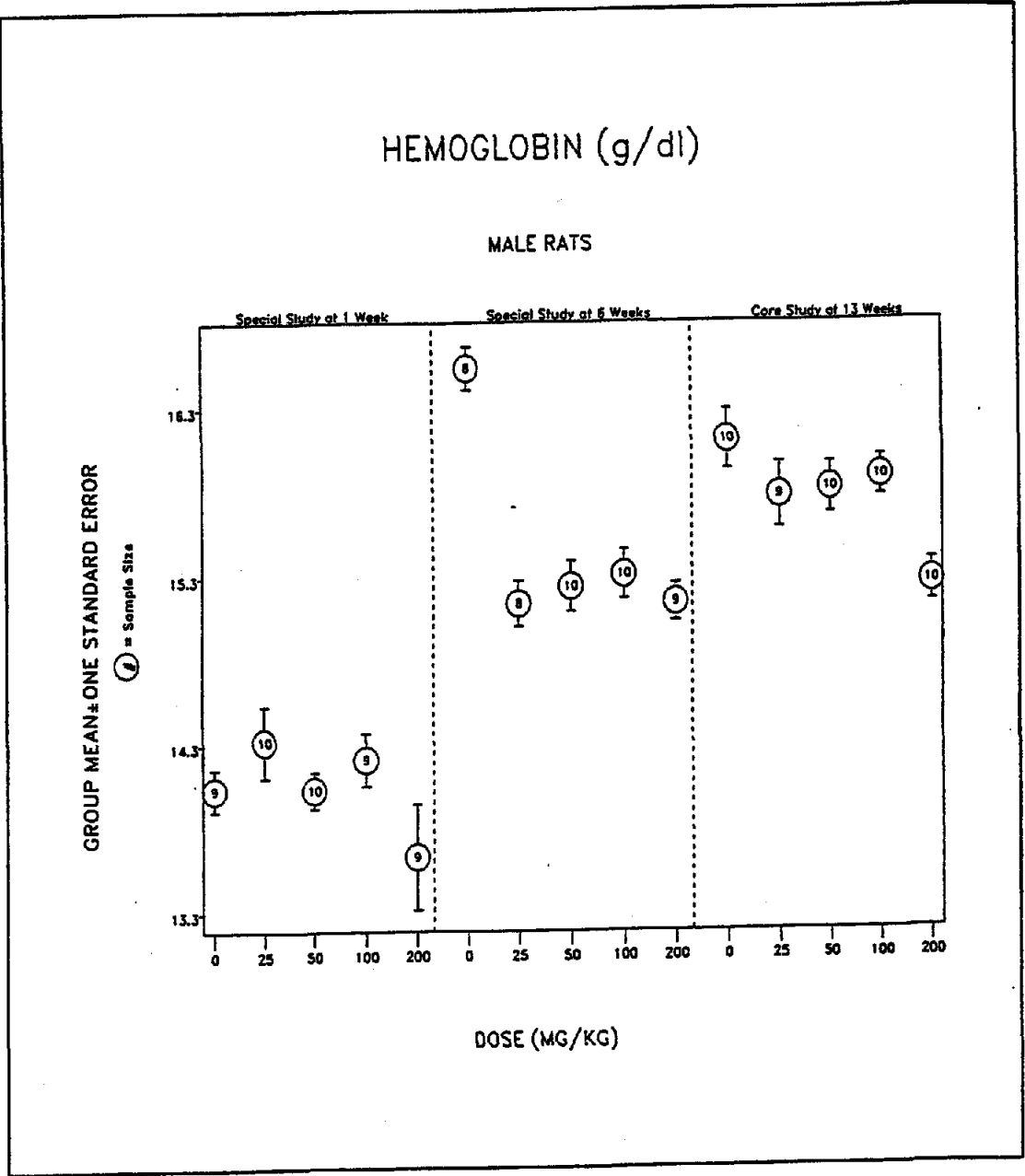
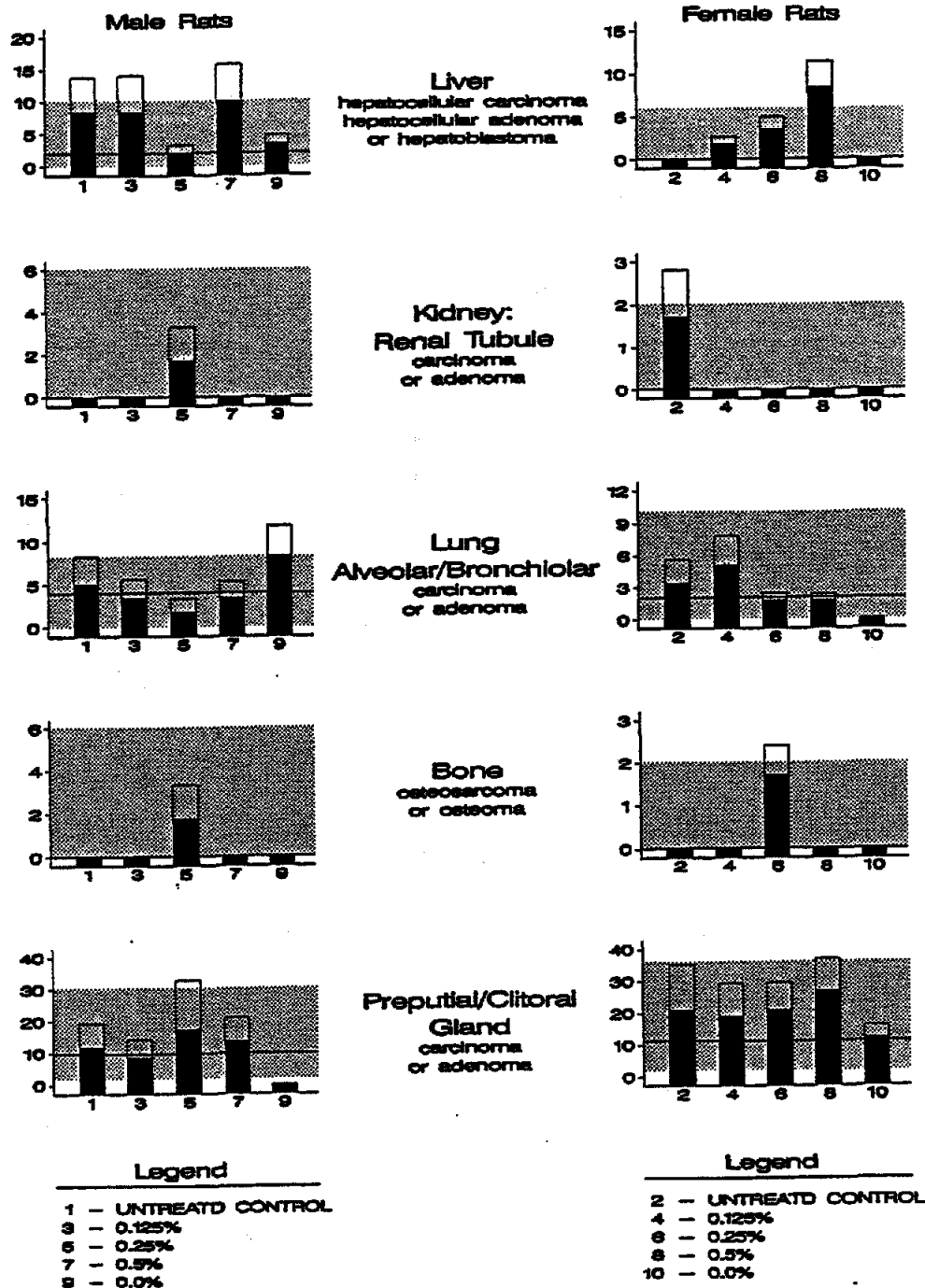


Figure 4

Tumor Incidence (%)

DOSED FEED (27 Control Studies)



Solid bar is observed incidence, hollow bar is survival-adjusted incidence
Shaded area shows range of incidence in control studies, line indicates median