CHAPTER 1

Introduction to Multiple-Plot Displays

1.1 Definition of Terms

Templates in SAS/GRAPH contain *panels* that define standard shapes, but none of these shapes are curved like the white spaces pictured in the flowchart template above. All panels in a SAS/GRAPH template are measured as percentages of the graphics output area. Panels can rest invisibly on top of each other. Think of a template as a collection of transparent, layered panels through which plots are displayed.

While the relationship between a template and its member panels is readily understood, the first three rows in Table 1.1 show there is no agreement among the experts about how to define the elements of a multiple-plot display. Therefore, the term *graph* is being used in this book to reference a collection of *plots*, and a *plot* is defined more broadly as the singular output from any SAS/GRAPH procedure except GREPLAY. Thus, output from GCHART, GPLOT, and GSLIDE are all defined as *plots* in the book, whereas output from PROC GREPLAY is excluded because it is a *template graph* of multiple plots.

Source	Single Display	Multiple Display
Tukey, John and Paul	Plot	Multiwindow Plot
Tufte	Graphic	Small Multiples
SAS/GRAPH	GRSEG	GRSEG
This book	Plot	Graph

Table 1.1.   *In this book single- and multiple-plot displays are defined as plots and graphs, respectively.*

Plots are displayed in the order in which they are created through panels separately ordered by panel number. The position of curves on a plot is determined by the axes definitions within a procedure and not by the values of the underlying panel coordinates.

The macros in this book always reserve the first panel in the upper-left corner for the first plot created in a program, whereas the penultimate panel in the lower-right corner gets the last plot in a

series of plots. The last panel in the template covers the entire graphics output area, and it is reserved for the grand title that describes the graph.

Figure 1.1 shows how the graphics elements just described are related to each other. While a graph is simply a page of plots, no clearly defined relationship exists between *plot* and *panel*. Instead, the title panel spans the graphics output area, covering all the other panels in the display, and it is possible for some panels not to be filled with plots when TREPLAY is invoked. Figure 1.1 also shows that it is not necessary for the rows in a graph to contain the same number of panels. Panels are created center-justified by default to compensate for any inequality in row totals. Technically, therefore, the sole panel in the first row of this diagram is defined as the first panel in the upper-left corner of the graph.

A few additional graphics elements are also depicted in Figure 1.1. A *margin* of an arbitrary size can be placed around the small plots in the graph, and each of the small plots can be *scaled* to a factor less than 1.0. Scaling separates panels so that plots stand out better. In effect, it defines an additional margin around each individual panel.
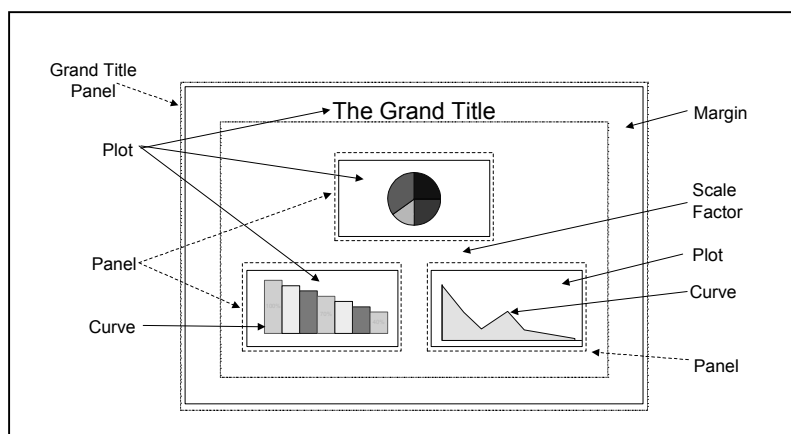

Figure 1.1   *Schematic of a graph showing the relationship of plots to panel displays. The four extension macros in this book take the drudgery out of creating panels through which plots are displayed. Any time the margin or the scale factor is changed in a graph, all panel coordinates have to be recalculated. This translates to 32 coordinates for the schematic above because each of the four panels has eight coordinates.*

More Information

For creative uses of template designs, see
  Brown, 1991.
  Carpenter and Shipp, 1995, Section 6.3, "Controlling Graphical Displays with
    Templates," pp. 154–184.
  Kalt, 1986.
  Kirk and Horney, 1998.

For a depiction of the graphical output area, see Carpenter, 1999, p. 17.

For definitions of terms abstracted from their context, see
  Tufte, 1983, pp. 14–15 and 170.
  Tukey, J., 1977, p. 129.
  Tukey, P. and J., 1981, pp. 231–232.

1.2 When Multiple-Plot Displays are Needed

When data contain many subpopulations that cannot be summarized by a single statistic, multiple-plot displays are the graph of choice for visual by-group comparisons.

The example that proves this assertion is taken from Appendix I of Kalbfleisch and Prentice (1980) and reproduced in *SAS/STAT User's Guide, Version 8*, Example 37.1 for the LIFETEST procedure (p.1832). Figure 1.2a is an enhanced reproduction of the graph that appears on page 1840 of *SAS/STAT User's Guide, Version 8*. If you work in the pharmaceutical industry, you will recognize the survival plot in Figure 1.2a as a primary vehicle for communicating information about a Phase III clinical trial.
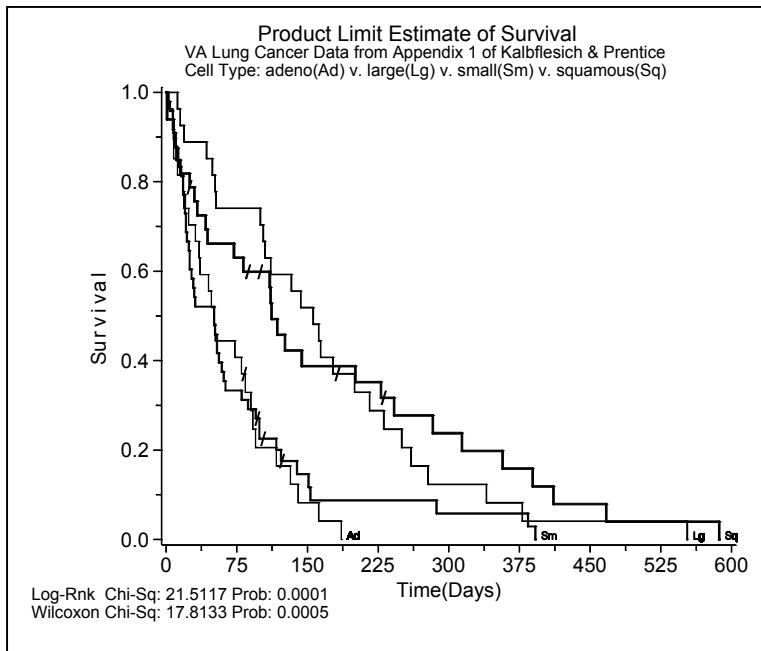
Figure 1.2a *Reproduction of Output 37.1.6: Graph of the Estimated Survival rates annotated with log-rank and Wilcoxon chi-square statistics, showing statistically significant results that can be attributed to the presence of non-overlapping strata.*

Visually, you can corroborate the significance of a chi-square statistic by looking for strata (curves) that are non-overlapping. In this case, the paired Adeno|Small and Large|Squamous strata are responsible for the statistically significant results (prob<0.05). The four strata are also spread far enough apart so that they can be fully tracked.

The situation becomes much more challenging when the reviewer wants to know if treatment affects survival. Figure 1.2b displays eight strata: two for each cell type/treatment combination. Again, the log-rank and Wilcoxon chi-square tests are statistically significant, but you would be hard-pressed to find all pairs of non-overlapping strata that contribute to the significant result. Also, Cell Type probably should be held constant before comparing treatment regimens. Figure 1.2c does this with an abbreviated display of two plots for the Large and Squamous Cell Types. Results for these more meaningful comparisons are insignificant.
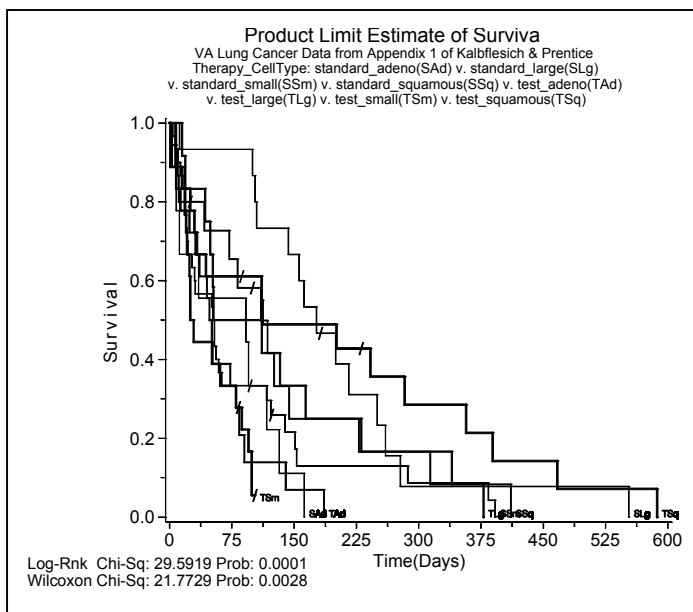

Figure 1.2b *All combinations of cell and treatment types appear in separate strata in one plot producing an unreadable graph of questionable value. There is no way to identify the non-overlapping strata that would contribute to a significant result.*
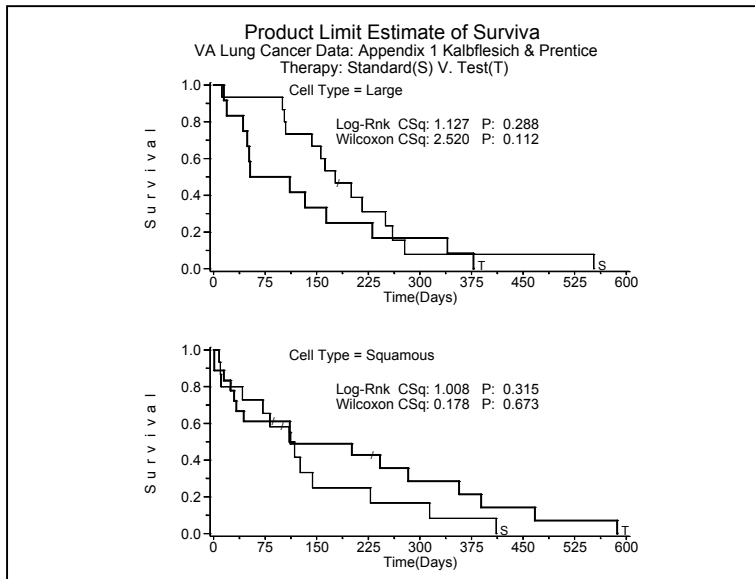
Figure 1.2c *The effect of treatment is clearly discernable in a multiple-plot display that compares two strata at a time. Results are not significant because the strata overlap. Additional two-strata graphs can be found in the Appendix.*

1.3 Data Sources Used in This Book Require Multiple-Plot Displays

Each of the data sources in this book is described in this section. Each description enumerates the subpopulations in the data that require their own separate plot.

Survival Data   (see Chapters 1 and 5)

As mentioned, the survival data are taken from Appendix I of Kalbfleish and Prentice (1980). They have been used to demonstrate an application of the LIFETEST procedure in Versions 6 and 8 of the *SAS/STAT User's Guide.* The response variable is survival time, and it is measured in days for a lung cancer patient. Censored events occur when no follow-up information is available for a patient who leaves the study prematurely. Censors are denoted by a forward slash (/) on the graphs. Additional covariates in the data besides *Cell Type* and *Therapy* are *Prior Therapy, Age, Diagnosis Time,* and *Performance Status* (*SAS/STAT User's Guide, Version 8*, p. 1831).

Mouse Tumor Data (see Chapters 2 through 5)

The mouse tumor data come from a research study conducted by Dr. Rudi Bao of the Fox Chase Cancer Center in Philadelphia, Pennsylvania. Four treatment regimens are applied to right and left flank tumors of a group of 32 mice. Tumor measurements taken at unequal time intervals are plotted as cubic regression curves for each mouse.

Format Optimization Study (see Chapter 4)

Format resolution times are recorded for different distributions of start values created by serial applications of the SAS RANUNI function. The start values in each distribution are sorted by frequency and then renamed to reflect their sort order (e.g., a start value of '1' also has the highest frequency). Because start values are renamed, binary, sequential, and embedded formats can be ranked by their resolution times.

Baseball Data (see Chapters 4 and 5)

The scatterplot matrix in Michael Friendly's *SAS System for Statistical Graphics* is redrafted as an upper triangular matrix to eliminate redundant transpositions and to demonstrate the flexibility of the TMPLT macro. The data set "contains variables that measure batting and fielding performance for 322 regular and substitute hitters in the 1986 year, their career performance statistics, and their salary at the start of the 1987 season" (Friendly, p. 598).

Shannon Scores for Amino Acid Diversity (see Chapter 3)

Cumulative distributions of Shannon scores (H) ± 1 standard deviations for numbered sites in replicates of the VAlpha, VBeta, and VKappa genes are displayed along with a schematic graph that shows how they are mapped to a gene backbone containing X,Y, and Z coordinates.

A review of the data dimensions for the graphs demonstrates that they are ideal candidates for multiple-plot displays. None of their subpopulations can be summarized by one statistic:

Survival Data: Each cell type contains two treatment regimens. Each treatment is a stratum in a survival plot.

Mouse Tumor Data: Each treatment group contains multiple mice. Each mouse has right and left tumor measurements that are plotted over time.

Format Optimization Study: Each format search type is tested with a different frequency distribution of start values. Each distribution is plotted with a step function.

Baseball Data: Each combination of variables generates a plot. Each plot is a scatter plot of discrete points.

Shannon Scores: Each graph contains multiple genes. Each gene contains many sites with Shannon scores.

See Also

Chapters 4 and 5 contain detailed descriptions of selected graphs.

Enlargements of some of the graphs can be found in the Appendix.

More Information

For a discussion of the role of graphics in data analysis, see Friendly, 1991, Chapter 1.

For a description of the baseball data, see Friendly, 1991, pp. 13–14, *577–581*, and *598–607*.

For a description of the Shannon scores, see Watts and Litwin, 1992.

For a description of the format optimization study, see Watts, 2001, "On the Relationship between Format Structure and Efficiency in SAS."