**Less is More: Two- and Three-Dimensional Graphics for Data Display**

Leland Wilkinson
SYSTAT, Inc. and Northwestern University
1800 Sherman Ave.
Evanston, IL 60201

Revised version published in *Behavior Research Methods, Instruments, & Computers, 26*, 1994, 172-176.

## Introduction

Whether in technical or in business publications, graphical displays seem to take two forms: the garish or the inscrutable. The icon of the garish is the 3-D pie chart. The icon of the inscrutable is the banner and stub table of means. There are cures for these afflictions, but not without price. Good design and clear presentation do not impress people. Garishness and inscrutability do. The displays in this paper are not "power graphics." But they communicate clearly.

We will look first at displays of the distribution of a single variable. Then we will examine two variable and multi-variable displays. This paper is not an exhaustive survey. Nor is it systematic. The topics chosen have been overlooked in more general discussions of graphic presentation. And the general theme is that, whenever possible, display the raw data.

## Single variable graphs

Figure 1 contains the most common single variable display: the histogram. The data are life expectancies in 17 countries for males and females, compiled by the World Health Organization. An advantage of the histogram is that data within the categories created by the bars can be counted. It is, in effect, a graphical tabulation. Its close relative, the bar chart, is a tabulation in which the bars are discrete, or separated. Bar charts are relatively easy to construct; the categories are already intrinsic to the data. Histograms are difficult. We must decide on the number of categories before constructing them. Figure 2 shows how the shapes of histograms can be seemingly arbitrarily manipulated by choosing different bar widths and sliding the base scale on the same data. There are guidelines for making intelligent choices for the bar widths (or, concomitantly, the number of bars), but not for their location on the base scale (Sturges, 1926; Doane, 1976; Scott, 1979). What seems obvious in elementary statistics books ("pick about 15 bars and fill them") is not. Viewers who may be aware that the number of bars can affect the shape of a histogram often don't realize that the location of the cutpoints can affect it more. More generally, statistics package users don't always understand that categorizing quantitative variables like age

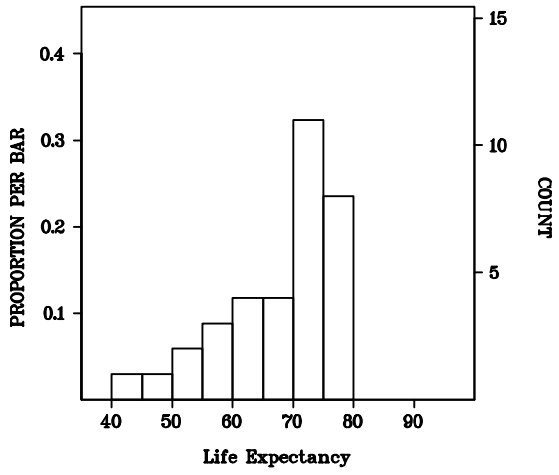


Figure 1. Histogram of life expectancies

can affect statistical conclusions. Deciding to make the lowest category boundary 32 instead of 40 can change the distribution of the data in the categories even when the category widths are held constant.

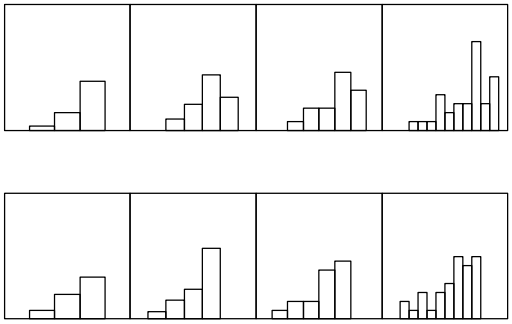

Figure 2. Histograms of life expectancies with different bar widths (horizontal dimension) and different locations of cutpoints on the same scale (vertical dimension)

An antidote is to tabulate the raw data. Instead of choosing cutpoints on a scale, we can take the most significant decimal digits of the data and display them together with the next digit. Figure 3 shows this display, called the stem-and-leaf diagram. It was invented by Tukey (1977) as a form of tally which could be done with paper and pencil. Actually, unlike many statistical and graphical procedures, programming it on a computer is more difficult than doing it by hand. Good stem-and-leaf programs make intelligent decisions about picking the digits to make the display compact. Notice that we can now see the raw data. The leftmost digit of each number appears to the left of the display. The next digit (regardless of how many trailing digits there are) is to the right. At the top, for example, there is one value (44). The next line shows another value (46). The third line shows two values (51, 54). By counting the "leaves" (digits on the right) we can tell how many values there are for each "stem" (digits on the left). There are 34 leaves in all, the total count in our sample.

```
4 4
4 6
5 14
5 557
6 1124
6 5679
7 00001122234
7 57888899
```

Figure 3. Stem-and-leaf diagram of life expectancy data

Digits look crude; histogram bars look somehow more mathematical and formal. But the histogram bars are nothing more than tallies. We can make the digits little squares and then the histogram and stem-and-leaf diagram would look the same. For large samples, the digits in the stem-and-leaf diagram can be reduced in size. When counting becomes difficult, we can add a count scale.

There is another way to display the density of a batch of data when we are less concerned with counting. Figure 4 illustrates this display: the kernel density (Silverman, 1986). The kernel density is immune to the scale shift problem. It is also not susceptible to the bar width problem because it has no bars. The shape of the smooth *can* be influenced by the choice of a smoothing window width, however. Changing this width can make the kernel density look more or less smooth. Like the histogram, however, there are statistical guidelines for choosing a width.

The drawback of the kernel estimator, however, is that the data values are concealed. We cannot count values or see their location, as with the stem-and-leaf diagram. Consequently, the kernel estimator should be a supplement to other density displays and not a replacement.

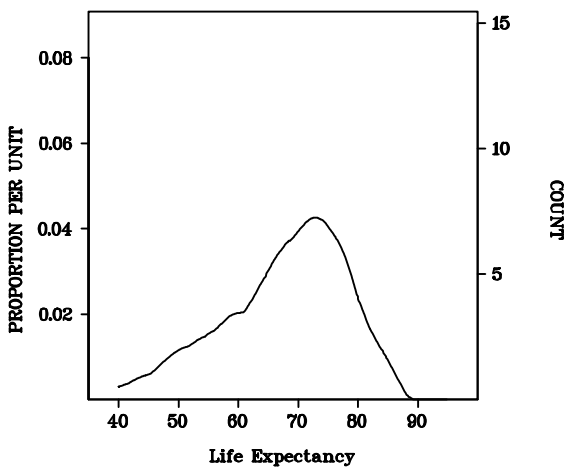


Figure 4. Kernel density of life expectancy data

There is another display which allows us to see both the raw data and the smooth: a dot-box plot. Tukey (1977) introduced the box or schematic plot along with the stem-and-leaf diagram. Its advantage is that the fractiles of the data, particularly the median and quartiles, can be seen. Its disadvantage is that it conceals the shape of the distribution. Bimodal and unimodal distributions can have the same box plot.

The dot-box plot solves this problem by displaying the box and data against the same scale. Figure 5 shows this plot for the life expectancy data. Notice that the dot values are symmetrically distributed about the center line. This type of dot plot (without the box) has been popular in the medical literature for several years.

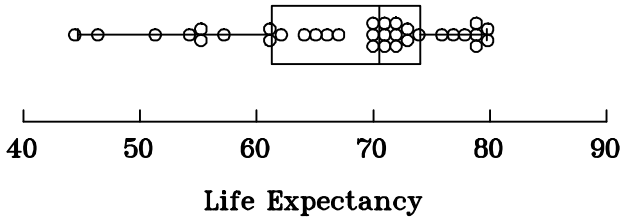


Figure 5. Box-dot plot of life expectancy data

**Two-variable graphs**

The most common two dimensional continuous variable data display is the scatterplot. Figure 6 shows an example for the life expectancy data. We have plotted the data for males against those for females. Enhancing scatterplots with smooths can sometimes reveal hidden structure. Like the dot-box plot, we can see the smooth *and* the raw data. Figure 7 shows the same scatterplot with a two dimensional kernel superimposed. This kernel reveals the skewness in the joint and marginal distributions of the data. The eye can focus on either aspect of the display.

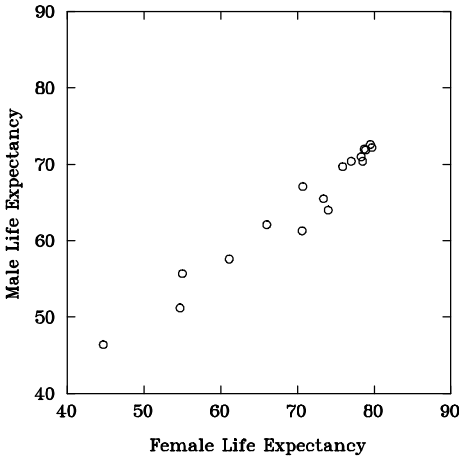


Figure 6. Scatterplot of male against female life expectancy

Unlike the kernel smooth for the histogram, the kernel is used here only to enhance perception of the data, not to conceal it. The contour lines are light enough so that the data remain visible.

Figure 8 shows how powerful this smoothing and data display can be. These data are birth and death rates per year per 100,000 people for 75 selected

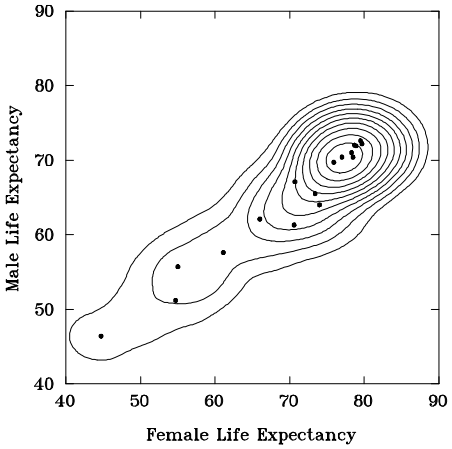


Figure 7. Scatterplot of male vs. female life expectancy with kernel smooth

countries. The bivariate kernel contours are superimposed to show the joint sample distributions. Selected points are labeled. The zero population growth line at the left of the plot separates countries like Hungary, which are losing population, from countries like Guatemala, which are gaining rapidly. This graph reveals a disturbing nonlinearity and bimodality in world health statistics. Developed nations show varying birth rates but relatively low death rates. Underdeveloped nations have extremely high birth rates and high death rates. Some graphs elude parsimonious mathematical modeling. This is an example.

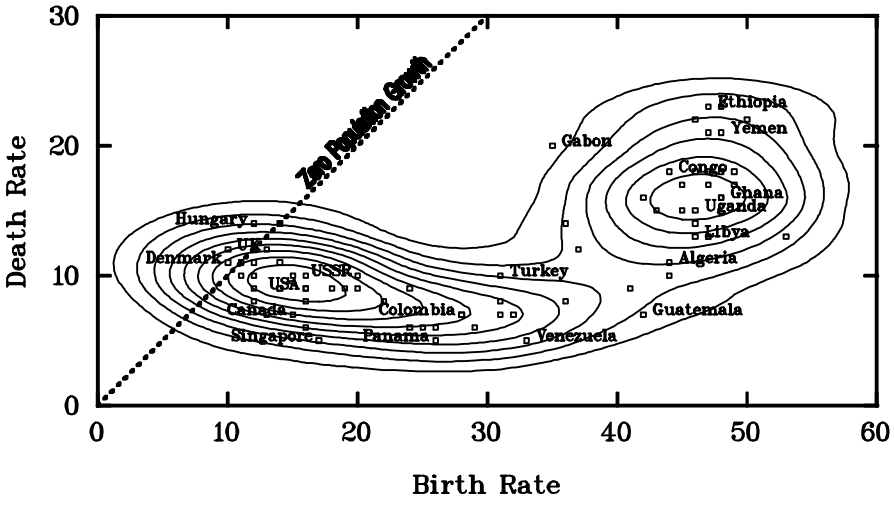


Figure 8. Bivariate kernel density of birth and death rates

**Multivariable graphs**

Most people think of 3-D displays when considering multivariable graphs. These are the popular graphs in computer magazines and they certainly sell software. There are even occasions when they can prove useful, particularly when the overall shape of a distribution or smoothing surface is of interest. As Becker and Cleveland (1991) have pointed out, however, statistical graphics are not the business of creating real life scenes. Scientific visualization is fashionable now and has extensive and important applications. In statistics, however, we gain more by displaying multivariate data directly rather than by attempting to smooth them into some recognizable scene.

One of the most useful statistical displays is the scatterplot matrix (SPLOM). Like tree displays, SPLOMS are easy for non-statisticians and people who have difficulty with spatial relationships to understand. They are simply arrays of scatterplots. By placing all possible scatterplots in a single display, SPLOMs help us to see overall structure.

Figure 10 shows a SPLOM of our birth and death data, with an additional variable - health

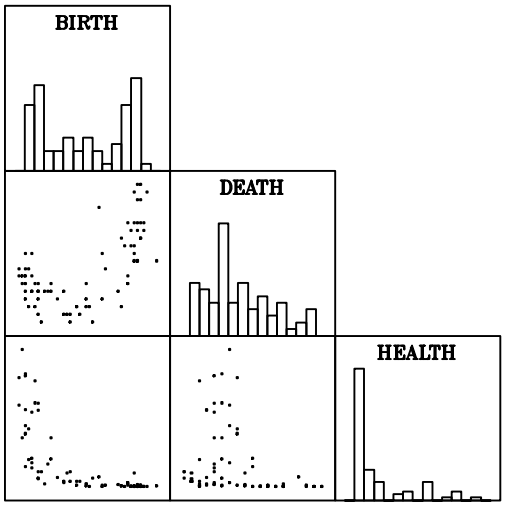


Figure 10. SPLOM with histograms.

expenditures for each of the countries in U.S. adjusted dollars. The histograms on the diagonal indicate that the HEALTH data should be logged to reduce the positive skewness. Figure 10 shows the same SPLOM after the transformation.

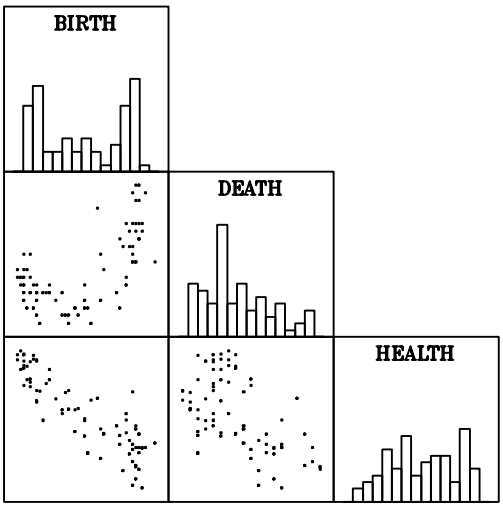


Figure 10. SPLOM of logged HEALTH

Now let's look at some enhancements of these scatterplots. Figure 11 shows the bivariate kernel densities superimposed on the same SPLOM. Now we see the bimodality apparent in Figure 8, but it enters the other cells as well. In addition, we have used stripe plots in the diagonal cells instead of histograms. Like dot plots, these density displays reveal the distribution of the actual data points.

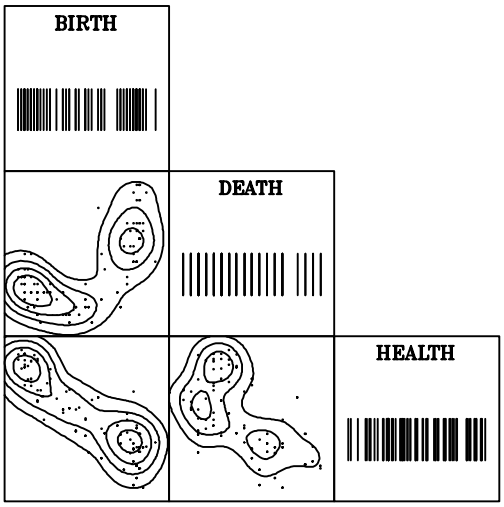


Figure 11. SPLOM with bivariate kernel

Other enhancements can be used to reveal different aspects of the bivariate structure. One of the more valuable trend enhancers is LOWESS (Cleveland, 1981). This nonlinear smoother is robust to outliers, so it is a good way to detect nonlinear trend in the bulk of the data. Figure 12 shows LOWESS curves superimposed on the same SPLOM. Notice the substantial nonlinearities. Journal editors would

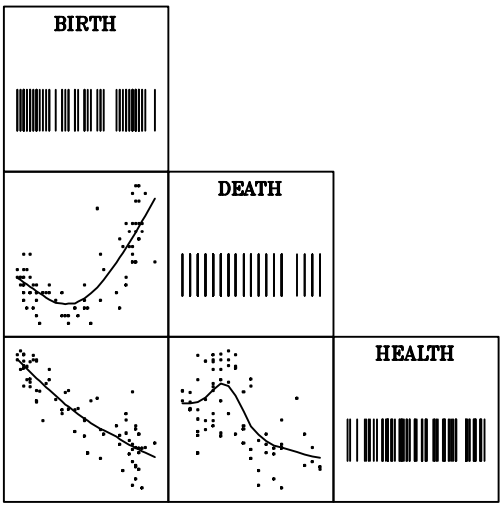


Figure 12. SPLOM with LOWESS

do well to require authors to submit SPLOMs whenever they perform analyses based on covariance or correlation matrices. This policy would improve replicability by showing that the data are normally distributed.

**Conclusion**

The graphs presented here are only a small sample of the kind which can be produced with a good statistical graphics package. All are black and white, although color has its uses. Particularly in presentations, color can be especially effective in distinguishing categories. Symbols in scatterplots can be drawn with different primary colors to reveal subgroups of the data, for example. In general, however, well designed black and white graphs can convey information succinctly and clearly. And if the data are displayed in the same graph whenever possible, it will be more difficult to deceive or convey the wrong impression.

References


Becker, R., and Cleveland, W.S. (1991). Take a broader view of scientific vixualization. *Pixel*, *2*, 42-44.

Cleveland, W.S. (1981). LOWESS: A program for smoothing scatterplots by robust locally weighted regression, *The American Statistician*, *35*, 54.

Doane, D.P. (1976). Aesthetic frequency classifications. *The American Statistician,*, *30*, 181-183.

Hartigan, J.A. (1975). *Clustering Algorithms*. New York: John Wiley & Sons.

Scott, D.W. (1979). Optimal and data-based histograms. *Biometrika*, *66*, 605-610.

Silverman, B.W. (1986). *Density estimation for statistics and data analysis*. New York: Chapman & Hall.

Sturges, H.A. (1926). The choice of a class interval. *Journal of the American Statistical Association, 21*, 65.

Tukey, J.W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.