3

GRAPHICAL METHODS

This chapter is about graphical methods: types of graphs and ways of encoding quantitative information on graphs. The methods allow us to analyze both the overall structure of the data and the detail of the data.

Section 3.1 discusses two methods, logarithms and residuals. These are general purpose tools that are useful in all areas of graphical data analysis.

Section 3.2 is about graphing one or more sets, or categories, of measurements of one quantitative variable. Suppose we have measurements of the brain weights of three groups of animals: gorillas, orangoutangs, and chimpanzees. In this example we have one quantitative variable, brain weight, and a categorical variable, animal species. The graphical methods of the section let us study and compare the data distributions: where the sets of data lie along the measurement scale.

Section 3.3 is about dot charts, which are used to show measurements of a quantitative variable in which each measurement has a label associated with it that we want to display on the graph. An example is the distances of the planets from the sun; each measured object, a planet, has a distance and a name. Several different forms of the dot chart are described; the different forms accommodate different measurement scales and different structures of the measurement labels.

Section 3.4 is about graphing two quantitative variables to study their relationship; for example, the methods could be used to study how brain weights of gorillas are related to their body weights.

In Section 3.5 the setting is similar to Section 3.4, but now there are two or more categories of measurements of two quantitative variables. For example, we might have measurements of brain weights and body weights of gorillas, orangoutangs, and chimpanzees. The section presents methods of superposition and juxtaposition of the different categories of data that allow us to study the relationship of the two variables and to identify the categories.

Section 3.6 deals with measurements of three or more quantitative variables; an example is measurements of blood pressure, heart rate, weight, height, age, and sodium intake for a group of people. Understanding such multidimensional data is difficult, but the use of graphical methods in the section can often increase our understanding.

Section 3.7 is about statistical variation. There is a general discussion of the empirical variation in data and the sample-to-sample variation of a statistic computed from data. Two-tiered error bars are introduced for showing sample-to-sample variation.

3.1 GENERAL METHODS: LOGARITHMS AND RESIDUALS

Logarithms

Logarithms are one of man's most useful inventions. They are indispensable in science and technology and are a vital part of graphical methods. Their usefulness has been amply illustrated earlier in the book — for improving resolution and for showing data where percents and factors are important.

In Figure 3.1, logarithms of the maximum amounts of solar radiation penetrating ocean water at various ocean depths are graphed against depth [88]. Until 1984 it was presumed that living things did not exist in the ocean below about 200 meters because of low light intensity. In 1984 scientists at the Smithsonian Institution in Washington, D.C. and the Harbor Branch Foundation in Florida discovered an alga at a depth of 268 meters in waters off the coast of San Salvador Island in the Bahamas. The filled circles in Figure 3.1 are measurements of radiation that the discoverers presented in their paper and the open circles are values that they extrapolated from the measured values. The line on the graph is the least squares line fitted to the measured values.

Logarithms are useful here because radiation changes by five powers of ten from about 10^3 at sea level to about 10^{-2} at 268 meters. Also, it is natural to use a log scale because we would expect attenuation of the solar radiation, if the transmission properties of the ocean water are relatively constant, to be multiplicative as a function of depth; if s is the radiation at sea level and f is the fraction of radiation remaining after passing through one centimeter of ocean water, then the radiation at a depth of one centimeter is $r(1) = fs$, at two centimeters is $r(2) = f^2 s$, and at d centimeters is $r(d) = f^d s$. On a log scale, radiation is

$$\log(r(d)) = d \log(f) + \log(s)$$

and is thus a linear function of d. Figure 3.1 shows such an attenuation process is commensurate with the log measurements, which are roughly

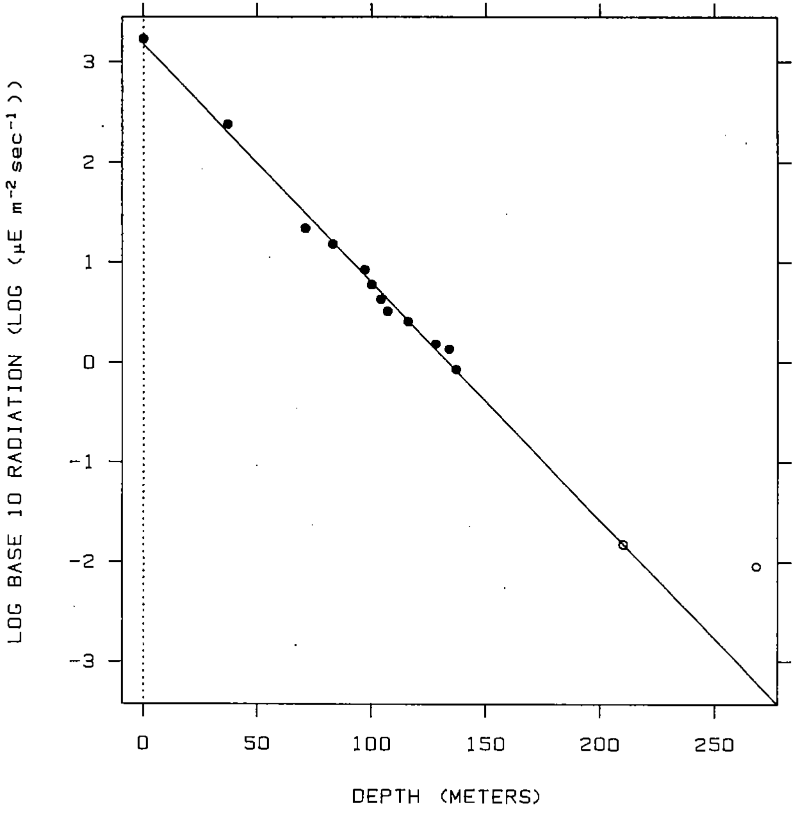

Figure 3.1 LOG BASE 10. Graphing data on a log base 10 scale is reasonable when the data go through many powers of 10, as on the vertical scale of this graph.

linear with depth. The extrapolated radiation value at 210 meters fits the pattern of the measured values, but the extrapolated value at 268 meters does not; either the ocean water changes its properties or there has been a faulty extrapolation.

Log Base 2 and Log Base e

Log base 10 is almost always used in scientific graphs for a log scale. This is much too limiting. Log base 2 and log base *e* (natural logarithms) should always be considered. Using a different base does not change the pattern of the points but changes only the values at the tick marks because the logarithm of one base is just a constant times the logarithm of another base. The relationship between log base *b* and log base *c* is

$$\log_c(x) = \log_b(x)/\log_b(c) .$$

Thus

$$\log_2(x) = \log_{10}(x)/\log_{10}(2)$$

and

$$\log_e(x) = \log_{10}(x)/\log_{10}(e) .$$

The choice of the base depends on the range of the data values that need to be visually compared. Suppose the data go through many powers of 10, as the radiation data in Figure 3.1 do. In such a case it is reasonable to use log base 10. But suppose the data range over two powers of 10 or less. This is the case in Figure 3.2; the data are the number of telephones in the United States from 1935 to 1970 [128, p. 783]. In such a case it is inevitable that equally spaced tick marks for log base 10 will involve fractional powers of 10, as Figure 3.2 illustrates. It is difficult to deal with such fractional powers. It is easy enough to remember $10^{0.5}$ is a little bigger than 3, but to keep many fractional powers of 10 in our heads and try to use them to study a graph is cumbersome. In such a situation it makes sense to convert to log base 2 as in Figure 3.3. It is easier to deal with powers of 2 than fractional powers of 10. For example, we can see that the number of phones increased by a factor of about 4 from 1935 to 1960.

On a log base 2 graph it is often helpful to label one scale line on a log scale and the other scale line in the original units of the data. This reduces the amount of mental conversion from the log scale to the original scale. This second scale, however, does not completely eliminate mental conversion. Suppose there is a datum at 2^7 and a datum at 2^{13}; the second is greater by a factor of 2^6. In order to evaluate this factor, we must know $2^6 = 64$. Remembering powers of 2 up to 2^{10} is easy:

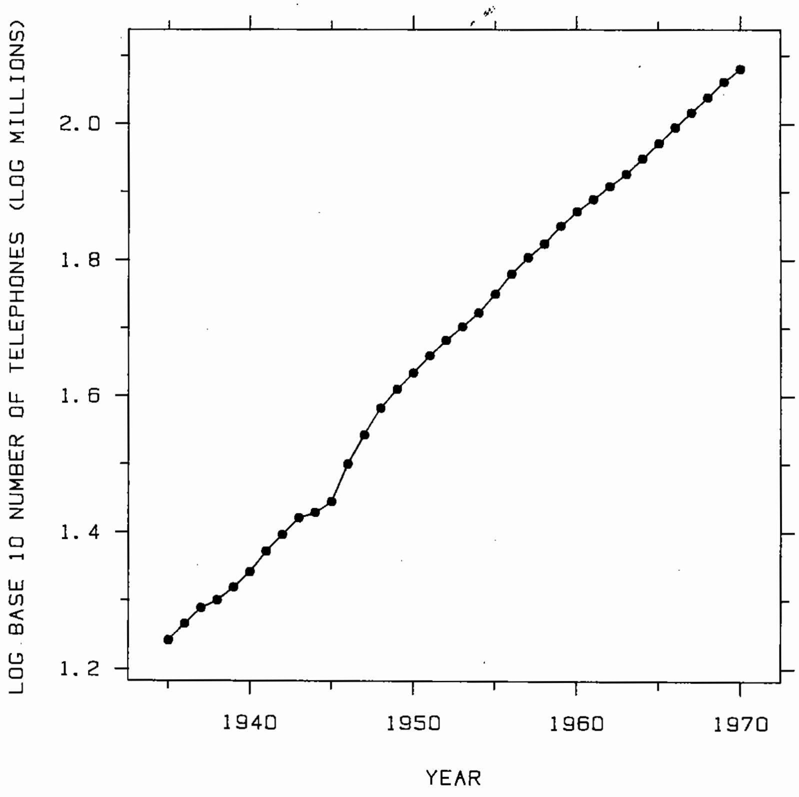

Figure 3.2 LOG BASE 10. The data, a time series of the number of telephones in the United States, are graphed on a log base 10 scale. When the data range through two or fewer powers of 10, the log base 10 scale is not as informative since we must deal with fractional powers of 10, as on this graph.

$$\begin{array}{llll} 2^1 = 2 & & 2^6 = 64 \\ 2^2 = 4 & & 2^7 = 128 \\ 2^3 = 8 & & 2^8 = 256 \\ 2^4 = 16 & & 2^9 = 512 \\ 2^5 = 32 & & 2^{10} = 1024 \ . \end{array}$$

The computer revolution has made it even easier to remember these powers. We can go even higher by using the computer scientists' trick: Let $k = 1000$ and approximate $1024 = 2^{10}$ by k so that

$$2^{14} = 2^4 \times 2^{10} \approx 16k = 16{,}000$$

and

$$2^{24} = 2^4 \times 2^{20} \approx 16k^2 = 16{,}000{,}000 \ .$$

Also, the following fractional powers of 2 are easy to remember because they are very close to simple numbers:

$$\begin{array}{l} 2^{0.3} = 1.231 \approx 1.25 \\ 2^{0.5} = 1.414 \approx 1.4 \\ 2^{0.6} = 1.516 \approx 1.5 \\ 2^{0.8} = 1.741 \approx 1.75 \ . \end{array}$$

A trick that can be used to keep the exponents on a log base 2 scale from getting too large — perhaps we can call it the statistical scientist's trick — is to take the original units to be in thousands, millions, or billions. For example, suppose the data range from 10^4 meters to 10^6 meters. The numbers on a log base two scale range from about 13 to 20. The trick is to think of the original units as kilometers; now the data range from 10 to 1000 kilometers and the numbers on the log base 2 scale range from about 3 to 10. This trick was employed in Figure 3.3, where the units of the vertical scale are log *millions* of telephones.

Logarithms base e are also useful because they have a wonderful property. Suppose u and v are values of the data. Let d be their difference on a natural log scale,

$$d = \log_e(v) - \log_e(u) \ .$$

Then if d lies between -0.25 to 0.25, the percent change in going from u to v, which is

$$100 \left(\frac{v-u}{u} \right) ,$$

is approximately equal to 100d%. In Figure 3.4 this approximation is illustrated with made-up data. A is larger than B by about 0.1 on the natural log scale, so A is about 10% larger than B; B is larger than C by about 0.25, so B is about 25% larger than C.

Let us see why this approximation works. Let

$$\frac{v}{u} = 1 + r$$

then the percent change in going from u to v is 100r%. Now

$$d = \log_e(v) - \log_e(u) = \log_e \left(\frac{v}{u} \right) = \log_e(1+r) \ .$$

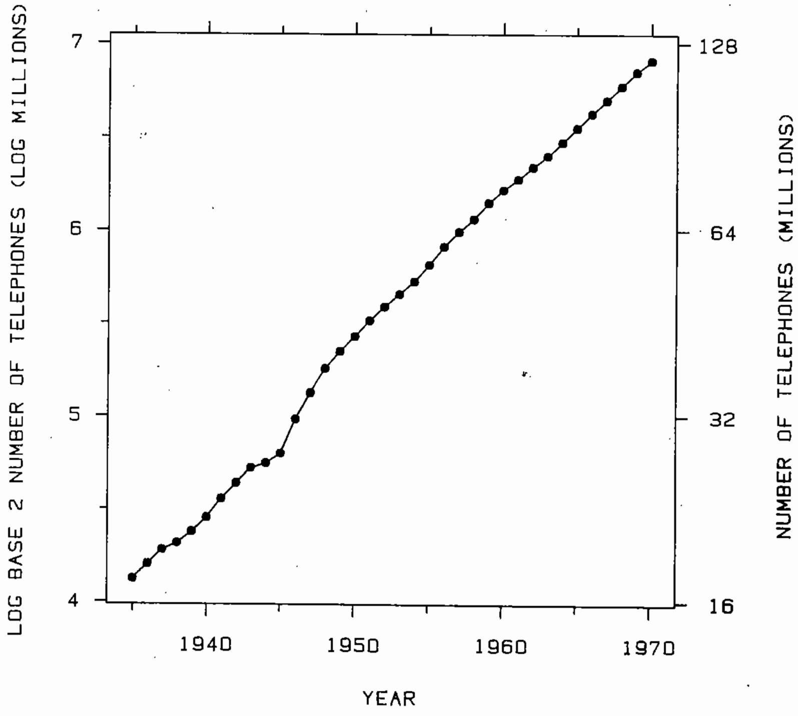

Figure 3.3 LOG BASE 2. When the data go through a small number of powers of 10, log base 2 often provides a useful scale. The left vertical scale line shows the data in log units and the right vertical scale line shows the original units.

But if d is small,

$$\log_e(1+r) \approx r$$

and therefore

$$d \approx r\ .$$

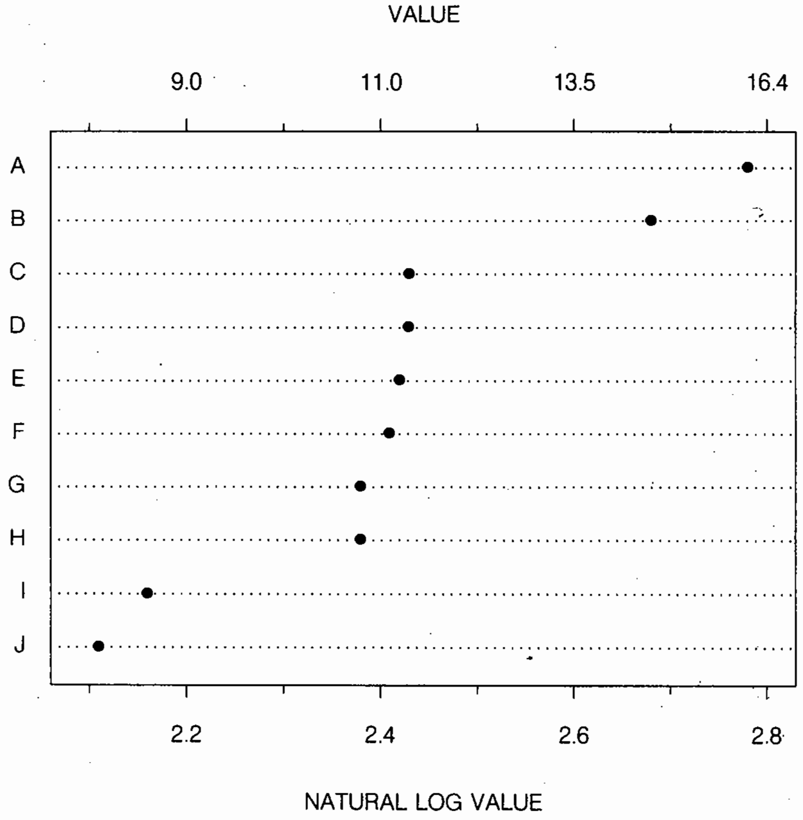

Figure 3.4 NATURAL LOGS. Logarithms base e are sometimes a good choice for a log scale. If two values on a natural log scale differ by d, where d is between -0.25 and 0.25, the percent difference of the values is to a good approximation $100d$%. On this graph, A is greater than B by about 0.1 log units, so A is about 10% bigger than B.

Here are several values of r and d:

$\log_e(1+0.05) = 0.049$	$\log_e(1-0.05) = -0.051$
$\log_e(1+0.1) = 0.095$	$\log_e(1-0.1) = -0.105$
$\log_e(1+0.15) = 0.140$	$\log_e(1-0.15) = -0.163$
$\log_e(1+0.2) = 0.182$	$\log_e(1-0.2) = -0.223$
$\log_e(1+0.25) = 0.223$	$\log_e(1-0.25) = -0.288.$

When d is greater than 0.25 or less than -0.25, the approximation is less accurate and is not as useful.

It is, of course, considerably harder to go back mentally to the original scale from a natural log scale than from base 2 or 10. We know readily what 2^3 and 10^3 are, but e^3 is harder. For this reason it is particularly important to use one scale line to show the original scale, as illustrated in Figure 3.4.

If all differences of the data on a natural log scale are between ±0.25, the approximation can be used, of course, for any two graphed values. This is illustrated in Figure 3.5. The conductivity of ocean water [88] is graphed against depth. The range of the data is about 0.15 natural log units, so no two measurements on the original scale differ by more than 15%.

Graphing Percent Change

When the maximum percent variation in the data is small, there is another way to graph the data that shows percent change between any two values. In Figure 3.6 the left vertical scale line shows the original data units and the right vertical scale shows percent change of conductivity from the sea-level value. The right scale shows that at 100 meters there is about a -5% change in conductivity from sea level and at 250 meters there is about a -15% change. Because these percent changes are small, the percent change in going from 100 meters to 250 meters is approximately $(-15\%) - (\ -5\%) = -10\%$. Thus, from the right vertical scale we can judge the percent change between *any* two values and not just between the baseline and another value; this approximation works well provided the percent changes of the two values from the baseline both lie between plus and minus 15%, which they always do in Figure 3.6. In general, a baseline value might be a value for some special condition, such as sea level in this example, or it might be the maximum value of the data, the minimum value, or a middle value.

Let us take a closer look at this approximation and why it works. Suppose b is the baseline value. Let $(1+s)b$ and $(1+t)b$ be two other values shown along the scale. The percent changes of the two values relative to the baseline are 100s% and 100t%, respectively; depending on the value of the baseline, s and t might be both positive, both negative, or have opposite signs. The percent change in going from $(1+s)b$ to $(1+t)b$ is

$$c = 100\left(\frac{(1+t)b - (1+s)b}{(1+s)b}\right) = 100\left(\frac{1+t}{1+s} - 1\right) .$$

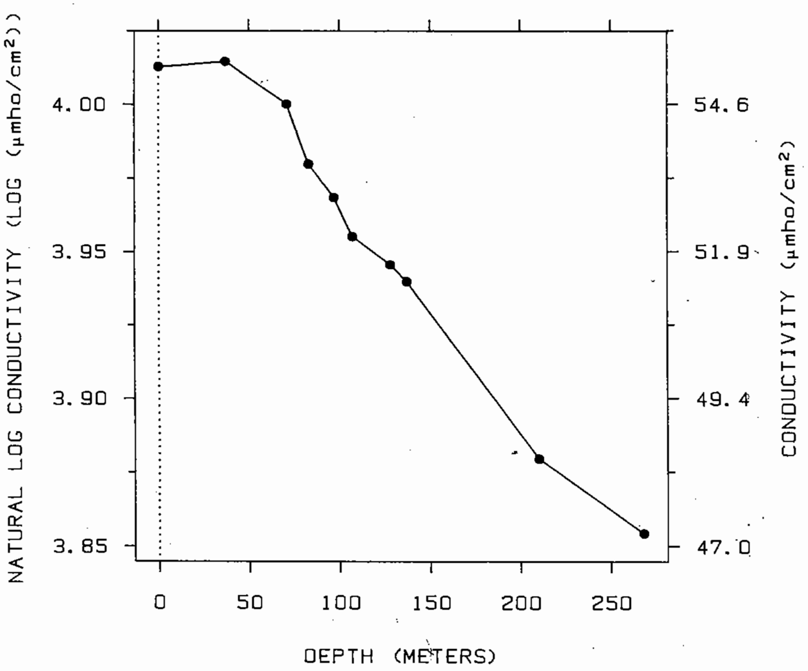

Figure 3.5 NATURAL LOGS. Conductivity is graphed on a natural log scale. Since the range of log conductivity is about 0.15, 100 times the difference of any two values can be interpreted as percent change.

If $|s|$ is small, then

$$\frac{1}{1+s} \approx 1 - s .$$

Thus

$$\frac{1+t}{1+s} \approx (1+t)(1-s) = 1 + t - s - ts .$$

If $|s|$ and $|t|$ are both small, then

$$ts \approx 0 .$$

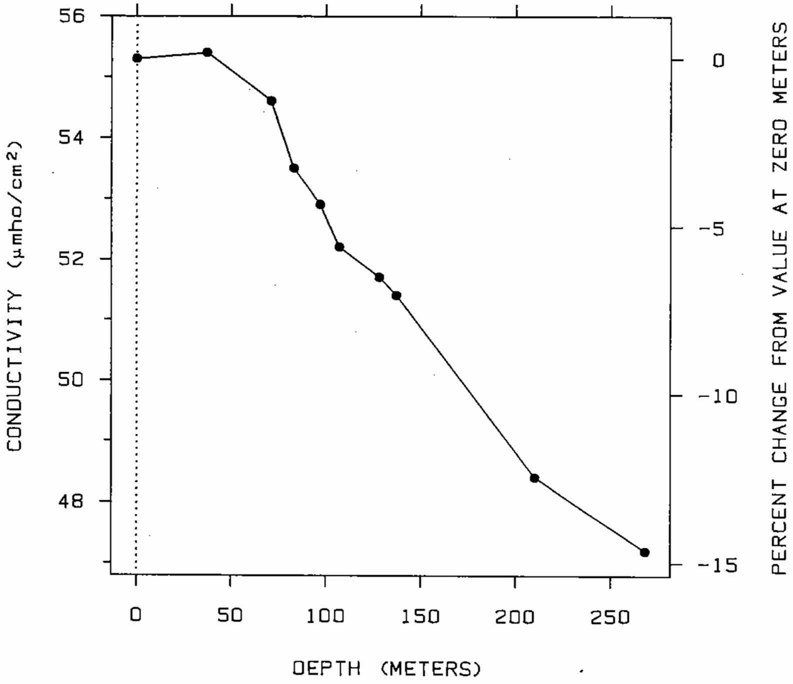

Figure 3.6 PERCENT CHANGE. The right vertical scale line shows percent change of conductivity from the sea-level value. Since the deviations from the baseline are all between ±15%, the right vertical scale line can be used to judge, to a good approximation, the percent change between any two values.

Thus

$$\frac{(1+t)}{(1+s)} \approx 1 + t - s \ .$$

This means that

$$c \approx 100t - 100s \ .$$

Here is one example of the approximation. Suppose the baseline is $b = 200$, and $u = 180$ and $v = 230$ are two other values. Then the change in u relative to b is -10% and the change in v relative to b is 15%. Thus the change in going from u to v is approximately 25%. The actual value, to one decimal place, is 27.8%.

Residuals

Figure 3.7 is a graph published in 1801 by William Playfair in his *Statistical Breviary* [109]. On the graph, the populations of 22 cities are encoded by the areas of the circles. Playfair, who was part statistical scientist and part political thinker, was the first person to study graphical data display and to experiment with graphical methods in a broad and serious way. In several brilliant strokes he invented many types of graphs that are in use today. His *Commercial and Political Atlas* of 1786 [108] and his subsequent publications contain time series graphs, bar charts, pie charts, and graphs with data encoded by circle areas and line lengths. However, some of Playfair's inventions did not work, as will be demonstrated in Chapter 4.

The graph in Figure 3.8 was made to see how accurately the circle areas of Playfair's graph encode the data; the analysis was inspired by the observation that the circle area for Turin is slightly less than that for Genoa, even though the population values recorded on the graph for these cities are equal. Let Y_i be the circle areas and let X_i be the populations. If the areas are to encode the data we should have

$$Y_i = KX_i \qquad \text{for } i = 1 \text{ to } 22 \ ,$$

which on a log scale is

$$\log_e(Y_i) = \log_e(X_i) + \log_e(K)$$

or

$$y_i = x_i + k \ .$$

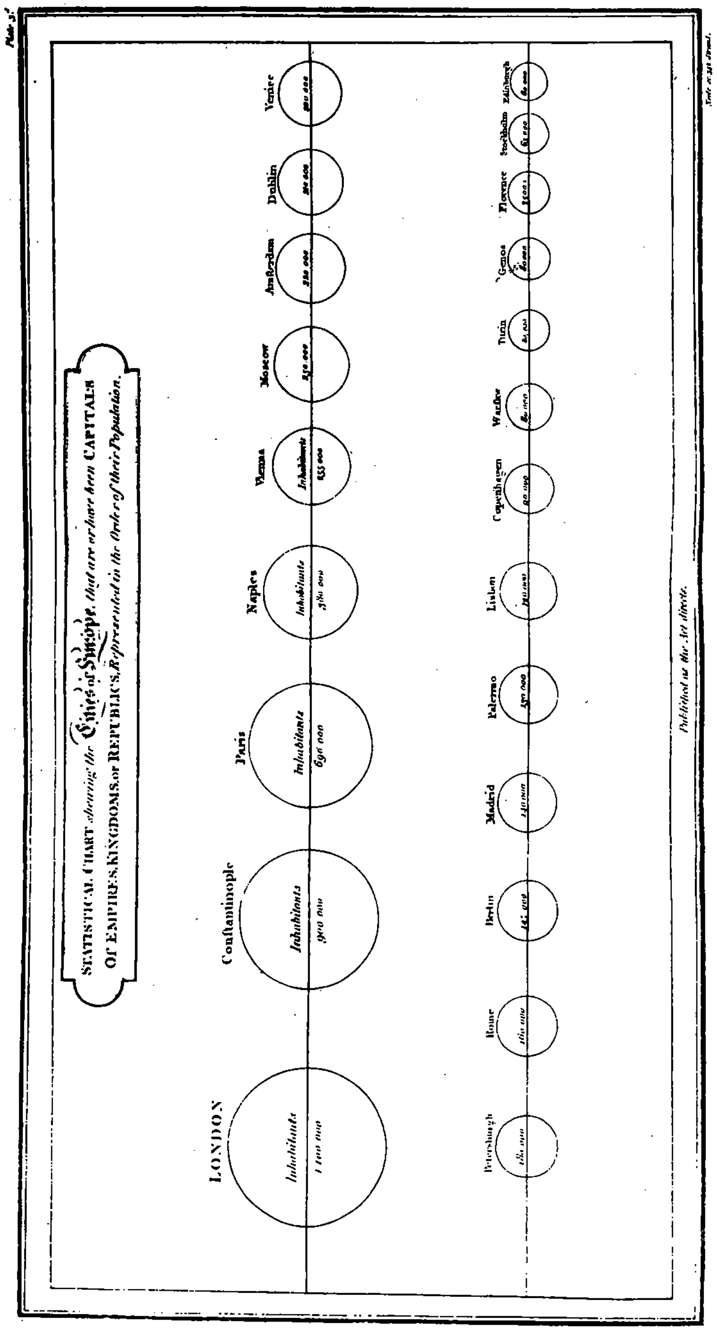

Figure 3.7 PLAYFAIR GRAPH. This graph, published by William Playfair in 1801, encodes the populations of European cities by circle areas.

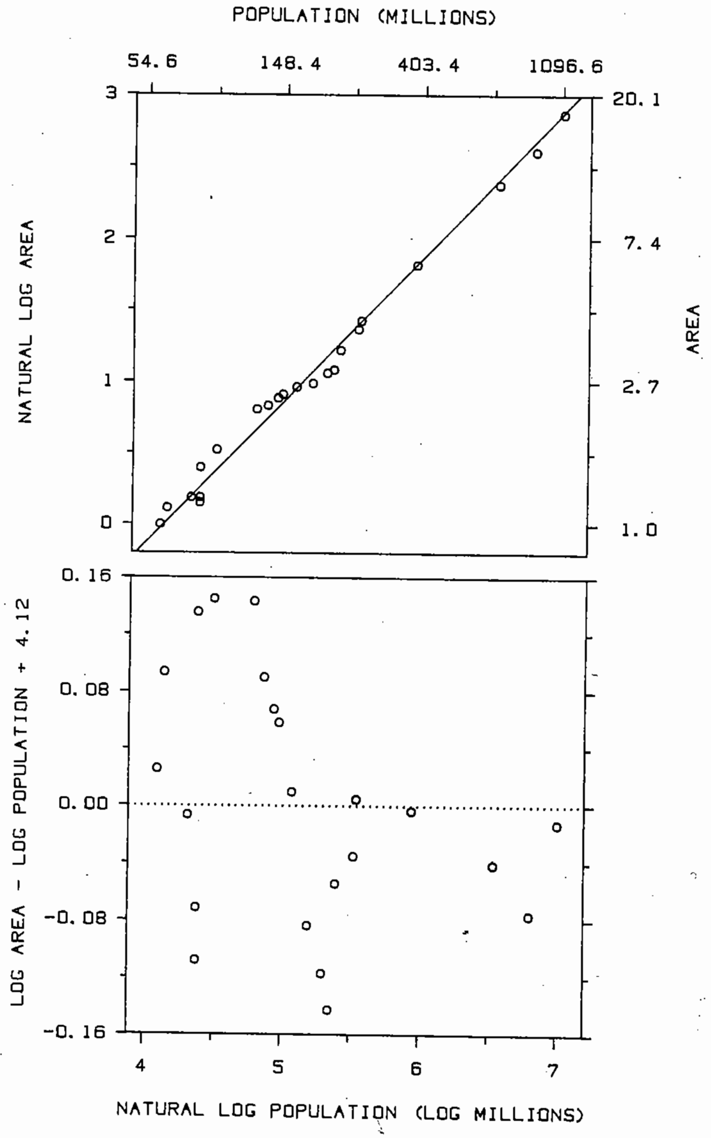

Figure 3.8 RESIDUALS. In the top panel, the areas of the circles in Playfair's graph are graphed against the populations, both on a natural log scale. The top panel shows that the points lie close to the line, but there is too little resolution to study the residuals, which are the vertical deviations. In the bottom panel the residuals are graphed against the populations and an interesting pattern in the deviations emerges.

In the top panel of Figure 3.8, y_i is graphed against x_i. Areas are relative to the area of the smallest circle in Figure 3.7, which shows the value for Edinburgh; that is, one unit of area on the vertical scale of the top panel in Figure 3.8 is equal to the area of the Edinburgh circle in Figure 3.7. Since

$$k = y_i - x_i \ ,$$

k was estimated by the mean of the 22 values of $y_i - x_i$; the estimate is -4.12. The line $y = x - 4.12$ is graphed in the top panel in Figure 3.8.

If the encoding by circle area were perfect, the points in the top panel of Figure 3.8 would lie exactly on the line. The vertical deviations of the points from the line, which are called *residuals*, tell us by how much the actual areas deviate from a perfect encoding. The values of the residuals are

$$y_i - (x_i - 4.12) = y_i - x_i + 4.12 \ , \quad \text{for } i = 1 \text{ to } 22 \ .$$

But it is difficult to assess the residuals because the points of the graph lie in a narrow band around the line, which results in poor resolution of the residuals.

The resolution of the residuals can be greatly improved by graphing them against x_i. This is done in the bottom panel of Figure 3.8. The residuals are now much more spread out since we have removed the overall linear effect. We can interpret the residuals as percent deviations, as discussed earlier in this section, because a log base e scale is used and because all residuals are between -0.25 and 0.25 log units. The largest residual is about 0.15, which means the area of the circle corresponding to the value is about 15% larger than the ideal area of the fitted line, and the smallest residual is about -0.15; thus the percent deviations of the actual areas from the ideal ones range between about -15% and 15%.

The graph of residuals in the bottom panel also shows an interesting pattern that is only barely discernible in the top panel. The residuals are not random as a function of the x_i but rather drift in a correlated way above and below zero. The tendency is for residuals corresponding to small populations to be positive and residuals for the larger populations to be negative; this means the circle areas for small populations tend to be too large and the circle areas for large populations tend to be too small. This drift in the residuals is curious.

If the deviation of the actual areas from an ideal encoding were a matter of measurement error, we would not expect, considering most mechanisms that might produce errors, to see the drift. More information about the production process and how the paper of the original graph has changed through time would be needed to solve the enigma.

Graphing residuals is an important method that has applications in all areas of graphical data analysis. We will look at several other examples.

Residuals can arise from comparing data with visual references other than fitted lines. The reference might be a curve, as in the top panel of Figure 3.9, which graphs made-up data. Judging the vertical deviations of the points from the curve is difficult because of the rapidly changing slope. (This issue of graphical perception is discussed in detail in Section 4 of Chapter 4.) The visual impression from the top panel is that the residuals are smaller on the right than on the left. The graph of residuals against x in the bottom panel shows that the opposite is the case.

Graphing residuals is also illustrated in Figure 3.10, again, by made-up data. Observations are compared to a theoretical value for each of eight groups. The two-tiered error bars show 50% and 95% confidence intervals for the observations. The residuals, which in this case are the data minus the theoretical values, are graphed in the bottom panel; the result is increased resolution of the deviations of the data from the theoretical values and a better comparison of where the theoretical values lie with respect to the confidence intervals for the data.

The Tukey Sum-Difference Graph

There is another situation where graphing residuals is helpful. Suppose y_i is graphed against x_i for $i = 1$ to n to see how close x_i and y_i are to one another. An example is shown in the top panel of Figure 3.11. The data on the vertical axis, y_i, are the logarithms of abundances of certain elements in rocks brought back from the moon's Mare Tranquillitatis by the Apollo 11 astronauts in 1969 [91, p. 27]. The data on the horizontal axis, x_i, are the logarithms of abundances of the same elements in basalt from the earth. The purpose of the graph is to see how the composition of the moon rocks compares with that of basalt.

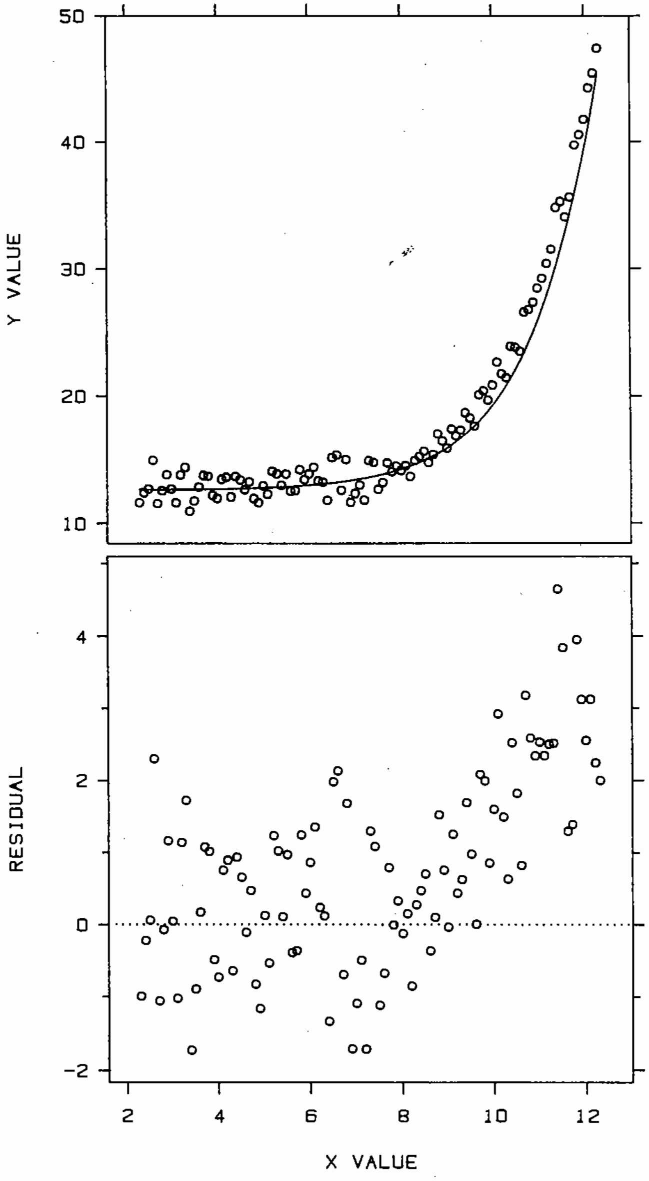

Figure 3.9 GRAPHING RESIDUALS. The visual impression from the top panel is that the vertical deviations of the points from the curve are greater for small x values than for large ones. The graph of residuals in the bottom panel shows the opposite is true.

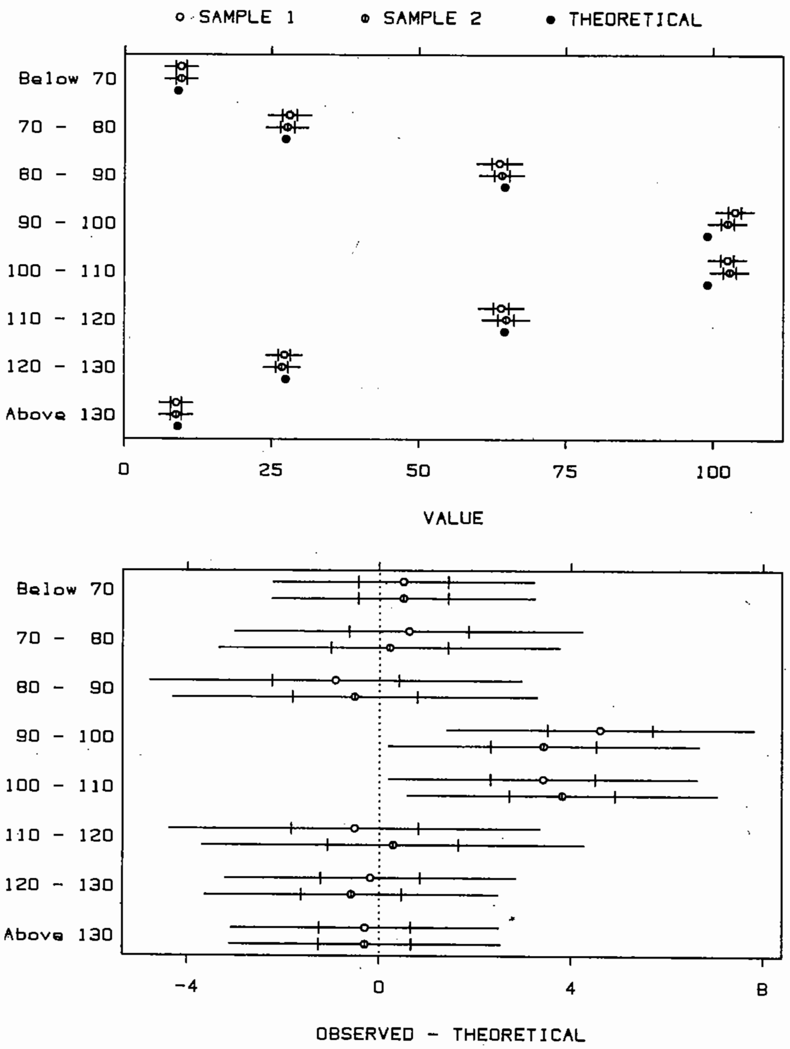

Figure 3.10 GRAPHING RESIDUALS. In the top panel made-up observations are compared with made-up theoretical values. The two-tiered error bars represent 50% and 95% confidence intervals. The residuals, which in this case are the data minus the theoretical values, are graphed in the bottom panel; the increased resolution allows us to compare more effectively where the theoretical values lie with respect to the confidence intervals.

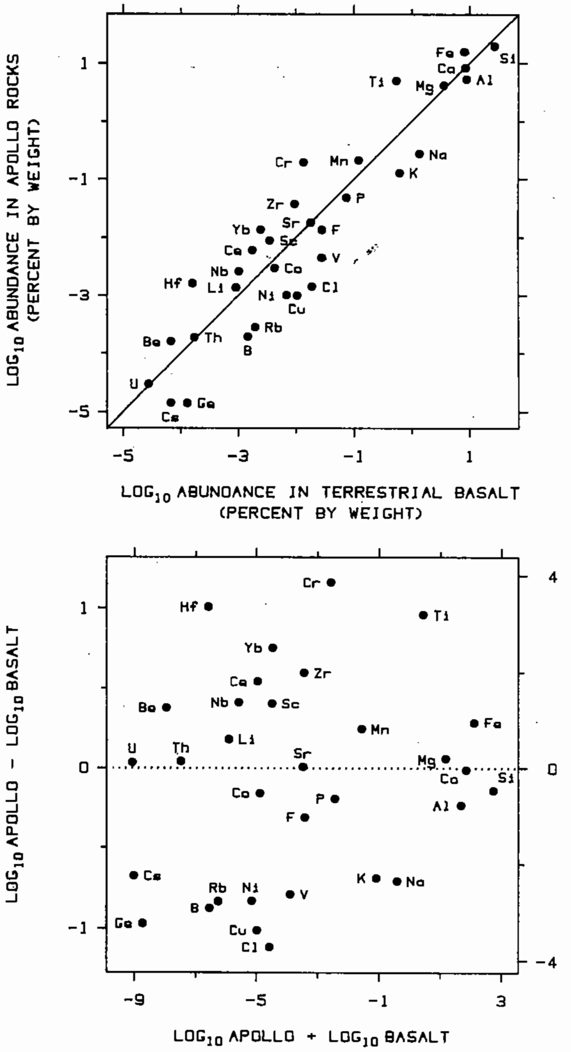

Figure 3.11 TUKEY SUM-DIFFERENCE GRAPH. In the top panel two sets of data with the same measurement scale are graphed to see how close the corresponding values are. The bottom panel is the Tukey sum-difference graph: $y_i - x_i$ is graphed against $y_i + x_i$. This graphical method, which is a 45° clockwise rotation of the top panel followed by an expansion of the vertical scale, allows us to study more effectively the deviations of the points from the line $y = x$.

In studying the composition data we would like to understand the values of $y_i - x_i$, the amounts by which the abundances differ. On the the top panel of Figure 3.11, $y_i - x_i$ is equal to the vertical deviation of the point (x_i, y_i) from the line $y = x$, and $x_i - y_i$ is equal to the horizontal deviation of (x_i, y_i) from the line. As with other graphs, however, it is difficult to assess the values of these deviations, or residuals, partly because the resolution of the residuals is poor.

In Figure 3.7, where we studied Playfair's data, our purpose was to see how the areas, y_i, *depend on* the population measurements, x_i. The variable y is a dependent variable and x is an independent variable. For the abundance data in the top panel of Figure 3.11 the situation is different. Neither variable is dependent or independent; we are seeking simply to see how the two variables are related, and by how much the abundances differ. We might look at residuals, in analogy with the Playfair data, by graphing $y_i - x_i$ against x_i. This, however, does not treat x_i and y_i equally, and we could just as well graph $y_i - x_i$ against y_i.

One way to graph $y_i - x_i$ that takes the equivalence of x_i and y_i into account is the *Tukey sum-difference graph*: $y_i - x_i$ is graphed against $y_i + x_i$, as illustrated in the bottom panel of Figure 3.11. The sum-difference graph can be thought of as the result of rotating the points in the top panel by 45° in a clockwise direction and then allowing the rotated points to expand in the vertical direction to fill the data region. To see this suppose

$$u_i = \frac{y_i + x_i}{\sqrt{2}}$$

and

$$v_i = \frac{y_i - x_i}{\sqrt{2}} .$$

If we graphed v_i against u_i and kept the number of data units per cm the same as on the graph of y_i vs. x_i, the points on the new graph would be exactly a 45° clockwise rotation of the points on the old one. The reader can rotate the book page 45° clockwise to see how the configuration of points on this new graph would appear. In the Tukey sum-difference graph there is no $\sqrt{2}$, which is a constant factor that does not affect the configuration of points; also, the number of data units per cm for $y_i - x_i$ is not forced to be the same as $y_i + x_i$, but rather the vertical scale is chosen so that the $y_i - x_i$ fill up the data region that is available.

In the bottom panel of Figure 3.11 the expansion of the scale for the $y_i - x_i$ now lets us assess these residuals far more effectively than in the top panel. The left vertical scale line shows differences of log base 10 abundances and the right vertical scale line shows differences of log base 2 abundances; the right vertical scale line helps us appreciate the factors since the differences vary only by about two powers of 10.

Figure 3.11 shows that titanium, which is one of the most abundant elements in both rock types, is higher in the moon rocks by about a factor of 10. Also, sodium is lower by a factor of about 5. This had already been discovered by the Surveyor spacecraft in 1967, which also measured composition in Mare Tranquillitatis. Surveyor landed on the moon, scooped up a lunar sample, measured abundances by alpha scattering, and sent the measurements back to earth as strings of zeros and ones. At the time, some doubted the reliability of the Surveyor results, in particular the high values of titanium and the low values of sodium. "Many doubting Thomases had to wait for the first Apollo landing on the Moon in July of 1969 to be convinced," wrote Anthony Turkevich, University of Chicago chemist and one of the developers of the Surveyor measurement methods [91, p. 23]. And convinced they were since the rock samples brought back by the Apollo missions showed the Surveyor measurements had been exceedingly accurate.

3.2 ONE OR MORE CATEGORIES OF MEASUREMENTS OF ONE QUANTITATIVE VARIABLE: GRAPHING DISTRIBUTIONS

Frequently, the goal of a data analysis is to study the distribution of one or more categories of measurements of a quantitative variable. That is, we want to study where the data for each category lie along the measurement scale.

An example of the study of distributions is shown in Figure 3.12. The data are from an experiment [51] on a special type of stereogram called a *random dot stereogram*, which was invented by Bela Julesz for studying visual perception [70, 71]. A viewer sees a three-dimensional object that is formed by a left and a right image, each of which has the appearance of tightly packed random dots. Typically, a viewer does not immediately see the object in such a stereogram, but after concentrating on the images for a while the object suddenly appears. The data in Figure 3.12 are the times taken by subjects to see a particular stereogram in which the viewed object was a spiral ramp pointing toward the viewer. Subjects were given varying types of prior information about what they were going to see, to determine if prior information can reduce appearance times. In Figure 3.12 there are two groups of measurements, where the grouping is based on the prior information given. The NV subjects received either *no* information or *verbal* information. The VV subjects received a combination of *verbal* and *visual* information. The NV group as whole received less prior information than the VV group, and the goal is to see if the distribution of the VV times is reduced compared with that for the NV times.

Point Graphs and Histograms

One standard way to show measurements of a variable or to compare different sets of measurements of a variable is to graph each set

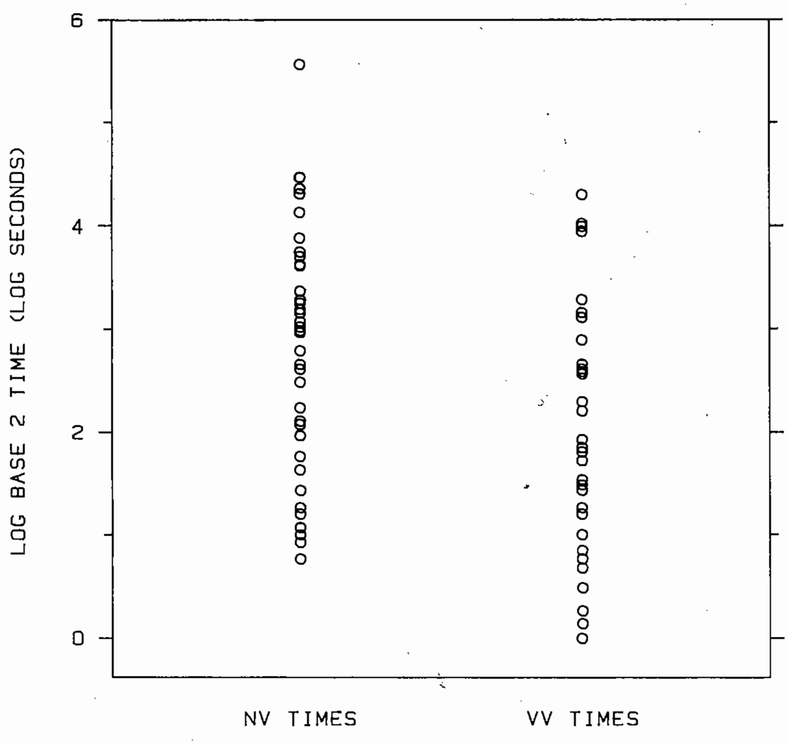

Figure 3.12 POINT GRAPH. The data are the times that two groups of people took to see a complex random dot stereogram. The goal is to compare the distributions of the two sets of data. In this figure two point graphs are used to make the comparison.

of values along a line. Such a *point graph* is used in Figure 3.12 to show the stereogram times.

Another standard method for studying distributions is the *histogram*, one of the staples of scientific graphics that has a long history going back at least to the 19th century. In Figure 3.13 the stereogram times are shown by histograms. The variable on the vertical scales is percent of counts — 100 times the number of counts in each interval divided by the total number of observations, which is 43 for the NV times and 35 for the VV times. Since the numbers of observations are different for the two groups, using the percent of counts in each interval rather than the counts themselves provides a more effective comparison of the two distributions.

A point graph is a reasonable display when the number of observations is not large. In Figure 3.12 we probably have reached the upper limit of the number of values that can be effectively shown without offsetting the plotting symbols in the horizontal direction to avoid overlap. When the number of values is large, or even moderate, the histogram is the better display to use. This is illustrated in Figure 3.14; the histogram shows redshifts of quasars from a catalog compiled by Adelaide Hewitt and Geoffrey Burbidge, two astronomers at the Kitt Peak National Observatory in Tucson, Arizona [59].

It should be remembered that a histogram reduces the information in the data. A measured value, such as redshift, is itself usually an interval of values because there is limited accuracy in measuring devices and because data are often rounded. When a histogram is made, the interval width of the histogram is generally greater than the data inaccuracy interval, so accuracy is lost. As we decrease the interval width of a histogram, accuracy increases but the appearance becomes more ragged until finally we have what amounts to a point graph. In most applications it makes sense to choose the interval width on the basis of what seems like a tolerable loss in the accuracy of the data; no general rules are possible because the tolerable loss depends on the subject matter and the goal of the analysis. (One exception to this statement is the very small fraction of cases in which the purpose of the histogram is to estimate a probability density rather than to simply show the data [44, 114]; this usage will not be treated here.)

Point graphs and histograms certainly do a good job of showing us individual distributions of data sets, but they generally do not provide *comparisons* of distributions that are as incisive as methods that will be described later in this section. From Figures 3.12 and 3.13 there is a suggestion that the VV times are less than the NV times — that is, that the increased prior information given to the VV group reduced viewing

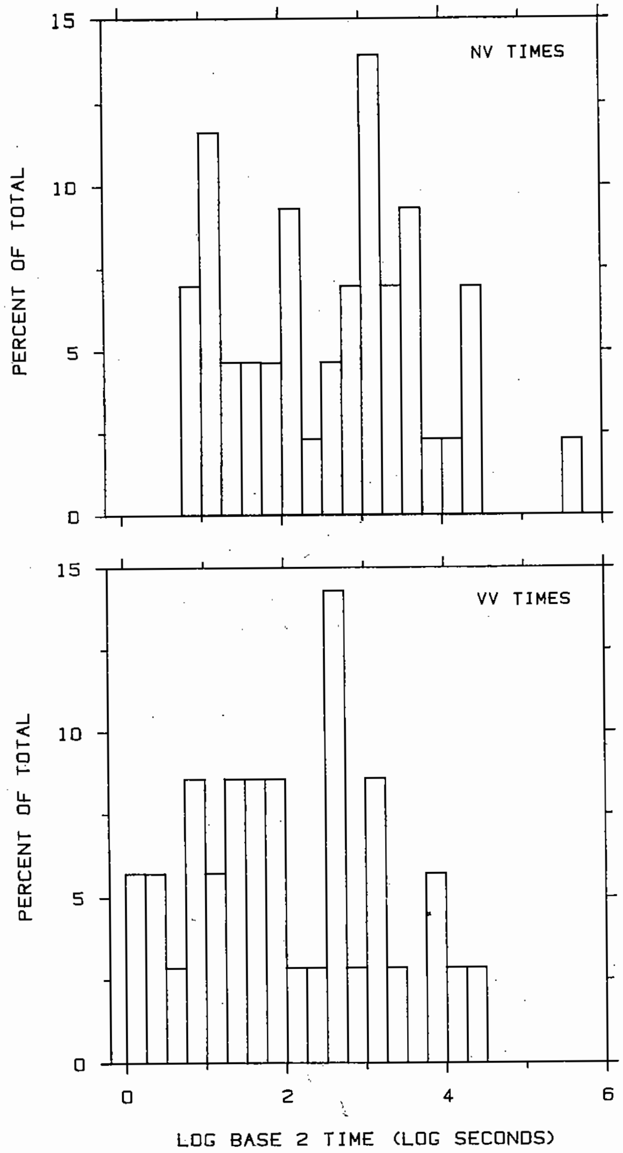

Figure 3.13 HISTOGRAM. Each histogram shows the percentage of values in intervals of equal length. The histogram does a good job of displaying each data set, but is usually not as effective for comparing distributions as other methods.

times — but the two graphs give us little quantitative information about the magnitude of the difference.

Percentile Graphs

Figure 3.15 shows *percentile graphs* of the two distributions of stereogram times. A pth *percentile* of a distribution is a number, q, such that approximately p percent of the values of the distribution are less

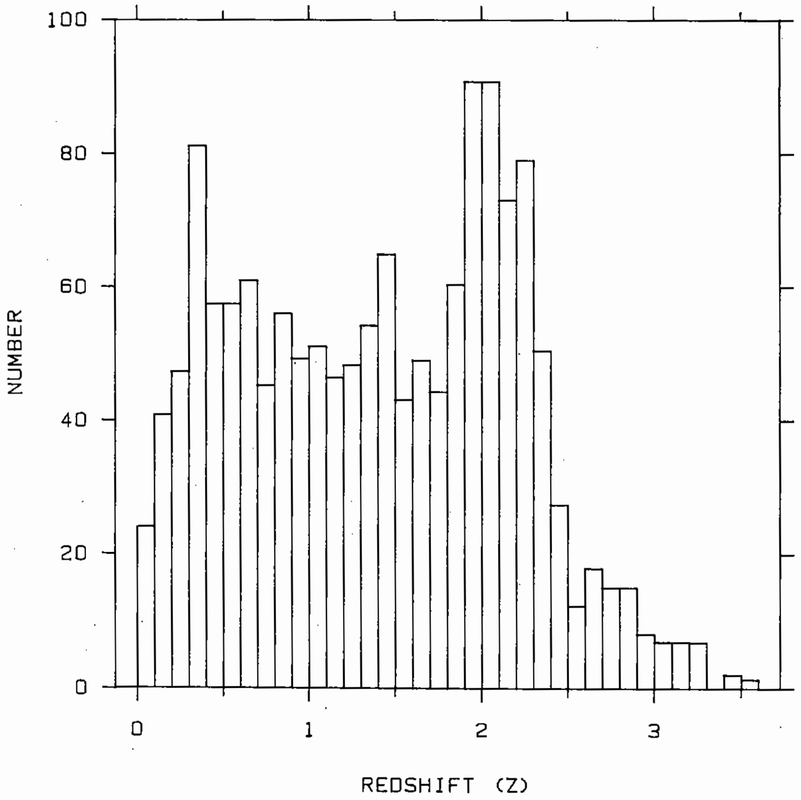

Figure 3.14 HISTOGRAM. In most applications it makes sense to choose the interval width on the basis of what seems like a tolerable loss in accuracy of the data. In this example the width is 0.1 units.

than or equal to q; p is the *p-value* of q. Suppose x_1 is the smallest observation in a data set, x_2 is the next to smallest, and so forth up to x_n, which is the largest observation. For example, if the data are

$$5 \quad 1 \quad 9 \quad 3 \quad 14 \quad 9 \quad 7$$

then

$$x_1 = 1 \quad x_2 = 3 \quad x_3 = 5 \quad x_4 = 7 \quad x_5 = 9 \quad x_6 = 9 \quad x_7 = 14 .$$

We will take x_i to be the p_ith percentile of the data where

$$p_i = 100 \, \frac{i-0.5}{n} .$$

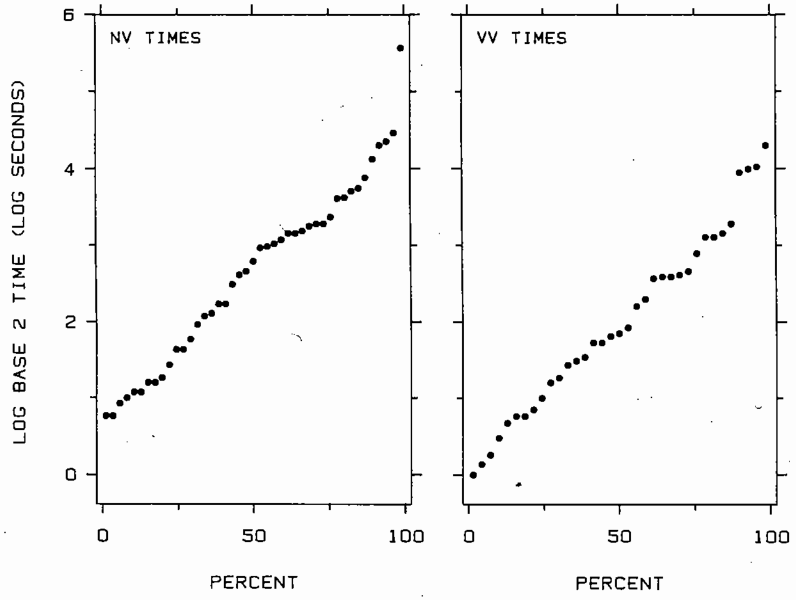

Figure 3.15 PERCENTILE GRAPH. On each panel, the data are graphed against their *p*-values. The *p*-value for an observation is very nearly the percentage of the data that is less than or equal to the observation; the observation is said to be the *p*th percentile.

For the above set of seven values

$$p_1 = 100(1 - 0.5) / 7 = 7.1$$

$$p_2 = 100(2 - 0.5) / 7 = 21.4$$

and so forth to

$$p_7 = 100(7 - 0.5) / 7 = 92.9 .$$

On the percentile graph each x_i is graphed against its p-value, p_i.

Subtracting 0.5 in the formula for the p-value of x_i is a convention in statistical science [21] and arises from the desire to make the definition of the percentile of a set of data as consistent as possible with the concept of the percentile of a theoretical probability distribution, such as the normal. One piece of heuristic reasoning that might satisfy some is the following: Suppose x_i is the result of rounding. When we count how many observations are less than or equal to x_i, we count only 1/2 for x_i itself, because there is a 50-50 chance that the actual value of the observation is less than or equal to x_i, the recorded value. But for percentile graphs the subtraction of 0.5 is a trivial issue that has little affect on the visual appearance of the display.

Percentile graphs are often more effective for comparing data distributions than point graphs or histograms because the p_i are shown, which means corresponding percentiles can be compared. For example, in Figure 3.15 we can easily see that the 50th percentile, or the *median*, of the NV times is slightly less than 3 $\log_2$ seconds; this median value can be compared with that of the VV times, which is about 2 $\log_2$ seconds. Comparing percentiles is usually the most informative way to compare two distributions; we will return to this point later.

Box Graphs

It is sometimes enough, in order to convey the salient features of the distribution of a set of data, to show just a summary of the data. One such summary, shown in Figure 3.16, is the Tukey *box graph* [125]. The five horizontal lines on each box graph portray five percentiles whose p-values, from bottom to top, are 10, 25, 50, 75, and 90. All values in the data set above the 90th percentile and below the 10th percentile are graphed, as on a point graph.

We need a rule to compute the percentiles that appear on the box graph. So far, we know only that x_i is the p_ith percentile. It is not always the case that the p_i will happen to include the numbers needed for the box graph. For the example introduced earlier, the x_i and p_i are

i	x_i	p_i
1	1	7.1
2	3	21.4
3	5	35.7
4	7	50
5	9	64.3
6	9	78.6
7	14	92.9

In this example, one of the x_i happened to be the 50th percentile, however, none of the other box graph percentiles appear. We can get other percentiles by linearly interpolating the x_i and p_i values.

Here is a simple way to do the linear interpolation. Let p be the p-value of the percentile. We want a value of v such that

$$100 \, \frac{v-0.5}{n} = p \; .$$

Solving for v we get

$$v = \frac{np}{100} + 0.5 \; .$$

If v turns out to be an integer then x_v is the pth percentile. However, v will often not be an integer. Let k be the integer part of v and let f be the fractional part; for example, if $v = 10.375$ then $k = 10$ and $f = 0.375$. The pth percentile using linear interpolation is

$$(1-f)x_k + f x_{k+1} \; .$$

Let us apply this to the computation of the 25th percentile for the above set of seven values.

$$v = \frac{7 \cdot 25}{100} + 0.5$$

$$= 2.25 \; .$$

The 25th percentile is

$$0.75 \, x_2 + 0.25 \, x_3$$

$$= 0.75 \cdot 3 + 0.25 \cdot 5$$

$$= 3.5 \; .$$

The interpolation rule always leads to a simple result for the 50th percentile; if n is odd, it is the middle observation, $x_{(n+1)/2}$ and if n is even, it is the average of the two middle observations, $x_{n/2}$ and $x_{n/2+1}$.

Box graphs have many strengths. One is that the chosen percentiles can be compared effectively. For example, in Figure 3.17 we can see easily that the 50th percentiles of the NV times and VV times differ by

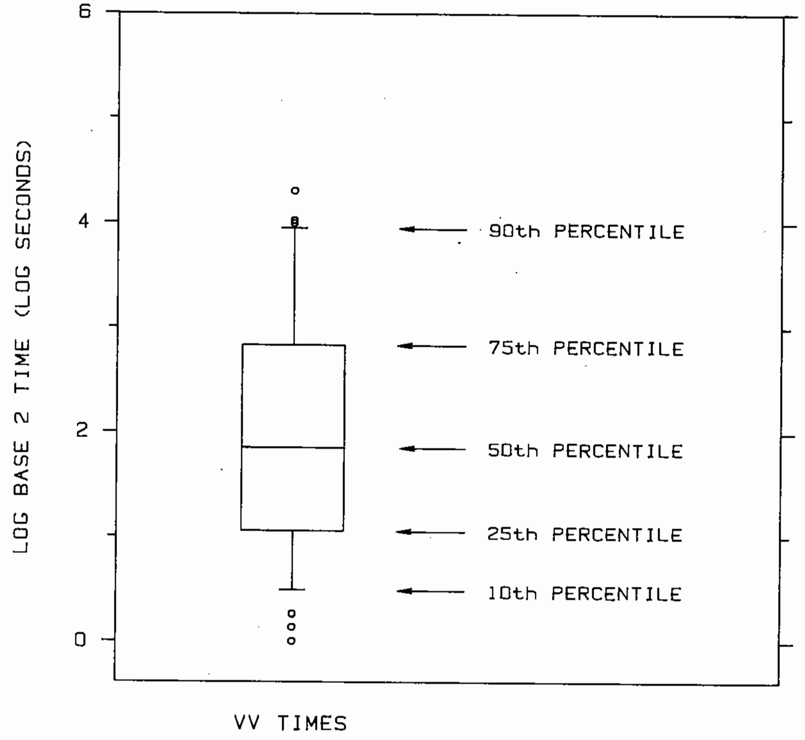

Figure 3.16 TUKEY BOX GRAPH. A box graph shows selected percentiles of the data, as illustrated in this figure. All values beyond the 10th and 90th percentiles are graphed individually as on a point graph.

roughly one $\log_2$ second, or a factor of 2. A second strength is that by graphing the large and small values, unusual values are not swept under the rug as they often are when the summary of the distribution consists of a sample mean and a sample standard deviation. (This point will be discussed further in Section 3.7.) Finally, box graphs can be used even when the number of distributions is not small.

In Figure 3.18 ten distributions are compared by box graphs. The data on the vertical axis are the payoffs from 254 runnings of the daily New Jersey Pick-It Lottery from May 22, 1975 to March 16, 1976 [102], just after the lottery began. In this game a player picks a three-digit number from 000 to 999. It costs 50¢ to bet on one number. Players who selected the winning number share the prize, which is half of the money bet on that day. Since the drawing of the winning number is random, so that all numbers are equally likely, the best strategy is to pick a number that few other people are likely to pick.

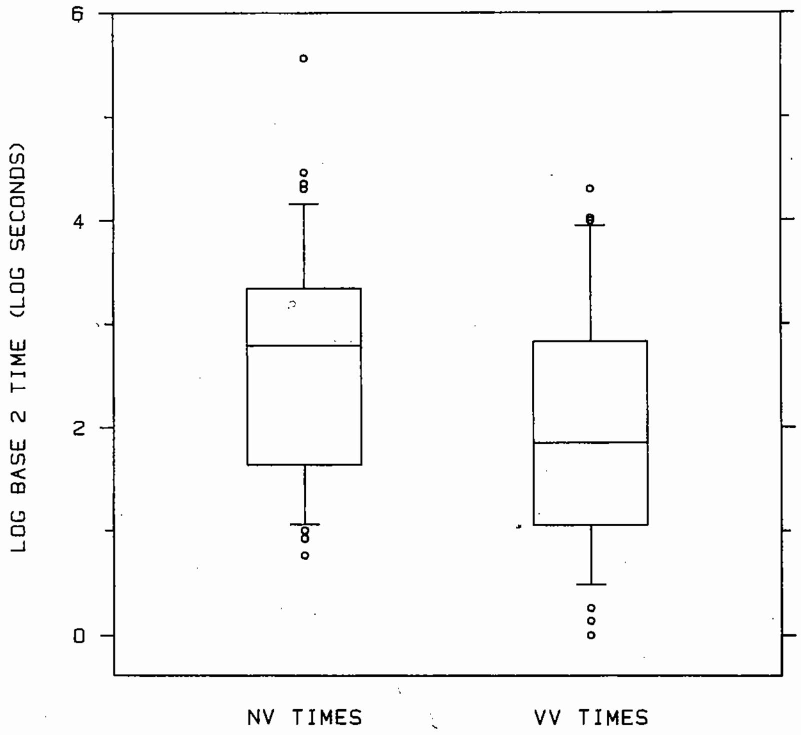

Figure 3.17 BOX GRAPH. Box graphs are an excellent way to compare distributions because they allow us to compare corresponding percentiles. In this example we see the 50th percentile of the NV times is greater than that for the VV times by about one $\log_2$ second, or a factor of 2.

The payoffs in Figure 3.18 have been divided into ten groups according to the winning number. The first group, labeled "0", is winning numbers from 000 to 099; the second group is 100 to 199; the third group is 200 to 299; and so forth. Thus the ten box graphs give a comparison of the ten distributions of payoffs.

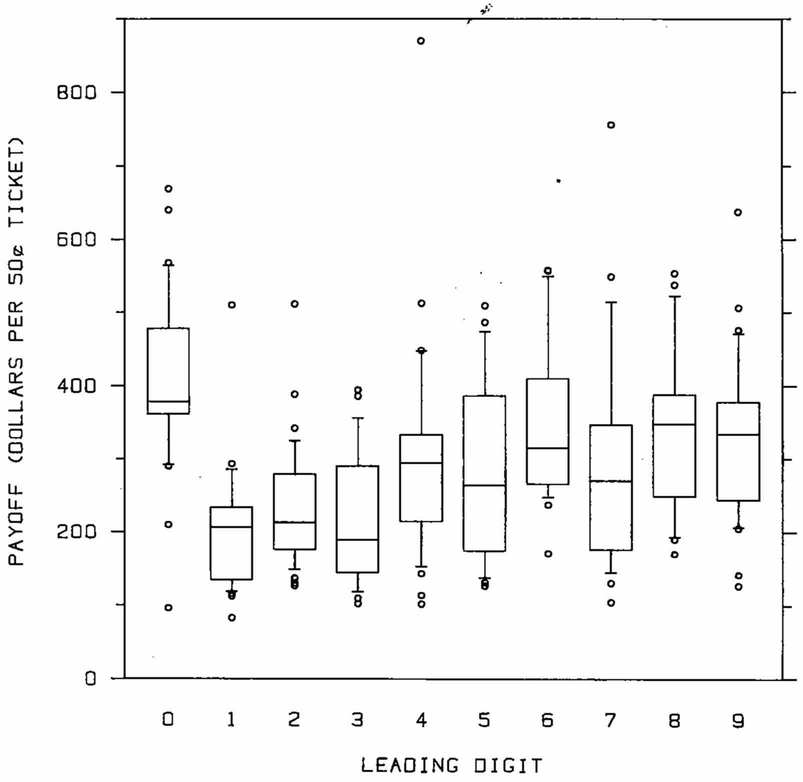

Figure 3.18 BOX GRAPH. The vertical scale is payoff of the New Jersey lottery, or numbers game, in which a player picks a three-digit number from 000 to 999. Winners share half of the pot. Each box graph shows the distribution of payoffs for all numbers with a particular leading digit. A leading digit of zero has the highest payoffs because fewer people tend to pick them. As the leading digit increases from one to nine the payoffs increase in a zigzag fashion, showing odd first digits are preferred to even.

Figure 3.18 has a clear message: the payoffs for numbers starting with zero tend to be high, which means bettors avoid them. One exception to this behavior is a zero-starting number with a payoff around $100, which is nearly the lowest value of all payoffs; in this case the winning number was 000, and it is not surprising that it was a popular one. There is an interesting trend in the remaining nine groups of numbers. The payoffs tend to increase in going from the smaller to the larger numbers, but in a zigzag fashion, suggesting that odd first digits are preferred to even.

If bettors' choices were uniformly distributed over all the numbers, the expected payoff would be $250 (not $500 since the state takes half of the money). However, the graph suggests that by the right choice of a number with a leading 0 we might be able to push the expected payoff above $500, the break-even point. Unfortunately, this is no longer true. Richard Becker and John Chambers showed that as time went along New Jersey Pick-It players caught on, the distribution of chosen numbers became more nearly uniform, and the maximum payoffs declined and rarely exceeded $500 [9].

The details of the box graph given in Figure 3.16 are not meant to create dogma. Variations are often sensible. Figure 3.16 is already a variation of the original method, which is called a *box plot* by its inventor, John Tukey [125]. In a particular application it might make sense to choose other percentiles or to eliminate the graphing of the individual large and small values or to draw the box graphs horizontally rather than vertically. Also, procedures other than linear interpolation can be used to compute percentiles. One simple rule is to select the x_i whose p_i comes closest to the p-value of the desired percentile. In the above example the 25th percentile would be 3 using this procedure, since its p-value, 21.4, is closest to 25. If n is not small, say n is greater than 50, linear interpolation and this procedure will usually give similar results.

Percentile Graphs with Summaries

The percentile graph and the box graph can be combined as in Figure 3.19 to form a *percentile graph with summary*. The horizontal reference lines show the five percentiles of the box graph; this allows us to compare these five percentiles with more visual efficiency than if the reference lines were not there.

Percentile Comparison Graphs

The *percentile comparison graph* was invented in 1966 by Martin Wilk and Ram Gnanadesikan [135]. It is not widely known in science and technology, but its use deserves to spread because of its enormous power for comparing two data distributions.

When distributions are compared, the goal is usually to rank the categories according to how much each has of the variable being measured; for the stereogram times we want to know which group took more time, and for the lottery data we are interested in finding the leading digits that give the highest payoffs.

The most effective way to investigate which of two distributions has more is to compare the corresponding percentiles. This was the insightful observation of Wilk and Gnanadesikan and their invention

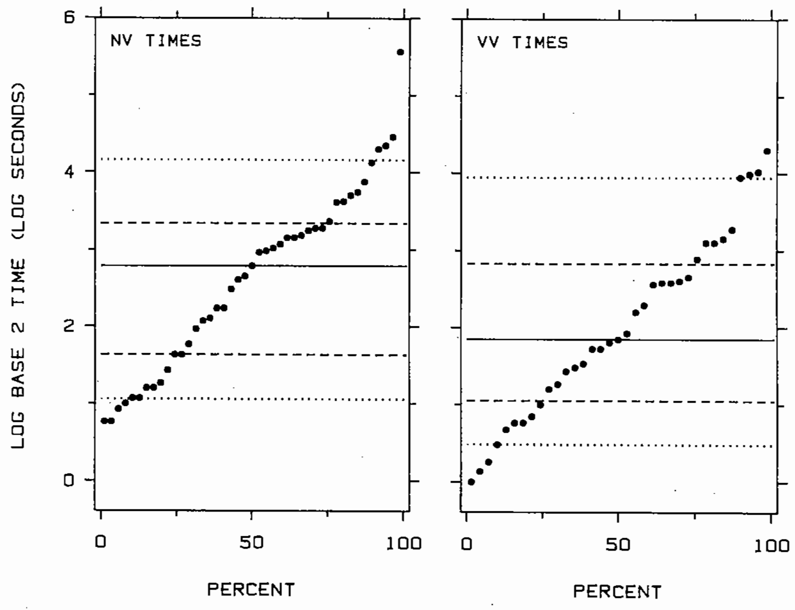

Figure 3.19 PERCENTILE GRAPH WITH SUMMARY. The five percentiles of the box graph are shown on a percentile graph by horizontal lines.

could not be more simple or elegant — graph the percentiles of one distribution against the corresponding percentiles of the other distribution. For example, we might graph the 50th percentile of the first data set against the 50th of the second data set, the 75th percentile of the first against the 75th percentile of the second, and so forth.

The top panel of Figure 3.20 is a percentile comparison graph; the two data sets are the scores of males and the scores of females on the verbal SAT test in 1983 [111]. There were 464,733 people in the males' data set and 497,809 in the females' data set. The highest possible score on the test is 800 and the lowest is 200. The following are the p-values of the percentiles of the distributions that are shown on the graph: 1 2 3 4 5 10 20 30 40 50 60 70 80 90 95 96 97 98 99. The point in the lower left corner of the data region is the 1st percentile for the males against the 1st percentile for the females, and the point in the upper right corner of the data region is the 99th percentile for the males against the 99th percentile for the females. The bottom panel of Figure 3.20 uses the Tukey sum-difference graph, discussed in Section 3.1, to give a clearer picture of the differences of the percentiles.

How do we make the percentile comparison graph? Suppose, first, that there is a moderate number of observations in the smaller of the two data sets, say no more than 50. Let $x_1,...,x_n$ be the first data set, ordered from smallest to largest, and let $y_1,...,y_m$ be the second set of data, also ordered.

Suppose $m = n$. Then y_i and x_i are both $100(i-0.5)/n$ percentiles of their respective data sets, so we would make the percentile comparison graph by graphing y_i against x_i. Thus in the $m = n$ case the graph is quite simple — we just graph the ordered values for one group against the ordered values of the other group.

Suppose $m < n$. Then y_i is the $100(i-0.5)/m$ percentile of the y data, so on the percentile comparison graph we graph y_i against the $100(i-0.5)/m$ percentile of the x data, which typically must be computed by interpolation. Thus in the case of an unequal number of observations in the two data sets, there are as many points on the graph as there are values in the smaller of the two data sets.

Figure 3.21 illustrates the unequal case; the display is a percentile comparison graph of the stereogram data: the 43 NV times and 35 VV times. There are 35 points on the graph. For example, the 9th VV time is $y_9 = 1.0 \log_2$ seconds; this is a percentile with p-value 24.3, and it is graphed against the 24.3 percentile of the NV times, which was computed by interpolating the 10th and 11th NV times, y_{10} and y_{11}; the interpolated value is $0.06\, y_{10} + 0.94\, y_{11} = 1.62 \log_2$ seconds.

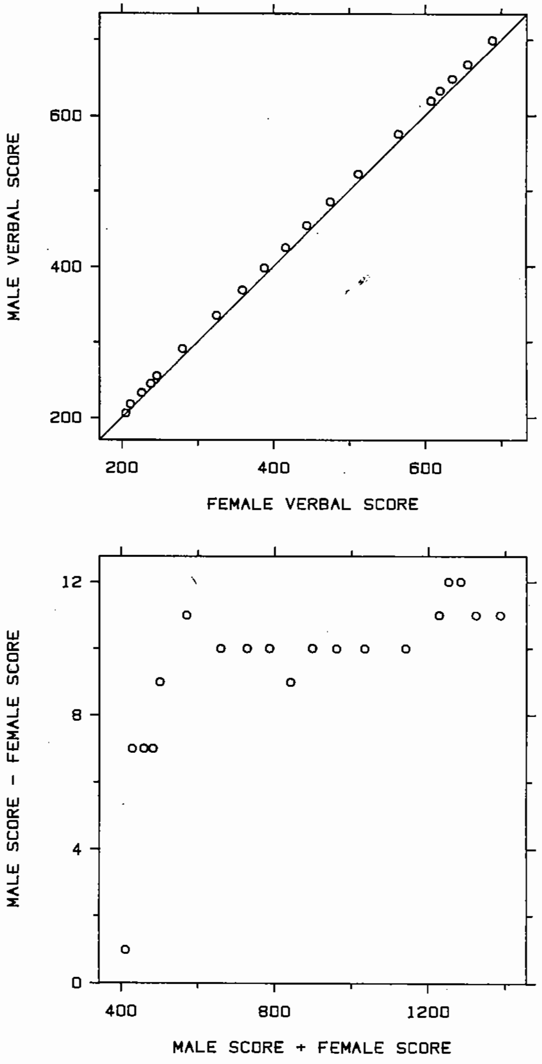

Figure 3.20 PERCENTILE COMPARISON GRAPH. The percentile comparison graph, illustrated in the top panel, is a simple but powerful tool for comparing two distributions. Percentiles from one distribution are graphed against corresponding percentiles from the other distribution. The data in this figure are scores of males and females on the verbal SAT test. The percentiles compared are 1, 2 ,..., 5; 10, 20 ,..., 90; and 95, 96 ,..., 99. The bottom panel is a Tukey sum-difference graph of the values in the top panel. The graph shows that throughout most of the range of the distribution, scores of males are about 10 points higher.

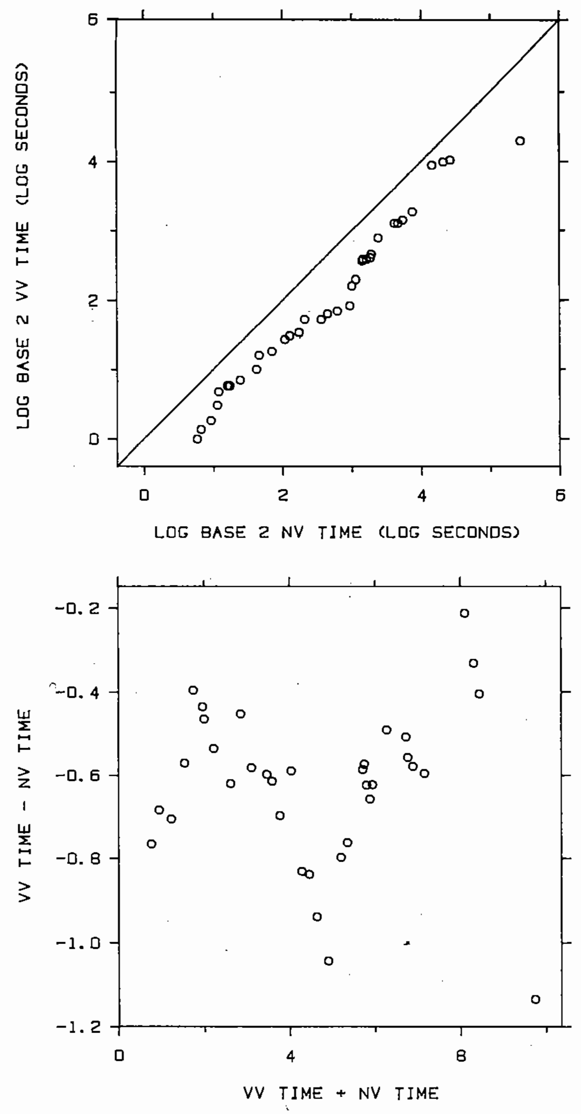

Figure 3.21 PERCENTILE COMPARISON GRAPH. In the top panel percentiles of the VV times are graphed against corresponding percentiles of the NV times. The bottom panel is a Tukey sum-difference graph. Throughout the entire range of the distribution the NV times are greater than the VV times; the average increase is about 0.6 $\log_2$ seconds, which is a factor of 1.5.

Suppose the smaller of the two data sets has a large number of values. For example, for the SAT data the smaller group, the males, has 464,899 values. We do not need, of course, to graph 464,899 points, because far fewer points can characterize the differences between the two distributions. In such a case a liberal helping of percentiles, with p-values ranging from close to 0 to close to 100, can be graphed against one another. In many cases, as few as 15 to 25 percentiles can adequately compare the two distributions. This procedure was used for the percentile comparison graph of the SAT scores in Figure 3.20.

The question of which of two distributions has more and by how much is a simple one whose answer can be complicated. The percentile comparison graph, by giving us a detailed comparison of the two distributions, can show whether the answer is simple or complicated, and if complicated, just what the complication is. This will be illustrated by several examples.

Figure 3.20 shows that the way in which the scores of males and females differ is relatively simple. Throughout most of the range of the distribution the males' percentiles are about 10 points higher than the females' percentiles, but at the very bottom end the difference tapers off. Thus a reasonable summary of the pattern of the points is a line parallel to the line $y = x$ with an equation $y = x + 10$. The comparison of the two distributions can be summarized by the simple statement, the males' scores are about 10 points higher throughout most of the range of the distributions.

Figure 3.22 is a percentile comparison graph of made-up test scores. The pattern is a line through the origin with equation $y = 0.8x$. Now it is not true that the corresponding percentiles differ by a constant amount as they did for the verbal SAT scores; now the high percentiles differ by more than the low ones. But because the general pattern is a line through the origin with slope 0.8, the percentage decrease of the males' scores is a fixed amount. That is, because the males' scores, y, are approximately related to the females' scores, x, by $y = 0.8x$, we have $(y-x)/x = -0.2$, which means the males' scores are approximately 20% lower throughout the range of the distribution.

If we were to take the logarithms of the values in Figure 3.22 the multiplicative pattern would be transformed into an additive pattern like Figure 3.20. In Figure 3.21, logarithms performed such a multiplicative-to-additive transformation for the stereogram times. The general pattern of the points in Figure 3.21 is a line, $y = x + k$, where k is about 0.6 $\log_2$ seconds. Had we graphed the points without taking logarithms the general pattern would have been a line through the origin with slope $2^{0.6} = 1.5$.

Figure 3.23 compares two other sets of hypothetical scores. The pattern of the data is a line with a slope less than 1; the line $y = x$ intersects this pattern at the 50th percentiles of the distributions. The 50th percentiles of the two groups are equal, but the distributions differ in a major way: the high scores for the females are higher than the high scores for the males, and the low scores for the males are higher than the low scores for the females. The two distributions are centered at the same place but the females' scores are more spread out.

Figure 3.24 also compares hypothetical scores. Throughout most of the range of the distribution, males and females are the same, but at the very top end the females have higher scores. That is, the exceptionally high scores for the females are better than the exceptionally high scores for the males.

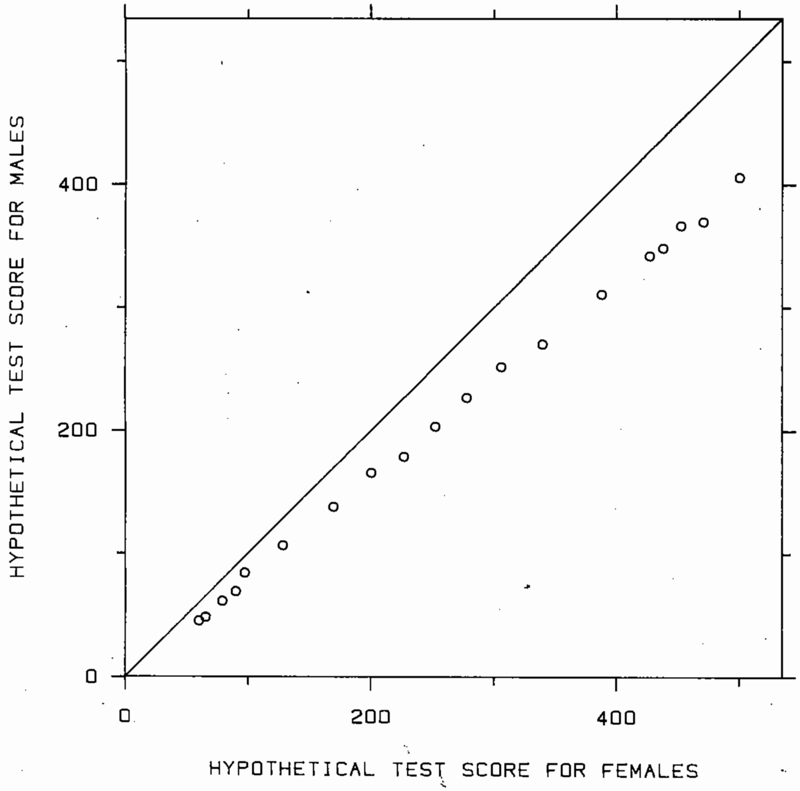

Figure 3.22 PERCENTILE COMPARISON GRAPH. The data are hypothetical test scores. Since the points lie close to a line through the origin with slope 0.8, scores of males are about 20% lower throughout most of the range of the distribution.

Figure 3.25 is back to real data: 1983 mathematics SAT scores for males and females [111]. The top panel compares the same percentiles that are compared for the verbal scores in Figure 3.20; the bottom panel is a Tukey sum-difference graph. From the 99th to the 50th percentile most of the percentiles for the males are 55 to 60 points higher than those for the females. But from the 50th percentile to the lowest percentiles the differences decrease from about 55 points to about 10

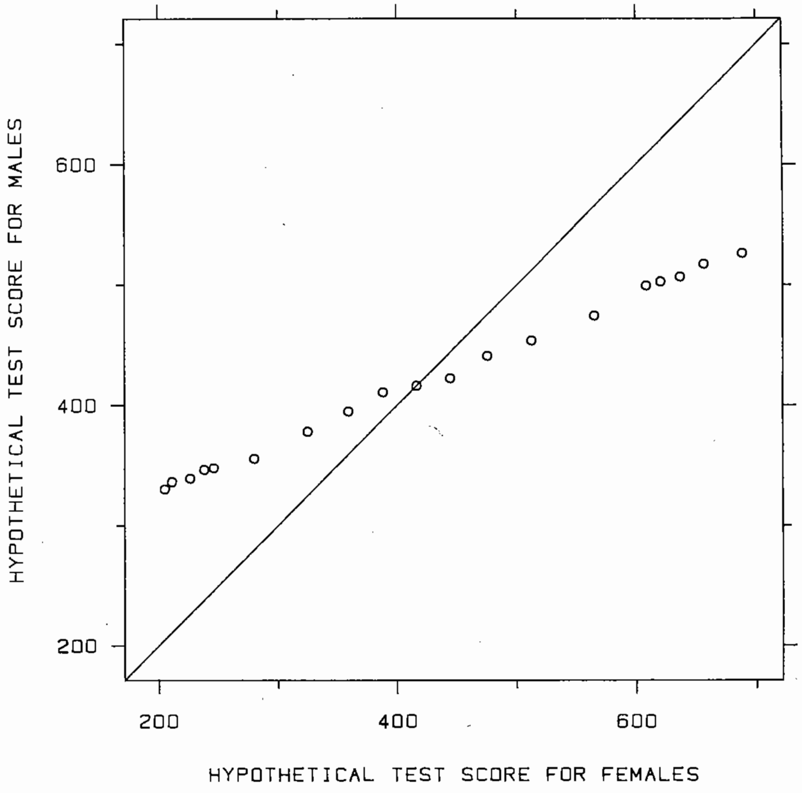

Figure 3.23 PERCENTILE COMPARISON GRAPH. The points lie close to a line that has slope less than one, and the 50th percentiles lie on the line $y = x$. Thus the 50th percentiles, or middles, of the two distributions are the same but the female scores are more spread out.

points. The way in which the scores of males and females differ is considerably more complicated than the simple linear patterns in some of the previous percentile comparison graphs.

Means are often used to characterize how two distributions differ, but this often misses important information or worse yet, misleads. The mean scores for the math test are 445 for the females and 493 for the males, a difference of 48. Using just the means misses the important fact that high scorers, middle scorers, and low scorers differ by different amounts. Data distributions can be complicated, and when they are, the percentile comparison graph can reveal the complication to us.

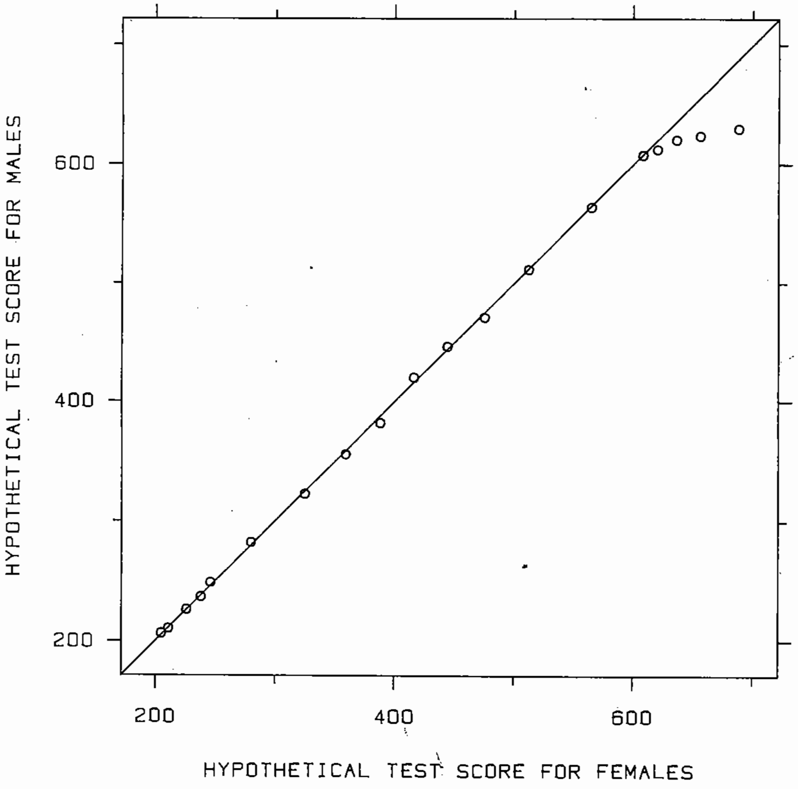

Figure 3.24 PERCENTILE COMPARISON GRAPH. Throughout most of the range of the distribution, male and female scores are nearly the same, but for the very highest percentiles, female scores are higher.

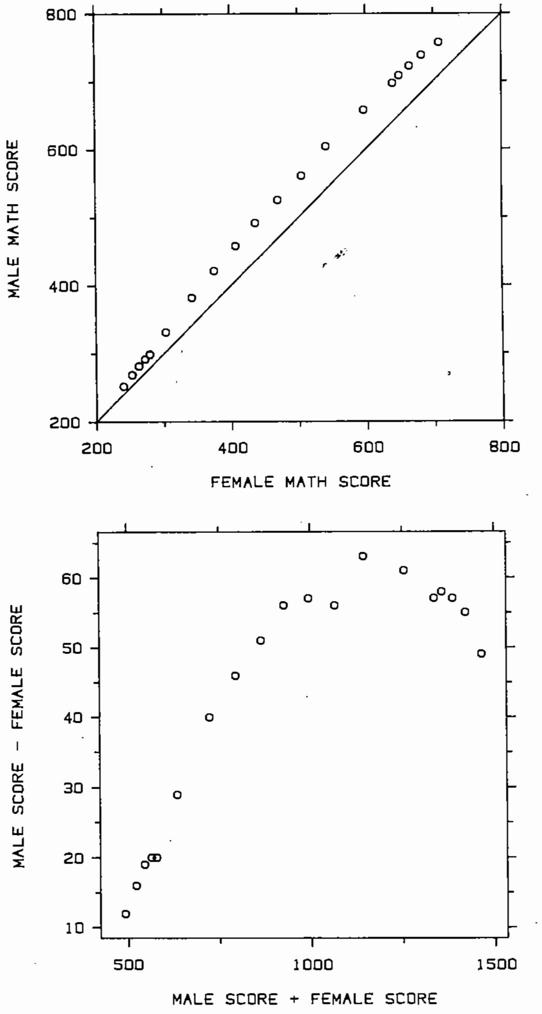

Figure 3.25 PERCENTILE COMPARISON GRAPH. The top panel is a percentile comparison graph of scores of males and females on the math SAT test. The same percentiles graphed in Figure 3.20 are graphed here. The bottom panel is a Tukey sum-difference graph of the values in the top panel. The graph shows that for the top half of the distributions, scores of males are typically 55 to 60 points higher, and that for the bottom half the difference ranges from 10 to 55 in going from the lowest percentiles to the 50th. The average scores, 445 for the females and 493 for the males, do not convey nearly as much information about how the two distributions differ.

3.3 ONE QUANTITATIVE VARIABLE WITH LABELS: DOT CHARTS

Ordinary Dot Charts

We often need to display measurements of a quantitative variable in which each value has a label associated with it. Figure 3.26 shows an example. The data are from a survey on the amount of use of graphs in 57 scientific publications [27]. For each journal, 50 articles from the period 1980-1981 were sampled. The variable graphed in Figure 3.26 is the fraction of space of the 50 articles devoted to graphs (not including legends) and the labels are the journal names. Figure 3.26 is a *dot chart*, a graphical method that was invented [28] in response to the standard ways of displaying labeled data — bar charts, divided bar charts, and pie charts — which usually convey quantitative information less well to the viewer than dot charts. (This is demonstrated in Section 4 of Chapter 4.)

When there are many values in the data set, as in Figure 3.26, the light dotted lines on the dot chart enable us to visually connect a graphed point with its label. When the number of values is small, as in Figure 3.27, the dotted lines can be omitted, since the visual connection can be performed without them.

The data in Figure 3.27 are the ratios of extragalactic to galactic energy in seven frequency bands [93], where energy is measured per unit volume. The frequencies in the seven bands increase in going from the top of the graph to the bottom. In five of the seven bands the galaxies have much higher intensities than the space between galaxies. One of these five bands is visible light; this should come as no surprise since on a clear night on the earth we can see galactic matter in the form of stars (or light reflected from a star by our moon) and only blackness in between. For microwaves and x-rays there is much more energy coming from outside the galaxies. The extragalactic microwave radiation, discovered by Nobel prize winners Arno Penzias and Robert Wilson of AT&T Bell Laboratories in 1965 [107], has an explanation: it is the remnant of the big bang that gave our universe its start. But the extragalactic x-ray radiation remains a mystery whose solution might also tell us something fundamental about the structure of the universe.

When they appear, the dotted lines on the dot chart are made light to keep them from being visually imposing and obscuring the large dots that portray the data. When we visually summarize the distribution of the data, the data dots stand out and the graph is a percentile graph,

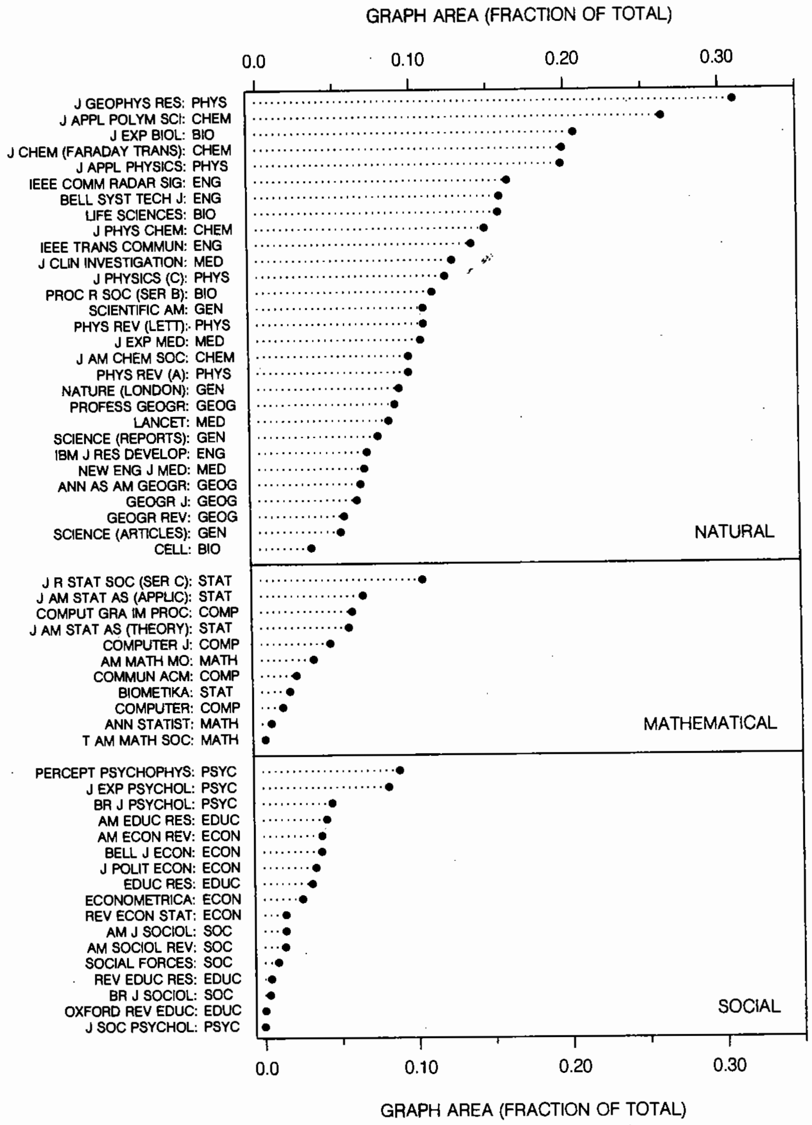

Figure 3.26 DOT CHART. A dot chart shows the fraction of space devoted to graphs for 57 scientific journals. The dot chart is a graphical method for data where each numerical value has a label. The dotted lines, which enable us to connect each value with its label, end at the data dots because the baseline is zero.

provided the data are ordered from smallest to largest. When we want to emphasize this distribution, a *p*-value scale can be put on the right vertical scale line as in Figure 3.28.

The data in Figure 3.28 are the per capita state taxes (sales, income, and fees for state services) in the 50 states of the U.S. during the fiscal year 1980 [137, p. 116]. The graph shows that state taxes vary by a factor of about 3. New Hampshire, the state where so many presidential candidates have gotten their start, or their finish, is clearly a state ready to listen to candidates who advocate lower taxes.

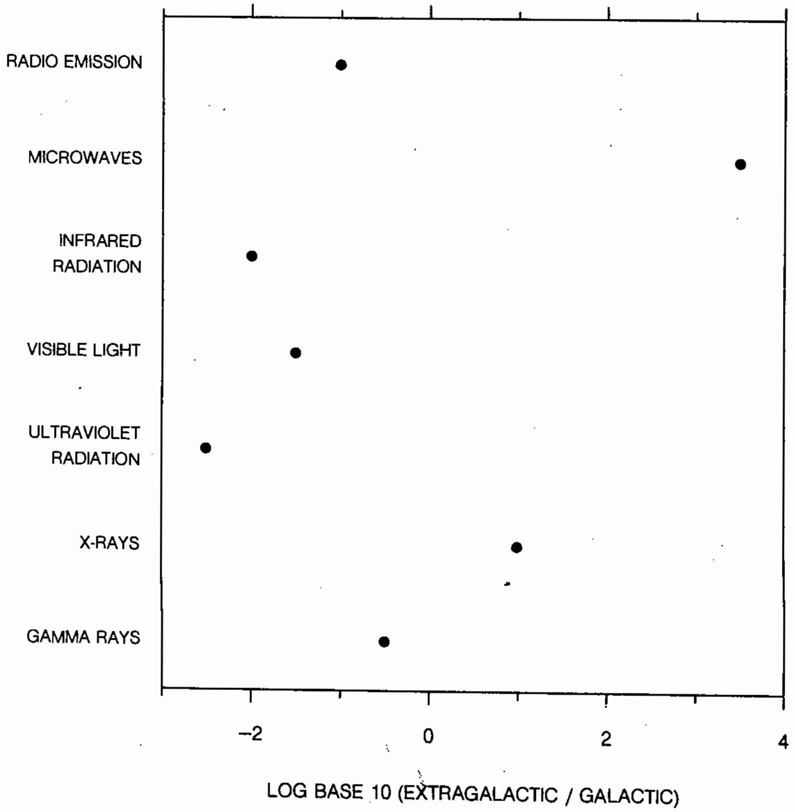

Figure 3.27 DOT CHART. The dotted lines are omitted because the labels and the numerical values can be visually connected without them.

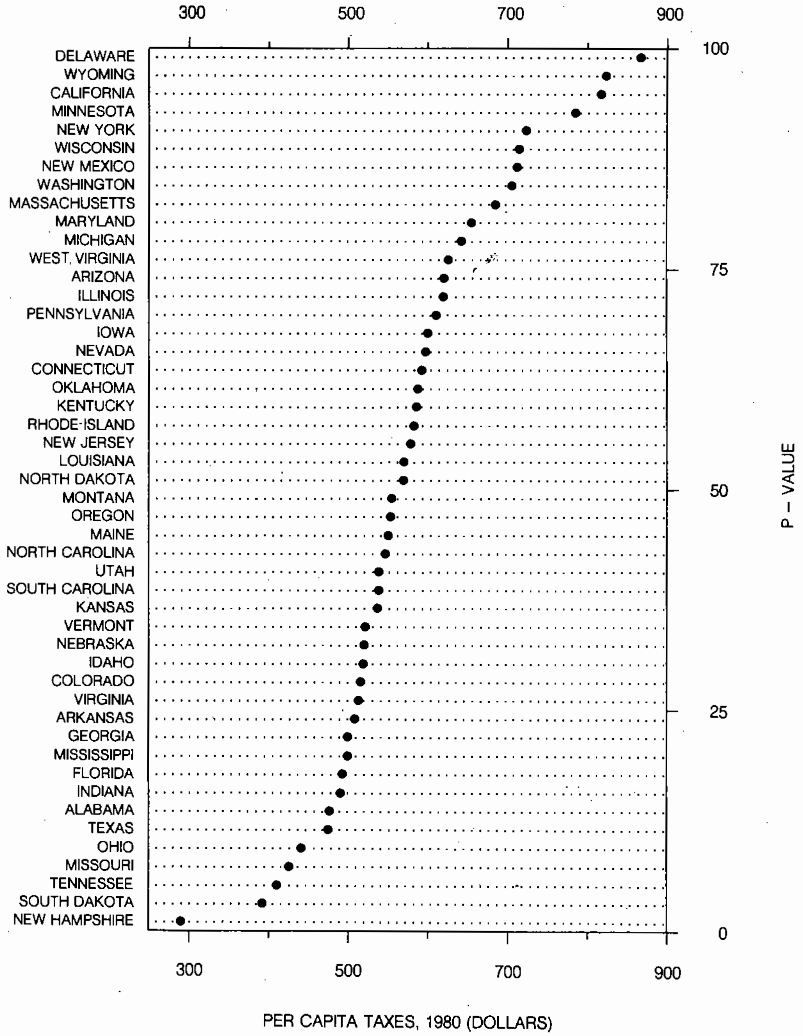

Figure 3.28 DOT CHART. When the data are ordered from smallest to largest, the dot chart provides a percentile graph; the *p*-values are shown by the right vertical scale line. The dotted lines go all of the way across the graph. The baseline is a number near 275, and if the dotted lines ended at the data dots, line length would encode taxes minus a number near 275 that has no significant meaning.

When there is a zero on the scale of a dot chart, or some other meaningful baseline value from which the dotted lines emanate, then the dotted lines can end at the data dots, as in Figure 3.26. The dotted lines should go across the graph when the baseline value has no particular meaning, as in Figure 3.28. Here is the reason. When the dotted lines stop at the data dots, there are two aspects of the graphical symbols that encode the quantitative information — the lengths of the dotted lines and the relative positions of the data dots along the common scale. The lengths of the dotted lines encode the magnitudes of the deviations from the baseline. In Figure 3.26 the baseline is zero, so line length encodes the fractional graph areas, which is perfectly reasonable. However, if the baseline value has no important meaning, the deviations have no meaning. Suppose that in Figure 3.28 the dotted lines ended at the data dots. Then line length would encode taxes minus a number around 275. Since this number has no significant meaning in this application, line length would be encoding meaningless values; changing line length would be wasted energy and might even have the potential to mislead. By making the dotted lines go across the graph in Figure 3.28, the portions between the left vertical scale line and the data dots are visually de-emphasized.

The dotted lines also should go across the graph when there is a scale break, as in Figure 3.29, which graphs speeds of animals [136]. If we stopped the dotted lines at the data dots in this figure, those that were broken by the scale break would not have any meaning, even though the baseline is meaningful.

Two different methods can be used to put scale breaks on dot charts. One, shown in Figure 3.29, is to use a vertical full scale break. A second method, shown in Figure 3.30, can be used when better resolution is needed on one or both sections of the scale; for example, the resolution of the scale for the slowest four animals is considerably better in Figure 3.30 than in Figure 3.29.

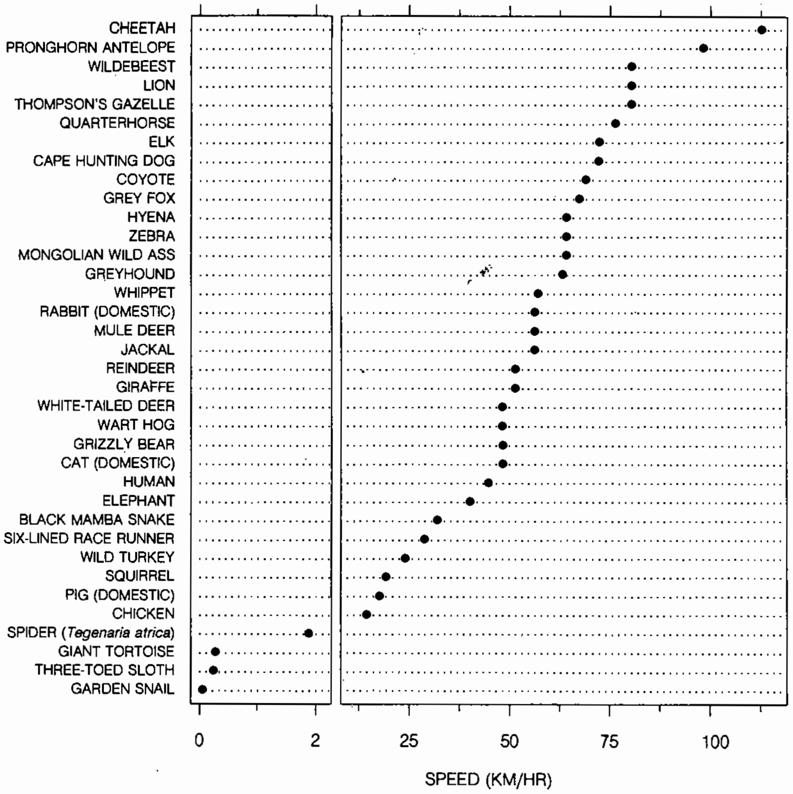

Figure 3.29 DOT CHART. A vertical full scale break is used on this dot chart. The dotted lines go all of the way across the graph since if they ended at the data dots, the lengths of those crossing the break would be meaningless.

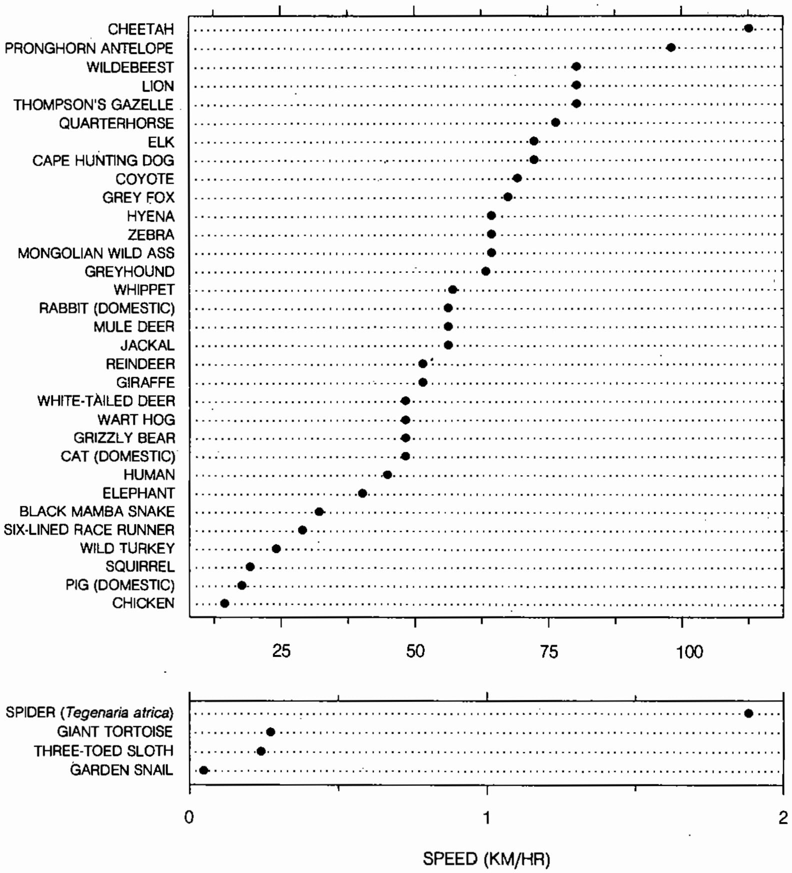

Figure 3.30 DOT CHART. This method can be used to break the scale of a dot chart when better resolution is needed on one or both panels formed by a vertical full scale break.

Two-Way, Grouped, and Multi-Valued Dot Charts

Figure 3.31 is a *two-way dot chart,* a method for showing labeled data that form a *two-way classification.* In this case the two-way data are the percentages of U.S. immigrants from six groups of nationalities during four time periods [76]. (The percentages add to 100% for each time period.) An observation is classified by the time period and the nationality group. Each column of the graph shows the values for one time period and each row shows one nationality. The graph portrays clearly the data's main event: The proportions for Europeans and Canadians have decreased through time and those for Asians and Latin Americans have increased.

Another way to show two-way data is by a *grouped dot chart.* In Figure 3.32 the immigration data are grouped by nationality group and in Figure 3.33 they are grouped by time period. The first grouped dot chart emphasizes the changes through time and the second emphasizes the mixture of nationality groups for each time period.

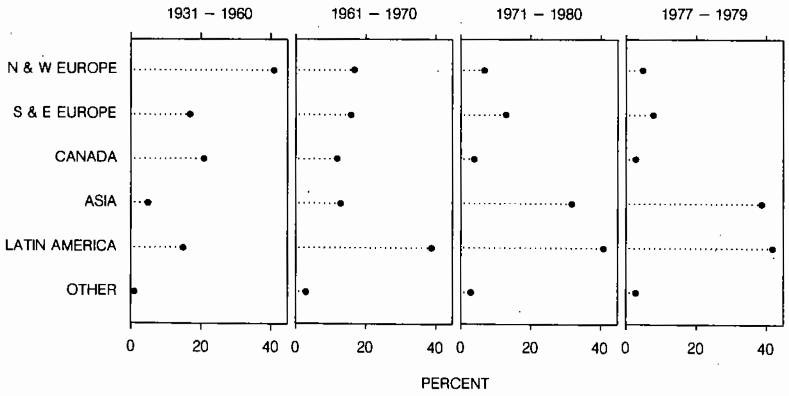

Figure 3.31 TWO-WAY DOT CHART. The two-way dot chart can be used to show data classified by two factors. In this example the data are the percentages of immigrants in six nationality categories for four time periods.

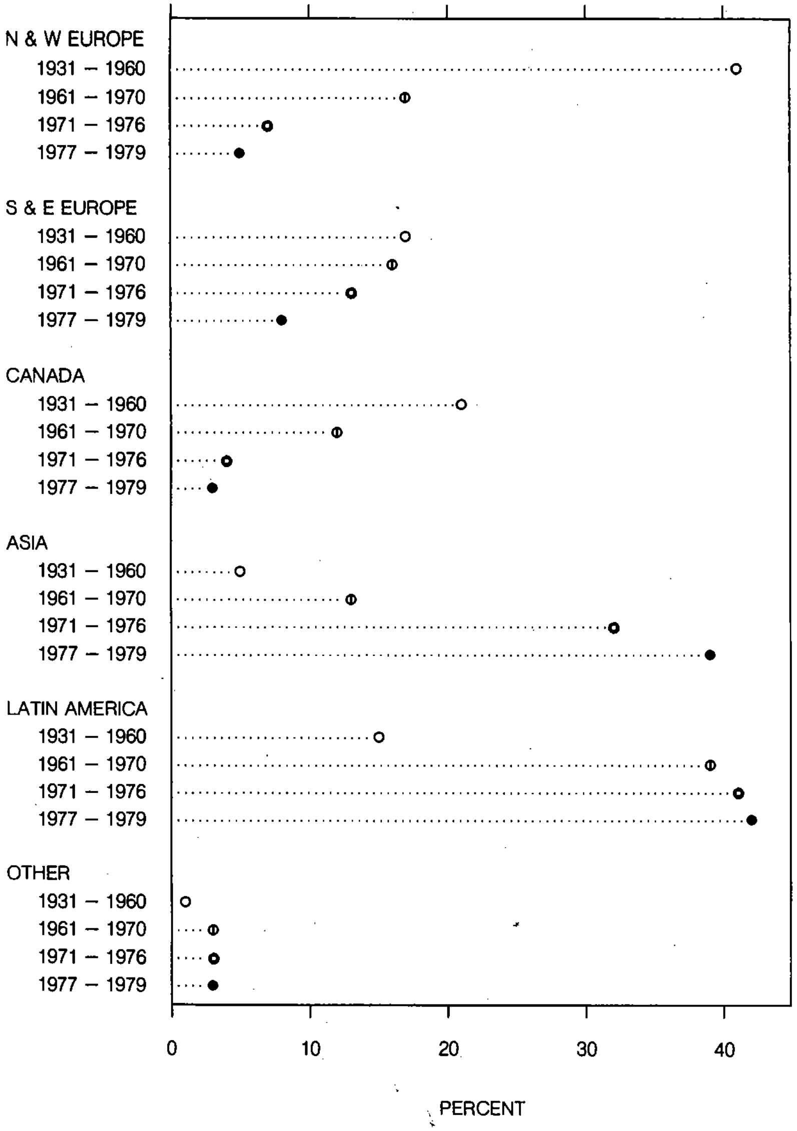

Figure 3.32 GROUPED DOT CHART. The immigration data are grouped by nationality. This emphasizes the time trend in the data for each nationality group.

A final way to show two-way data, provided one of the two groupings has a small number of categories, is the *multi-valued dot chart* in Figure 3.34. The data are the immigration percentages for just the first and last time periods.

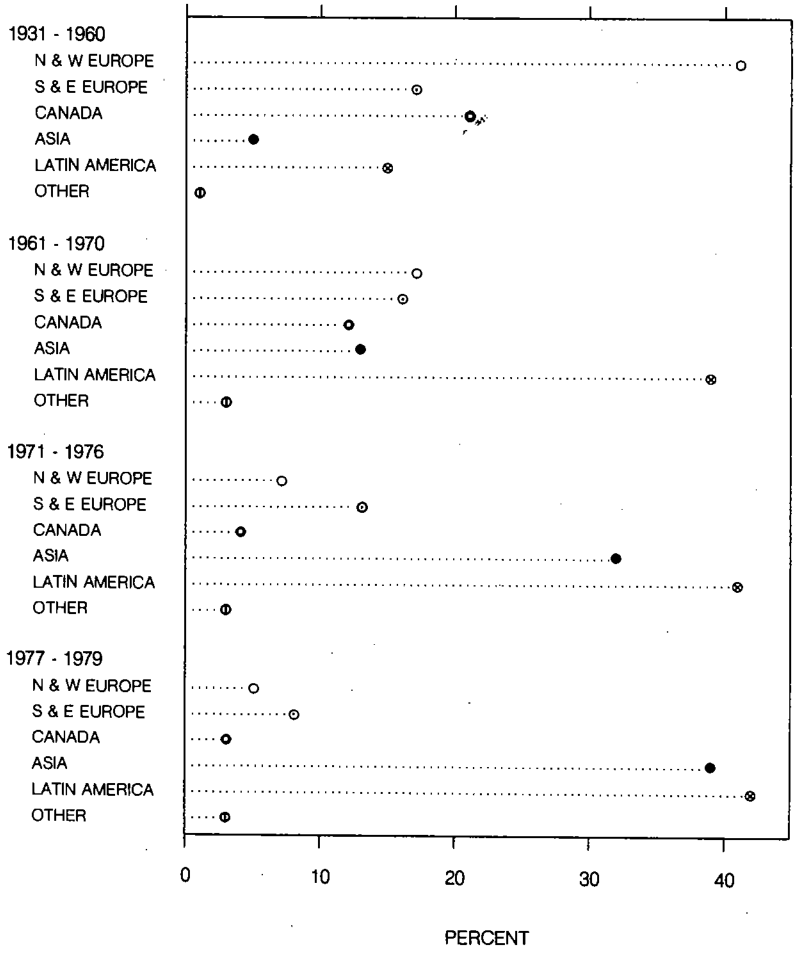

Figure 3.33 GROUPED DOT CHART. The immigration data are grouped by time. This emphasizes the mixture of nationality groups for each time period.

3.4 TWO QUANTITATIVE VARIABLES

Many scientific investigations are aimed at discovering how two quantitative variables are related. An example is measurements of caloric intake and blood sugar levels for a group of people, where the purpose is to discover how the two variables are related. In a two-variable study we often want to find out how one, the *dependent variable*, depends on the other, the *independent variable*. For example, we might want to know how blood sugar depends on caloric intake. This section is about graphing two quantitative variables.

Overlap: Logarithms, Residuals, Moving, Sunflowers, Jittering, and Circles

In Section 2 of Chapter 2 it was pointed out that a recurring problem of graphing two variables is overlapping plotting symbols, which is caused by graph locations of different values being identical or very close. When overlap occurs, different plotting symbols can obscure one another and we can lose an appreciation of the values of the data. Methods that help avoid a loss of visual distinguishability will now be described.

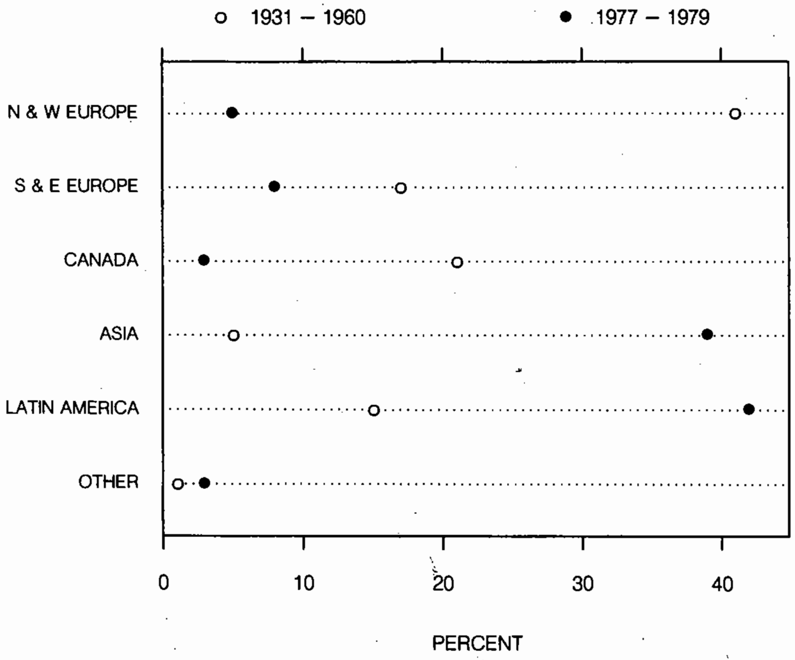

Figure 3.34 MULTI-VALUED DOT CHART. The dot chart is multi-valued because there is more than one value on each line.

When both scales of a two-variable graph have poor resolution, severe overlap can occur. This is illustrated in Figure 3.35, which shows brain weights and body weights of 27 animal species [113, p. 39]. The values of each variable are skewed to the right, that is, most of the data are squashed together near the origin and a few values stretch out toward the high end of the scale. In Section 1 of this chapter and in Section 4 of Chapter 2 we have seen that taking logarithms and graphing residuals are two methods that can improve resolution; for this

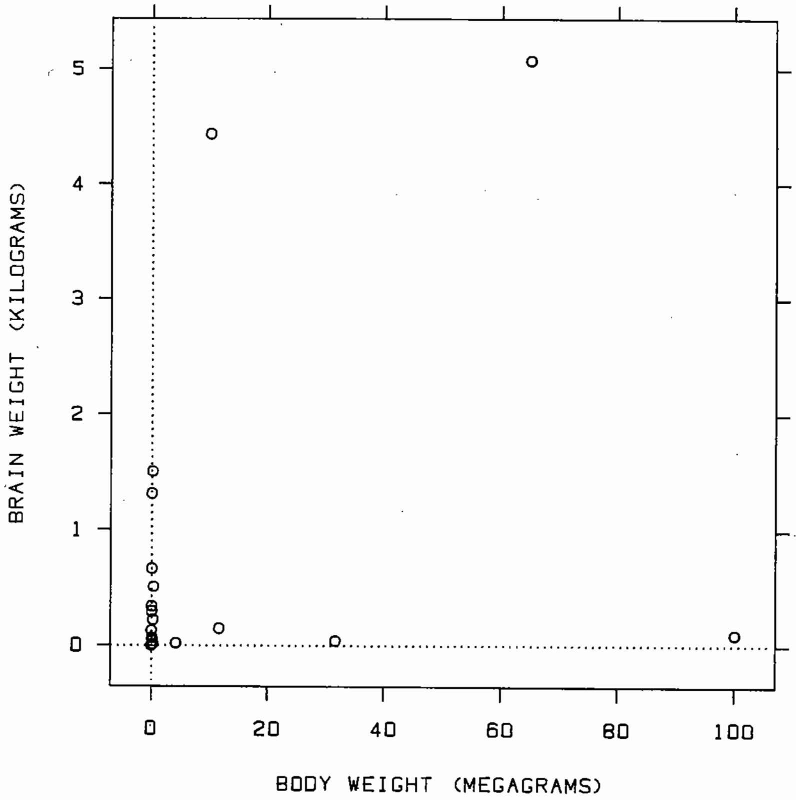

Figure 3.35 OVERLAP. Overlapping plotting symbols must be visually distinguishable. If the resolution along both scales of a two-variable graph is poor because the measurements are skewed, overlap can cause problems, as on this graph.

reason these two methods can reduce or eliminate overlap. For Figure 3.35, logarithms solve the problem; in Figure 3.36 logarithms are graphed and now there is no overlap.

The top panel of Figure 3.37 shows data on magnetic moments and beta decays of mirror nuclei [19]. Theory suggests that the variable on the vertical scale is linearly related to the variable on the horizontal scale, and the data support the theory since the points lie close to the line on the graph, which was fitted using least squares on all but three of the points. Plotting symbols on the graph overlap because the data are squashed together along the line. Graphing residuals in the bottom panel of 3.37 improves the resolution and nearly eliminates the overlap.

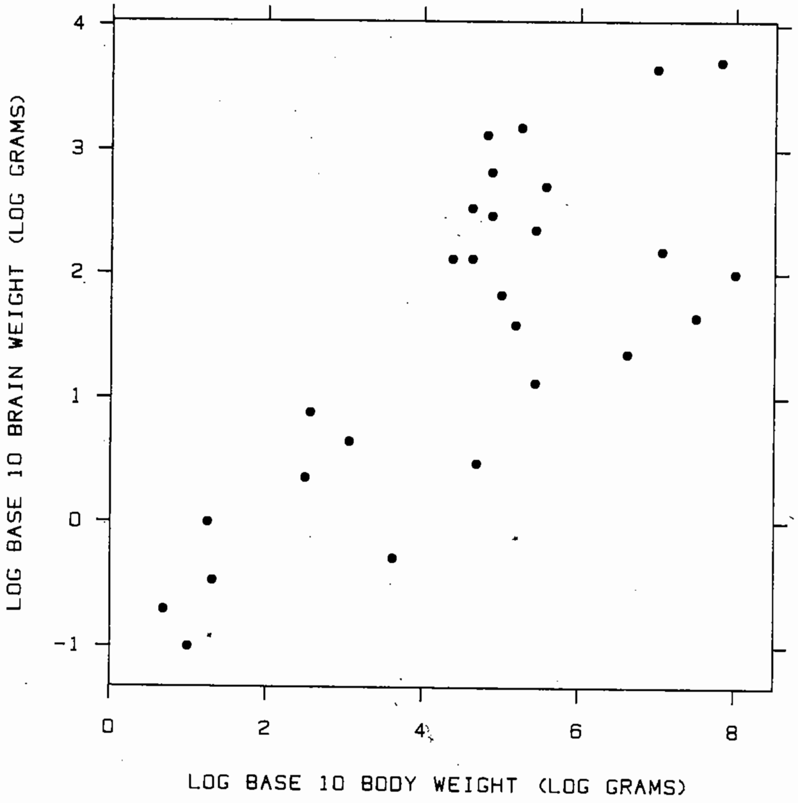

Figure 3.36 LOGARITHMS. The logarithms of the data in Figure 3.35 are graphed and now there is no overlap. Taking logs will often alleviate the overlap caused by skewed positive data.

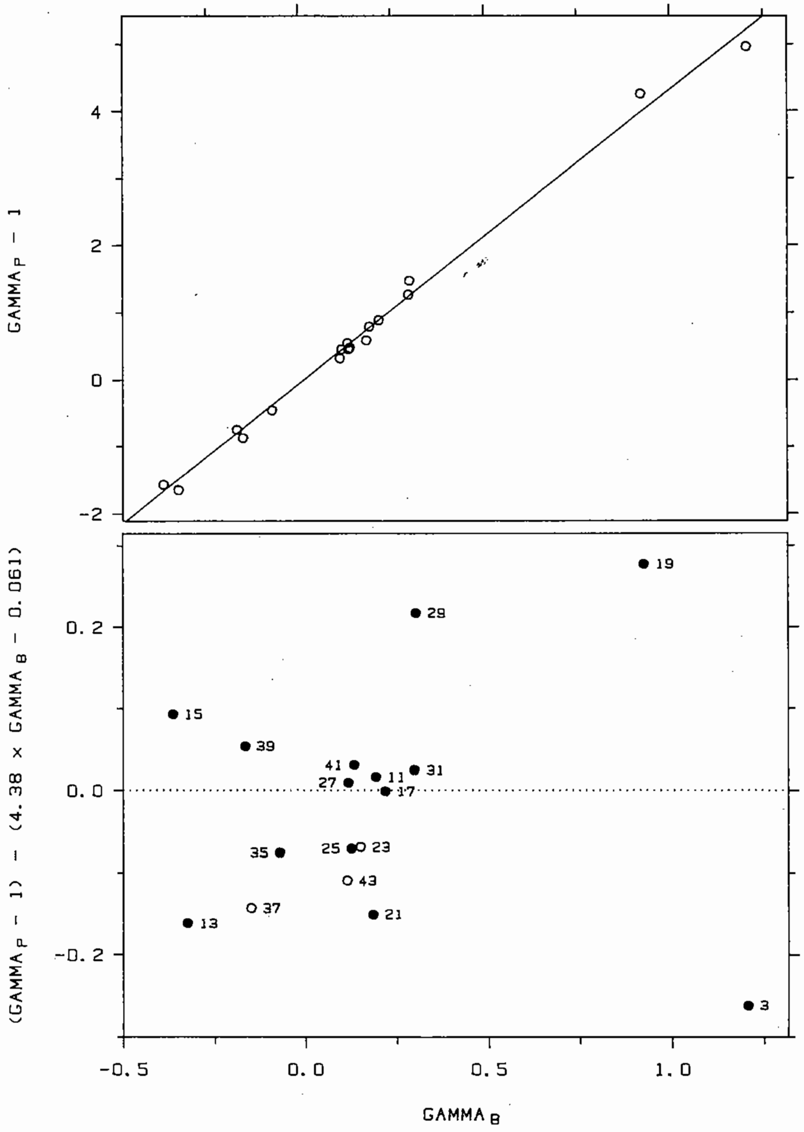

Figure 3.37 RESIDUALS. Graphing residuals is another way to reduce overlap. The data in the top panel are squashed together along the line. Graphing residuals in the bottom panel improves the resolution and nearly eliminates the overlap.

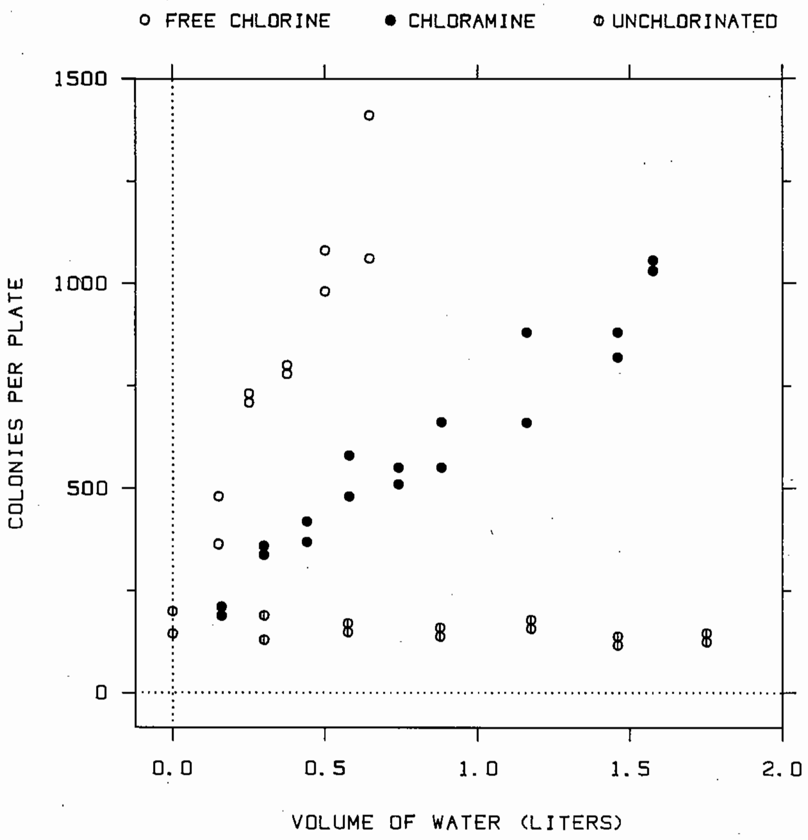

Figure 3.38 MOVING. If the number of overlapping plotting symbols is small, the graph locations of the points can be altered slightly to reduce the overlap. On this graph, symbols that just touch one another have been moved vertically.

Information now can be added to the graph; the numbers by the points are mass numbers and points graphed with unfilled circles are those omitted in the least squares fit. The two panels of Figure 3.37 show far more about the data than the top panel alone.

Another method for fighting overlap that works well if the number of overlapping symbols is small, is to move slightly the graph locations of certain points. This has been done in Figure 3.38; the data are from an experiment on the production of mutagens in drinking water [23]. Any symbol that touches another has had its actual location altered slightly. It is, of course, important to mention this movement if the graph is used to communicate quantitative information to others.

Sunflowers are a graphical method that can relieve exact and partial overlap [34]. They are illustrated with geological data in Figure 3.39 [25] and with data on graphical perception in Figure 3.40 [35]. A dot by itself means one point. A dot with line segments (petals) means more than one point; the number of petals indicates the number of points. The method is helpful when there is exact overlap or when many points are crowded into a small region. For the data in Figure 3.40 there is exact overlap; for Figure 3.39 the overlap is not exact, but points are very close to one another. When there are a large number of points on the graph there is a need for a sunflower algorithm: partition the data region into squares, count the number of points in each square, use sunflowers to show the counts, and position them in the centers of the squares.

The data in Figure 3.40 are from a perceptual experiment that will be discussed in detail in Section 3 of Chapter 4 [35]. Subjects judged the distances of four points — A, B, C, and D — from a line and recorded the percents that the B, C, and D distances were of the A distance. The true percents for B, C, and D were 52.5%, 47.5%, and 57.5% respectively. Figure 3.40 graphs the judged percents for D against the percents for B for 126 subjects. The graph was made to see if the judgments are correlated, an important issue whose answer affected the way the data were analyzed. The graph shows clearly that there is a large amount of

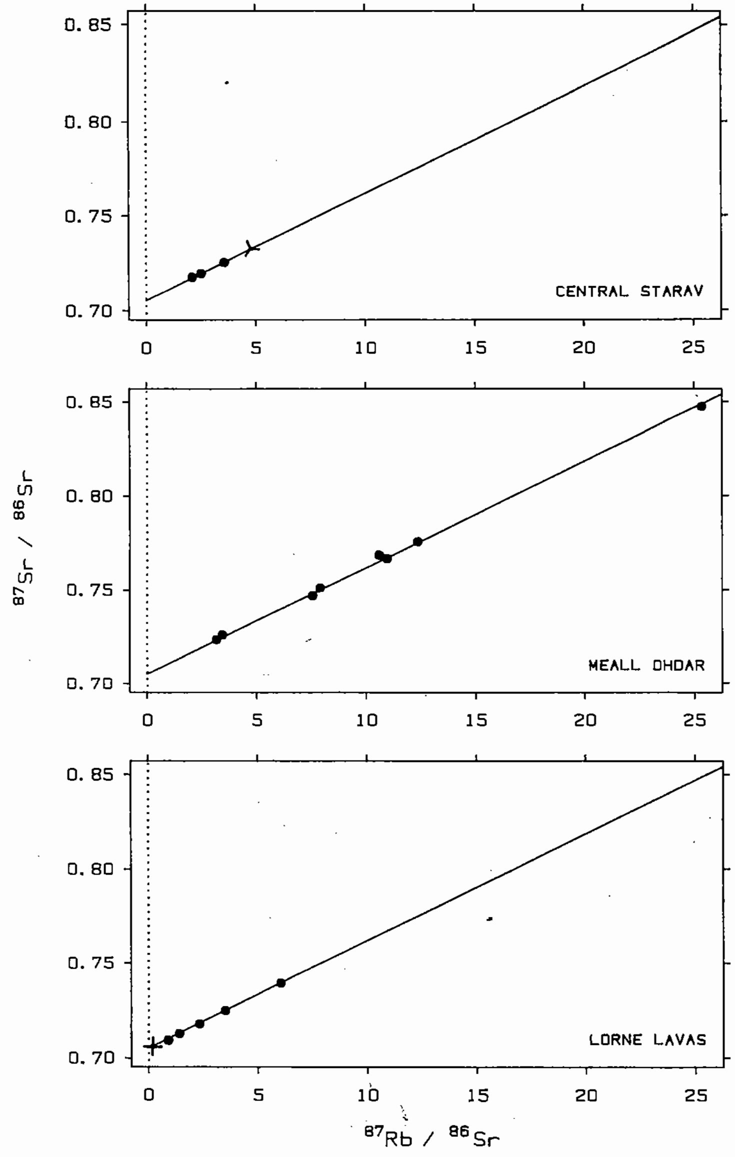

Figure 3.39 SUNFLOWERS. Each symbol with lines emanating from a dot is a sunflower. The number of petals (lines) is the number of data points at or near the center of the sunflower; sunflowers can be used to solve the overlap problem.

correlation. There is substantial overlap of the graph locations because answers tended to be multiples of 5. Figure 3.41 shows a scatterplot of the judgments with the overlap problem ignored, and only 51 points appear; not showing the multiplicity is misleading.

Another solution for exact overlap of graph locations is *jittering*: adding a small amount of random noise to the data before graphing [21]. This is illustrated in Figure 3.42 for the perception data. Jittering is a simpler remedy than sunflowers, but does not help, as sunflowers can, when resolution is degraded by a large number of partially overlapping symbols.

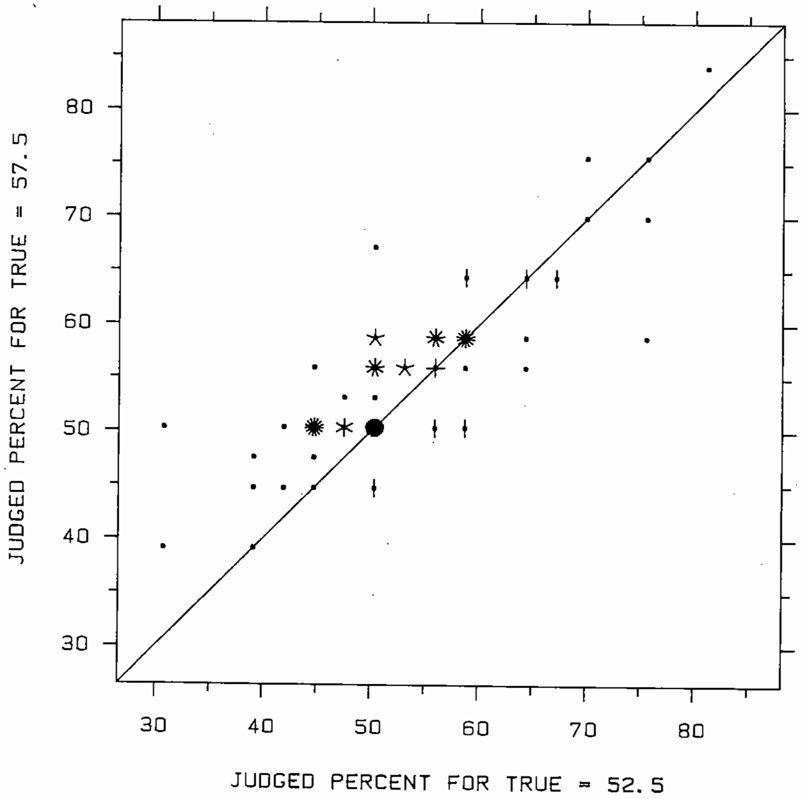

Figure 3.40 SUNFLOWERS. The sunflowers in this example alleviate exact overlap in the data.

If there is only partial overlap and no exact overlap, using an unfilled circle as the plotting symbol can improve the distinguishability of individual points [34]. This is illustrated in Figure 3.43. Circles can tolerate substantial partial overlap and still maintain their individuality. (Examples outside the graph domain are the symbol of the Olympics and the three-ring sign for Ballantine beer.) The reason is that distinct circles intersect in regions that are visually very different from circles. Squares, rectangles, and triangles do not share this property and degrade more rapidly.

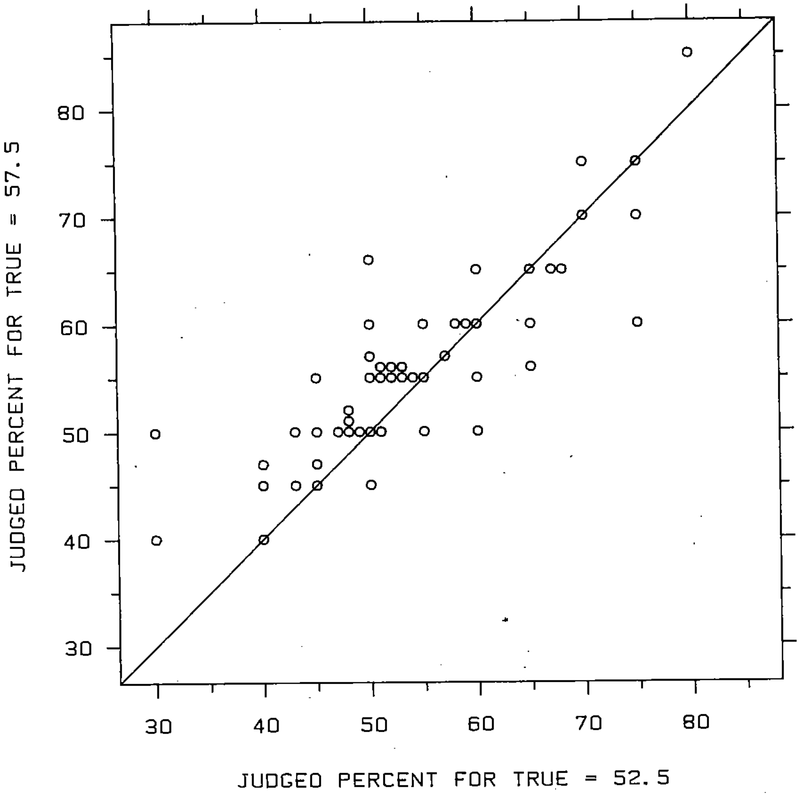

Figure 3.41 OVERLAP. This graph shows the result of graphing the data in Figure 3.40 and ignoring the overlap. Not indicating the multiplicity is misleading.

Box Graphs for Summarizing Distributions of Repeat Measurements of a Dependent Variable

Suppose the data consist of many repeat measurements of a dependent variable, *y*, for each of several different levels of an independent variable, *x*. One way to graph such data is illustrated in Figure 3.44. Each box graph portrays 25 values of the dependent variable for each of 11 distinct values of the independent variable; the center of the box graph is positioned horizontally at the value of the independent variable. In a sense we are back to the setting of

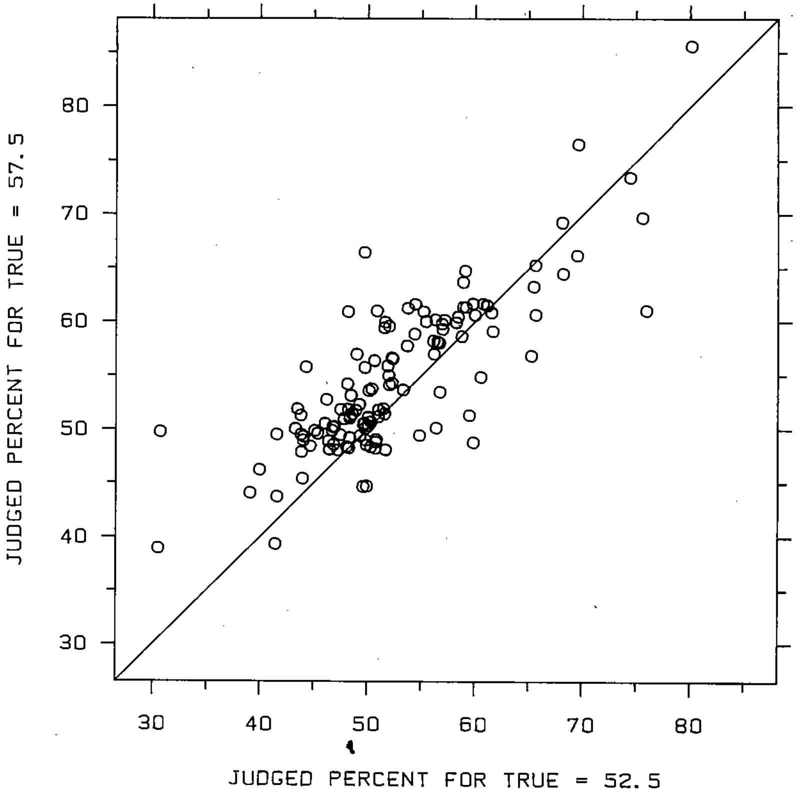

Figure 3.42 JITTERING. Another way to fight exact overlap is to add a small amount of random noise to the data. Now all of the data from Figure 3.41 can be seen.

Section 3.2 on graphing distributions, since the goal is to see how the distribution of the measurements of the dependent variable changes as the independent variable changes.

The data in Figure 3.44 are from an interesting experiment in bin packing [11]: k numbers, called weights, are randomly picked from the interval zero to u, where u is a positive number less than or equal to one; for the data in Figure 3.44, u was 0.8. The there are bins of size one and the object is to pack the weights into those bins; no overflowing is allowed, and we can use as many bins as necessary, but the goal is to use as few as possible. Unfortunately, to do this in an optimal manner is an NP-complete problem, which means that for anything but very small values of k the computation time is enormous. Fortunately, there are heuristic algorithms which, while not optimal, do an extremely good job of packing. Mathematicians and computer scientists had studied the worst-case behavior of bin packing [37] but there came a point where many appreciated that average behavior was an important issue as well; algorithms can be studied profitably by probing them with inputs, sometimes randomly generated, and using graphs and statistical methods to study the results [11].

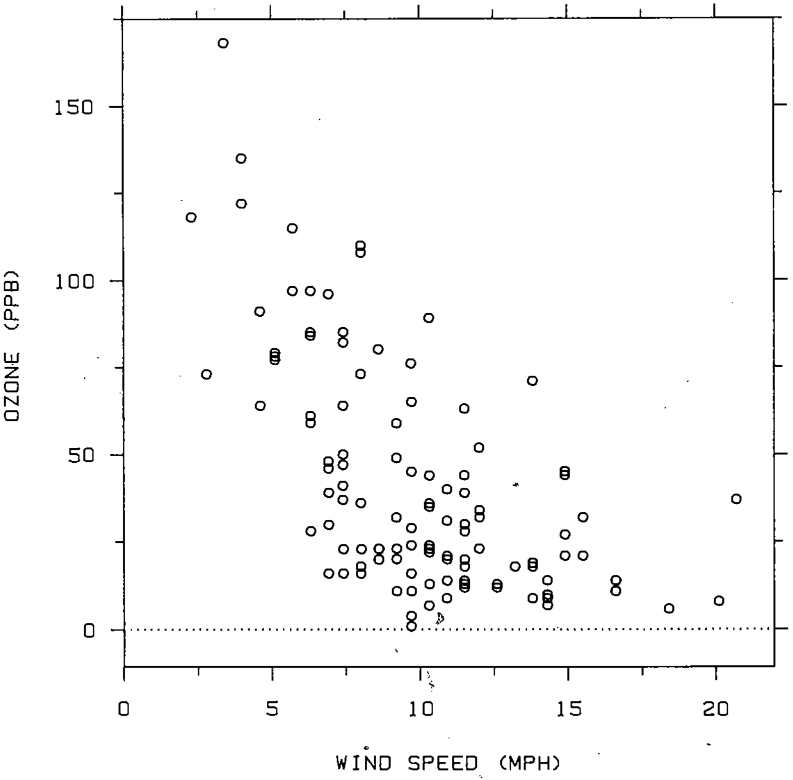

Figure 3.43 CIRCLES. Unfilled circles are good plotting symbols since they tend to maintain their individuality when there is partial overlap.

In Figure 3.44 the horizontal scale shows the number of weights, k, on a log base 10 scale. k varies from 125 to 128,000 by steps of a factor 2; that is, the first number is 125, the second is 250, and so forth up to $125 \times 2^{10} = 128{,}000$. There were 25 runs of the bin packing procedure for each value of k; for each run, k weights were chosen randomly from the interval 0 to 0.8 and a packing carried out. The

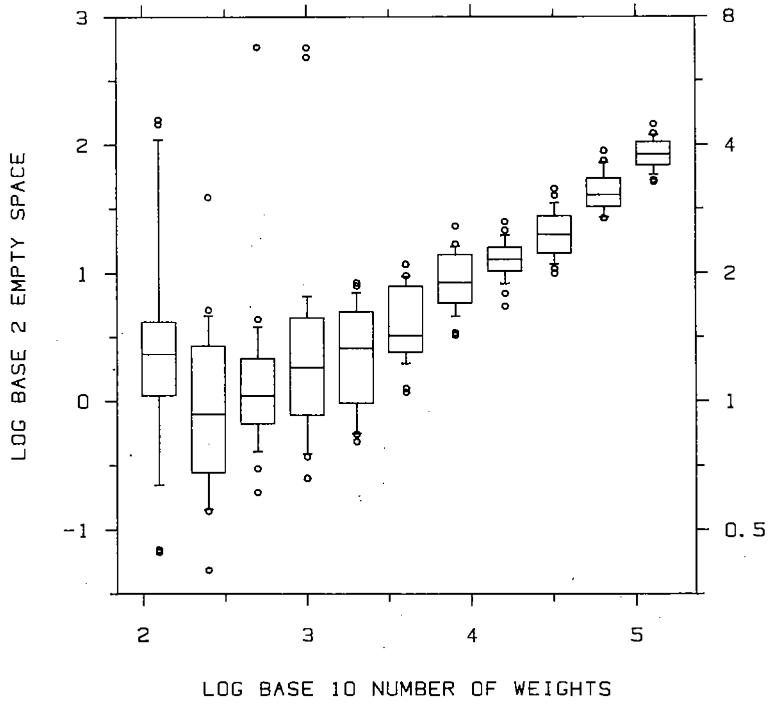

Figure 3.44 BOX GRAPHS FOR REPEAT MEASUREMENTS OF A DEPENDENT VARIABLE. The purpose of the graph is to see how the dependent variable, the variable on the vertical axis, depends on the independent variable, the variable on the horizontal axis. For each value of the independent variable there are 25 measurements of the dependent variable; the distribution of these 25 values is summarized by a box graph.

algorithm used to do the packing was *first fit decreasing*: The weights are ordered from largest to smallest and are packed in that order. For each weight the first bin is tried; if it has room, the weight is inserted and if not the second bin is tried; if the second bin has room, the weight is inserted and if not, the third bin is tried; the algorithm proceeds in this way until a bin with room, possibly a completely empty one, is found. The vertical scale in Figure 3.44 is the logarithm base 2 of the amount of empty space in the bins that have at least one weight. Since the bin size is one, this amount of empty space is equal to the number of bins used minus the sum of the weights. In this example, empty space is the dependent variable and number of weights is the independent variable.

What does Figure 3.44 show us about bin packing? One thing is that the first-fit-decreasing algorithm is very efficient. The amount of empty space is never greater than 8 in these runs. For runs of size 128,000 the performance is superlative; the median empty space is about 4 even though the sum of the weights in this case averages 128,000 × 0.8/2 = 51,200. The figure also shows that median log empty space grows nonlinearly with log number of weights, although the pattern becomes linear for large numbers of weights. This latter result is predicted by a theorem about the asymptotic behavior of empty space [12]. Figure 3.44 also shows that for the smaller numbers of weights there are outliers: values that are large compared to the majority of the values.

Strip Summaries Using Box Graphs

Box graphs can be used even when there are no repeat measurements of the dependent variable by grouping the data according to the values of the independent variable. This grouping is illustrated in Figure 3.45. The data have been divided into five groups by vertical strips with as nearly an equal number of observations in each strip as possible. In Figure 3.46 box graphs summarize the distributions of the y values for the five strips. Each box graph is centered, horizontally, at the median of the x values for its strip.

The data in this example, which were also graphed in Section 2 of Chapter 2, are from an experiment on 144 hamsters in which their lifetimes and the fractions of their lifetimes they spent hibernating were measured [89]. The objective of the experiment was to see how lifetime depends on hibernation. Figure 3.46 shows that as fraction of lifetime spent hibernating increases, the distribution of lifetime increases.

Smoothing: Lowess

One hypothesis suggested by Figure 3.46 is that hamster DNA parcels out a fixed amount of nonhibernation hours; a hamster gets only so much awake time, and if it hibernates longer, it lives longer by the same amount, but otherwise there is no effect on lifetime. Suppose ℓ = lifetime and p = fraction of lifetime spent hibernating. If this hypothesis is true then $(1-p)\ell$, the amount of time spent not hibernating does not depend on $p\ell$, the amount of time spent hibernating.

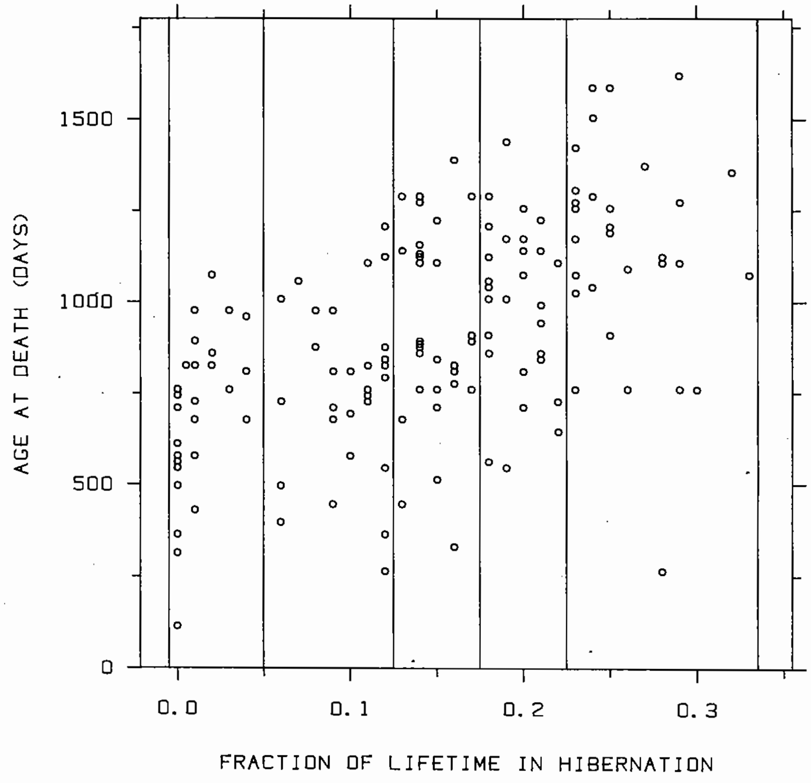

Figure 3.45 DEPENDENT-INDEPENDENT VARIABLE DATA. Age at death is graphed against fraction of lifetime spent hibernating for 144 hamsters. The data have been divided into five strips with nearly equal numbers of points, in preparation for the graph in Figure 3.46.

Figure 3.47 is a graph of time spent not hibernating against time spent hibernating. It shows that, overall, the hypothesis is false; increased hibernation time results in increased nonhibernation time. But how would we describe the dependence? Is there a linear or nonlinear dependence? With a graph of just the (x_i, y_i) values it is hard to answer these questions.

We could study the dependence by strip summaries with box graphs, but Figure 3.48 shows another method: a smooth curve put through the points. For each point, (x_i, y_i), on the graph there is a

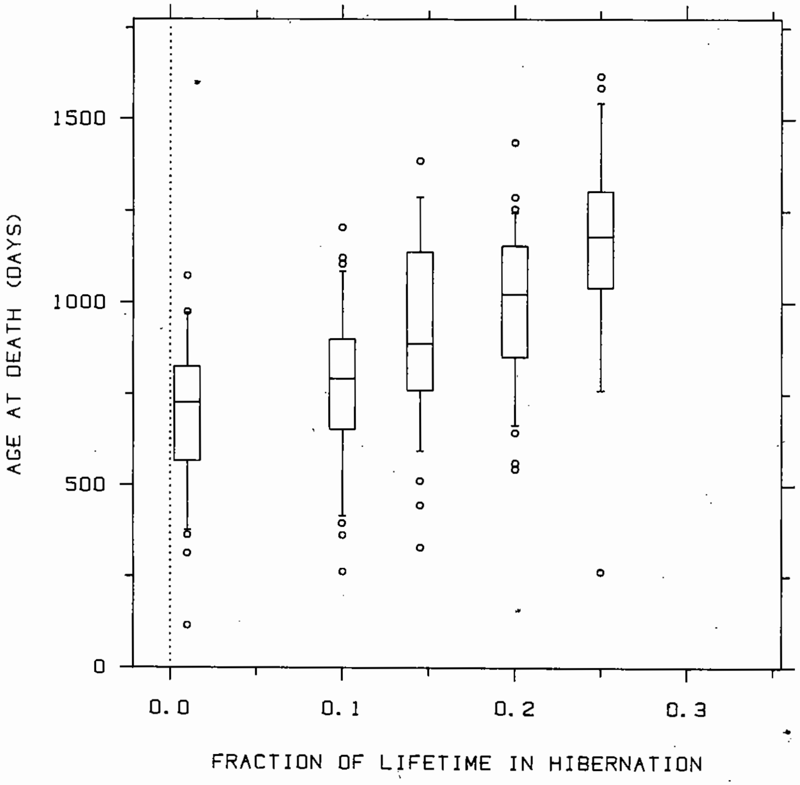

Figure 3.46 STRIP SUMMARIES USING BOX GRAPHS. The distribution of the y values of the points in each of the five vertical strips of Figure 3.45 is shown by a box graph. Each box graph is centered, along the horizontal scale, at the median of the x values of the points in the strip.

smoothed value, $(x_i, \hat{y}_i)$. $\hat{y}_i$ is the *fitted value* at x_i. The curve is graphed by connecting successive smoothed values, moving from left to right, by lines. The purpose of the curve is to summarize the *middle* of the distribution of y for each value of x. Thus the curve is performing the same task as the medians of the box graphs in strip summaries; if we took a narrow vertical strip, the curve should describe the middle of the distribution of the y values in the strip. Statistical scientists call this a *regression curve*, a misnomer since there is nothing regressive about it at all. The method used to compute the smoothed values will be discussed later.

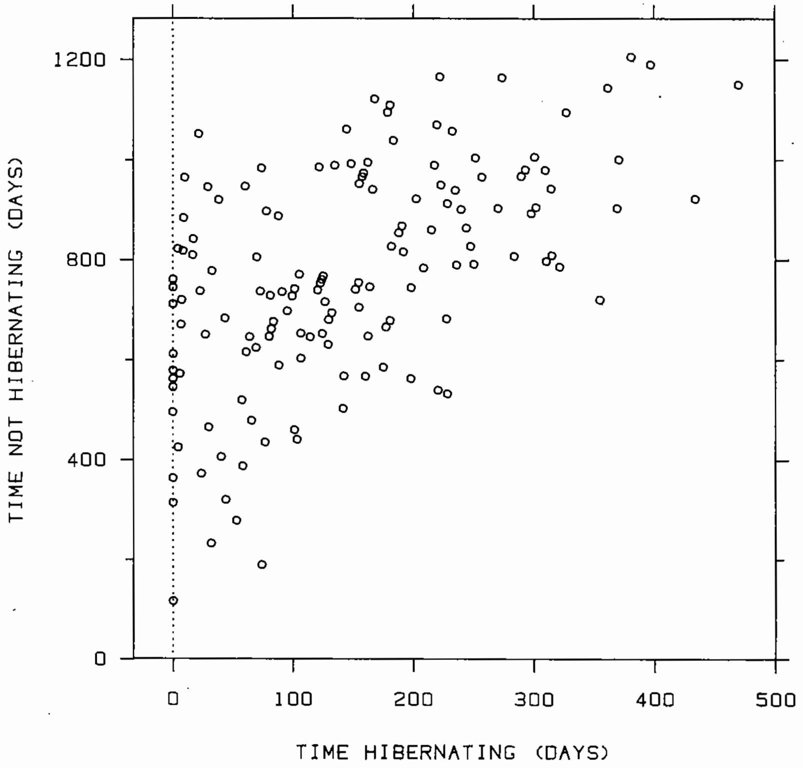

Figure 3.47 DEPENDENT-INDEPENDENT VARIABLE DATA. The total time spent not hibernating is graphed against the time spent hibernating for the 144 hamsters. There appears to be a dependence of y on x but it is difficult to assess the nature of the dependence from the graph.

The smooth curve shows that there is some truth to the hypothesis stated earlier. While there is, overall, an increase in nonhibernation lifetime as hibernation increases, the response is in fact constant until the amount of hibernation is above 100 days. From 100 days and above, the effect is nearly linear and the slope is about 1, so each minute spent hibernating beyond 100 days produces on the average about one extra minute of nonhibernation lifetime. We have been assuming that there is a causal mechanism, but this is reasonable in view of current biological information [89].

The curve in Figure 3.48 was produced by a data smoothing procedure called *robust locally weighted regression* [26]; the name of the procedure is often shortened to *lowess* (locally-weighted scatterplot smoother). The user must choose a smoothness parameter f, which is a number between 0 and 1. As f increases, the smooth curve becomes smoother. In Figure 3.48 the value of f is 0.5 and in Figure 3.49 it is 0.25. Lowess is very computing intensive, but there is a fast, efficient computer program that carries it out [110].

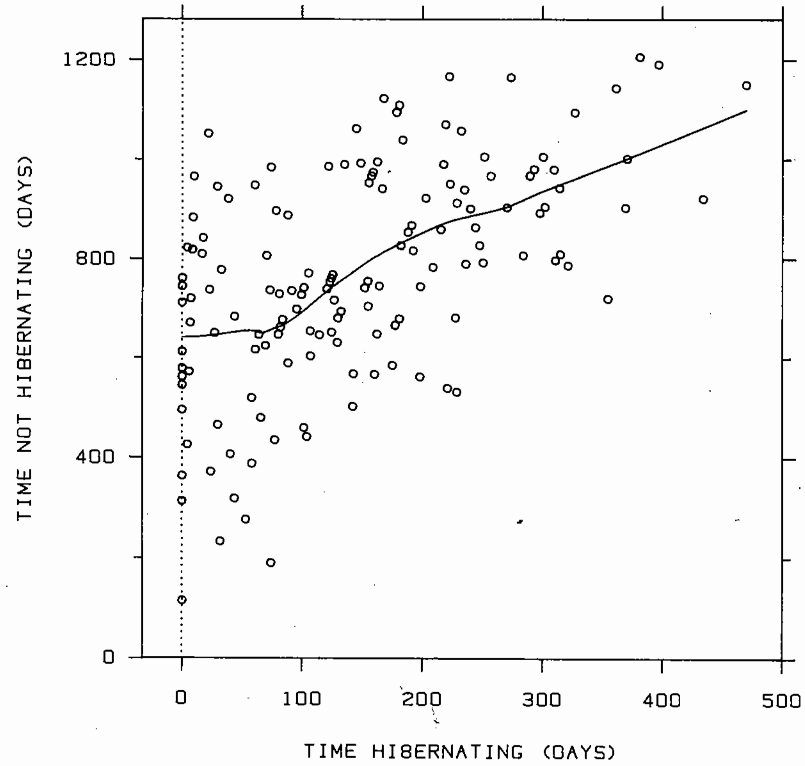

Figure 3.48 LOWESS. The smooth curve, which was computed by a procedure called lowess, summarizes how y depends on x. For each point, (x_i, y_i), on the graph, lowess produces a smoothed value, $(x_i, \hat{y}_i)$. The curve is graphed by connecting successive smoothed values, moving from left to right, by straight lines.

Choosing f requires some judgment for each application. In most applications an f that works well is usually between 0.5 and 0.8. The goal is to try to choose f to be as large as possible to get as much smoothness as possible without distorting the underlying pattern in the data.

Residuals, useful in so many situations, can help in choosing f. This will be illustrated with an example. Figure 3.50 is a graph of the air pollutant ozone against wind speed for 111 days in New York City

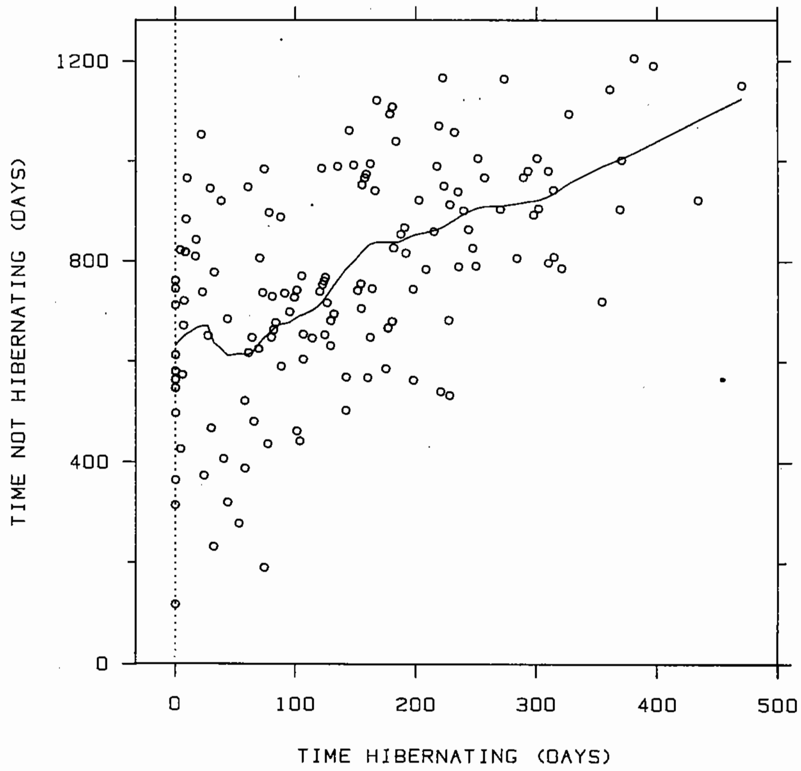

Figure 3.49 LOWESS. The smoothness of the lowess curve depends on a smoothness parameter, f, which varies between 0 and 1. As f increases the curve becomes smoother. In Figure 3.48, $f = 0.5$ and in this figure, $f = 0.25$.

from May 1 to September 30 of 1973. From this graph we can see that the general pattern is for ozone to decrease as wind speed increases because of the increased ventilation of air pollution that higher wind speeds bring. However, it is difficult to see more precise aspects of the pattern, for example, whether there is a linear or nonlinear decrease.

The top panel of Figure 3.51, which has a lowess curve with $f = 0.8$, suggests the decrease is nonlinear. How do we know the lowess curve is not distorting the pattern? Since we cannot discern easily the pattern when a lowess curve is absent we cannot expect to assess easily how well lowess is doing. The solution is to graph $y_i - \hat{y}_i$ against x_i, add a lowess smoothing to this graph of residuals, and see if there is an effect. This is illustrated in the bottom panel of Figure 3.51.

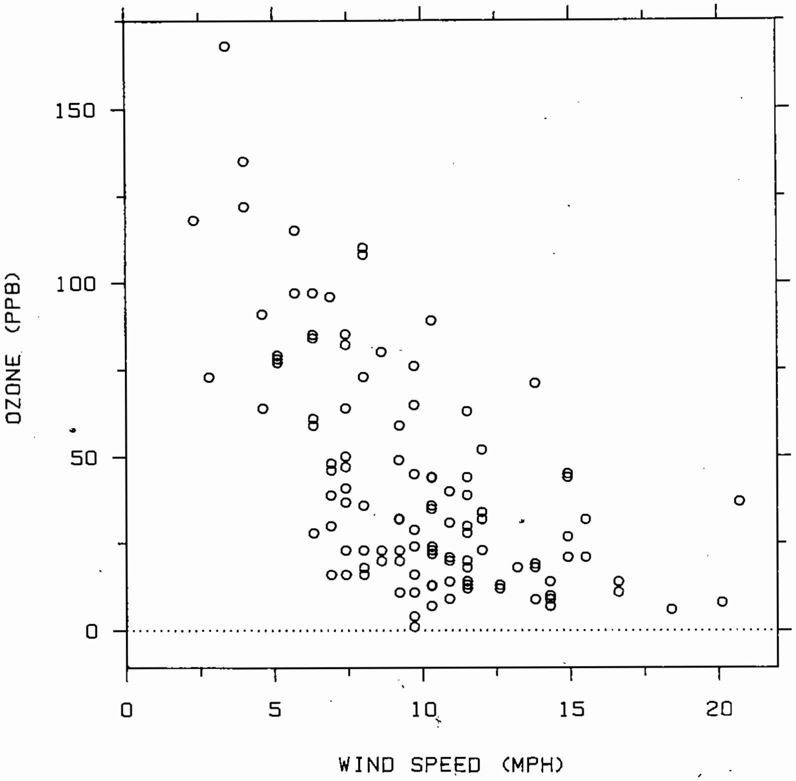

Figure 3.50 DEPENDENT-INDEPENDENT VARIABLE DATA. The data are daily measurements of ozone and wind speed for 111 days. It is difficult to see the nature of the dependence of ozone on wind speed.

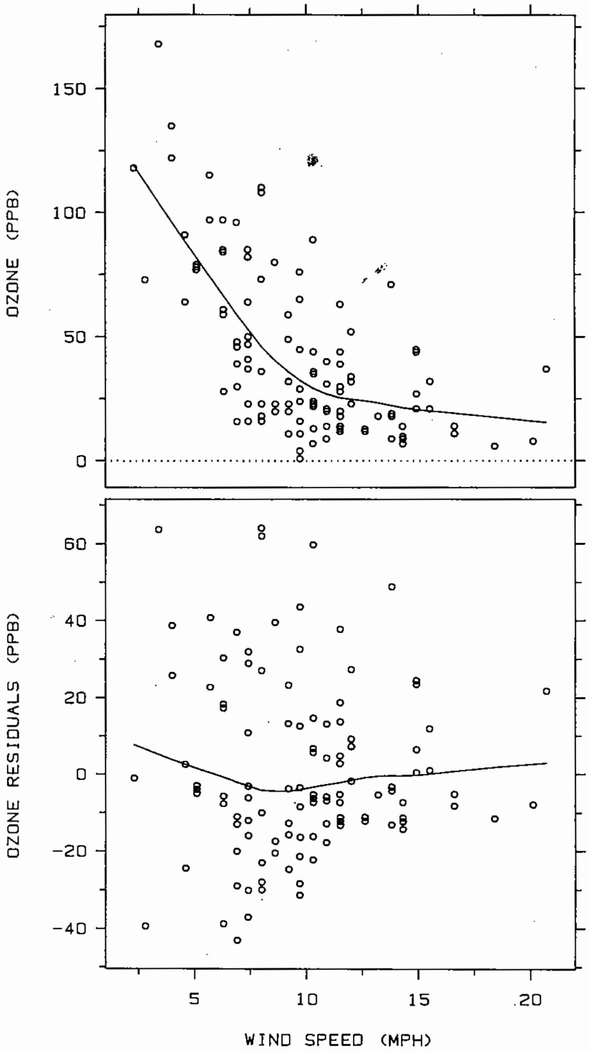

Figure 3.51 CHECKING LOWESS. On the top panel the graph from Figure 3.50 now has a lowess curve with $f = 0.8$. It is difficult to assess visually whether lowess is correctly depicting the dependence. On the bottom panel the residuals, $y_i - \hat{y}_i$, are graphed against x_i, and a lowess curve is superposed; the curve suggests there is a small dependence of the residuals on x_i, which means f is too large in the smoothing of the top panel.

The lowess curve suggests that there is some dependence of the residuals on x_i. This should not happen; the curve should be nearly a horizontal line since the residuals should be variation in y_i not explainable by x_i. The problem is that the lowess smoothing in the top panel has missed part of the pattern because f is too large, and this missed part has gone into the residuals.

In Figure 3.52, f has been reduced to 0.5. The curve on the graph of the residuals is now reasonably close to a horizontal line, so the amount of smoothing for the curve in the top panel is not too great.

This method of graphing and smoothing residuals is a one-sided test: it can show us when f is too large but sets off no alarm when f is too small. One way to keep f from being too small is to increase it to the point where the residual graph just begins to show a pattern, and then use a slightly smaller value of f.

Lowess is quite detailed and mathematical, and a full discussion of how it works would sidetrack us too much. In the remainder of this section a brief description will be given; the details can be found in the source [26] or in [21]. Suppose the x_i are ordered from smallest to largest so that x_1 is the smallest and x_n is the largest. For each pair of values, (x_i,y_i), lowess produces a fitted value, $\hat{y}_i$. Figure 3.53 shows how the fitted value is computed at one x_i. Look at the upper left panel. The data, which are made up, are shown by the unfilled circles; the value of x_i at which the fitted value is to be computed is x_6, which is marked by the vertical dotted line. The value of f is 0.5 in this example; it is multiplied by 20, the number of observations, which gives the number 10. We now pick from among the x_i the 10th closest x_i to x_6, which is x_1. (x_6 itself is included in this count.) A vertical strip, depicted by the solid vertical lines, is defined by putting the left boundary of the strip at x_1 and the right boundary on the other side of x_6 at the same distance from x_6 as x_1.

Look at the lower left panel. A weight function, $w(x)$, is defined. The points, (x_i, y_i) for $i = 1$ to n, are assigned weights $w(x_i)$. Notice that (x_6,y_6) has the largest weight; moving away from x_6 the weight function decreases and becomes zero at the boundaries of the strip.

Look at the upper right panel. A line is fitted to the points of the graph using *weighted* least squares with weight $w(x_i)$ at (x_i, y_i). This means that (x_6,y_6) plays the largest role in determining the line and the role played by other points decreases as their x values increase in distance from x_6. Points on and outside the strip boundary play no role at all. The fitted value, $\hat{y}_6$, is the y-value of the line at $x = x_6$. The point $(x_6,\hat{y}_6)$ is depicted by the filled circle.

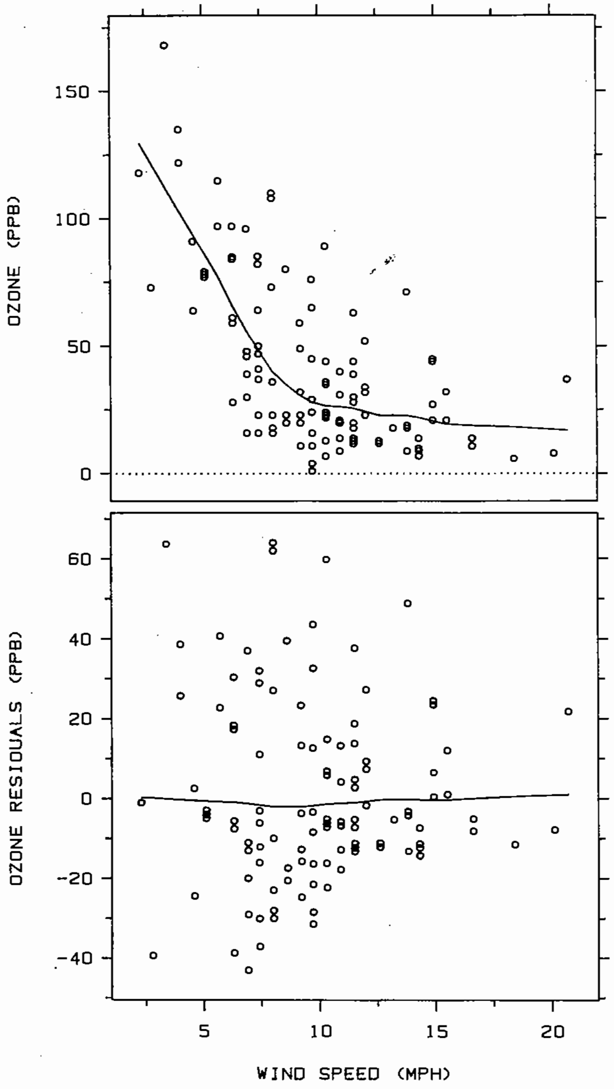

Figure 3.52 CHECKING LOWESS. On the top panel the value of f for lowess has been reduced to 0.5 since Figure 3.51 suggests $f = 0.8$ is too large. The bottom panel shows no dependence of the residuals on x_i, which suggests the lowess curve with $f = 0.5$ is not distorting the pattern of the dependence of ozone on wind speed.

Look at the lower right panel. The result of the previous operations is the one lowess smoothed value, $(x_6, \hat{y}_6)$, shown by the filled circle. The same operations are carried out for each point, (x_i, y_i), on the graph.

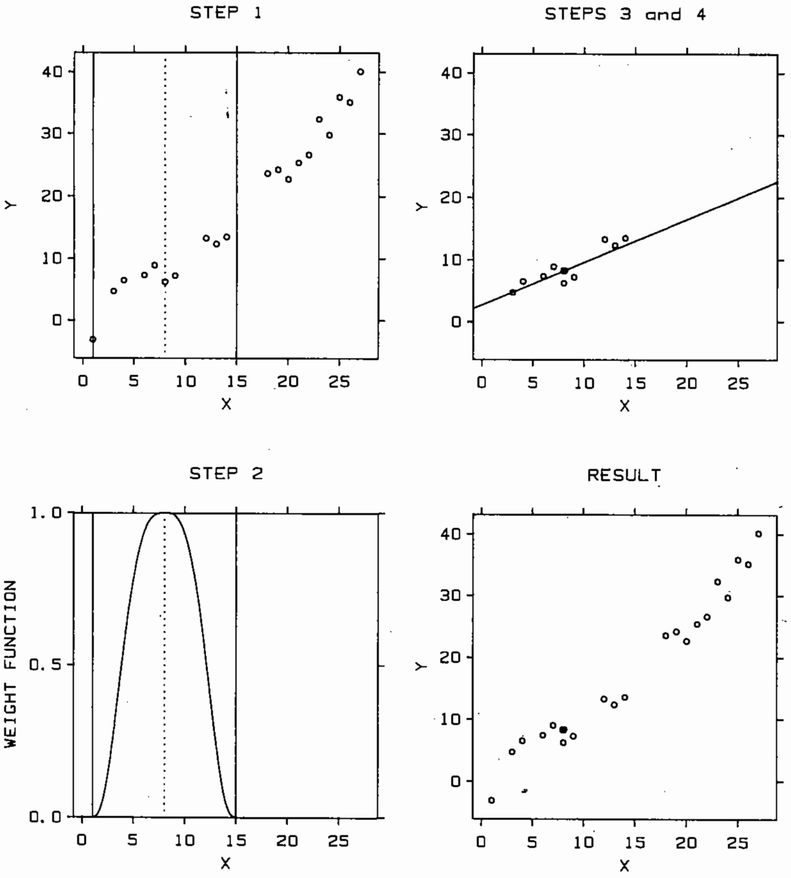

Figure 3.53 HOW LOWESS WORKS. The graphs show how the fitted value at x_6 is computed. (Top Left) f, which is 0.5, is multiplied by 20, the number of points, which gives 10. A vertical strip is defined around x_6 so that the boundary is at the 10th nearest neighbor. (Bottom Left) Weights are defined for the points using the weight function. (Top right) A line is fitted using weighted least squares. The value of the line at x_6 is the lowess fitted value, $\hat{y}_6$. (Bottom right) The result is one value of lowess, shown by the filled circle. The computation is repeated for each point on the graph.

Figure 3.54 shows the sequence of operations for the rightmost point, (x_{20}, y_{20}). The right boundary of the strip does not appear in the two left panels because it is beyond the right extreme of the horizontal scale line.

There is another piece to the lowess algorithm. What has been described is the locally weighted regression part of robust locally weighted regression. There is also a *robustness* part. Suppose the data contain one or more *outliers*; an outlier is a point, (x_i, y_i), with an

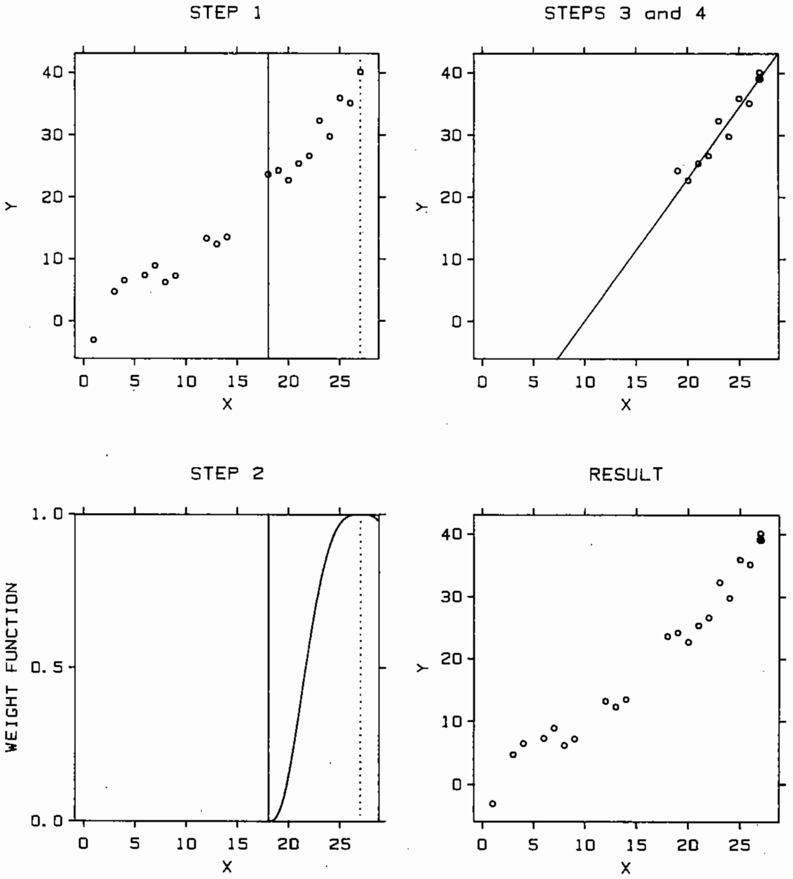

Figure 3.54 HOW LOWESS WORKS. The computation of the lowess fitted value at x_{20} is illustrated.

unusually large or small value of y_i compared with other points in a vertical strip around x_i. The upper panel of Figure 3.55 shows an example. The unfilled circles are the data and one point, (x_{11}, y_{11}), has a y value that is much larger than the y values of points whose x values are close to x_{11}. Carrying out lowess as described above yields the filled circles; the outlier has distorted the fitted values in the neighborhood of x_{11} so that the general pattern of the data is no longer described.

Lowess has a robustness feature in which, after a first smoothing as described above, outliers are identified and downweighted in a second smoothing. This identification, downweighting, and resmoothing can be done any number of times, although two times is almost always sufficient. The result of the full lowess algorithm, including the robustness part, is shown in the lower panel of Figure 3.55. Now the smoothed values describe the behavior of the majority of the data.

Time Series: Connected, Symbol, Connected Symbol, and Vertical Line Graphs

A *time series* is a set of measurements of a variable through time. Figure 3.56 shows an example. The data are yearly values, from 1868 to 1967, of the aa index [96], which measures the magnitude of fluctuations in the earth's magnetic field. The index is the average of measurements of geomagnetic fluctuations at observatories in Australia and England that are roughly antipodal: at opposite ends of an earth diameter. Figure 3.56 shows there has been an increase in the overall level of the aa index from 1900 to 1967. The solar wind causes fluctuations in the aa index from 1900 to 1967. The solar wind causes fluctuations in the earth's magnetic field, so the increase in the index suggests that the solar wind has increased during this century [49]. Figure 3.56 also shows the aa index has a cycle of about 11 years. This is the same as the sunspot cycle; increased sunspots are associated with increased solar activity and therefore an increased solar wind, but interestingly, the sunspots do not show an increase in their overall level, as the aa index does.

A time series is a special case of the broader dependent-independent variable category. Time is the independent variable. One important property of most time series is that for each time point of the data there is only a single value of the dependent variable; there are no repeat measurements. Furthermore, most time series are measured at equally-spaced or nearly equally-spaced points in time. These special properties invite special graphical methods which, as will be illustrated at the end of this section, are relevant for any situation with a single-valued dependent variable and an equally-spaced independent variable.

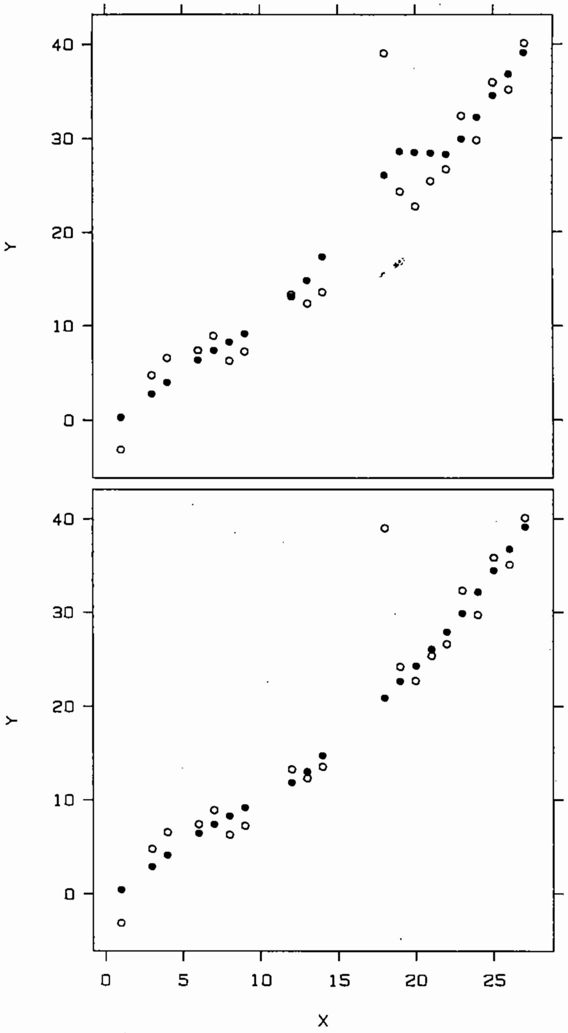

Figure 3.55 HOW LOWESS WORKS. Lowess has a robustness feature that prevents outliers from distorting the smoothed values. (Top panel) The open circles are the points of the graph; there is one outlier between $x = 15$ and $x = 20$. The smoothed values for lowess without the robustness feature, which are shown by the filled circles, have been distorted in the neighborhood of the outlier. (Bottom panel) The filled circles are from lowess with the robustness feature; now the smoothed values follow the general pattern of the data.

There are many ways to graph a time series. Figure 3.56 is a *connected symbol graph* since symbols together with lines connecting successive points in time are used. Figure 3.57 is a *symbol graph* because just the symbols are used, and Figure 3.58 is a *connected graph* because just the lines are used. Figure 3.59 is called a *vertical line graph* for the obvious reason.

Each of these four methods of graphing a time series has its data sets for which it provides the best portrayal. For the aa data the best one is the connected symbol graph. The symbol graph does not give a clear portrayal of the cyclic behavior, because we cannot perceive the order of the series over short time periods of several years, which makes seeing the 11-year cycle difficult. In the words of spectrum analysis, we cannot appreciate the high and middle frequency behavior of the series on the symbol graph.

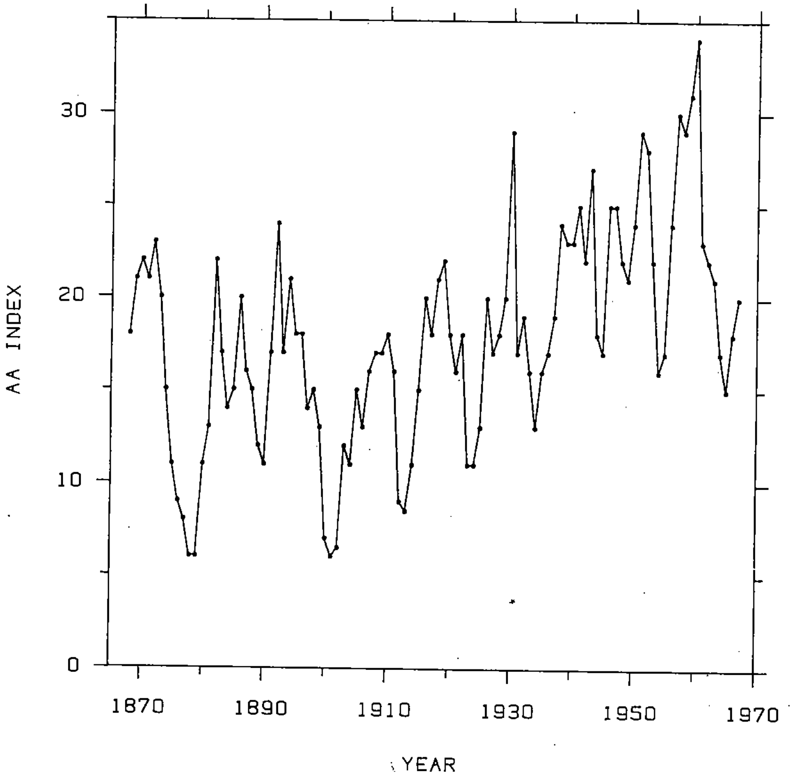

Figure 3.56 CONNECTED SYMBOL GRAPH. The time series shown on the graph is the yearly average of the aa index: measurements of the magnitudes of fluctuations in the earth's magnetic field. A connected symbol graph, which allows us to see the individual data points and the ordering through time, reveals an 11-year cycle and a trend from 1900 to 1967.

On the connected graph in Figure 3.58 the individual data points are not unambiguously portrayed. For example, it is clear that there is an unusual peak in the observations around 1930, but it is hard to decide if the peak is a single outlier for one year or is supported by a rise and fall of a few values. On the connected symbol graph, and the other graphs, it is clear that the peak consists of one value.

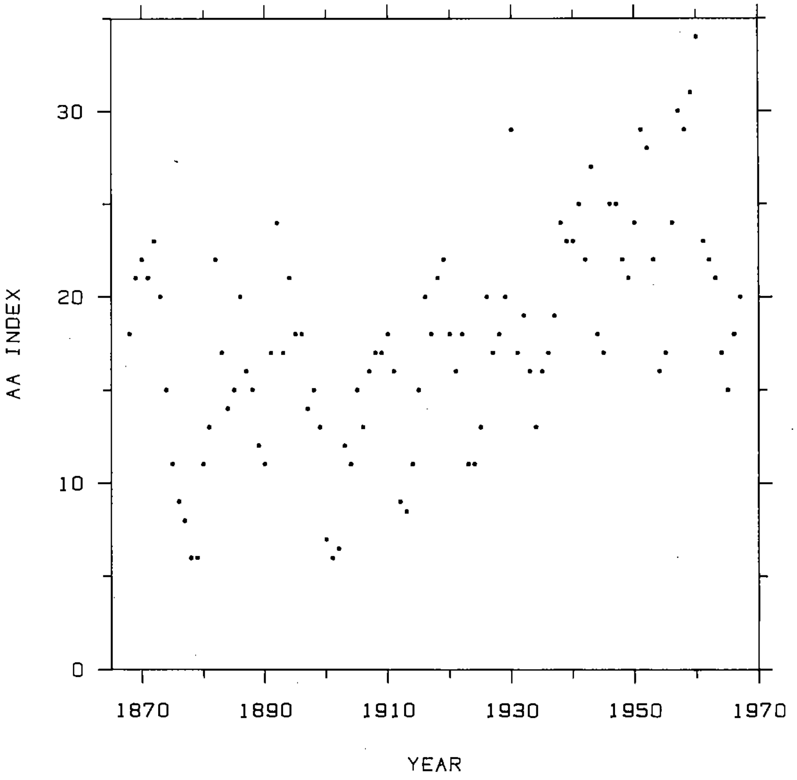

Figure 3.57 SYMBOL GRAPH. A symbol graph of the aa data does not reveal the 11-year cycle.

On the vertical line graph in Figure 3.59 there is an unfortunate asymmetry: The peaks of the 11-year cycle stand out more clearly than the troughs. There is also a disconcerting visual phenomenon: Our visual system cannot simultaneously perceive the peaks and the troughs. This is what psychophysicists call a figure-ground effect [55, pp. 10-11]; for example, there is a famous black and white drawing where if you focus on the black, you see profiles of two faces looking at one another and if you focus on the white, you see a vase, but both cannot be simultaneously perceived [55, p. 11].

There is, however, a place for vertical line, connected, and symbol graphs. A symbol graph of a time series is appropriate if what we want to convey is the long-term trend, that is, the low frequency behavior. In such a case it is not necessary to perceive the exact time order over short time intervals. Figure 3.60 is an example. The data are the daily ozone measurements we have seen before. One very low ozone value, an outlier on the log scale, has been omitted in Figure 3.60. In this example the day-to-day movement of ozone is less interesting than the trend, so the symbol graph is used. A lowess curve with $f = 0.5$ is superposed to help us see the trend.

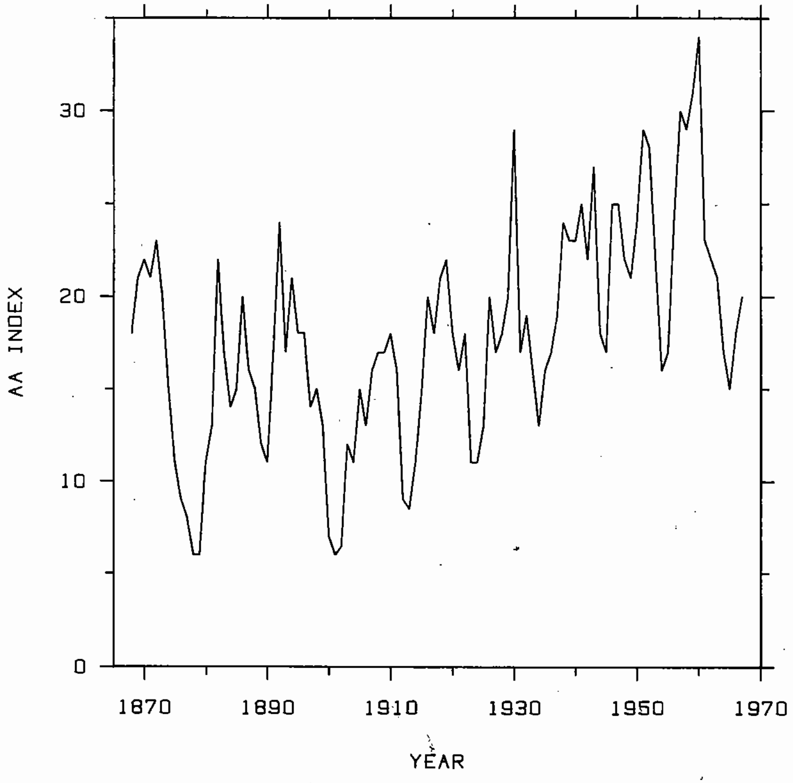

Figure 3.58 CONNECTED GRAPH. The connected graph does not reveal the positions of the aa measurements. It is not possible to determine if the peak around 1930 consists of one or many values.

A connected graph is appropriate when the time series is smooth, so that perceiving individual values is not important. A vertical line graph is appropriate when it is important to see individual values, when we need to see short-term fluctuations, and when the time series has a large

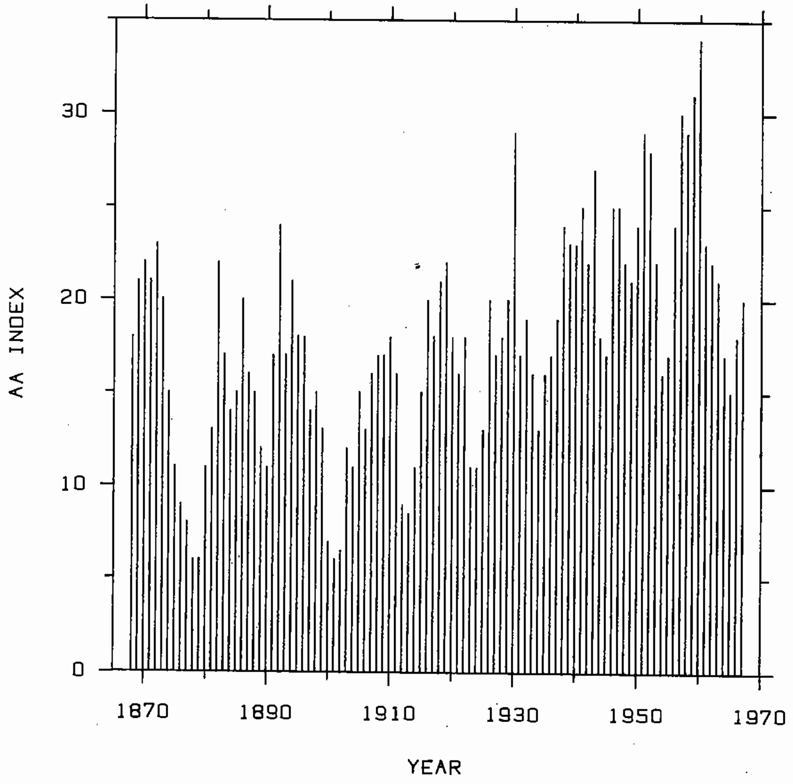

Figure 3.59 VERTICAL LINE GRAPH. On this graph the peaks stand out more clearly than the troughs.

number of values; the use of vertical lines allows us to pack the series tightly along the horizontal axis. The vertical line graph, however, usually works best when the vertical lines emanate from a horizontal line through the center of the data and when there are no long-term trends in the data.

Figure 3.61 is the graph of CO_2 and its components that was discussed in detail in Section 2 of Chapter 1. A connected graph is used for the two top panels because the data are smooth and seeing

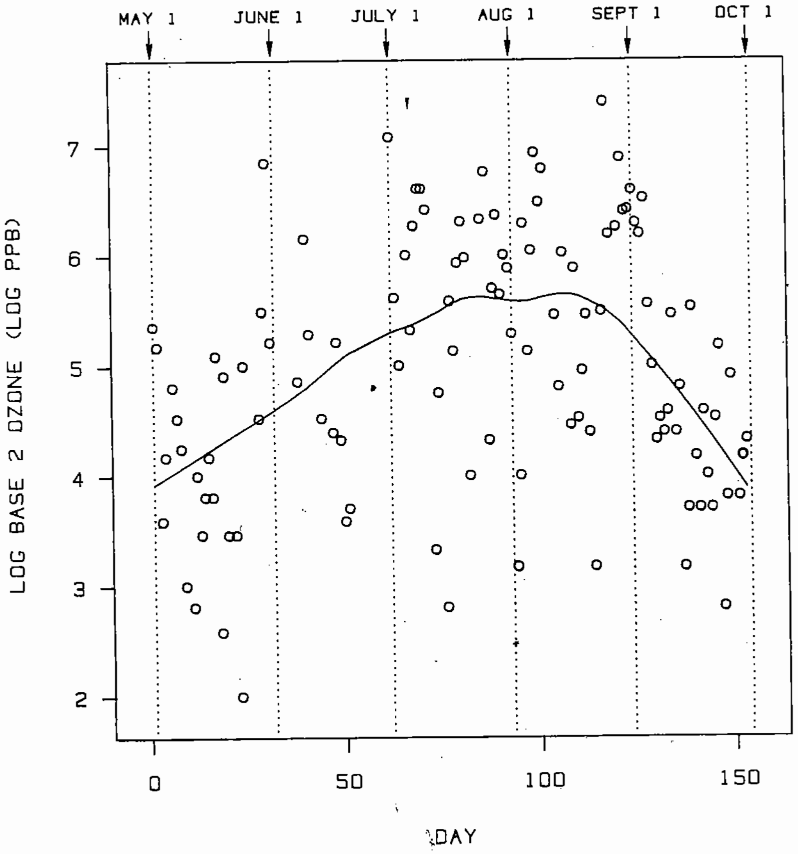

Figure 3.60 SYMBOL GRAPH. A symbol graph is appropriate for a time series when the goal is to show the long-term trend in the series, but not high frequency behavior. On this symbol graph a lowess curve is superposed to help assess the trend.

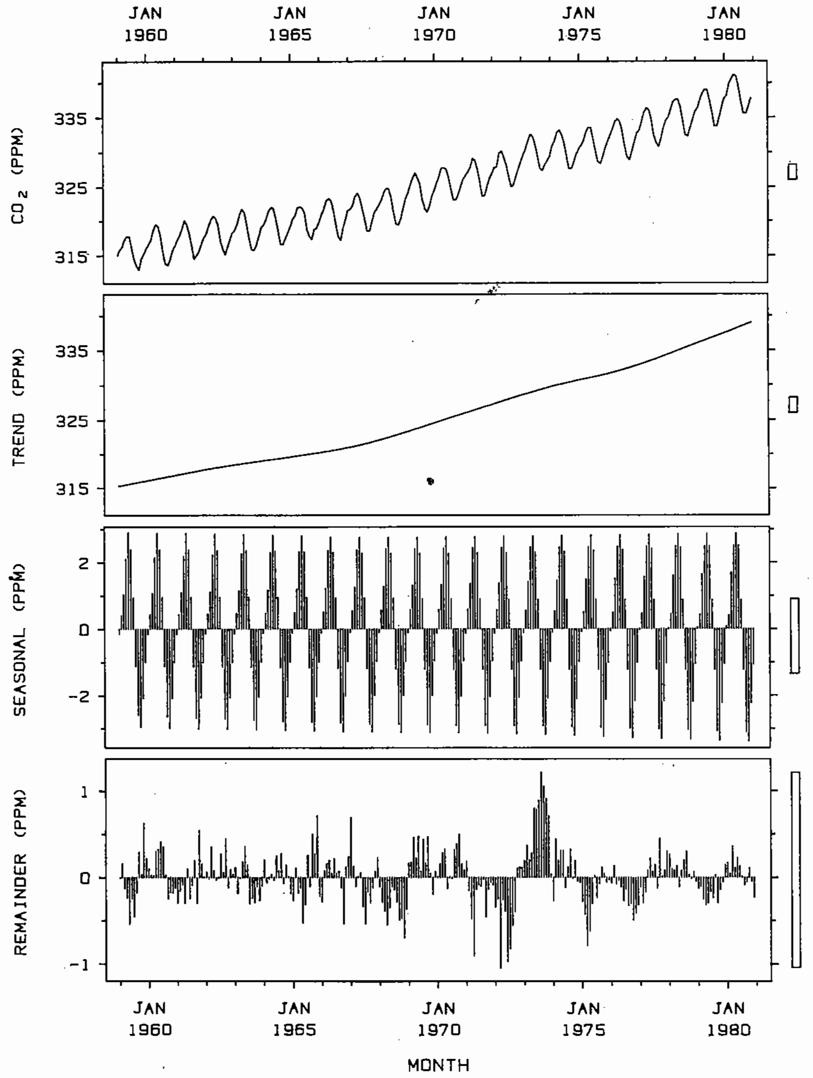

Figure 3.61 CONNECTED AND VERTICAL LINE GRAPHS. The graph shows the monthly average CO_2 concentrations from Mauna Loa and the three components. Connected graphs are used in the top two panels since it is not important to see individual values. Vertical line graphs are used in the bottom two panels since it is important to see individual values and to assess behavior over short periods of time and since each series has many values.

individual values was not judged important. A vertical line graph, emanating from zero, is used for the two bottom panels because it is important in this application to see the individual monthly values and to assess behavior over short time periods and because the time series is long.

Time Series: Seasonal Subseries Graphs

Figure 3.62 shows a *seasonal subseries graph,* a graphical method that was invented in 1980 to study the behavior of a seasonal time series or the seasonal component of a seasonal time series [36]. The data in Figure 3.62 are the seasonal component of the CO_2 series in Figure 3.61. In this example it is important to study how the individual monthly subseries are changing through time; for example, we want to analyze the behavior of the January values through time. We cannot make a graphical assessment from Figure 3.61 since it is not possible to focus on the values for a particular month; the graphical method in Figure 3.62 makes it possible.

In the seasonal subseries graph, the January values of the seasonal component are graphed for successive years, then the February values are graphed, and so forth. For each monthly subseries the mean of the values is portrayed by a horizontal line. The individual values of the subseries are portrayed by the vertical lines emanating from the horizontal line. In Figure 3.62 the January subseries is the first group of values on the left, the February subseries is the next group of values, and so forth. The graph allows an assessment of the overall pattern of the seasonal, as portrayed by the horizontal mean lines, and also of the behavior of each monthly subseries. Since all of the monthly subseries are on the same graph we can readily see whether the change in any subseries is large or small compared with the overall pattern of the seasonal component.

Figure 3.62 shows interesting features. The first is the overall seasonal pattern, with a May maximum and an October minimum. This pattern has long been recognized and is due to the earth's vegetation (See the discussion in Section 2 of Chapter 1.) The second feature is the patterns in the individual monthly subseries. Subseries near the yearly maximum tend to be increasing; the biggest year-to-year increases occur during the months March and April. Subseries near the yearly minimum tend to be decreasing; the biggest year-to-year decreases occur during the months September and October. The net effect, of course, is that the seasonal oscillations are increasing.

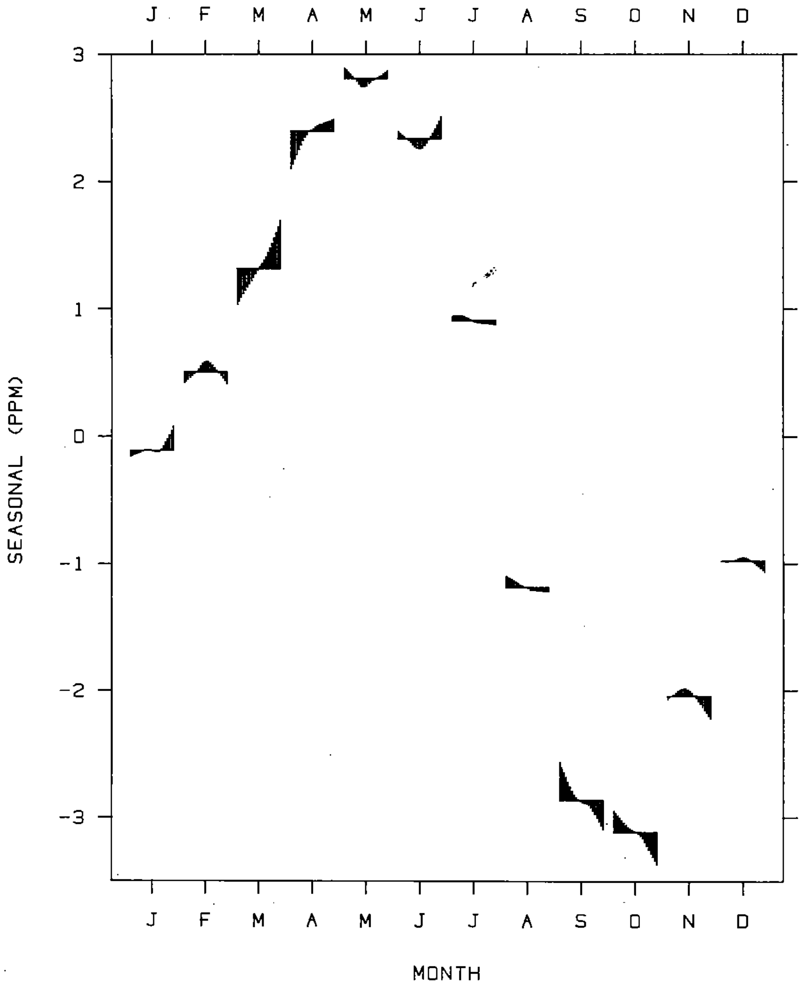

Figure 3.62 SEASONAL SUBSERIES GRAPH. The seasonal component from Figure 3.61 is graphed. First the January values are graphed for successive years, then the February values, and so forth. For each monthly subseries the mean of the values is portrayed by the horizontal line. The values of each monthly subseries are portrayed by the ends of the vertical lines. Now we can see the average seasonal change and the behavior of the individual monthly subseries. Monthly subseries near the yearly maximum tend to be increasing and those near the minimum tend to be decreasing.

An Equally-Spaced Independent Variable with a Single-Valued Dependent Variable

When two-variable data have a single-valued dependent variable and equally-spaced values of the independent variable, the methods of graphing a time series that were just discussed can be considered. Figure 3.63 shows one example. The dependent variable is an estimate of the spectrum of the aa index, and the independent variable is frequency, measured in cycles per year. There are 101 estimates of the spectrum at 101 frequencies spaced 0.005 cycles/year apart. Since the spectrum estimate is a smooth function of frequency, a connected graph was used.

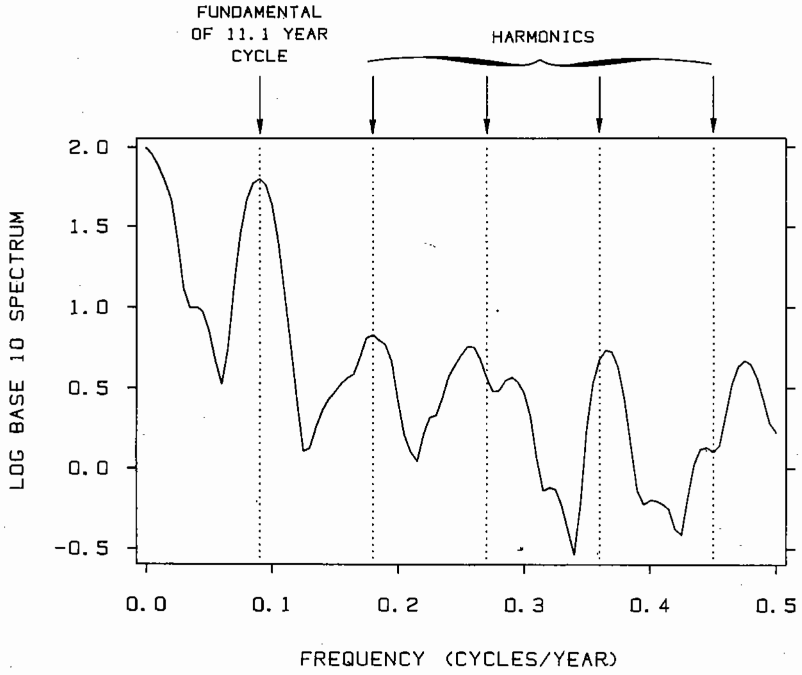

Figure 3.63 SINGLE-VALUED DEPENDENT VARIABLE WITH EQUALLY-SPACED INDEPENDENT VARIABLE. Data with a single-valued dependent variable and an equally spaced independent variable can be graphed using connected, symbol, connected symbol, and vertical line graphs. In this example the dependent variable is an estimate of the spectrum of the aa index and the independent variable is frequency. A connected graph is used to show the data.

The rise in the spectrum near zero frequency is just the trend observed earlier in the graph of the series against time. Heading toward higher frequencies, the first peak, whose frequency is marked with a vertical reference line, provides an estimate of the average fundamental frequency of the cycles in the aa index; the estimated frequency is 0.09 cycles/year, which has a period of 11.1 years. The first four harmonics (multiples) of this fundamental are also marked by reference lines. It seems likely that the peaks in the spectrum near the first three harmonics are also a result of the 11.1 year cycle.

The spectrum in this example was estimated by the following procedure: subtract the mean; multiply by a full cosine taper [15, ch. 5]; compute the squared modulus of the Fourier transform; smooth with a boxcar window with five raw spectrum values per estimate.

Step Function Graphs

A *step function graph* is appropriate when the dependent variable is constant over intervals of the independent variable. Figure 3.64 is a step-function graph that shows the weight of the Hershey Bar over a time period of about 20 years. In his essay, "Phyletic Size Decrease in

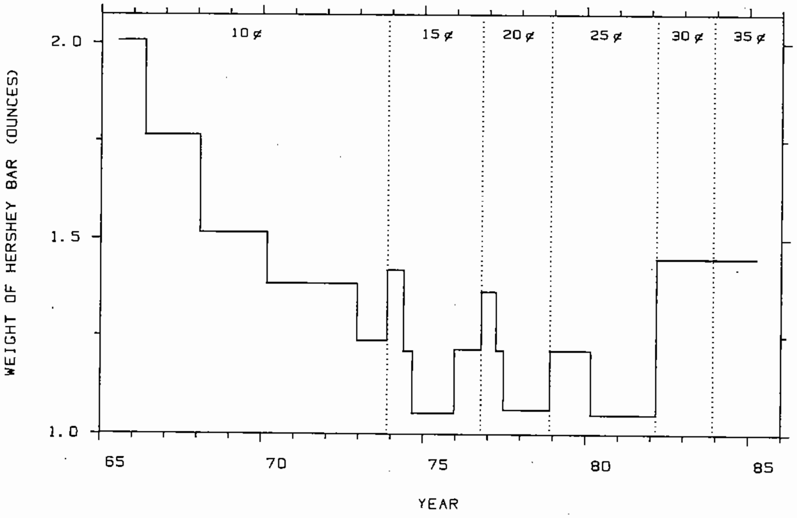

Figure 3.64 STEP FUNCTION GRAPH. The weight of the Hershey Bar is graphed against time. A step function graph is appropriate when the dependent variable is constant over intervals of the independent variable.

Hershey Bars," Stephen Jay Gould showed that the history from 1965 generally has been one of a decline in size of a bar with a fixed price, followed by a sudden rise in both price and size, and then followed by a gradual decline in size [54, pp. 313-319]. This is illustrated in Figure 3.64, which includes some additional data not available to Gould at the time.

With the additional data, we can now see something else quite striking in Figure 3.64. It appears that one ounce is a reflecting barrier below which bar weight will not drop. Maybe the barrier is psychological. Hershey executives might see one ounce as the last line of defense, and fear that were bar weight to drop below it, there would be nothing to stop its ultimate extinction. But what will happen when the United States converts to the metric system? One ounce is 28.35 grams. The human mind puts special emphasis on simple numbers, and a new psychological barrier of 25 grams may take over.

Figure 3.64 seems to beg for a new graph using just the same data as the old one, but graphed in a new way. The idea, which arises from the field of economics, is that what really counts is the cost (price) per unit of weight. In other words, how much does one bite of a Hershey Bar cost? Figure 3.65 shows the cost per ounce through time, again using a step function graph, which reveals the real law of nature: the inexorable rise in price per mouthful through time. The changing size is just a way of helping to obey the law, and we can see a critical fact not apparent in the first step function graph — every *price increase,* except the change from the 25¢ bar to the 30¢, was in fact an increase in cost per ounce. During the time period of this data there were only two points in time when cost per ounce decreased — once, when the price rose to 30¢, and once, in 1975, when the weight increased but the price stayed constant.

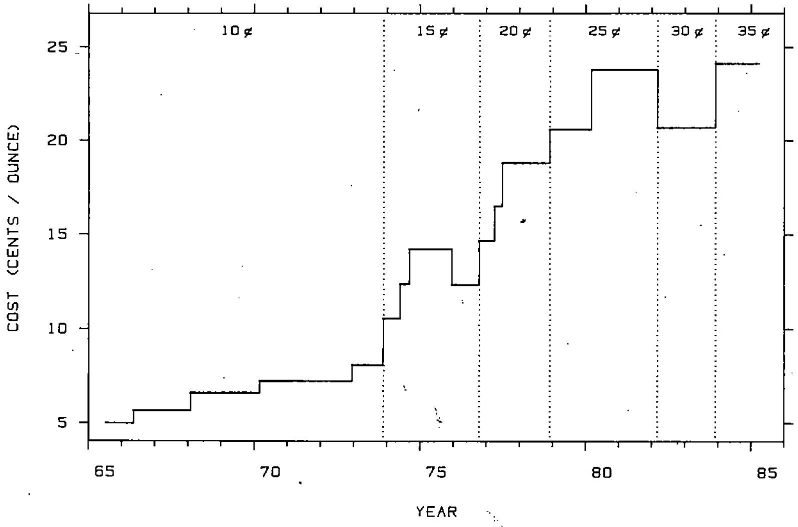

Figure 3.65 STEP FUNCTION GRAPH. The cost per ounce of the Hershey Bar is graphed against time. There are only two points in time when cost per ounce decreased.

3.5 TWO OR MORE CATEGORIES OF MEASUREMENTS OF TWO QUANTITATIVE VARIABLES: SUPERPOSITION AND JUXTAPOSITION

This section is about graphical methods for two or more categories, or groups, of measurements of two quantitative variables. We saw in Section 2 of Chapter 2 that graphing different data sets can be a challenge. If we *superpose* them in the same data region, we must be sure that the graphical elements portraying each of the data sets can be visually discriminated from the graphical elements showing the other data sets. If we graph them on *juxtaposed* data regions we want to be able to compare the different data sets as readily as possible. This section discusses graphical methods for achieving these goals.

Superposed Plotting Symbols

Figure 3.66 has four superposed data sets. The measurements are from the survey of graphs in scientific publications discussed in Section 3.3 [27]. For a large number of scientific journals, measurements were made of the fraction of space each journal devoted to graphs (not including legends) and the fraction of space each journal devoted to graph legends. Figure 3.66 is a graph of log (legend area/graph area) against log (graph area) for 46 journals. The ratio of legend area to graph area is a rough measure of the amount of legend explanation given to graphs. The letters encode four journal categories:

Biological — biology, medicine

Physical — physics, chemistry, engineering, geography

Mathematical — mathematics, statistics, computer science

Social — psychology, economics, sociology, education.

One advantage of the letters is that it is easy to remember the groups, and looking back and forth between the graph and the key is not necessary. But a serious disadvantage of the letters is that they do not provide high visual discrimination with one another; it is hard, compared with other encoding schemes, to perceive the points for a particular group as a whole, mentally filtering out the points of other groups.

Figures 3.67 and 3.68 present two other methods for encoding the four categories. To make the ensuing discussion about visual discrimination more meaningful, look at each of Figures 3.66 to 3.68 and try to see the points of each category as a unit as if the other points were not there.

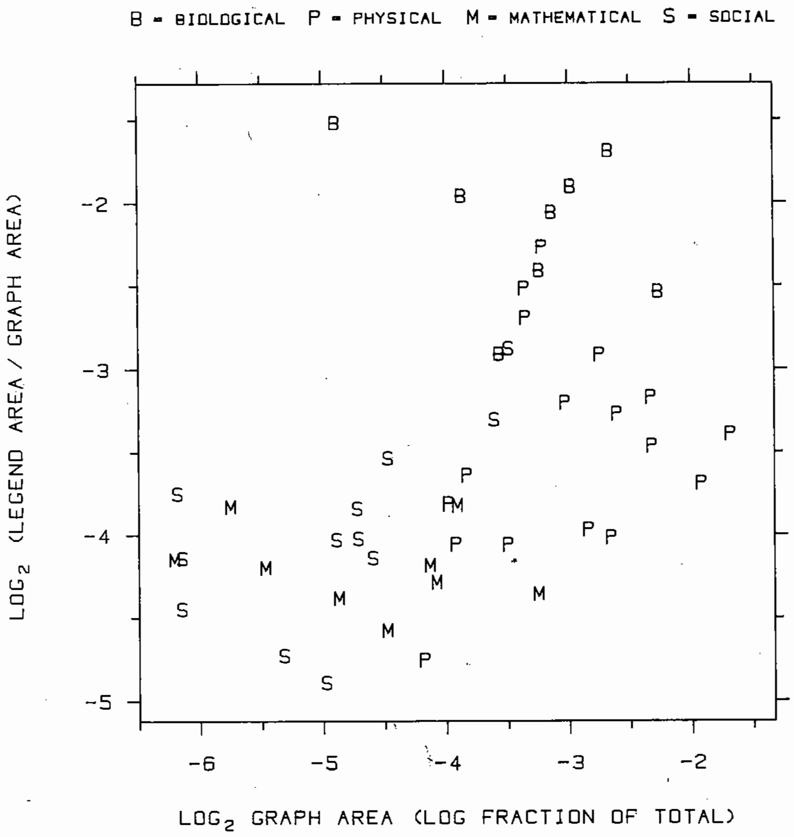

Figure 3.66 SUPERPOSED SYMBOLS. Four categories of measurements of two variables are graphed. The letters encoding the four categories do not provide high visual discrimination of the four sets of points.

The encoding scheme in Figure 3.67, commonly used in science and technology, is different geometric shapes; the visual discrimination appears somewhat greater than for the letters in Figure 3.66. It is harder to remember the category associated with a shape than with a letter, but this is a minor point. In Figure 3.68, four types of circle fill are used to encode the categories. Theoretical and experimental evidence from the field of visual perception suggests that different methods of fill should provide high visual discrimination [34]. In fact, Figure 3.68 appears to provide better discrimination than the other two figures.

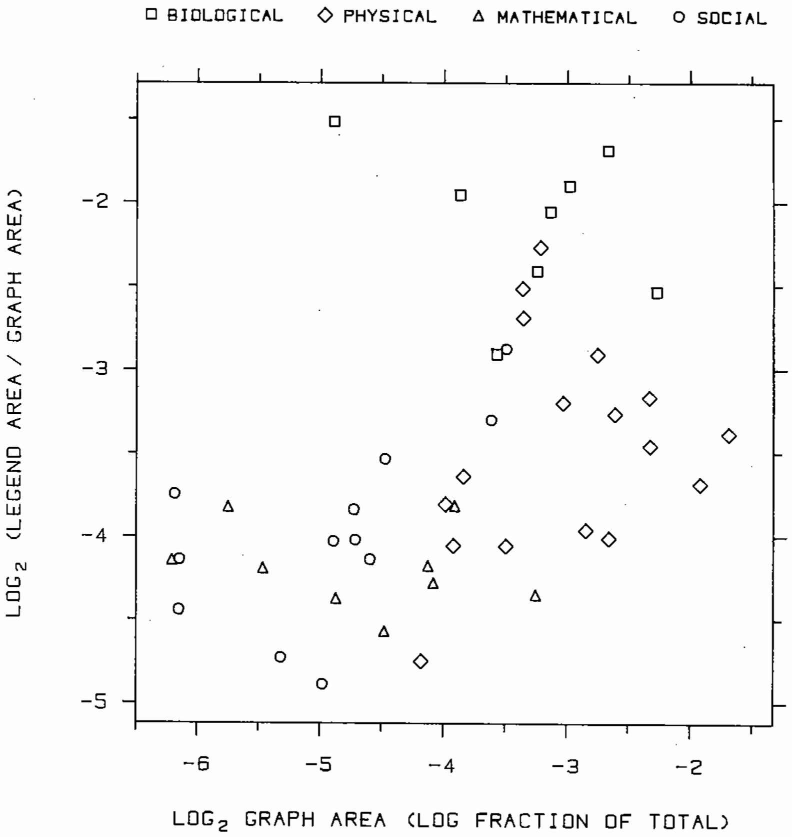

Figure 3.67 SUPERPOSED SYMBOLS. The data of Figure 3.66 are graphed with the categories encoded by differently shaped plotting symbols. This provides somewhat greater visual discrimination than using the letters.

Figure 3.68 shows two interesting phenomena: social science journals and mathematical science journals tend to use graphs less than the other two categories, and the biological science journals tend to have more in the figure legends. The second phenomenon is probably due to the tendency in biological journals to put experimental procedures in figure legends.

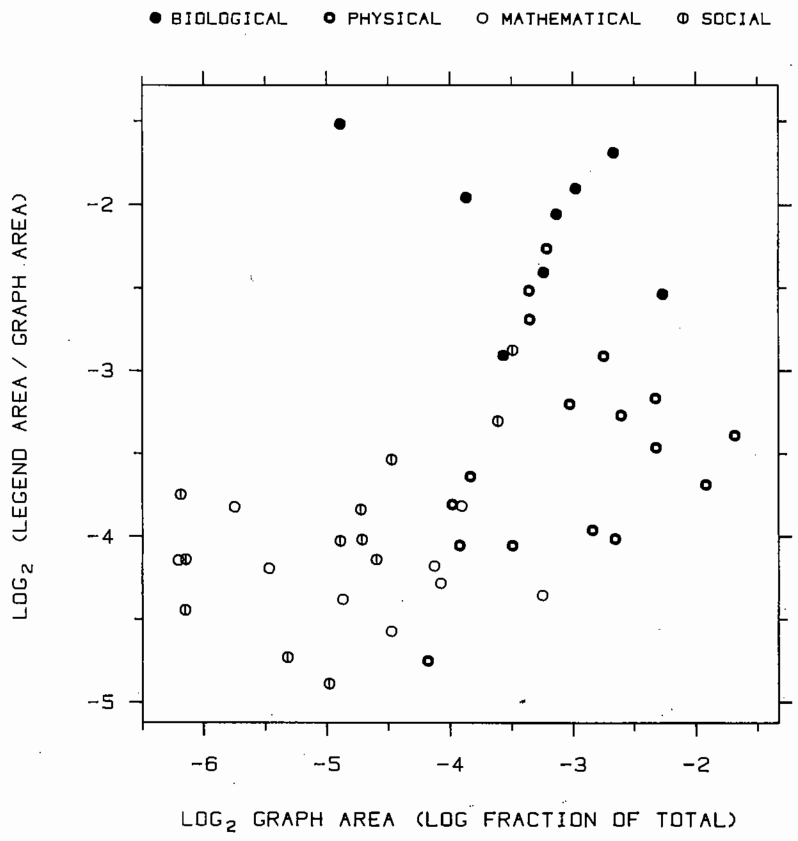

Figure 3.68 SUPERPOSED SYMBOLS. The data of Figure 3.66 are graphed with the categories encoded by circles with different methods of fill. This provides the highest visual discrimination of the methods shown in Figures 3.66 to 3.68.

The encoding scheme in Figure 3.68 works well if there is not much overlap of the plotting symbols. When there is overlap, the solid portions of the symbols can form uninterpretable blobs. In such a case we must attempt to use symbols that provide as much visual discrimination as possible, subject to the constraint that the symbols tolerate overlap. The constraint seems to restrict us to symbols consisting of curves and lines, with no solid parts, and with a minimum of ink. One encoding scheme that does reasonably well is shown in Figure 3.69.

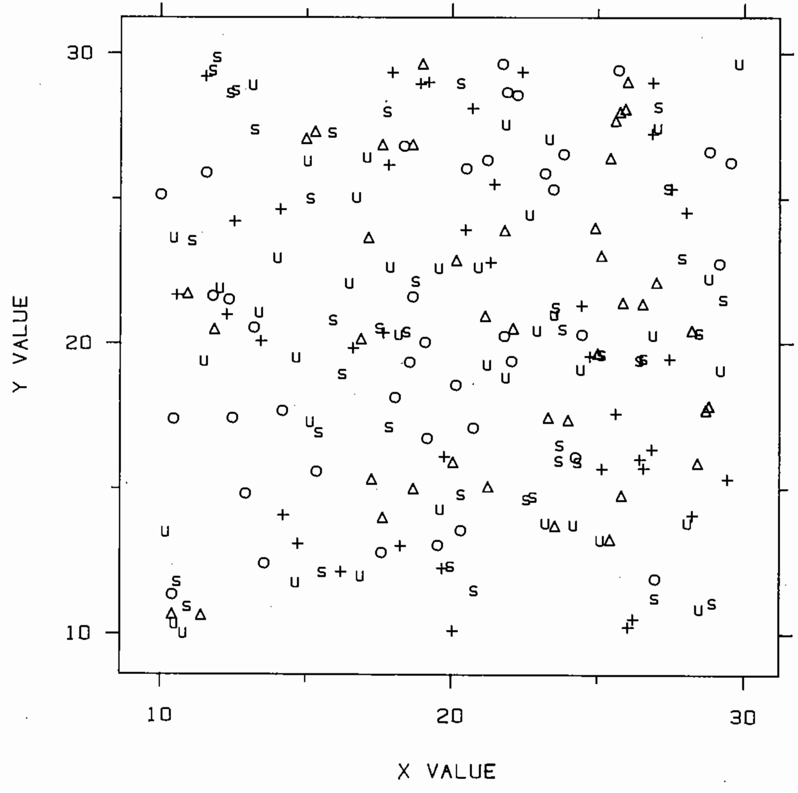

Figure 3.69 SUPERPOSED SYMBOLS. The plotting symbols used on this graph provide fair visual discrimination and can tolerate overlap.

Figure 3.70 shows two sets of plotting symbols, one in each row. The top set is to be used when there is little overlap, and the bottom set is to be used when overlap causes problems with the top set. For each set, the suggestion is to use the first two symbols on the left if there are two categories, the first three symbols on the left if there are three categories, and so forth.

Superposed Curves in Black and White

Sometimes superposed data sets come in the form of superposed curves, as in Figure 3.71. The data will be described shortly. Often, we can make each curve solid and still have the requisite visual discrimination. If at the intersection of two curves, the slopes of the curves are very different, our eyes have no trouble visually tracking each curve. For example, in Figure 3.71, *5 gt airburst* and *3 gt* are easy to follow at their crossing between 100 and 150 days. But if two curves come together with similar slopes, they can lose their identity; *5 gt* and *5 gt airburst* almost do this at their intersection just after 100 days.

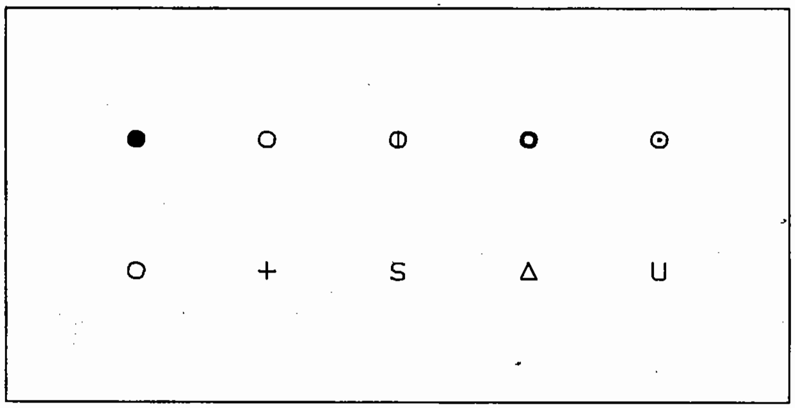

Figure 3.70 PLOTTING SYMBOLS. The top set can be used when there is little overlap, and the bottom can be used when overlap causes problems with the top set. The first two symbols on the left are to be used when there are two categories, the first three symbols are to be used when there are three categories, and so forth.

If solid curves lose their identity we can switch to different curve types, as in Figure 3.72. The goal, as it was for symbols, is to choose curves that have high visual discrimination. We want to see each curve effortlessly and as a whole and not have to visually trace it out as we do a secondary road on an automobile map. Figure 3.73 is a palette of curve types that shows the variety possible from dots and dashes.

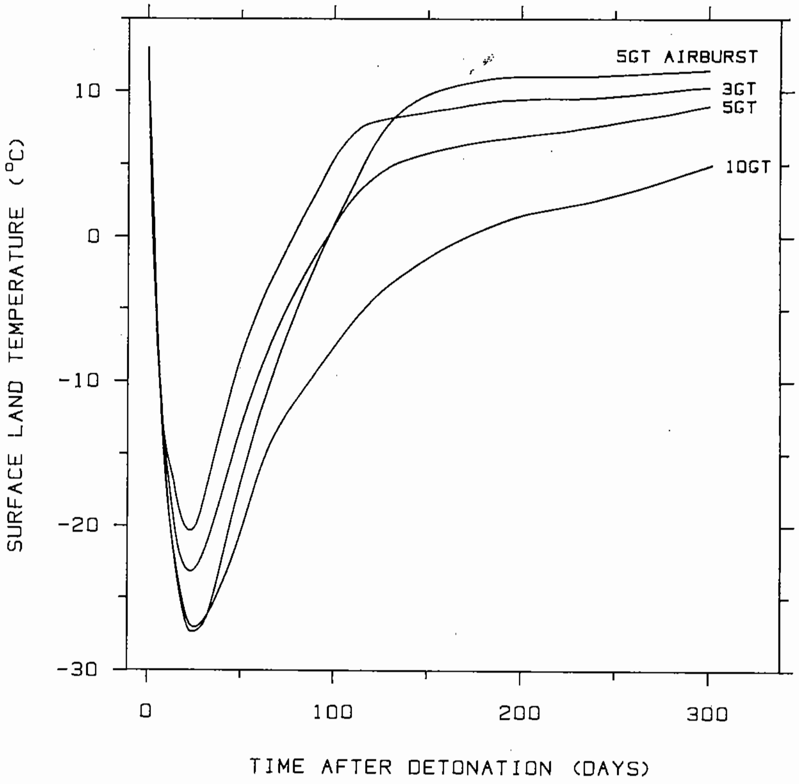

Figure 3.71 SUPERPOSED CURVES. Superposed curves need to be visually discriminated. In this case the behavior of the data is simple enough that each curve is visually distinct.

Juxtaposition

Sometimes the only solution for visual discrimination of different data sets is to give up superposition and use juxtaposition of two or more panels. This is illustrated in Figure 3.74.

Figure 3.74 shows model predictions of temperature in the Northern Hemisphere following different types of nuclear exchanges [127]. The temperatures following major exchanges drop precipitously due to soot from conflagrations of cities and forests and due to dust from soil and vaporization of earth and rock. The soot and dust substantially reduce radiation from the sun which, in turn, causes the temperature to drop.

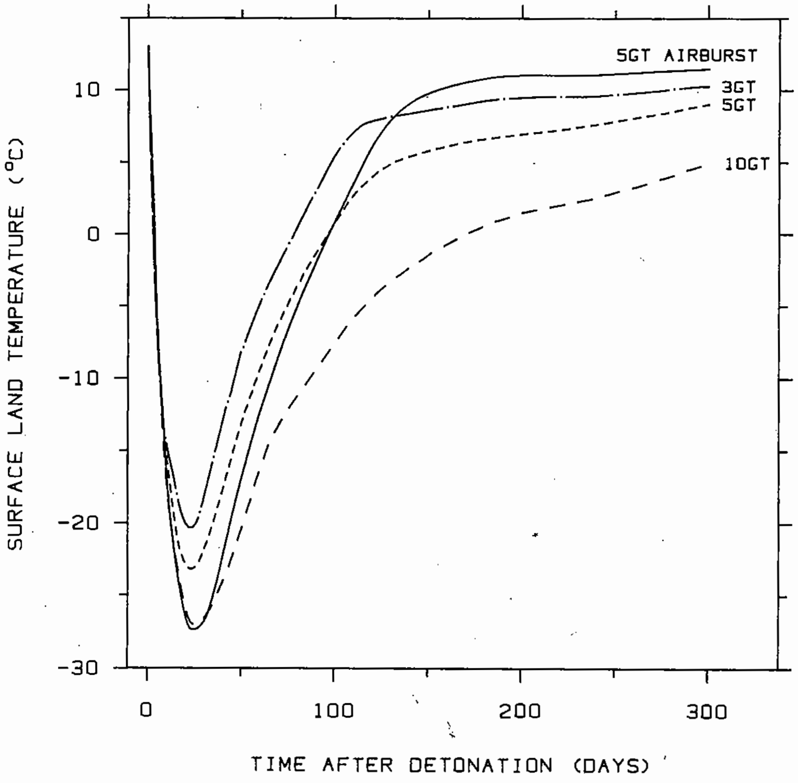

Figure 3.72 SUPERPOSED CURVES. Visual discrimination can be increased by different curve types.

The temperatures are computed from physical models that describe the creation of particles, the production of radiation, convection, and a script for the nuclear war. The panels in Figure 3.74 are different exchange scenarios, which are explained in Table 3.1. The total world nuclear arsenal of strategic weapons is 17 gigatons (gt), which is roughly equal to 10^6 Hiroshima bombs.

Table 3.1 NUCLEAR EXCHANGE SCENARIOS.

Code	Description
10 gt	10 gt exchange.
5 gt	5 gt exchange.
5 gt air	5 gt airburst in which all weapons are detonated above ground.
5 gt dust	5 gt exchange with only the effects of dust included, but not fires.
3 gt	3 gt exchange.
3 gt silo	3 gt exchange aimed only at missile silos.
1 gt	1 gt exchange.
0.1 gt city	0.1 gt exchange aimed only at major cities.

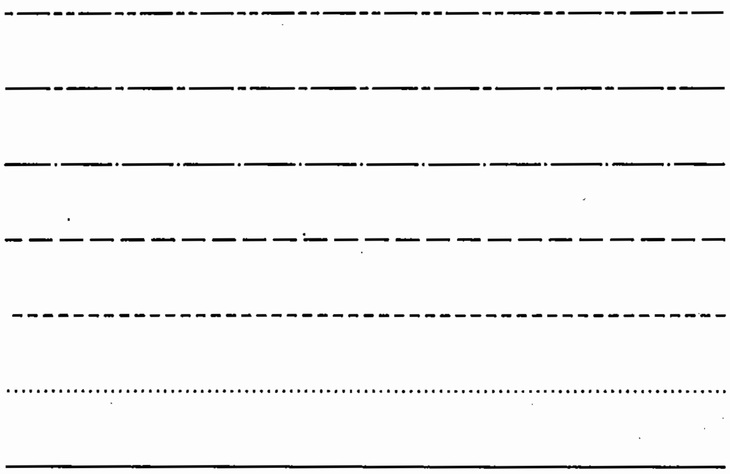

Figure 3.73 CURVE TYPES. Dots, dashes, and combinations provide a variety of patterns for graphing curves.

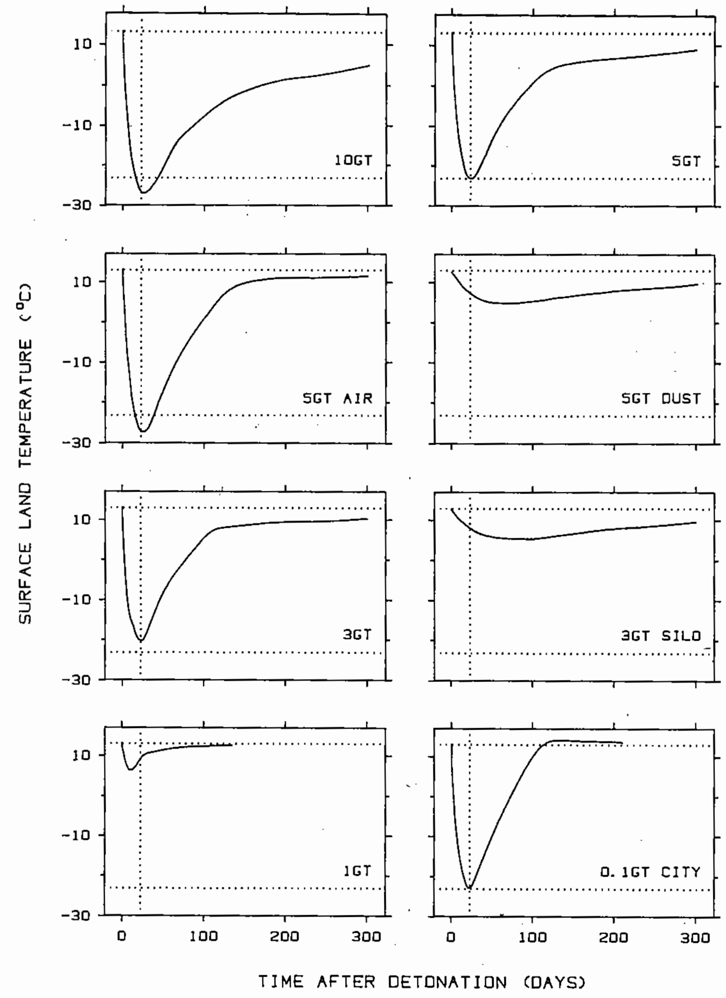

Figure 3.74 JUXTAPOSITION. Each curve shows averaged Northern Hemisphere temperature following a nuclear war. The scenarios of the war are different for different panels. On this graph the different data sets are juxtaposed. Comparisons of the curves are enhanced by the strategically placed reference lines: the upper horizontal reference line on each panel shows the current average ambient Northern Hemisphere temperature, the lower horizontal reference line shows the minimum temperature for the *5 gt* exchange, and the vertical reference line shows the time of the *5 gt* temperature minimum.

Juxtaposition is needed for this temperature data. Superposition results in the tangle of Figure 3.75. We could attempt to improve the graph by using different curve types, but no black and white method appears to reduce the clutter substantially. Actually, it is not necessary to settle for one extreme or the other; we might have attempted four juxtaposed panels, each with two curves superposed.

When it works, superposition is better than juxtaposition because it allows a more incisive comparison of the values of the different data sets. For example, in Figure 3.75 we can see very clearly that the

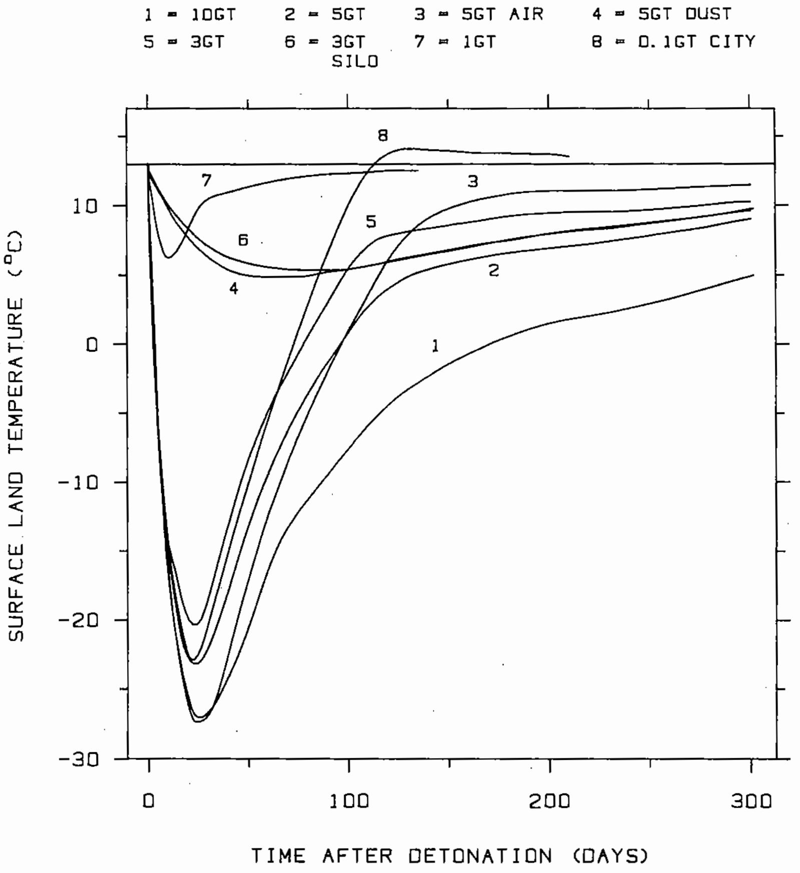

Figure 3.75 SUPERPOSITION. The curves of Figure 3.74 cannot be easily visually discriminated when they are superposed.

minimum for each scenario occurs at about the same time and we can effectively compare the values of the minima; the problem in this example is that it is not easy to see which curve goes with which scenario.

In giving up superposition for juxtaposition we decrease our ability to compare the values of different data sets in order to increase our ability to discriminate the data sets. However, we can employ a method that greatly improves our ability to compare the values on different juxtaposed panels — *strategically place the same lines or curves on all panels to serve as visual references.* For example, in Figure 3.74 the lower horizontal reference line on all panels is the value of the *5 gt* minimum; this line allows us to compare the temperature minima more effectively. The top horizontal reference line is the Northern Hemisphere average ambient temperature; this line helps us to judge the progress each curve makes in getting back to normal conditions. The vertical reference line shows the time of the *5 gt* minimum; this line provides a more effective comparison of the times of the minima.

Figure 3.74 does a good job of showing the temperature profiles. The major exchanges result in a rapid drop to around −25 °C and then a slow recovery lasting many months. The *0.1 gt city* attack has such a strong effect because of the tremendous concentration of combustible materials in urban areas.

Visual references on juxtaposed panels can take many different forms: lines, curves, or plotting symbols. We will now give two more examples to show how varied the nature of the visual reference can be.

Figure 3.76 is a graph of brain weights and body weights for four categories of species [40]. Juxtaposition is necessary because superposition results in so much overlap that visual resolution of the four groups is impossible whatever (black and white) method is tried. The same three lines are drawn on each panel. The top line shows the major axis of the primate point cloud, the middle line shows the major axes of the bird and nonprimate mammal point clouds, and the lower line is for the fish. These three lines help us to compare the relative positions of the four point clouds. All three lines have slope 2/3, because brain weights tend to be related to body weights to the 2/3 power; the reason for this relationship is discussed in Section 3 of Chapter 1.

In Figure 3.77 the four lines have the same slopes but the intercepts are different. The data are the logarithms of the winning times for four track events at the Olympics from 1900 to 1984 [22, 138, p. 833]. The lines were fit to the data using least squares; the slope was held fixed but the intercept was allowed to vary from one data set to the next.

Because the number of units per cm is the same on the four vertical scales, the lines on the four panels have the same angle with the horizontal. In this example the lines help us to see that the decrease in the log running times has been nearly linear and that the slopes for the four data sets are the same. This means that the overall percent decrease since 1900 has been about the same for the four races.

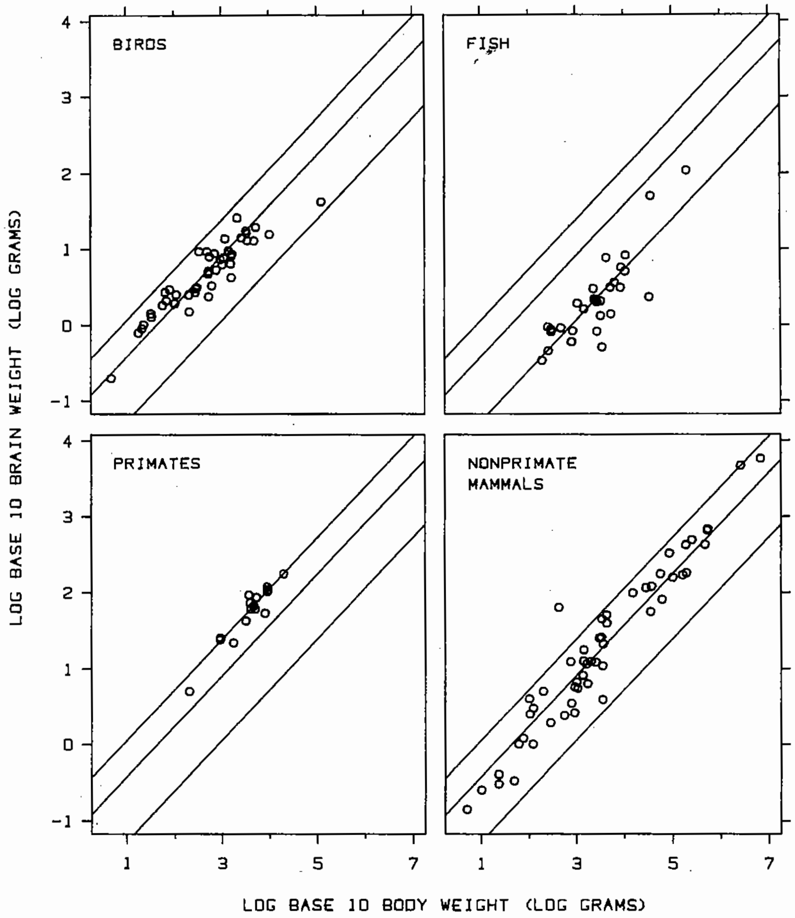

Figure 3.76 JUXTAPOSITION. Log brain weights are graphed against log body weights for four categories of species. The same three reference lines are drawn on the four panels. Each line has slope 2/3; the top line describes the primates, the middle line describes the birds and nonprimate mammals, and the bottom line describes the fish. These strategically placed lines enhance our ability to compare data on different panels.

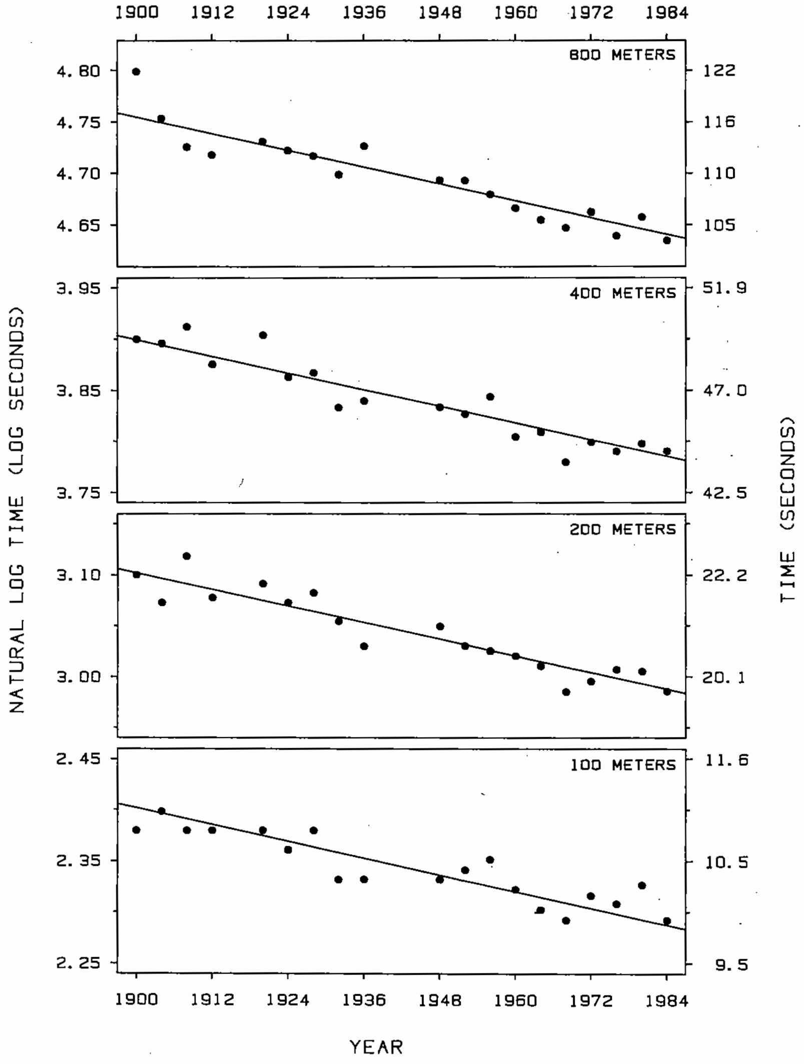

Figure 3.77 JUXTAPOSITION. The graph shows the logarithms of the winning times at the Olympics in four races. The vertical scales on the four panels have the same number of log seconds per cm. The four lines on the panels have the same slope, determined by a least squares fit. Since logarithms are graphed and since the points nearly follow lines with the same slope, we can conclude that the percent decrease in the running times is roughly constant through time and the constant is the same for all four races.

Figure 3.78 shows curves used as references. The data are from an experiment on graphical perception [33] that will be discussed in Section 3 of Chapter 4. A group of 51 subjects judged 40 pairs of values on bar charts and the same 40 pairs on pie charts; each judgment consisted of studying the two values and visually judging what percent the smaller was of the larger. The left panel of Figure 3.78 shows the 40 average judgment errors (averaged across subjects) graphed against the true percents for the 40 pie chart judgments. The right panel shows the same variables for the bar chart judgments. Lowess curves, described in Section 3.4, were fit to each of the two data sets; both curves are graphed on each panel and serve as visual references to help us compare the average errors for the two types of charts.

Color

If color is available we do not as frequently need to give up superposition and use juxtaposition. Our visual system does a marvelous job of discriminating different colors. In Figures 3.79 and 3.80 superposition in black and white is used for two sets of data that we have seen earlier in the chapter. We cannot effectively discriminate the different data sets. Color is used in Plates 1 and 2, which follow page 212, and discrimination is considerably enhanced.

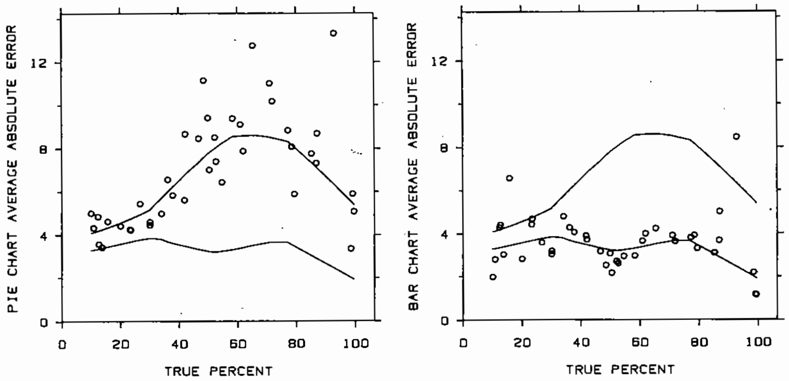

Figure 3.78 JUXTAPOSITION. The graph compares pie chart and bar chart judgment errors of 51 subjects. Two curves show how the bar chart errors and the pie chart errors depend on the true percent being judged. Graphing the two curves on both panels helps us to compare the two sets of data.

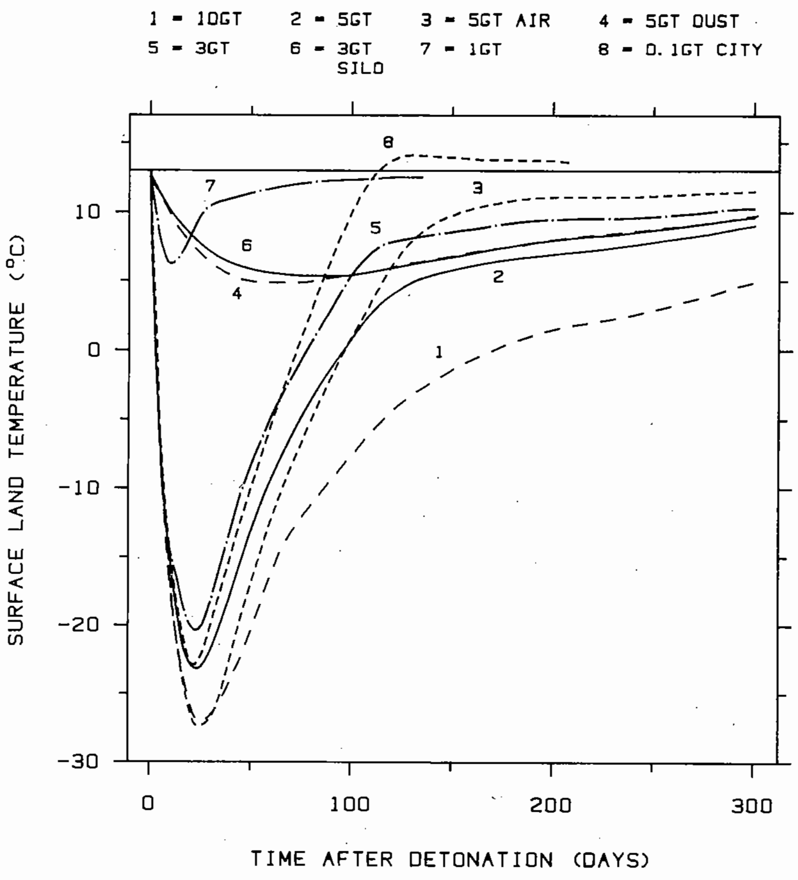

Figure 3.79 COLOR. Color is a good methrod for providing visual discrimination. The eight curves are not as easy to discriminate as they are in the color encoding in Plate 1, which follows page 212.

Many people will find the colors in Plates 1 and 2 unesthetic, garish, and clashing. This was done on purpose to maximize the visual discrimination. Pleasing colors that blend well tend not to provide as good visual discrimination.

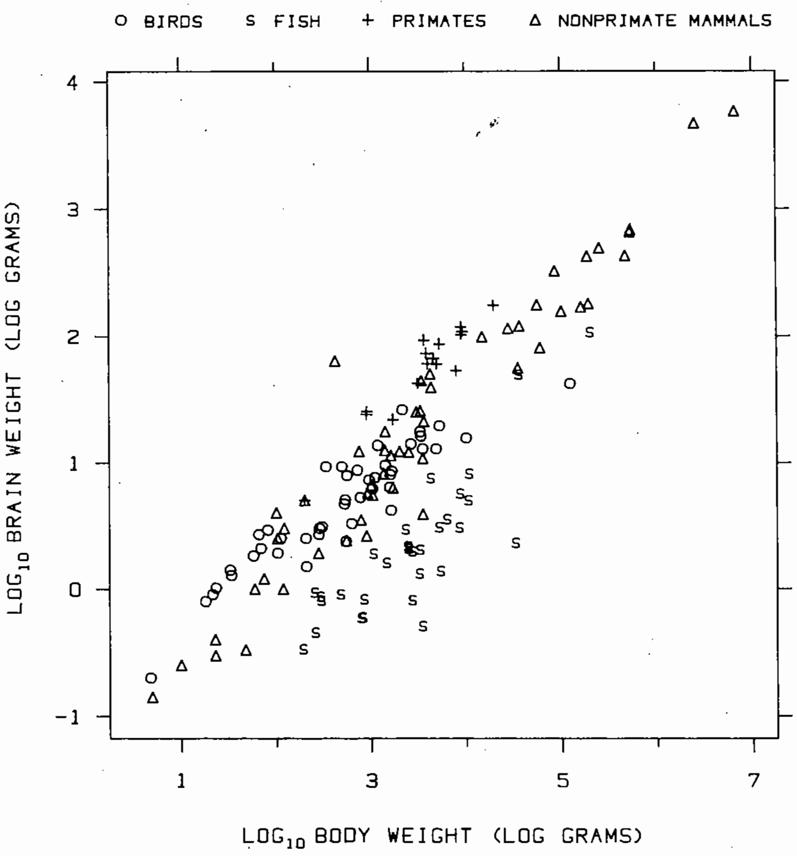

Figure 3.80 COLOR. The four categories of data cannot be easily discriminated. Discrimination is greatly enhanced by color encoding in Plate 2.

3.6 THREE OR MORE QUANTITATIVE VARIABLES

Science and technology would be far simpler if data, like the people of Edwin A. Abbott's *Flatland* [1], always stayed in two dimensions. Unfortunately, data can live in three, four, five or any number of dimensions. Consider, for example, measurements of temperature, humidity, barometric pressure, percentage cloud cover, solar radiation intensity, and wind speed at a particular location at noon on 100 different days. The data on these six variables consist of 100 points in a six-dimensional space. How are we to graph them to understand the complex relationships? How are we to peer into this six-dimensional space and see the configuration of points?

Graphs are two-dimensional. If there are only two variables — for example, just temperature and humidity in our meteorological data set — then the data space is two-dimensional, and a Cartesian graph of one variable against the other shows the configuration of points. As soon as data move to even three variables and three dimensions we must be content with attempting to infer the multidimensional structure by a two-dimensional medium. In this section, some methods for doing this are described.

Framed-Rectangle Graphs

Figure 3.81 is a *framed-rectangle graph* [33], which can be used to show how one variable depends on two others. The data are the per capita debts in dollars of the 48 continental states of the U.S. in 1980 [137, p. 116]. Each value is portrayed by a solid rectangle inside a frame that has tick marks halfway up the vertical sides. The frames are the same size, which helps us judge the relative magnitudes of the values by providing a common visual reference. For geographical data, such as those in Figure 3.81, the framed-rectangle graph conveys the values far more efficiently and accurately to the human viewer than the very common statistical map [97, pp. 282-288] in which the data are encoded by shading the geographical units, which in this example are the states. Issues of graphical perception such as this are the topic of Chapter 4.

The data in Figure 3.81 are three-dimensional; geographical location needs two dimensions and debt is the third. Furthermore, we are in the dependent-independent variable case because the goal is to see how debt depends on geographical location.

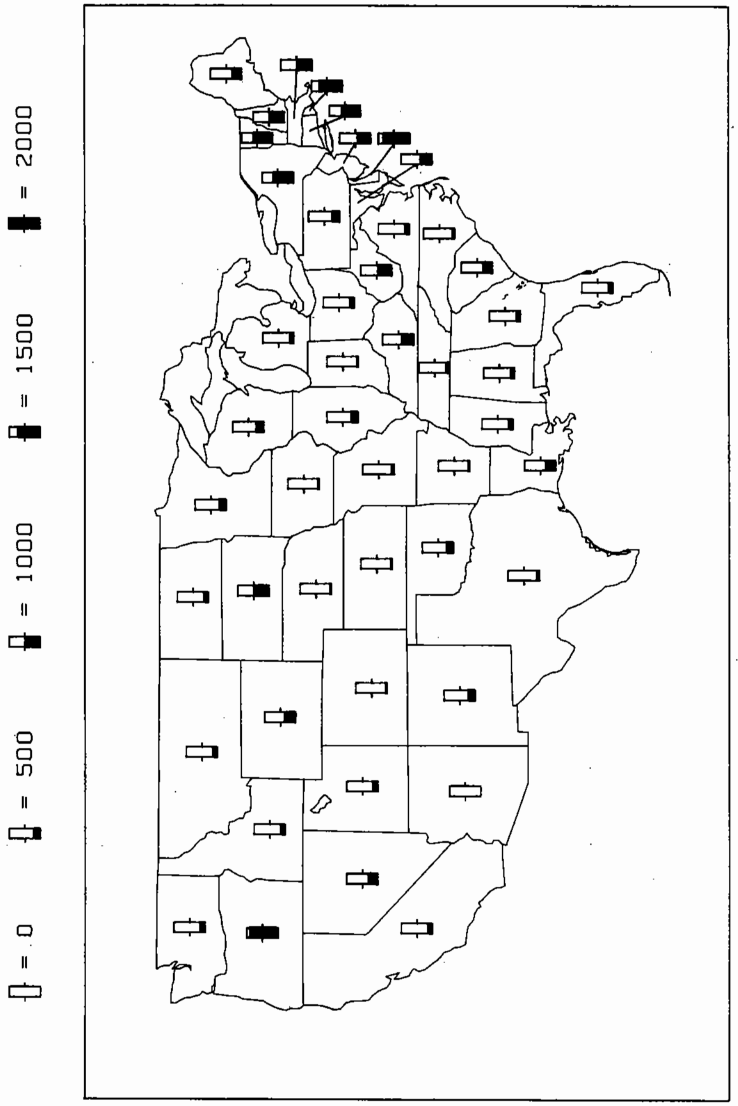

Figure 3.81 FRAMED-RECTANGLE GRAPH. The data are the per capita debts in dollars of the 48 continental states of the United States in 1980. The frames of the framed rectangles help us to judge the values of the data by providing a common visual reference.

The framed-rectangle graph can be useful in any situation where we want to see how measured values of one variable, z, depend on values of two others, x and y. However, since the framed rectangles cannot withstand overlap, the method is helpful only when the number of observations is small or moderate and when there is not too much crowding of the (x,y) values in any one region of the plane.

Figure 3.81 shows that the middle Atlantic states and New England are the regions of the country where the states are least afraid to go into debt. States in the South and in the West tend to be more restrained in their indebtedness, although Oregon, with its near $2000 per capita debt, leads the country and is a striking anomaly.

Scatterplot Matrices

An award should be given for the invention of the *scatterplot matrix*, but the inventor (or inventors) is unknown — an anonymous donor to the world's collection of graphical methods. Early drafts of *Graphical Methods for Data Analysis* [21] contain the first written discussion of the idea, but it was in use before that. The inventor may not have fully appreciated the significance of the method or may have thought the idea too trivial to bring it forward, but its simple, elegant solution to a difficult problem is one of the best graphical ideas around.

Suppose the multidimensional data consist of k variables, so that the data points lie in a k-dimensional space. One way to study the data is to graph each pair of variables; since there are $k(k-1)/2$ pairs, such an approach is practical only if k is not too large. But just making the $k(k-1)/2$ graphs of each variable against each other, without any coordination, often results in a confusing collection of graphs that are hard to integrate, both visually and cognitively.

The important idea of the scatterplot matrix is to arrange the graphs in a matrix with shared scales. An example is shown in Figure 3.82. There are four variables: wind speed, temperature, solar radiation, and concentrations of the air pollutant, ozone. The data, from a study of the dependence of ozone on meteorological conditions [18], are measurements of the four variables on 111 days from May to September of 1973 at sites in the New York City metropolitan region. There is one measurement of each variable on each day; so the data consist of 111 points in a four-dimensional space. (The details of the measurements are the following: solar radiation is the amount from 0800 to 1200 in the frequency band 4000-7700Å; wind speed is the average of values at 0700 and 1000; temperature is the daily maximum; and ozone is the average of values from 0800 to 1200.)

Each panel of the matrix is a scatterplot of one variable against another. For the three graphs in the second row of Figure 3.82, the vertical scale is ozone, and the three horizontal scales are solar radiation, temperature, and wind speed. So the graph in position (2,1) in the matrix — that is, the second row and first column — is a scatterplot of ozone against solar radiation; position (2,3) is a scatterplot of ozone against temperature; position (2,4) is a scatterplot of ozone against wind speed.

The upper right triangle of the scatterplot matrix has all of the $k(k-1)/2$ pairs of graphs, and so does the lower right triangle; thus

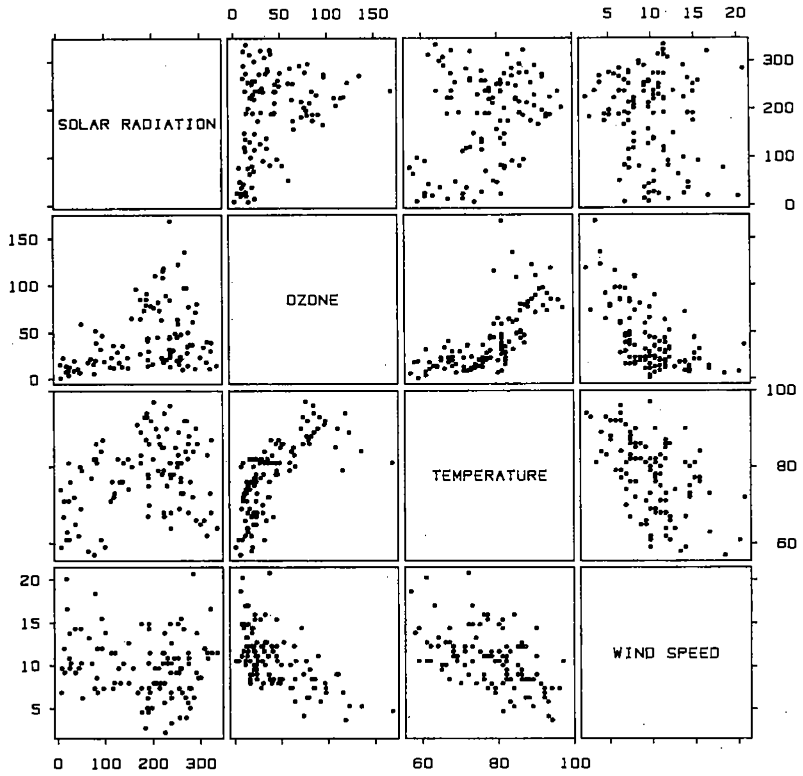

Figure 3.82 SCATTERPLOT MATRIX. The data are measurements of solar radiation, ozone, temperature, and wind speed on 111 days. Thus the measurements are 111 points in a four-dimensional space. The graphical method in this figure is a scatterplot matrix: all pairwise scatterplots of the variables are aligned into a matrix with shared scales.

altogether there are $k(k-1)$ panels and each pair of variables is graphed twice. For example, in Figure 3.82 the (1,3) panel is a graph of solar radiation on the vertical scale against temperature on the horizontal scale, and the (3,1) panel has the same variables but with the scales reversed.

The most important feature of the scatterplot matrix is that we can visually scan a row, or a column, and see one variable graphed against all others with the three scales for the one variable lined up along the horizontal, or the vertical. This is the reason, despite the redundancy, for including both the upper and lower triangles in the matrix. Suppose that in Figure 3.82 only the lower left triangle were present. To see temperature against everything else we would have to scan the first two graphs in the temperature row and then turn the corner to see wind speed against temperature; the three temperature scales would not be lined up, which would make visual assessment more difficult.

Space and resolution quickly become a problem with the scatterplot matrix; the method of construction in Figure 3.82 reduces the problem somewhat. The labels of the variables are inside the main-diagonal boxes so that the graph can expand as much as possible. The tick mark labels for the horizontal scales, as well as for the vertical scales, alternate sides so that labels for successive scales do not interfere with one another. And the panels have been squeezed tightly together, allowing just enough space to provide visual separation.

The scatterplot matrix in Figure 3.82 reveals much about the ozone and meteorological data. Ozone is a secondary air pollutant; it is not emitted directly into the atmosphere but rather is a product of chemical reactions that require solar radiation and emissions of nitric oxide and hydrocarbons from smoke stacks and automobiles. For ozone to get to very high levels, stagnant air conditions are also required.

It is no surprise then to see a relationship between solar radiation and ozone in panel (2,1), but the nature of the relationship is enlightening. There is an upper envelope in the form of an inverted "V". For low values of solar radiation, high values of ozone never occur. The major reason is that the photochemical reactions that produce ozone need a minimum amount of solar radiation. The (2,1) panel also shows that when solar radiation is between 200 and 300 Langleys, ozone can be either high or low. If we scan across the ozone row to panels (2,3) and (2,4) it becomes clear that the high ozone days are those with high temperatures and low wind speeds — stagnant days. Overall, there is a strong association between wind speed and ozone and between temperature and ozone. Both wind speed and temperature are measures of stagnancy; as wind speed decreases or as temperature

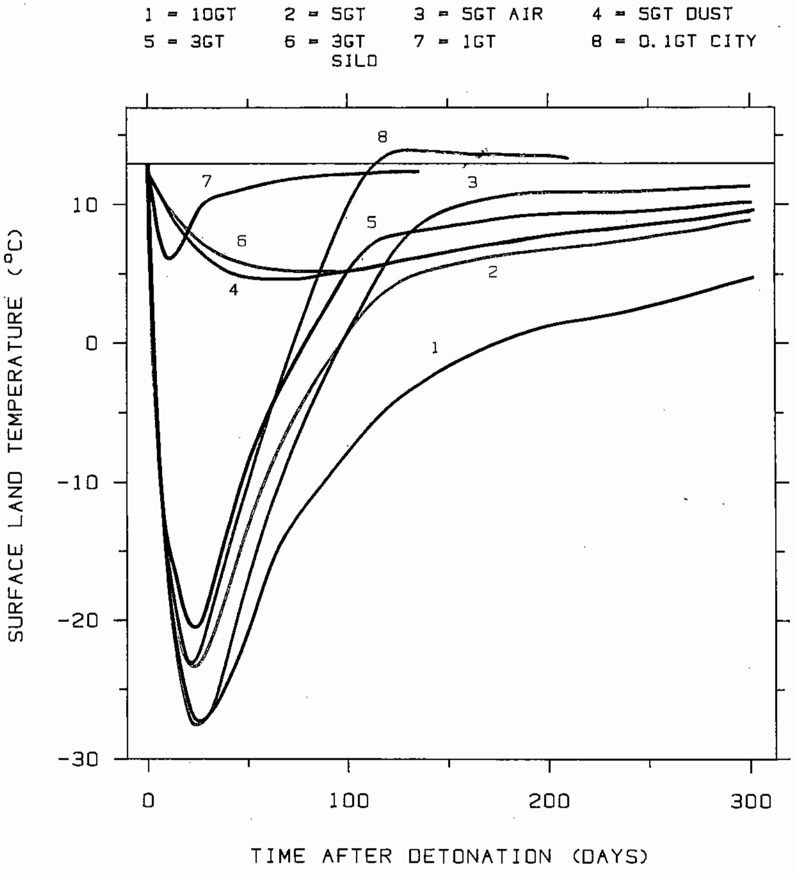

Plate 1. Color provides good discrimination of the different data sets. Compare with Figure 3.79.

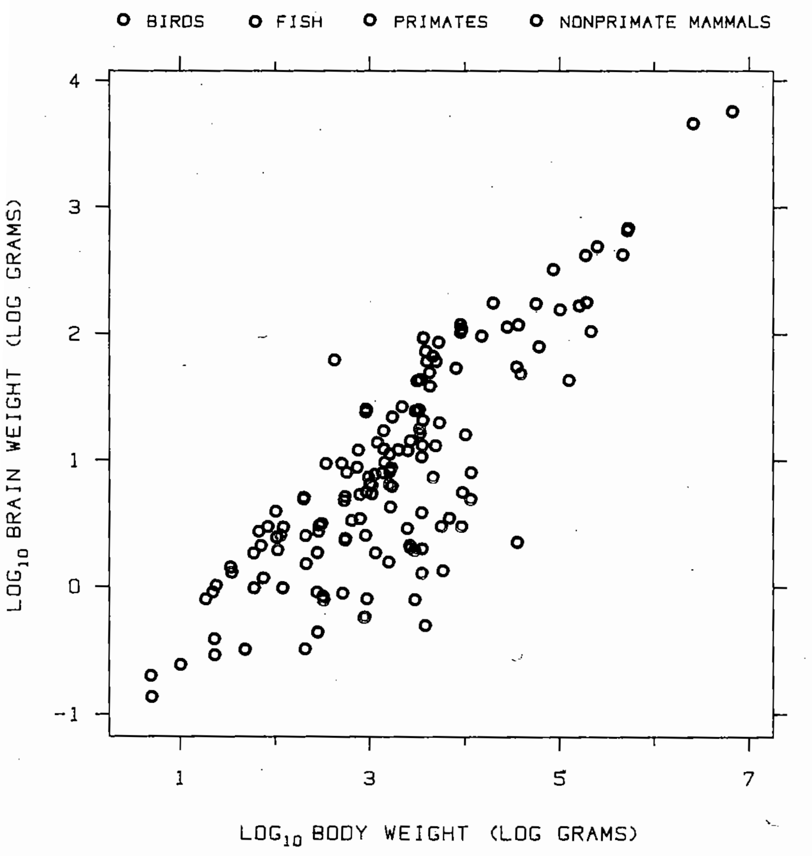

Plate 2. Color provides good discrimination of the different data sets. Compare with Figure 3.80.

increases, conditions become more stagnant and ozone rises. But the (3,4) panel shows that wind speed and temperature are related and thus are measuring stagnancy, to some extent, in the same way.

Panel (2,1) shows that for the very highest levels of solar radiation, ozone does not get high. Panels (3,1) and (4,1) show why. For the very highest levels of solar radiation, wind speed tends not to be low and temperature tends not to be high. In fact, there is a type of feedback mechanism at work here. The very highest levels of solar radiation at ground level can occur only on the brisk days with no air pollution, because when the pollution is present, the sun's rays are attenuated by particles in the air that form as part of the photochemistry.

Clearly, the scatterplot matrix has revealed much to us about the ozone and meteorological data.

A View of the Future: High-Interaction Graphical Methods

The computer graphics revolution has brought us into a new arena for graphing data. This does not mean simply that the ideas, methods, and principles of this book can be implemented in powerful, yet easy-to-use software systems, although that is surely true. It means more. Modern computer graphics has given us a new type of methodology: *high-interaction methods.* A person sitting in front of a computer screen now can have a high degree of interaction with a graph, changing it, even in a continuous way in real time, by using a physical device such as a light pen, a mouse, a graphics tablet, or even a finger. This capability gives us more than just a fast, convenient way to iterate to a single graph, just the way we want it. The changing of the graphical image on the screen can itself give information and be a graphical method, and we can see in just a few seconds what amounts to dozens of static graphs. There are many ways to change the graphical image on the screen, and they are all graphical methods.

Brushing a scatterplot matrix is a high-interaction graphical method that was invented in 1984 for analyzing multidimensional data [10]. Only a small part of the system will be described here; the reader should appreciate that it is no small challenge to describe a high-interaction computer graphical method, with dynamic elements that change in real time, on the static pages of a book.

Brushing a scatterplot matrix, as the name suggests, is based on the scatterplot matrix. This is illustrated in Figure 3.83 where three variables are graphed. The data are from an industrial experiment [43, p. 155] in which three measurements were made on each of thirty rubber specimens; the measurements are hardness, tensile strength, and

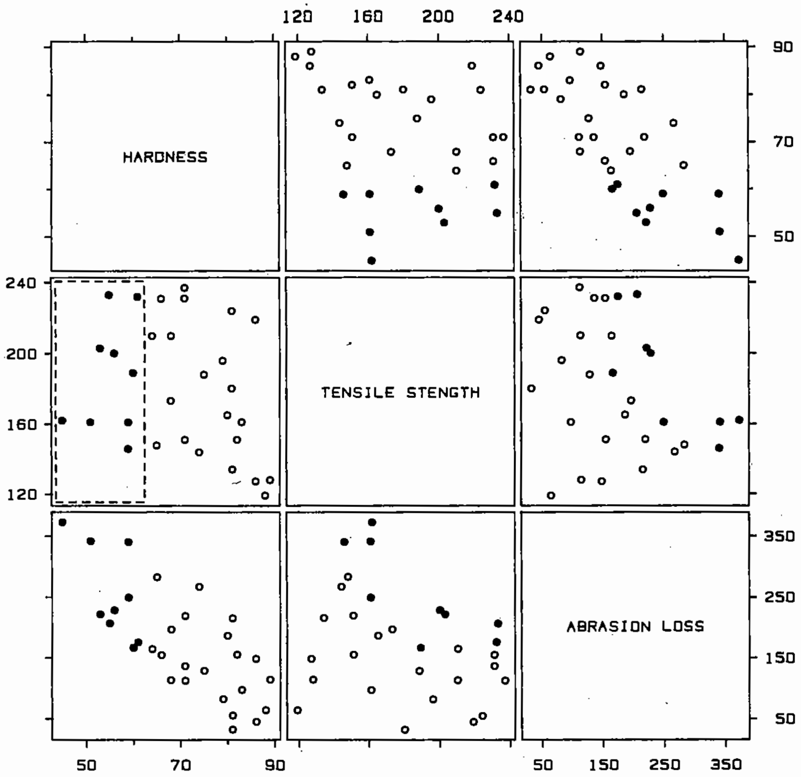

Figure 3.83 BRUSHING A SCATTERPLOT MATRIX: A HIGH-INTERACTION GRAPHICAL METHOD. High-interaction computer graphics is ushering in a new era in graphical methods for data analysis. This display appears on the screen of a graphics terminal. The brush is the dashed rectangle on the (2, 1) panel. Points selected by the brush are highlighted on all panels. The brush is moved by the user moving a mouse; as the brush moves, different points are selected and the highlighting changes instantaneously. In this figure points with low values of hardness are selected. The (3, 2) panel shows that for hardness held fixed to low values, abrasion loss depends nonlinearly on tensile strength.

abrasion loss, which is the amount of rubber rubbed off by an abrasive material. The goal of the experiment was to determine how abrasion loss depends on tensile strength and hardness, and in the original analysis, abrasion loss was modeled as a linear function of hardness and tensile strength [43, ch. 7]. In a later analysis, using some involved graphical statistical methods [21], it was discovered that abrasion loss depends nonlinearly on tensile strength, although it does depend linearly on hardness. Brushing a scatterplot matrix, however, gives us a very simple way of seeing the nonlinearity.

The principal high-interaction object in brushing is the *brush*: a rectangle on the screen, which is shown by dashed lines on the (2, 1) panel of Figure 3.83. The user moves the brush around the screen by moving a mouse, a physical device connected to the display terminal. The mouse is also used to change the size and shape of the brush.

Figure 3.84 shows one hardware configuration on which the brushing idea has been implemented. The young man in the front is holding a three-button mouse; the user moves the mouse on the table, which causes the brush to move on the screen. The high-interaction graphics code runs on the terminal, a Teletype 5620, but the preliminary data structuring is done on a supermicro, an AT&T 3B2 computer, which is underneath the display terminal.

Figure 3.83 shows the result of brushing when the *highlight* operation has been selected by a pop-up menu. The data in this example consist of 30 points in a three-dimensional space. Each panel in the figure is a projection of the points onto a plane. When the brush encloses graphed values on one panel it is in a sense selecting a subset of the points in three dimensions; the graphed values of these points are highlighted on all panels by graphing them using filled circles. As the brush is moved, different values are enclosed and the highlighting changes instantaneously. For example, in Figure 3.85 the brush has moved to the right on the (2,3) panel and different points are highlighted.

Let us now consider what this highlighting has shown us about the rubber data. In Figure 3.83 the brush was positioned so that points with low values of hardness are highlighted. Look at panel (3,2). The highlighted points are a graph of abrasion loss against tensile strength for low values of hardness; in other words, we see the dependence of abrasion loss on tensile strength with hardness held fixed, or nearly so. The highlighted points show that for hardness held to low values there is a nonlinear dependence of abrasion loss on tensile strength.

In Figure 3.85 middle values of hardness are selected. On the (3,2) panel the highlighted points show that for hardness held to middle levels the dependence of abrasion loss on tensile strength is again nonlinear and the pattern — a drop followed by a leveling out of the effect — is similar to that with hardness held to low values. The pattern emerges, although a little less crisply, in Figure 3.86, where hardness is held to high values.

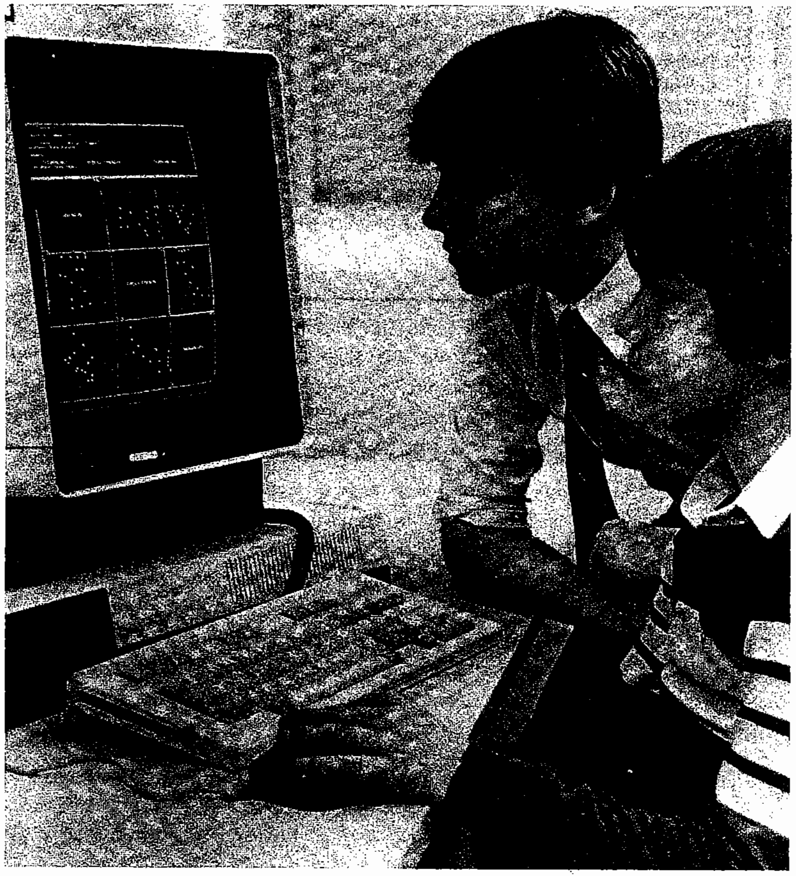

Figure 3.84 MOUSE, TERMINAL, AND COMPUTER. The young man in the front is holding the mouse, the device used to control the size and shape of the brush and to move it around the screen. The high-interaction graphics code runs on the terminal, a Teletype 5620, but the preliminary data structuring is done on a supermicro, an AT&T 3B2 computer, which is underneath the display terminal.

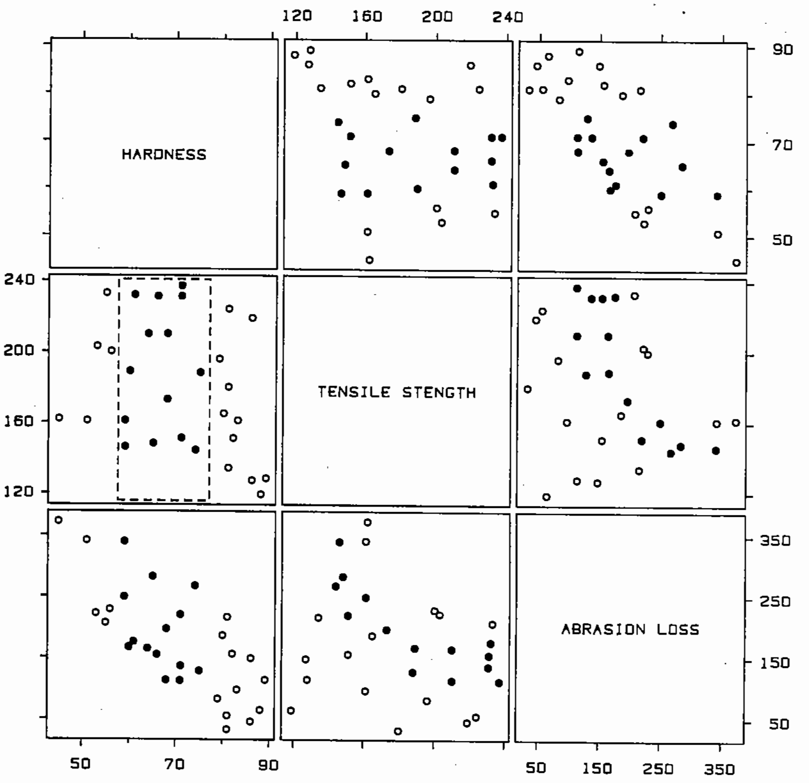

Figure 3.85 BRUSHING. Middle values of hardness have been selected. The highlighted values on the (3, 2) panel show that for hardness held fixed to middle levels, the dependence of abrasion loss on tensile strength is nonlinear.

The brushing has let us see easily the nonlinearity in these data. High-interaction graphical methods are now a reality. Graphical methods for data analysis have entered a new era.

3.7 STATISTICAL VARIATION

Measurements vary. Even when all controllable variables are kept constant, measurements vary because of uncontrollable variables or measurement error. One of the important functions of graphs in science and technology is to show the variation.

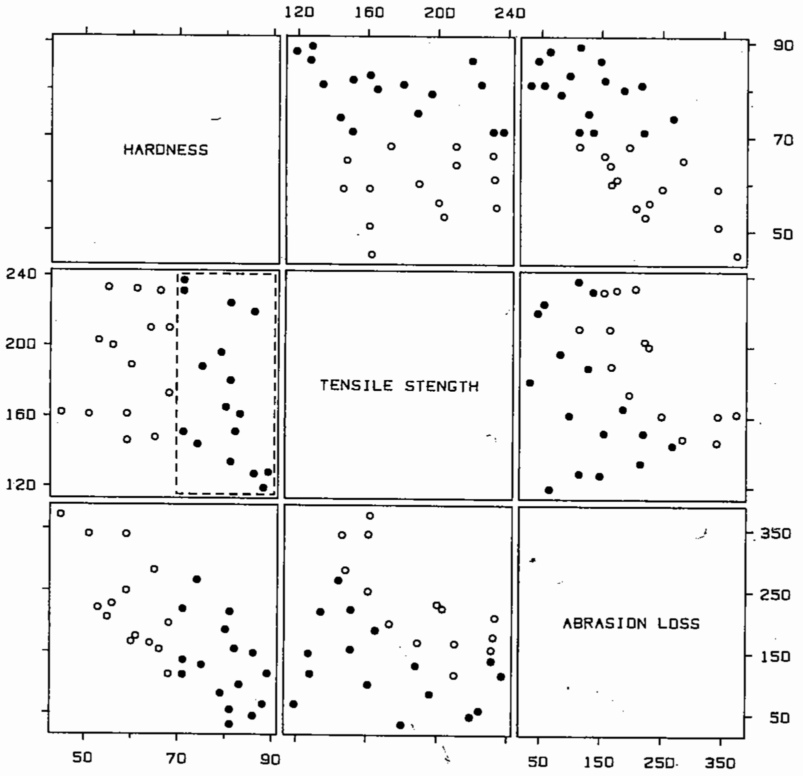

Figure 3.86 BRUSHING. High values of hardness have been selected. The highlighted values on the (3, 2) panel also suggest the nonlinearity.

Empirical Distribution of the Data

There are two very different domains of showing variation. One is to show the actual variation in the measurements, that is, to show the values of the data. This is the empirical distribution of the data that was discussed in Section 3.2. Figure 3.87 is an example from Section 3.4 — the bin-packing data. For each value of the *x* variable, the empirical distribution of the 25 values of the *y* variable is shown by a box graph.

When the goal is to convey just the empirical distribution of the data and not to make *formal* statistical inferences about a population distribution from which the data might have come, we can use the graphical methods for showing data distributions that were discussed in Section 3.2. The box graphs in Figure 3.87 are an example.

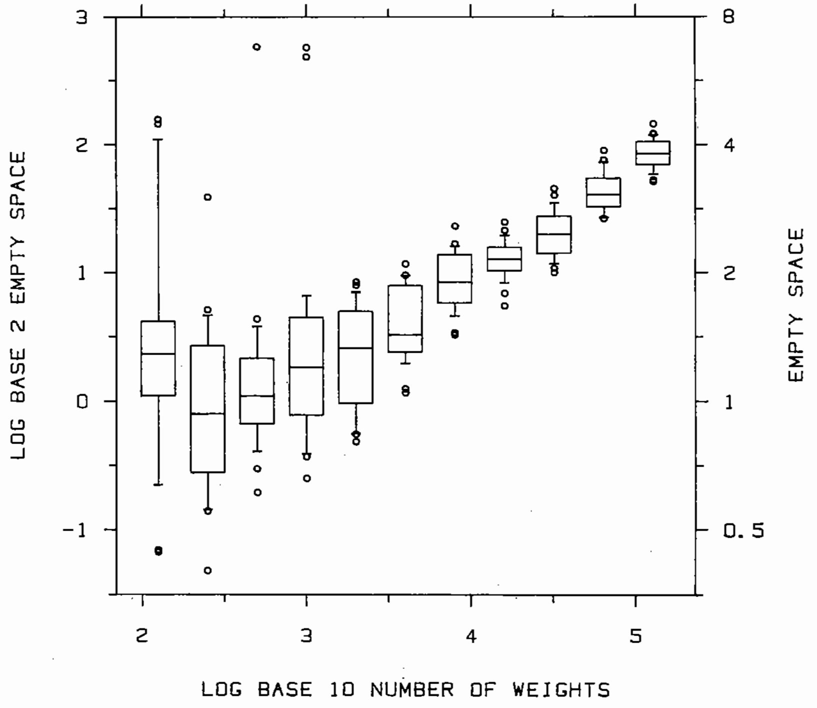

Figure 3.87 SHOWING EMPIRICAL VARIATION. For each value of log number of weights there are 25 measurements of log empty space whose distribution is summarized by a box graph.

Another method for showing the variation in the data, one that is very common in science and technology, is to use a plotting symbol and error bars to portray the *sample mean* and the *sample standard deviation.* Suppose the values of the data are $x_1,...,x_n$ then the sample mean is

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

and the sample standard deviation is

$$s = \left[\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right]^{1/2} .$$

Figure 3.88 uses a filled circle and error bars to show the mean plus and minus one sample standard deviation for each of the 11 data sets of the bin packing example. This graph does a poor job of conveying the variation in the data. The means show the centers of the distributions, but the standard deviations give us no sense of the upper and lower limits of the sample and camouflage the outliers: the unusually high values of empty space that occur for low numbers of weights. The box graphs in Figure 3.87 do a far better job of conveying the empirical variation of the data.

This result — the mean and sample standard deviation doing a poor job of conveying the distribution of the data — is frequently the case, because without any other information about the data, the sample standard deviation tells us little about where the data lie. This is further illustrated in Figure 3.89. The top panel shows four sets of made-up data. The four sets have the same sample size, the same sample mean, and the same sample standard deviation, but the behavior of the four empirical distributions is radically different. The means and sample standard deviations in the bottom panel do not capture the variation of the four data sets.

There is an exception to this poor performance of the sample standard deviation. If the empirical distribution of the data is well approximated by a normal probability distribution then we know approximately what percentage of the data lies between the mean plus and minus a constant times s. For example, approximately 68% lies between $\bar{x} \pm s$, approximately 50% lies between $\bar{x} \pm 0.67\,s$, and approximately 95% lies between $\bar{x} \pm 1.96\,s$. However, empirical distributions are often not well approximated by the normal. The normal distribution is symmetric, but real data are often skewed to the right. The normal distribution does not have wild observations, but real data often do.

One approach to showing the empirical variation in the data might be to check how well the empirical distribution is approximated by a normal, and then use the mean and sample standard deviation to summarize the distribution if the approximation is a good one. For example, one method for checking normality is a normal probability plot [21]. If the goal were to make inferences about the population distribution then checking normality is a vital matter and well worth the effort, as will be discussed shortly. But going through the trouble of checking normality, when the *only* goal is to show the empirical variation in the data, is often needless effort. The direct, easy, and rapid approach to showing the empirical variation in the data is to show the

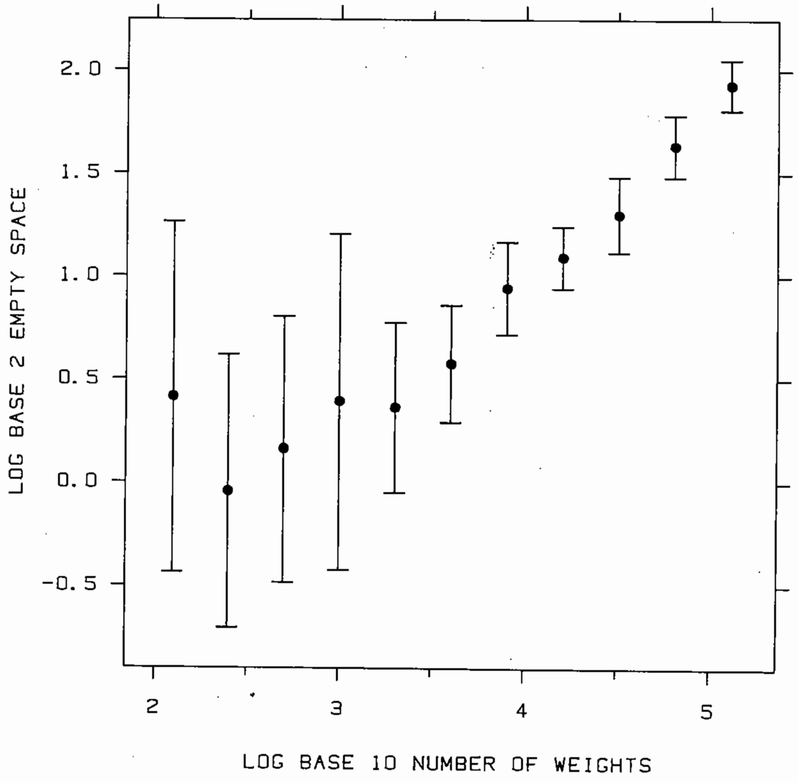

Figure 3.88 MEANS AND SAMPLE STANDARD DEVIATIONS. Showing just means and sample standard deviations is often a poor way to convey the variation in the data. This example shows means and sample standard deviations for the 11 sets of data graphed in Figure 3.87. The outliers in the data are not conveyed.

data. This means using graphical methods such as box graphs and percentile graphs to show the empirical distribution of the data. Thus, after this long discussion we have been led to the following circular advice: If the goal is to show the data, then show the data.

Sample-to-Sample Variation of a Statistic

The second domain of variation is the *sample-to-sample variation of a statistic*. Let us consider a simple but common sampling situation. Suppose we have a random sample of measurements, x_i for $i = 1$ to n, from a population distribution. Suppose we are interested in making inferences about the mean, μ, of the population distribution. The population mean can be estimated by the sample mean, $\bar{x}$, of the data.

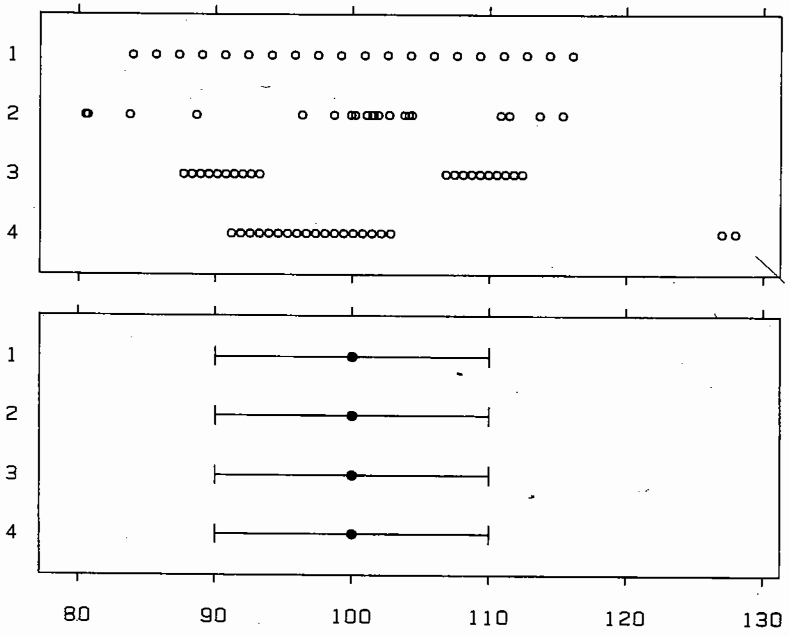

Figure 3.89 FAILURE OF MEANS AND SAMPLE STANDARD DEVIATIONS. Means and sample standard deviations cannot characterize the wide variety of distributions that data can have. Four sets of data are graphed in the top panel and their means and sample standard deviations are graphed in the bottom panel. The four distributions have the same numbers of observations, the same means, and the same sample standard deviations, but the distributions are very different.

The sample mean is a *statistic*, a numerical value based on the sample, and if we took a new sample of size n, $\bar{x}$ would be different; the variation in $\bar{x}$ from one sample of size n to the next is the sample-to-sample variation in $\bar{x}$.

$\bar{x}$ also has a population distribution and the sample-to-sample variation in $\bar{x}$ is characterized by it. Suppose σ is the standard deviation of the population distribution of the data, then the standard deviation of the population distribution of $\bar{x}$ is $\sigma/\sqrt{n}$. As n gets large this standard deviation gets small, the population distribution of $\bar{x}$ closes in on μ, and $\bar{x}$ varies less and less from sample to sample. The standard deviation of the mean, like μ, is unknown but it can be estimated; since s, the sample standard deviation, is an estimate of σ, $\sigma/\sqrt{n}$ can be estimated by $s/\sqrt{n}$, which is often called the *standard error of the mean*, although *estimated standard deviation of the sample mean* is a more complete name.

One-Standard-Error Bars

The current convention in science and technology for portraying sample-to-sample variation of a statistic is to graph error bars to portray plus and minus one standard error of the statistic, just the way the sample standard deviation is used to summarize the empirical variation of the data.

Figure 3.90 shows statistics from experiments on graphical perception that will be discussed in more detail in the next chapter. Subjects in the three experiments made graphical judgments that can be grouped into seven types. The types for each experiment are described by the labels in Figure 3.90. For each judgment type in each experiment a statistic was computed that measures the absolute error; the statistic is averaged across all subjects and across all judgments of that type made in the experiment. The filled circles in Figure 3.90 graph the statistics. The subjects in each experiment are thought of as a random sample from the population of subjects who can understand graphs. If we took new samples of subjects, the statistics shown in Figure 3.90 would vary. The error bars in Figure 3.90 show plus and minus one standard error of the statistics. (The statistics in this example are not means; the standard errors are computed from a formula that is more complicated than that for the standard error of the mean [35], however, we do not need to be concerned with the formula here.)

Now the critical point is the following: A standard error of a statistic has value only insofar as it conveys information about *confidence intervals*. The standard error by itself conveys little. It is confidence intervals that convey the sample-to-sample variation of a statistic.

In some cases confidence intervals are formed by taking plus and minus a multiple of the standard error. For example, suppose the x_i are a sample from a normal population distribution, suppose the statistic is $\bar{x}$, and suppose our purpose is to estimate the mean, μ, of the population

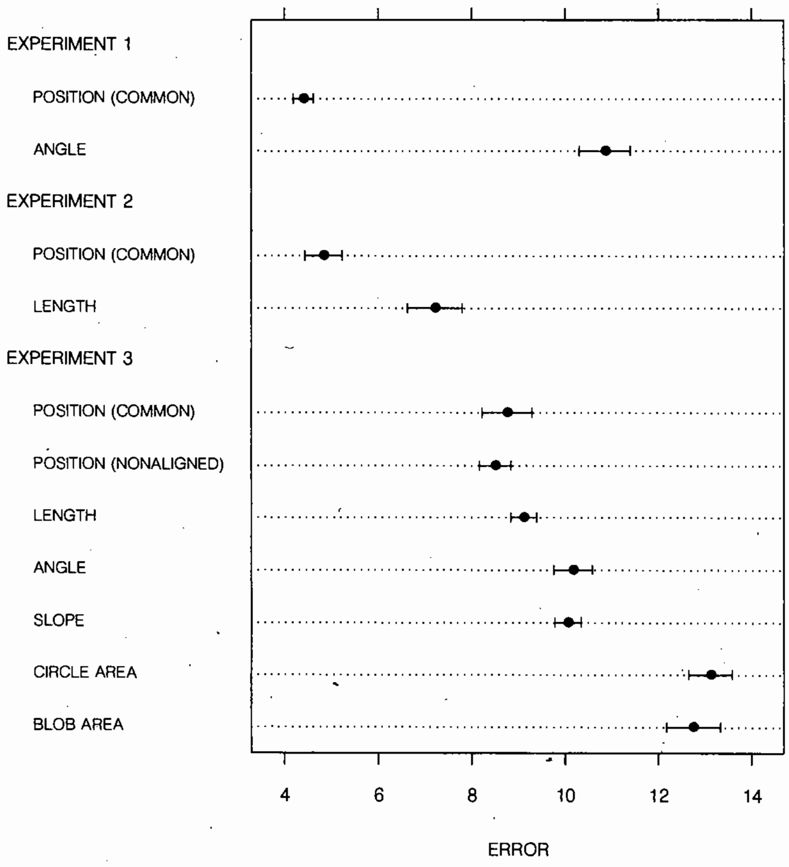

Figure 3.90 ONE-STANDARD-ERROR BARS TO SHOW SAMPLE-TO-SAMPLE VARIATION. The filled circles show statistics from experiments on graphical perception. Each error bar, conforming to the convention in science and technology, shows plus and minus one standard error. The interval formed by the error bars is a 68% confidence interval, which is not a particularly interesting interval. One standard error bars are probably a naive translation of the convention for numerical reporting of sample-to-sample variation.

distribution. Let $t_d(\alpha)$ be a number such that the probability between $-t_d(\alpha)$ and $t_d(\alpha)$ for a t-distribution with d degrees of freedom is α. Then the interval

$$\bar{x} - t_{n-1}(\alpha)s/\sqrt{n} \quad \text{to} \quad \bar{x} + t_{n-1}(\alpha)s/\sqrt{n}$$

is a $100\alpha\%$ confidence interval for the mean. In other words, μ is in the above interval for $100\alpha\%$ of the samples of size n drawn from the population distribution. This confidence interval is just the sample mean plus and minus a constant times the standard error of the mean. If n is about 60 or above, the t distribution is very nearly a normal distribution. This means

$$t_{n-1}(0.5) \approx 0.68 \qquad t_{n-1}(0.67) \approx 1 \qquad t_{n-1}(0.95) \approx 1.96\ ,$$

so in this case $\bar{x} \pm s/\sqrt{n}$ is approximately a 68% confidence interval, $\bar{x} \pm 0.67\, s/\sqrt{n}$ is approximately a 50% interval, and $\bar{x} \pm 1.96\, s/\sqrt{n}$ is approximately a 95% interval.

There are other sampling situations, however, where confidence intervals are *not* based on standard errors. For example, if the x_i are from an exponential distribution, then confidence intervals for the population mean are based on the sample mean, but they do not involve the standard error of the mean [86, p. 103].

How did it happen that the solidly entrenched convention in science and technology is to show one standard error on graphs? In some cases plus and minus one standard error has no useful, easy interpretation. True, in many cases plus and minus one standard error is a 68% confidence interval; Figure 3.90 is one example. Is a 68% confidence interval interesting? Are confidence intervals thought about at all when error bars are put on graphs?

It seems likely that the one-standard-error bar of graphical communication in science and technology is a result of the convention for numerical communication. If we want to communicate sample-to-sample variation numerically in cases where confidence intervals are based on standard errors, then it is reasonable to communicate the standard error and let the reader do some arithmetic, either mentally or otherwise, to get confidence intervals. A reasonable conjecture is that this numerical convention was simply brought to graphs. But the difficulty with this translation is that we are visually locked into what is shown by the error bars; it is hard to multiply the bars visually by some constant to get a desired visual confidence interval on the graph. Another difficulty, of course, is that confidence intervals are not always based on standard errors.

Two-Tiered Error Bars

Figure 3.91 uses *two-tiered error bars* to convey sample-to-sample variation. For each statistic the ends of the inner error bars, which are marked by the short vertical lines, are a 50% confidence interval; the ends of the outer error bars a 95% confidence interval. When confidence intervals are quoted numerically in scientific writings the level is almost always a high one such as 90%, 95%, or 99%; the outer interval in the two-tiered system simply reflects this practice. The inner interval of 50% gives a middle range for the sample-to-sample variation of the statistic that is analogous to the box of a box graph.

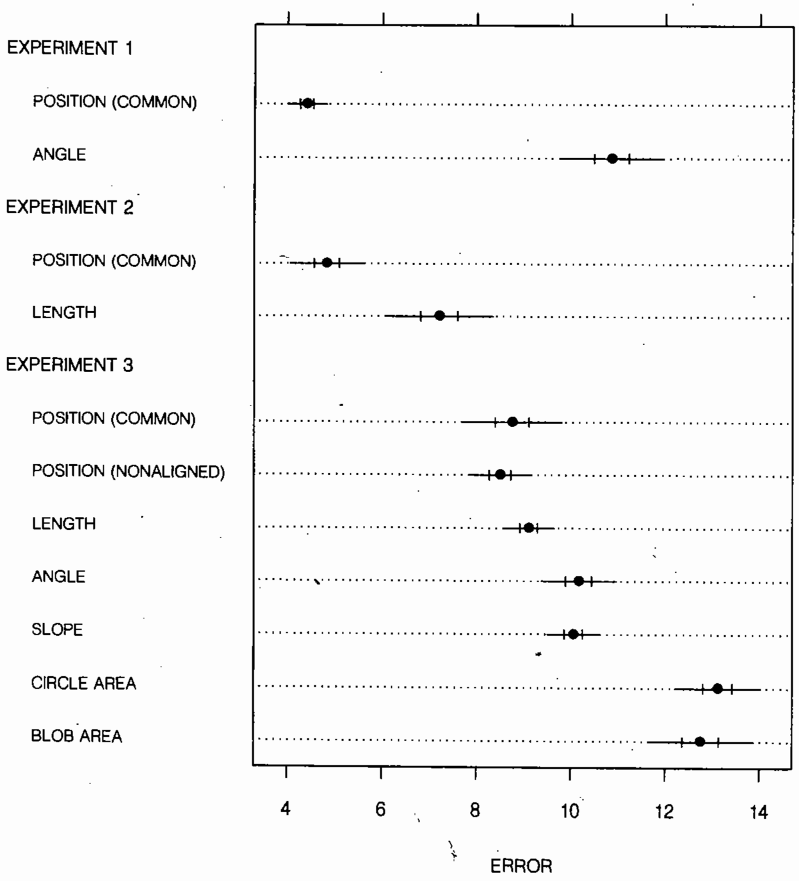

Figure 3.91 TWO-TIERED ERROR BARS. The outer error bars are 95% confidence intervals and the inner error bars are 50% confidence intervals. The goal in this method is to show confidence intervals and not standard errors, although for some statistics, confidence intervals happen to be formed from multiples of standard errors.

Two-tiered error bars are suggested as a replacement for one-standard-error bars. The most important aspect is that the goal is to show confidence intervals and not standard errors. Even when confidence intervals are based on standard errors, the two-tiered error bars are more sensible since they convey more cogent confidence interval information. The details of the two-tiered system are not meant to create dogma, but rather to encourage thought about what is shown. Variations should occur; for example, if an interval of very high confidence is desired, the ends of the outer bars could represent a 99.9% interval.