Chapter 3

Empirical Distributions and Exploratory Data Analysis

3.1. BACKGROUND

One very important application of statistical ideas in meteorology and climatology is in making sense of a new set of data. Ultimately, the goal is to extract insight about the processes underlying the generation of the numbers. As mentioned in Chapter 1, meteorological observing systems and computer models, supporting both operational and research efforts, produce torrents of numerical data. Many of these large data sets are easily available through the internet, on such websites as www.ncdc.noaa.gov, www.ecmwf.int/en/forecasts/datasets, www.data.gov, and www.data.gov.uk. Even when working with a rela tively small data set, it can be a significant task just to get a feel for a new batch of numbers, and to begin to make some sense of them.

Broadly speaking, this activity is known as *Exploratory Data Analysis*, or EDA. Its systematic use increased substantially following Tukey's (1977) pathbreaking and very readable book of the same name. EDA methods draw heavily on a variety of graphical tools to aid in the comprehension of the large batches of numbers that may confront an analyst. Graphics are a very effective means of compressing and summarizing data, portraying much in little space, and exposing unusual features of a data set. Sometimes unusual data points result from errors in recording or transcription, and it is well to know about these as early as possible in an analysis. Sometimes the unusual data are valid and may turn out to be the most interesting and informative parts of a data set.

Many EDA methods were designed originally to be applied by hand, with pencil and paper, to small (up to perhaps 200-point) data sets. Modern computing capabilities have greatly broadened the scope of statistical graphics, a large variety of which are easily available (e.g., R Development Core Team, 2017; Theus and Urbanek, 2009).

3.1.1. Robustness and Resistance

Many of the classical techniques of statistics work best when fairly stringent assumptions about the nature of the data are met. For example, it is often assumed that data will follow the familiar bell-shaped curve of the Gaussian distribution (Section 4.4.2). Classical procedures can behave very badly (i.e., produce quite misleading results) if their assumptions are not satisfied by the data to which they are applied.

The assumptions of classical statistics were not made out of ignorance, but rather out of necessity. Invocation of simplifying assumptions in statistics, as in other fields, has allowed progress to be made

Statistical Methods in the Atmospheric Sciences. https://doi.org/10.1016/B978-0-12-815823-4.00003-1

through the derivation of elegant analytic results, which are relatively simple but powerful mathematical formulas. As has also been the case in many quantitative fields, the advent of cheap computing power has freed the data analyst from sole dependence on such results, by allowing alternatives requiring less stringent assumptions to become practical. This does not mean that the classical methods are no longer useful. However, it is much easier to check that a given set of data satisfies particular assumptions before a classical procedure is used, and good alternatives are computationally feasible in cases where the classical methods may not be appropriate.

Two important properties of EDA methods are that they are *robust* and *resistant*. Robustness and resistance are two aspects of reduced sensitivity to assumptions about the nature of a set of data. A robust method is not necessarily optimal in any particular circumstance, but performs reasonably well in most circumstances. For example, the sample average is the best characterization of the center of a set of data if it is known that those data follow a Gaussian distribution. However, if those data are decidedly non-Gaussian (e.g., if they are a record of extreme rainfall events), the sample average may yield a misleading characterization of their center. In contrast, robust methods generally are not sensitive to particular assumptions about the overall nature of the data.

A resistant method is not unduly influenced by a small number of outliers, or "wild data." As indicated previously, such points often show up in a batch of data through errors of one kind or another. The results of a resistant method change very little if a small fraction of the data values are changed, even if they are changed drastically. In addition to not being robust, the sample average is not a resistant characterization of the center of a data set, either. Consider the small set {11, 12, 13, 14, 15, 16, 17, 18, 19}. Its average is 15. However, if instead the set {11, 12, 13, 14, 15, 16, 17, 18, 91} had resulted from a transcription error, the "center" of the data (erroneously) characterized using the sample average would be 23. Resistant measures of the center of a batch of data, such as those to be presented later, would be changed little or not at all by the substitution of "91" for "19" in this simple example.

3.1.2. Quantiles

Many common summary measures rely on the use of selected sample *quantiles* (also known as *fractiles*). Quantiles and fractiles are essentially equivalent to the more familiar term, *percentile*. A sample quantile, q_p, is a number having the same units as the data, which exceeds that proportion of the data given by the subscript p, with $0 < p < 1$. The sample quantile q_p can be interpreted approximately as that value expected to exceed a randomly chosen member of the data set, with probability p. Equivalently, the sample quantile q_p would be regarded as the $p \times 100$th percentile of the data set.

The determination of sample quantiles requires that a batch of data first be arranged in order. Sorting small sets of data by hand presents little problem. Sorting larger sets of data is best accomplished by computer. Historically, the sorting step presented a major bottleneck in the application of robust and resistant procedures to large data sets. Today the sorting can be done easily using either a spreadsheet or data analysis program on a desktop computer, or one of many sorting algorithms available in collections of general-purpose computing routines (e.g., Press et al., 1986).

The sorted, or ranked, data values from a particular sample are called the *order statistics* of that sample. Given a set of data $\{x_1, x_2, x_3, x_4, x_5, \ldots, x_n\}$, the order statistics for this sample would be the same numbers, sorted in ascending order. These sorted values are conventionally denoted using parenthetical subscripts, that is, by the set $\{x_{(1)}, x_{(2)}, x_{(3)}, x_{(4)}, x_{(5)}, \ldots, x_{(n)}\}$. Here the ith smallest of the n data values is denoted $x_{(i)}$.

Certain sample quantiles are used especially often in the exploratory summarization of data. The *median*, or $q_{0.5}$, or 50th percentile, is most commonly used. This is the value at the center of the data set, in the sense that equal proportions of the data fall above and below it. If a data set contains an odd number of values, the median is simply the middle order statistic. If there are an even number, however, the data set has two middle values. In this case the median is conventionally taken to be the average of these two middle values. Formally,

$$q_{0.5} = \begin{cases} x_{([n+1]/2)} & , n \text{ odd} \\ \dfrac{x_{(n/2)} + x_{([n/2]+1)}}{2} & , n \text{ even} \end{cases}. \tag{3.1}$$

The *quartiles*, $q_{0.25}$ and $q_{0.75}$, are almost as commonly used as the median. Usually these are called the lower (LQ) and upper quartiles (UQ), respectively. They are located half-way between the median, $q_{0.5}$, and the extremes, $x_{(1)}$ and $x_{(n)}$. In typically colorful terminology, Tukey (1977) calls $q_{0.25}$ and $q_{0.75}$ the "*hinges*," apparently imagining that the data set can be folded first at the median, and then at the quartiles. The quartiles are thus the two medians of the half-data sets between $q_{0.5}$ and the extremes. If n is odd, these half-data sets each consist of $(n+1)/2$ points, and both include the median. If n is even these half-data sets each contain $n/2$ points, and do not overlap. The upper and lower *terciles*, $q_{0.333}$ and $q_{0.667}$, separate a data set into thirds, although sometimes the term tercile is used also to refer to any of the three equally sized portions of the data set so defined. Other quantiles that also are used frequently enough to be named are the four *quintiles*, $q_{0.2}$, $q_{0.4}$, $q_{0.6}$, and $q_{0.8}$; the *eighths*, $q_{0.125}$, $q_{0.375}$, $q_{0.625}$, and $q_{0.875}$ (in addition to the quartiles and median); and the nine *deciles*, $q_{0.1}$, $q_{0.2}$, …, $q_{0.9}$.

Example 3.1. Computation of Common Quantiles

If there are $n = 9$ data values in a batch of data, the median is $q_{0.5} = x_{(5)}$, or the fifth largest of the nine. The lower quartile is $q_{0.25} = x_{(3)}$, and the upper quartile is $q_{0.75} = x_{(7)}$.

If $n = 10$, the median is the average of the two middle values, and the quartiles are the single middle values of the upper and lower halves of the data. That is, $q_{0.25}$, $q_{0.5}$, and $q_{0.75}$ are $x_{(3)}$, $[x_{(5)}+x_{(6)}]/2$, and $x_{(8)}$, respectively.

If $n = 11$ then there is a unique middle value, but the quartiles are determined by averaging the two middle values of the upper and lower halves of the data. That is, $q_{0.25}$, $q_{0.5}$, and $q_{0.75}$ are $[x_{(3)}+x_{(4)}]/2$, $x_{(6)}$, and $[x_{(8)}+x_{(9)}]/2$, respectively.

For $n = 12$ both quartiles and the median are determined by averaging pairs of middle values; $q_{0.25}$, $q_{0.5}$, and $q_{0.75}$ are $[x_{(3)}+x_{(4)}]/2$, $[x_{(6)}+x_{(7)}]/2$, and $[x_{(9)}+x_{(10)}]/2$, respectively. ◇

Estimating quantiles other than the median and quartiles involves use of more elaborate formulas, several choices for which are available (Hyndman and Fan, 1996).

3.2. NUMERICAL SUMMARY MEASURES

Some simple robust and resistant numerical summary measures are available that can be used without hand plotting or computer graphic capabilities. Often these will be the first quantities to be computed from a new and unfamiliar set of data. The next three subsections describe numerical summary measures of *location*, *spread*, and *symmetry*. Location refers to the central tendency or general magnitude of the data values. Spread denotes the degree of variation or dispersion around the center. Symmetry describes the balance with which the data values are distributed about their center. Asymmetric data tend to spread more either on the high side (have a long right tail) or the low side (have a long left tail). These three types

of numerical summary measures correspond to the first three statistical moments of a data sample, but the classical measures of these moments (i.e., the sample mean, sample variance, and sample coefficient of skewness, respectively) are neither robust nor resistant.

3.2.1. Location

The median, $q_{0.5}$, is the most common robust and resistant measure of central tendency. Consider again the data set {11, 12, 13, 14, 15, 16, 17, 18, 19}. The median and mean are both 15. If, as noted before, the "19" is replaced erroneously by "91," the *mean*

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3.2}$$

(=23) is very strongly affected, illustrating its lack of resistance to outliers. The median is unchanged by this common type of data error.

The *trimean* is a slightly more complicated measure of location, which takes into account more information about the magnitudes of the data. It is a weighted average of the median and the quartiles, with the median receiving twice the weight of each of the quartiles:

$$\text{Trimean} = \frac{q_{0.25} + 2q_{0.5} + q_{0.75}}{4}. \tag{3.3}$$

The *trimmed mean* is another resistant measure of location, whose sensitivity to outliers is reduced by removing a specified proportion of the largest and smallest observations. If the proportion of observations omitted at each end of the data distribution is α, then the α-trimmed mean is

$$\bar{x}_\alpha = \frac{1}{n-2k}\sum_{i=k+1}^{n-k} x_{(i)}, \tag{3.4}$$

where k is an integer rounding of the product αn, the number of data values "trimmed" from each tail. The trimmed mean reduces to the ordinary mean (Equation 3.2) for $\alpha = 0$.

Other methods of characterizing location can be found in Andrews et al. (1972), Goodall (1983), Rosenberger and Gasko (1983), and Tukey (1977).

3.2.2. Spread

The *Interquartile Range* (IQR) is the most common, and simplest, robust and resistant measure of spread (also known as dispersion or scale). The IQR is simply the difference between the upper and lower quartiles:

$$\text{IQR} = q_{0.75} - q_{0.25}. \tag{3.5}$$

The IQR is a good index of the spread in the central part of a data set, since it simply specifies the range of the central 50% of the data. The fact that it ignores the upper and lower 25% of the data makes it quite resistant to outliers. This quantity is sometimes also called the *fourth-spread*.

It is worthwhile to compare the IQR with the conventional measure of scale of a data set, the *sample standard deviation*

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}. \tag{3.6}$$

The square of the sample standard deviation, s^2, is known as the *sample variance*. Because of the square root in Equation 3.6, the standard deviation has the same physical dimensions as the underlying data. The standard deviation is neither robust nor resistant. It is very nearly just the square root of the average squared difference between the data points and their sample mean. (The division by $n-1$ rather than n often is done in order to compensate for the fact that the x_i are closer, on average, to their sample mean than to the true population mean: dividing by $n-1$ exactly counters the resulting tendency for the sample standard deviation to be too small, on average.) Even one very large data value will be felt very strongly because it will be especially far away from the mean and that difference will be magnified by the squaring process. Consider again the set {11, 12, 13, 14, 15, 16, 17, 18, 19}. The sample standard deviation is 2.74, but it is greatly inflated to 25.6 if "91" erroneously replaces "19." It is easy to see that in either case IQR = 4.

The IQR is very easy to compute, but it does have the disadvantage of not making much use of a substantial fraction of the data. The *median absolute deviation* (MAD) is a more complete, yet reasonably simple alternative. The MAD is easiest to understand by imagining the transformation $y_i = |x_i - q_{0.5}|$. Each transformed value y_i is the absolute value of the difference between the corresponding original data value and the median. The MAD is then just the median of the transformed (y_i) values:

$$\text{MAD} = median\,|x_i - q_{0.5}|. \tag{3.7}$$

Although this process may seem a bit elaborate at first, a little thought illustrates that it is analogous to computation of the standard deviation, but using operations that do not emphasize outlying data. The median (rather than the mean) is subtracted from each data value, any negative signs are removed by the absolute value (rather than squaring) operation, and the center of these absolute differences is located by their median (rather than their mean).

The *trimmed variance* is a still more elaborate measure of spread. The idea, as for the trimmed mean (Equation 3.4), is to omit a proportion of the largest and smallest values and compute the analog of the sample variance (the square of Equation 3.6)

$$s_\alpha^2 = \frac{1}{n-2k}\sum_{i=k+1}^{n-k}\left(x_{(i)} - \overline{x}_\alpha\right)^2. \tag{3.8}$$

Again, k is the nearest integer to αn, and squared deviations from the consistent trimmed mean (-Equation 3.4) are averaged. The trimmed variance is sometimes multiplied by an adjustment factor to make it more consistent with the ordinary sample variance, s^2 (Graedel and Kleiner, 1985).

Other measures of spread can be found in Hosking (1990) and Iglewicz (1983).

3.2.3. Symmetry

The *sample skewness coefficient* is the conventional moments-based measure of symmetry in a batch of data,

$$\gamma = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^3}{s^3}. \tag{3.9}$$

This statistic is neither robust nor resistant. The numerator is similar to the sample variance, except that the average is over cubed deviations from the mean. Thus the sample skewness coefficient is even more sensitive to outliers than is the standard deviation. The average cubed deviation in the numerator is divided by the cube of the sample standard deviation in order to standardize and nondimensionalize the skewness coefficient, so that comparisons of skewness among different data sets are more meaningful.

Notice that cubing differences between the data values and their mean preserves the signs of these differences. Since the differences are cubed, the data values farthest from the mean will dominate the sum in the numerator of Equation 3.9. If there are a few very large data values, the sample skewness will tend to be positive. Therefore batches of data with long right tails are referred to both as right skewed and positively skewed. Data that are physically constrained to lie above a minimum value (such as precipitation or wind speed, both of which must be nonnegative) are often positively skewed. Conversely, if there are a few very small (or large negative) data values, these will fall far below the mean. The sum in the numerator of Equation 3.9 will then be dominated by a few large negative terms, so that the sample skewness coefficient will tend to be negative. Data with long left tails are referred to as left skewed or negatively skewed. For essentially symmetric data, the skewness coefficient will be near zero.

The *Yule-Kendall index*,

$$\gamma_{YK} = \frac{(q_{0.75} - q_{0.5}) - (q_{0.5} - q_{0.25})}{\text{IQR}} = \frac{q_{0.25} - 2q_{0.5} + q_{0.75}}{\text{IQR}}, \tag{3.10}$$

is a robust and resistant alternative to the sample skewness. It is computed by comparing the distance between the median and each of the two quartiles. If the data are right skewed, at least in the central 50% of the data, the distance to the median will be greater from the upper quartile than from the lower quartile. In this case the Yule-Kendall index will be greater than zero, consistent with the usual convention of right skewness being positive. If the positive skewness is strong enough that $q_{0.25} = q_{0.50}$, then $\gamma_{YK} = 1$. Conversely, left-skewed data will be characterized by a negative Yule-Kendall index, and if the negative skewness is sufficiently strong that $q_{0.75} = q_{0.50}$, then $\gamma_{YK} = -1$. Analogously to Equation 3.9, division by the interquartile range nondimensionalizes γ_{YK} (i.e., scales it in a way that the physical dimensions, such as meters or millibars, cancel) and thus improves its comparability between data sets.

Alternative measures of skewness can be found in Brooks and Carruthers (1953) and Hosking (1990).

3.2.4. Kurtosis

Extending Equation 3.9 by increasing the exponent leads to the coefficient of *kurtosis*,

$$\kappa = \frac{\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})^4}{s^4} - 3. \tag{3.11}$$

Although this summary statistic is often characterized as reflecting "flatness" or "peakedness" of a data distribution, it is really a measure of the weights of the upper and lower tails of a distribution relative to the distribution center (Westfall, 2014). Distributions for which $\kappa > 0$ are termed *leptokurtic* and have relatively heavy tails. Distributions for which $\kappa < 0$ are termed *platykurtic* and have relatively light tails. The subtraction of 3 in Equation 3.11 is a convention that allows comparison with the kurtosis of the

Gaussian, or bell-curve distribution, for which $\kappa = 0$. Accordingly, Equation 3.11 is often also known as *excess kurtosis*.

3.3. GRAPHICAL SUMMARY DEVICES

Numerical summary measures are quick and easy to compute and display, but they can express only a small amount of detail. In addition, their visual impact is limited. Graphical displays for exploratory data analysis have been devised that require only slightly more effort to produce.

3.3.1. Stem-and-Leaf Display

The *stem-and-leaf display* is a very simple but effective tool for producing an overall view of a new set of data. At the same time it provides the analyst with an initial exposure to the individual data values. In its simplest form, the stem-and-leaf display groups the data values according to their all-but-least significant digits. These values are written in either ascending or descending order to the left of a vertical bar, constituting the "stems." The least significant digit for each data value is then written to the right of the vertical bar, on the same line as the more significant digits with which it belongs. These least significant values constitute the "leaves."

Figure 3.1a shows a stem-and-leaf display for the January 1987 Ithaca maximum temperatures in Table A.1. The data values are reported to whole degrees and range from 9°F to 53°F. The all-but-least significant digits are thus the tens of degrees, which are written to the left of the bar. The display is built up by proceeding through the data values one by one, and writing its least significant digit on the appropriate line. For example, the temperature for 1 January is 33°F, so the first "leaf" to be plotted is the first "3" on the stem of temperatures in the 30s. The temperature for 2 January is 32°F, so a "2" is written to the right of the "3" just plotted for 1 January.

The initial stem-and-leaf display for this particular data set is a bit crowded, since most of the values are in the 20s and 30s. In cases like this, better resolution can be obtained by constructing a second plot, like that in Figure 3.1b, in which each stem has been split to contain only the values 0–4 or 5–9. Sometimes the opposite problem will occur, and the initial plot is too sparse. In that case (if there are at least three significant digits), replotting can be done with stem labels omitting the two least significant digits. Less stringent groupings can also be used. Regardless of whether or not it may be desirable to split or

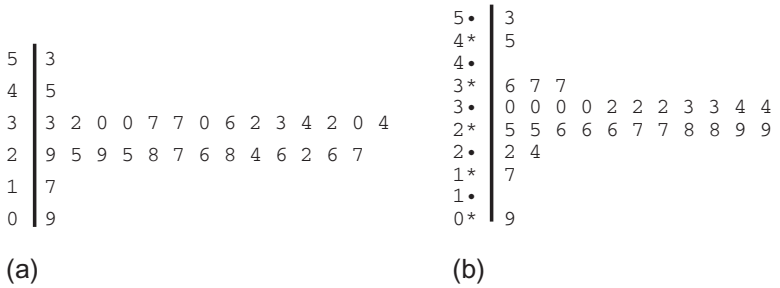


FIGURE 3.1 Stem-and-leaf displays for the January 1987 Ithaca maximum temperatures in Table A.1. The plot in panel (a) results after a first pass through the data, using the 10s as "stem" values. Optionally, a bit more resolution is obtained in panel (b) by creating separate stems for least-significant digits from 0 to 4 (•), and from 5 to 9 (*). At this stage it is also easy to sort the data values before rewriting them.

consolidate stems, it can be useful to rewrite the display with the leaf values sorted, as has also been done in Figure 3.1b.

The stem-and-leaf display is much like a quickly plotted histogram of the data, placed on its side. In Figure 3.1, for example, it is evident that these temperature data are reasonably symmetrical, with most of the values falling in the upper 20s and lower 30s. Optionally, sorting the leaf values also facilitates extraction of quantiles of interest. In this case it is easy to count inward from the extremes in Figure 3.1b to find that the median is 30, and that the two quartiles are 26 and 33.

It can happen that there are one or more outlying data points that are far removed from the main body of the data set. Rather than plot many empty stems, it is usually more convenient to just list these extreme values separately at the upper and/or lower ends of the display, as in Figure 3.2. This display is of data of wind speeds in kilometers per hour (km/h) to the nearest tenth. Merely listing two extremely large values and two values of calm winds at the top and bottom of the plot has reduced the length of the display by more than half. It is quickly evident that the data are strongly skewed to the right, as often occurs for wind data.

The stem-and-leaf display in Figure 3.2 also reveals something that might have been missed in a tabular list of the daily data. All the leaf values on each stem are the same. Evidently a rounding process has been applied to the data, knowledge of which could be important to some subsequent analyses. In this case the rounding process consists of transforming the data from the original units (knots) to km/h. For example, the four observations of 16.6 km/h result from original observations of 9 knots. No values on the 17 km/h stem would be possible, since observations of 10 knots transform to 18.5 km/h.

3.3.2. Boxplots

The *boxplot*, or *box-and-whisker plot*, is a very widely used graphical tool introduced by Tukey (1977). It is a simple plot of five sample quantiles: the minimum, $x_{(1)}$; the lower quartile, $q_{0.25}$; the median, $q_{0.5}$; the upper quartile, $q_{0.75}$; and the maximum, $x_{(n)}$. Using these five numbers, the boxplot essentially presents a quick sketch of the distribution of the underlying data.

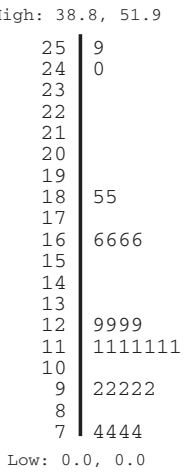


FIGURE 3.2 Stem-and-leaf display of 0100 local-time wind speeds (km/h) at the Newark, New Jersey, Airport during December 1974. Very high and very low values are written outside the plot itself to avoid having many blank stems. The striking grouping of repeated leaf values suggests that a rounding process has been applied to the original observations. *From Graedel and Kleiner (1985).*

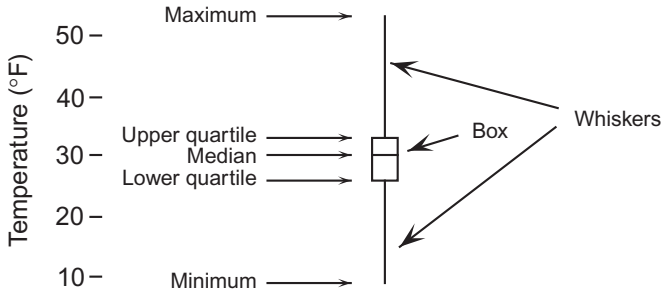


FIGURE 3.3 A simple boxplot, or box-and-whiskers plot, for the January 1987 Ithaca maximum temperature data. The upper and lower ends of the box are drawn at the quartiles, and the bar through the box is drawn at the median. The whiskers extend from the quartiles to the maximum and minimum data values. The arrows and associated labels are shown here only to define the plot attributes and are not included in practice.

Figure 3.3 shows a boxplot for the January 1987 Ithaca maximum temperature data in Table A.1. The box in the middle of the diagram is bounded by the upper and lower quartiles, and thus locates the central 50% of the data. The bar inside the box locates the median. The whiskers extend away from the box to the two extreme values.

Boxplots can convey a surprisingly large amount of information at a glance. It is clear from the small range occupied by the box in Figure 3.3, for example, that the data are concentrated quite near 30°F. Being based only on the median and the quartiles, this portion of the boxplot is highly resistant to any outliers that might be present. The full range of the data is also apparent at a glance. Finally, we can see easily that these data are nearly symmetrical, since the median is near the center of the box, and the whiskers are of comparable lengths.

3.3.3. Schematic Plots

A shortcoming of the boxplot is that information about the tails of the data distribution is highly generalized. The whiskers extend to the highest and lowest values, but there is no information about the distribution of data points within the upper and lower quarters of the data. For example, although Figure 3.3 shows that the highest maximum temperature is 53°F, it gives no information as to whether this is an isolated point (with the remaining warm temperatures cooler than, say, 40°F) or whether the warmer temperatures are more or less evenly distributed between the upper quartile and the maximum.

It is often useful to have some idea of the degree of unusualness of the extreme values. The *schematic plot*, which was also originated by Tukey (1977), is a refinement of the boxplot that presents more detail in the tails. The schematic plot is identical to the boxplot, except that extreme points deemed to be sufficiently unusual are plotted individually. Just how extreme is sufficiently unusual depends on the variability of the data in the central part of the sample, as reflected by the IQR. A given extreme value is regarded as being less unusual if the two quartiles are far apart (i.e., if the IQR is large), and more unusual if the two quartiles are near each other (the IQR is small).

The dividing lines between less- and more-unusual points are known in Tukey's idiosyncratic terminology as the "*fences*." Four fences are defined: inner and outer fences, above and below the data, according to

$$\begin{aligned}
\text{Upper outer fence} &= q_{0.75} + 3\,IQR \\
\text{Upper inner fence} &= q_{0.75} + \frac{3\,IQR}{2} \\
\text{Lower inner fence} &= q_{0.25} - \frac{3\,IQR}{2} \\
\text{Lower outer fence} &= q_{0.25} - 3\,IQR.
\end{aligned} \tag{3.12}$$

Thus the two outer fences are located three times the distance of the interquartile range above and below the two quartiles. The inner fences are midway between the outer fences and the quartiles, being 1.5 times the distance of the interquartile range away from the quartiles.

In the schematic plot, points within the inner fences are called "inside." The range of the inside points is shown by the extent of the whiskers. Data points between the inner and outer fences are referred to as being "outside" and are plotted individually in the schematic plot. Points above the upper outer fence or below the lower outer fence are called "far out" and are plotted individually with a different symbol. These automatically generated boundaries, while somewhat arbitrary, have been informed by Tukey's experience and intuition. The resulting differences from the simple boxplot are illustrated in Figure 3.4. In common with the boxplot, the box in a schematic plot shows the locations of the quartiles and the median.

Example 3.2. Construction of a Schematic Plot

Figure 3.4 is a schematic plot for the January 1987 Ithaca maximum temperature data. As can be determined from Figure 3.1, the quartiles for these data are 33°F and 26°F, and the IQR = 33 − 26 = 7°F. From this information it is easy to compute the locations of the inner fences at 33 + (3/2)(7) = 43.5°F and 26 − (3/2)(7) = 15.5°F. Similarly, the outer fences are 33 + (3)(7) = 54°F and 26 − (3)(7) = 5°F. The dashed lines locating the fences are normally not included in schematic plots, but have been shown in Figure 3.4 for clarity.

The two warmest temperatures, 53°F and 45°F, are larger than the upper inner fence, and are shown individually by circles. The coldest temperature, 9°F, is less than the lower inner fence, and is also plotted individually. The whiskers are drawn to the most extreme temperatures inside the fences, 37°F

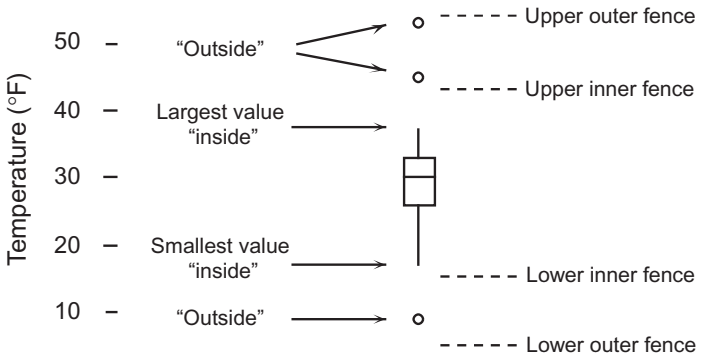


FIGURE 3.4 A schematic plot for the January 1987 Ithaca maximum temperature data. The central box portion of the figure is identical to the boxplot of the same data in Figure 3.3. The three values outside the inner fences are plotted separately. None of the values are beyond the outer fences, or "far out." Notice that the whiskers extend to the most extreme "inside" data values, and not to the fences. Dashed lines locating the fences are shown here for clarity, but are not normally included in a schematic plot. The arrows, dashed lines, and associated labels are shown only to define the plot attributes and are not included in practice.

and 17°F. If the warmest temperature had been 55°F rather than 53°F, it would have fallen outside the outer fence (far out) and would have been plotted individually with a different symbol. This separate symbol for the far out points is often an asterisk. ◇

One important use of schematic plots or boxplots is simultaneous graphical comparison of several batches of data. This use of schematic plots is illustrated in Figure 3.5, which shows side-by-side schematic plots for all four of the batches of temperature data in Table A.1. Of course it is known in advance that the maximum temperatures are warmer than the minimum temperatures, and comparing their schematic plots brings out this difference quite strongly. Apparently, Canandaigua was slightly warmer than Ithaca during this month, and more strongly so for the minimum temperatures. The Ithaca minimum temperatures were evidently more variable than the Canandaigua minimum temperatures. For both locations, the minimum temperatures are more variable than the maximum temperatures, especially in the central parts of the distributions represented by the boxes. The location of the median in the upper ends of the boxes in the minimum temperature schematic plots suggests a tendency toward negative skewness, as does the inequality of the whisker lengths for the Ithaca minimum temperature data. The maximum temperatures appear to be reasonably symmetrical for both locations. Note that none of the minimum temperature data are outside the inner fences, so that boxplots of the same data would be identical.

3.3.4. Other Boxplot Variants

Two variations on boxplots or schematic plots suggested by McGill et al. (1978) are sometimes used, particularly when comparing side-by-side plots. The first is to plot each box width proportional to $\sqrt{n}$. This simple variation allows plots for data having larger sample sizes to stand out and give a stronger visual impact.

The notched boxplot or schematic plot is a second variant. The boxes in these plots resemble hourglasses, with the constriction, or waist, located at the median. The lengths of the notched portions of the box differ from plot to plot, reflecting estimates of preselected confidence limits (Chapter 5) for the

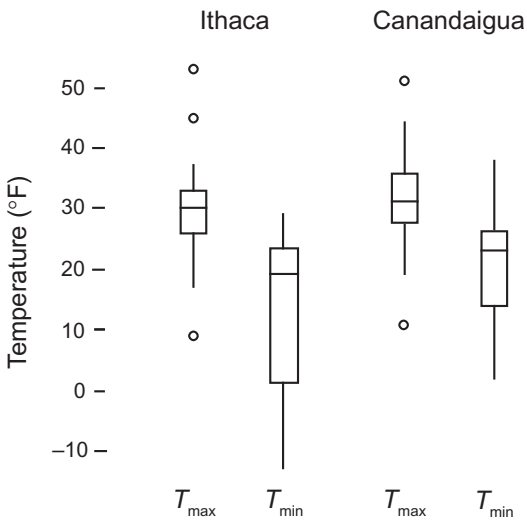


FIGURE 3.5 Side-by-side schematic plots for the January 1987 temperatures in Table A.1. The minimum temperature data for both locations are all "inside," so the schematic plots are identical to ordinary boxplots.

median. The details of constructing these intervals are given in Velleman and Hoaglin (1981). Combining both of these techniques, that is, constructing notched, variable-width plots, is straightforward. If the notched portion needs to extend beyond the quartiles, however, the overall appearance of the plot can begin to look a bit strange (an example can be seen in Graedel and Kleiner, 1985). An alternative to notching is to add shading or stippling in the box to span the computed interval, rather than deforming its outline with notches.

3.3.5. Histograms

The *histogram* is a very familiar graphical display device for representing the distribution of a single batch of data. The range of the data is divided into class intervals or *bins*, and the number of values falling into each interval is counted. The histogram then consists of a series of rectangles whose widths are defined by the class limits implied by the binwidths, and whose heights depend on the number of values in each bin. Example histograms are shown in Figure 3.6. Histograms quickly reveal such attributes of the data distribution as location, spread, and symmetry. If the data are multimodal (i.e., more than one "hump" in the distribution of the data), this is quickly evident as well.

Usually the widths of the bins are chosen to be equal. In that case the heights of the histogram bars are simply proportional to the numbers of counts. The vertical axis can be labeled to give either the number of counts represented by each bar (the absolute frequency) or the proportion of the entire sample represented by each bar (the relative frequency). More properly, however, it is the areas of the histogram bars (rather than their heights) that are proportional to probabilities. This point becomes important if the histogram bins are chosen to have unequal widths, or when a parametric probability function (Chapter 4) is to be superimposed on the histogram.

The main issue to be confronted when constructing a histogram is choice of the binwidth. Intervals that are too wide will result in important details of the data being masked (the histogram is too smooth). Intervals that are too narrow will result in a plot that is irregular and difficult to interpret (the histogram is

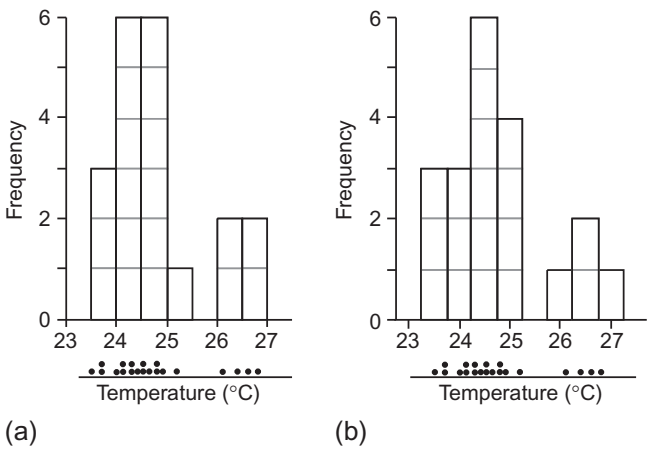


FIGURE 3.6 Histograms of the June Guayaquil temperature data in Table A.3, illustrating differences that can arise due to arbitrary shifts in the horizontal placement of the bins. Neither of these plots in panels (a) or (b) is more "correct" than the other. This figure also illustrates that each histogram bar can be viewed as being composed of stacked "building blocks" (gray) equal in number to the number of data values in the bin. Dotplots below each histogram locate the original data.

too rough). In general, narrower histogram bins are justified by larger data samples, but the nature of the data also influences the choice. A good approach to selecting the binwidth, w, is to begin by computing

$$w \approx \frac{c\,IQR}{n^{1/3}}, \tag{3.13}$$

where c is a constant in the range of perhaps 2.0–2.6. Results given in Scott (1992) indicate that $c = 2.6$ is optimal for Gaussian (bell-shaped) data, and that smaller values are more appropriate for skewed and/or multimodal data.

The initial binwidth computed using Equation 3.13, or arrived at according to any other rule, should be regarded as just a guideline or rule of thumb. Other considerations also will enter into the choice of the binwidth, such as the practical desirability of having the class boundaries fall on values that are natural with respect to the data at hand. (Computer programs that plot histograms must use rules such as that in Equation 3.13, and one indication of the care with which the software has been written is whether the resulting histograms have natural or arbitrary bin boundaries.) For example, the January 1987 Ithaca maximum temperature data has IQR = 7°F, and $n = 31$. A binwidth of 5.7°F would be suggested initially by Equation 3.13, using $c = 2.6$ since the schematic plot for these data (Figure 3.5) look at least approximately Gaussian. A natural choice in this case might be to choose 10 bins of width 5°F, yielding a histogram looking much like the stem-and-leaf display in Figure 3.1b.

3.3.6. Kernel Density Smoothing

One interpretation of the histogram is as a nonparametric estimator for the underlying probability distribution from which the data have been drawn. "Nonparametric" means that fixed mathematical forms of the kind presented in Chapter 4 are not assumed. However, the alignment of the histogram bins on the real line is an arbitrary choice, and construction of a histogram requires essentially that each data value is rounded to the center of the bin into which it falls. For example, in Figure 3.6a the bins have been aligned so that they are centered at integer temperature values ±0.25°C, whereas the equally valid histogram in Figure 3.6b has shifted these by 0.25°C. The two histograms in Figure 3.6 present somewhat different impressions of the data, although both indicate bimodality that can be traced (through the asterisks in Table A.3) to the occurrence of El Niño. Another, possibly less severe, difficulty with the histogram is that the rectangular nature of the histogram bars presents a rough appearance and appears to imply that any value within a given bin is equally likely.

Kernel density smoothing is an alternative to the histogram that does not require arbitrary rounding to bin centers, and which presents a smooth result. The application of kernel smoothing to the empirical frequency distribution of a data set produces the *kernel density estimate*, which is a nonparametric alternative to the fitting of a parametric probability density function (Chapter 4). It is easiest to understand kernel density smoothing as an extension of the histogram. As illustrated in Figure 3.6, after rounding each data value to its bin center the histogram can be viewed as having been constructed by stacking rectangular building blocks above each bin center, with the number of blocks equal to the number of data points in each bin. In Figure 3.6 the distribution of the data is indicated below each histogram in the form of *dotplots*, which locate each data value with a dot, and indicate instances of repeated data with stacks of dots.

The rectangular building blocks in Figure 3.6 each have area equal to the binwidth (0.5°F), because the vertical axis is just the raw number of counts in each bin. If instead the vertical axis had been chosen so the area of each building block was $1/n$ (=1/20 for these data), the resulting histograms would be

TABLE 3.1 Some Commonly Used Smoothing Kernels

Name	$h(t)$	Support [t for Which $h(t)>0$]	$1/\sigma_k$
Quartic (Biweight)	$(15/16)\,(1-t^2)^2$	$-1<t<1$	$\sqrt{7}$
Triangular	$1-\lvert t\rvert$	$-1<t<1$	$\sqrt{6}$
Quadratic (Epanechnikov)	$(3/4)\,(1-t^2)$	$-1<t<1$	$\sqrt{5}$
Gaussian	$(2\pi)^{-1/2}\exp[-t^2/2]$	$-\infty<t<\infty$	1

quantitative estimators of the underlying probability distribution, since the total histogram area would be 1 in each case, and total probability must sum to 1.

Kernel density smoothing proceeds in an analogous way, using characteristic shapes called *kernels*, that are generally smoother than rectangles. Table 3.1 lists four commonly used smoothing kernels, and Figure 3.7 shows their shapes graphically. These are all nonnegative functions with unit area, that is, $\int h(t)\,dt=1$ in each case, so each is a proper probability density function (discussed in more detail in Chapter 4). In addition, all are centered at zero. The *support* (value of the argument t for which $h(t)>0$) is $-1<t<1$ for the triangular, quadratic, and quartic kernels and covers the entire real line for the Gaussian kernel. The kernels listed in Table 3.1 are appropriate for use with continuous data (taking on values over all or some portion of the real line). Some kernels appropriate to discrete data (able to take on only a finite number of values) are presented in Rajagopalan et al. (1997).

Instead of stacking rectangular kernels centered on bin midpoints (which is one way of looking at histogram construction), kernel density smoothing is achieved by stacking kernel shapes, equal in number to the number of data values, with each stacked element being centered at the data value it represents. Of course in general kernel shapes do not fit together like building blocks, but kernel density smoothing is achieved through the mathematical equivalent of stacking, by adding the heights of all the kernel functions contributing to the smoothed estimate at a given value, x_0,

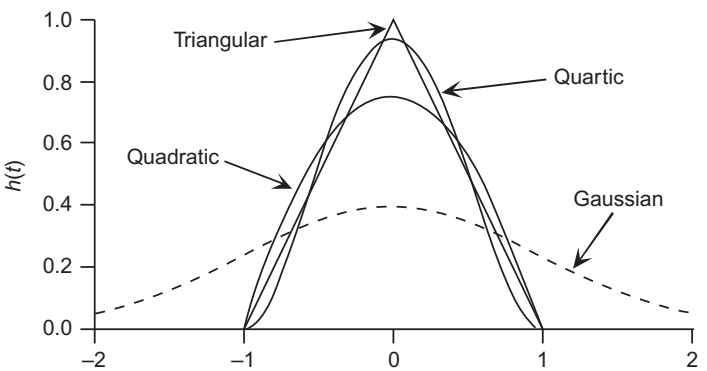


FIGURE 3.7 The four commonly used smoothing kernels defined in Table 3.1.

$$\hat{f}(x_0) = \frac{1}{nw} \sum_{i=1}^{n} h\left(\frac{x_0 - x_i}{w}\right). \tag{3.14}$$

The argument within the kernel function indicates that each of the kernels employed in the smoothing (corresponding to the data values x_i close enough to the point x_0 that the kernel height is not zero) is centered at its respective data value x_i and is scaled in width relative to the shapes as plotted in Figure 3.7 by the smoothing parameter w. Consider, for example, the triangular kernel in Table 3.1, with $t=(x_0-x_i)/w$. The function $h[(x_0-x_i)/w]=1-|(x_0-x_i)/w|$ is an isosceles triangle with support (i.e., nonzero height) for $x_i-w<x_0<x_i+w$, and the area within this triangle is w, because the area within $1-|t|$ over the support interval $-1<t<1$ is 1 and its base has been expanded (or contracted) by a factor of w. Therefore in Equation 3.14 the kernel heights stacked at the value x_0 will be those corresponding to any of the x_i at distances closer to x_0 than w. In order for the area under the entire function in Equation 3.14 to integrate to 1, which is desirable if the result is meant to estimate a probability density function, each of the n kernels to be superimposed should have area $1/n$. This is achieved by dividing each $h[(x_0-x_i)/w]$, or equivalently dividing their sum, by the product nw.

The choice of kernel type is usually less important than choice of the smoothing parameter. The Gaussian kernel is intuitively appealing, but it is computationally slower both because of the exponential function calls, and because its infinite support leads to all data values contributing to the smoothed estimate at any x_0 (none of the n terms in Equation 3.14 are ever zero). On the other hand, all the derivatives of the resulting function will exist, and nonzero probability is estimated everywhere on the real line, whereas these are not characteristics of probability density functions estimated using the other kernels listed in Table 3.1.

Example 3.3. Kernel Density Estimates for the Guayaquil Temperature Data

Figure 3.8 shows kernel density estimates for the June Guayaquil temperature data in Table A.3, corresponding to the histograms in Figure 3.6. The four probability density estimates have been constructed using the quartic kernel and four choices for the smoothing parameter w, which increase from panels (a) through (d). The role of the smoothing parameter is analogous to that of the histogram binwidth, also called w, in that larger values result in smoother shapes that progressively suppress details. Smaller values result in more irregular shapes that reveal more details, including more of the sampling variability. Figure 3.8b, plotted using $w=0.6$, also shows the individual kernels that have been summed to produce the smoothed density estimate. Since $w=0.6$ and the support of the quartic kernel is $-1<t<1$ (see Table 3.1) the width of each of the individual kernels in Figure 3.8b is 1.2°C. The five repeated data values 23.7, 24.1, 24.3, 24.5, and 24.8 (compare the dotplots at the bottom of Figure 3.6) are represented by the five taller kernels, the areas of which are each $2/n$. The remaining 10 data values are unique, and their kernels each have area $1/n$. ◇

Comparing the panels in Figure 3.8 emphasizes that a good choice for the smoothing parameter w is critical. Silverman (1986) suggests that a reasonable initial choice for use with the Gaussian kernel could be

$$w = \frac{\min\left\{0.9\,s,\ \frac{2}{3}IQR\right\}}{n^{1/5}}, \tag{3.15}$$

where s is the standard deviation of the data. Equation 3.15 indicates that less smoothing (smaller w) is justified for larger sample sizes n, although w should not decrease with sample size as quickly as does the

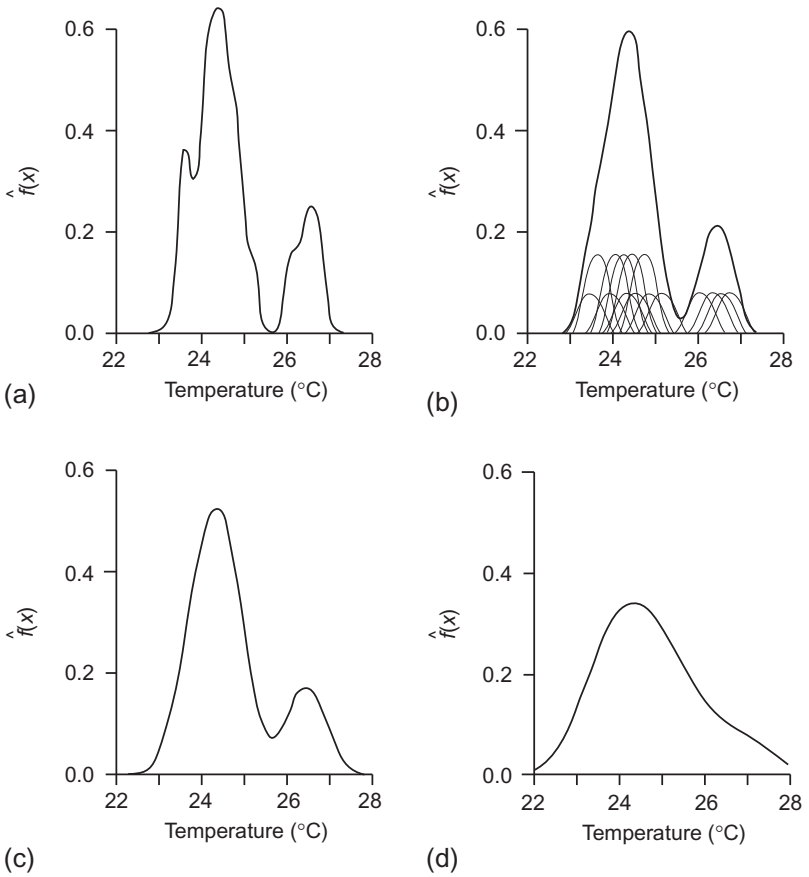


FIGURE 3.8 Kernel density estimates for the June Guayaquil temperature data in Table A.3, constructed using the quartic kernel and (a) $w=0.3$, (b) $w=0.6$, (c) $w=0.92$, and (d) $w=2.0$. Also shown in panel (b) are the individual kernels that have been added together to construct the estimate. These same data are shown as histograms in Figure 3.6.

histogram binwidth (Equation 3.13). Since the Gaussian kernel is intrinsically broader than the others listed in Table 3.1 (compare Figure 3.7), smaller smoothing parameters are appropriate for these, in proportion to the reciprocals of the kernel standard deviations (Scott, 1992), which are listed in the last column of Table 3.1. For the Guayaquil temperature data, $s=0.98$ and IQR $=0.95$, so 2/3 IQR is smaller than $0.9s$, and Equation 3.15 yields $w=(2/3)(0.95)/20^{1/5}=0.35$ for smoothing these data with a Gaussian kernel. But Figure 3.8 was prepared using the more compact quartic kernel, whose standard deviation is $1/\sqrt{7}$, yielding an initial choice for the smoothing parameter $w=(\sqrt{7})(0.35)=0.92$ (Figure 3.8c).

When kernel smoothing is used for an exploratory analysis or construction of an esthetically pleasing data display, a recommended smoothing parameter computed according to Equation 3.15 will often be the starting point for a subjective choice following some exploration through trial and error, and this process may even enhance the exploratory data analysis. In instances where the kernel density estimate will be used in subsequent quantitative analyses it may be preferable to estimate the smoothing parameter objectively using cross-validation methods similar to those presented in Chapter 7 (Scott, 1992; Sharma et al., 1998; Silverman, 1986). Adopting the exploratory approach, both $w=0.92$

(Figure 3.8c) and $w = 0.6$ (Figure 3.8b) appear to produce reasonable balances between display of the main data features (here, the bimodality related to El Niño) and suppression of irregular sampling variability. Figure 3.8a, with $w = 0.3$, is too rough for most purposes, as it retains irregularities that can probably be ascribed to sampling variations, and (almost certainly spuriously) indicates zero probability for temperatures near 25.5°C. On the other hand, Figure 3.8d is clearly too smooth, as it suppresses entirely the bimodality in the data.

Kernel smoothing can be extended to bivariate, and higher dimensional, data using the product-kernel estimator

$$\hat{f}(\boldsymbol{x}_0) = \frac{1}{n\, w_1\, w_2 \cdots w_K} \sum_{i=1}^{n} \left[\prod_{k=1}^{K} h\left(\frac{x_{0,k} - x_{i,k}}{w_k} \right) \right] .. \tag{3.16}$$

Here there are K data dimensions, $x_{0,k}$ denotes the point at which the smoothed estimate is produced in the kth of these dimensions, and the uppercase *pi* indicates multiplication of factors analogously to the summation of terms indicated by an uppercase sigma. The same (univariate) kernel $h(\bullet)$ is used in each dimension, although not necessarily with the same smoothing parameter w_k. In general the multivariate smoothing parameters w_k will need to be larger than for the same data smoothed alone (i.e., for a univariate smoothing of the corresponding kth variable in the vector $\boldsymbol{x}$) and should decrease with sample size in proportion to $n^{-1/(K+4)}$. Equation 3.16 can be extended to include also nonindependence of the kernels among the K dimensions by using a multivariate probability density (e.g., the multivariate normal distribution described in Chapter 12) for the kernel (Scott, 1992, Sharma et al., 1998, Silverman, 1986).

Kernel density estimates can be combined with boxplots to produce an informative graphic known as the *violin plot* (Hintze and Nelson, 1998). A violin plot usually consists of a central boxplot, with kernel density estimates based on the same underlying data plotted symmetrically on both sides of the boxplot. Figure 3.9 shows violin plots for the same maximum and minimum temperature data portrayed as

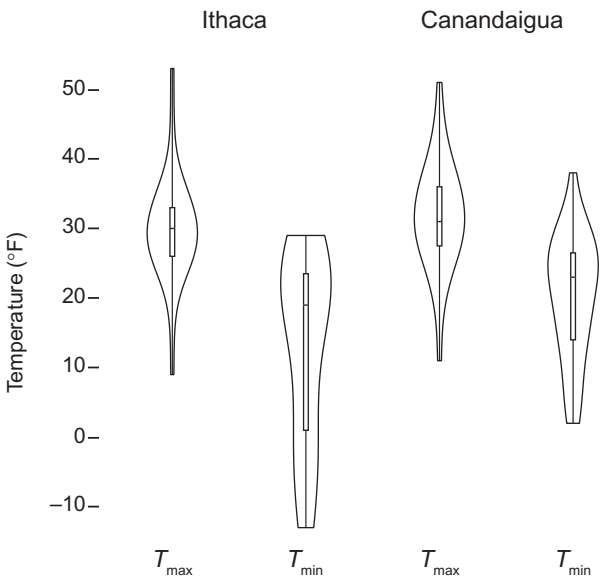


FIGURE 3.9 Violin plots for the January 1987 temperature data in Table A.1, which can be compared to the schematic plots for the same data shown in Figure 3.5.

schematic plots in Figure 3.5. Violin plots show more detail about the data distribution than the simpler boxplots alone, while the boxplots on the midlines allow identification of the median and quartiles of the smoothed distribution. Distributional features which might not be evident from a boxplot alone can be discerned in a violin plot. For example, the violin plot for a bimodal distribution would tend to take on the shape of the musical instrument for which it is named. The violin plot for the Ithaca minimum temperatures in Figure 3.9 begins to suggest this attribute.

Finally, note that kernel smoothing can be applied in settings other than estimation of probability distribution functions. When estimating a general smoothing function, which is not constrained to integrate to 1, a smoothed value of a function $y=f(x)$ at any point x_0 can be computed using the *Nadaraya-Watson kernel-weighted average*,

$$f(x_0) = \frac{\sum_{i=1}^{n} h\left(\frac{x_0 - x_i}{w}\right) y_i}{\sum_{i=1}^{n} h\left(\frac{x_0 - x_i}{w}\right)}, \tag{3.17}$$

where y_i is the raw value at x_i of the response variable to be smoothed. For example, Figure 3.10 shows mean numbers of tornado days per year, based on daily tornado occurrence counts in 80×80 km grid squares, for the period 1980–1999. The figure was produced after a three-dimensional smoothing using a Gaussian kernel, smoothing in time with $w = 15$ days, and smoothing in latitude and longitude with $w = 120$ km. The figure allows a straightforward interpretation of the underlying data, which in raw form are very erratic in both space and time.

More on kernel smoothing methods can be found in Chapter 6 of Hastie et al. (2009).

3.3.7. Cumulative Frequency Distributions

The *cumulative frequency distribution* is a display related to the histogram. It is also known as the *empirical cumulative distribution function*. The cumulative frequency distribution is a two-dimensional plot in which the vertical axis shows cumulative probability estimates associated with data values on the horizontal axis. That is, the plot represents relative frequency estimates for the probability that an arbitrary or random future datum will not exceed the corresponding value on the horizontal axis. Thus the cumulative frequency distribution is like the integral of a histogram with arbitrarily narrow binwidth. Figure 3.11 shows two empirical cumulative distribution functions, illustrating that they are step functions with probability jumps occurring at the data values. Just as histograms can be smoothed using kernel density estimators, smoothed versions of empirical cumulative distribution functions can be obtained by integrating the result of a kernel smoothing.

The vertical axes in Figure 3.11 show the empirical cumulative distribution function, $p(x)$, which can be expressed as

$$p(x) \approx \Pr\{X \leq x\}. \tag{3.18}$$

The notation on the right side of this equation can be somewhat confusing at first, but is standard in statistical work. The uppercase letter X represents the generic random variable or the "arbitrary or random future" value referred to in the previous paragraph. The lowercase x, on both sides of Equation 3.18, represents a specific value of the random quantity. In the cumulative frequency distribution, these specific values are plotted on the horizontal axis.

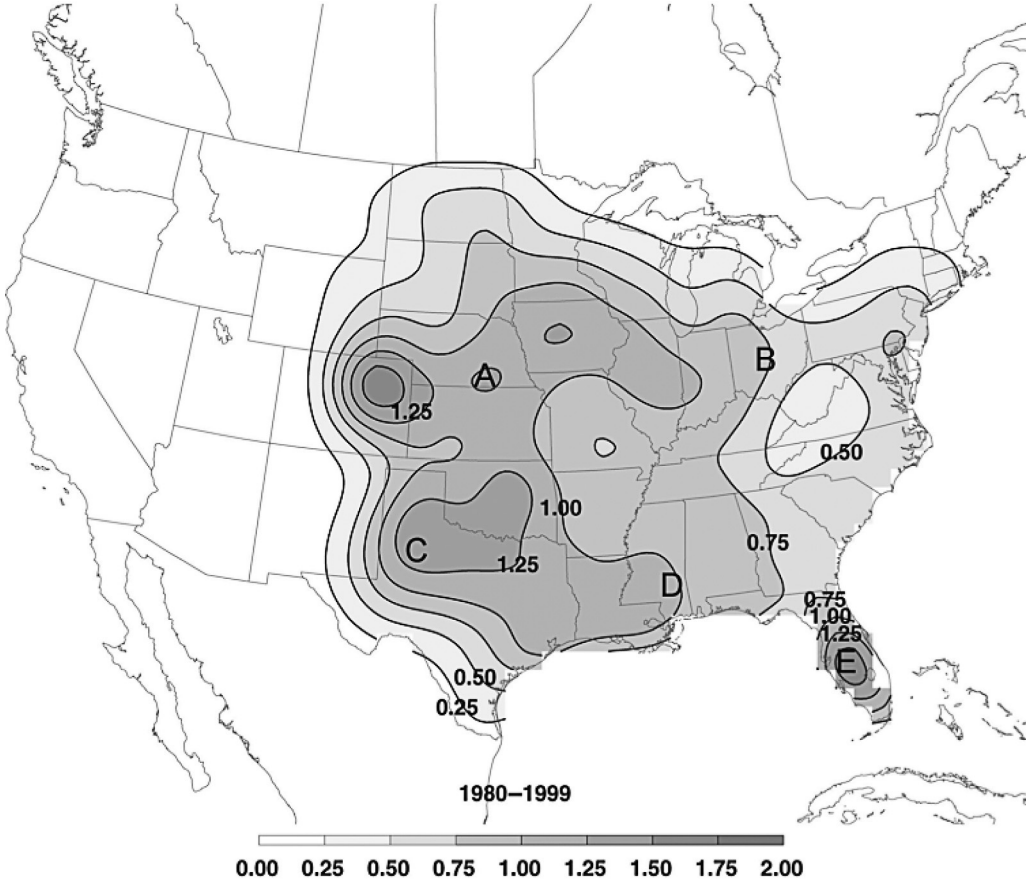


FIGURE 3.10 Mean numbers of tornado days per year in the United States, as estimated using a three-dimensional (time, latitude, longitude) kernel smoothing of daily, 80×80 km gridded tornado occurrence counts. *From Brooks et al. (2003). © American Meteorological Society. Used with permission.*

In order to construct a cumulative frequency distribution, it is necessary to estimate $p(x)$ using the ranks, i, of the order statistics, $x_{(i)}$. In the literature of hydrology these estimates are known as *plotting positions* (e.g., Harter, 1984), reflecting their historical use in graphically comparing the empirical distributions with candidate parametric functions (Chapter 4) that might be used to represent them. There is a substantial literature devoted to equations that can be used to calculate plotting positions, and thus to estimate cumulative probabilities from data sets. Most are particular cases of the formula

$$p\left(x_{(i)}\right) = \frac{i-a}{n+1-2a} \quad ,0 \le a \le 1, \tag{3.19}$$

in which different values for the constant a result in different plotting position estimators, some of which are presented in Table 3.2. The names in this table relate to authors who proposed the various estimators and not to particular probability distributions that may be named for the same authors.

Several of the plotting positions in Table 3.2 are motivated by characteristics of the *sampling distributions* of the cumulative probabilities associated with the order statistics. The notion of a

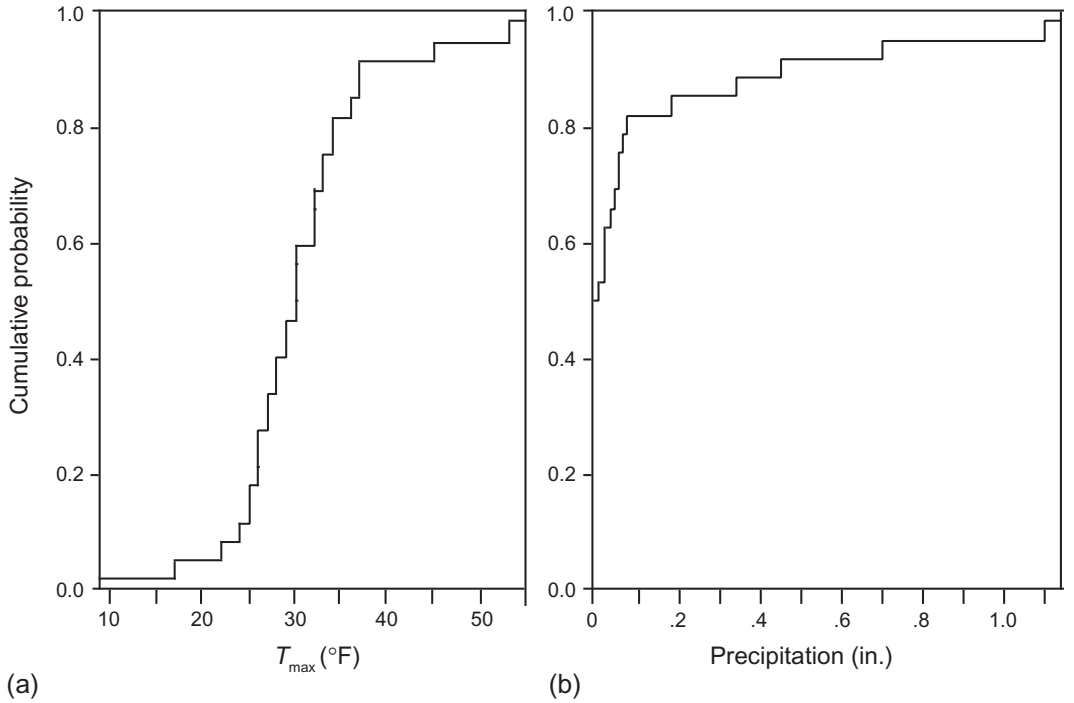


FIGURE 3.11 Empirical cumulative frequency distribution functions for the January 1987 Ithaca maximum temperature (a), and precipitation data (b). The S-shape exhibited by the temperature data is characteristic of reasonably symmetrical data, and the concave downward character exhibited by the precipitation data is characteristic of data that are skewed to the right.

TABLE 3.2 Some Common Plotting Position Estimators for Cumulative Probabilities Corresponding to the *i*th Order Statistic, $x_{(i)}$, and the Corresponding Values of the Parameter *a* in Equation 3.18

Name	Formula	*a*	Interpretation
Weibull	$i/(n+1)$	0	mean of sampling distribution
Benard & Bos-Levenbach	$(i-0.3)/(n+0.4)$	0.3	approximate median of sampling distribution
Tukey	$(i-1/3)/(n+1/3)$	1/3	approximate median of sampling distribution
Cunnane	$(i-2/5)/(n+1/5)$	2/5	subjective choice, commonly used in hydrology
Gringorten	$(i-0.44)/(n+0.12)$	0.44	consonance with the Gumbel distribution (Equation 4.67)
Hazen	$(i-1/2)/n$	1/2	midpoints of *n* equal intervals on [0, 1]
Gumbel	$(i-1)/(n-1)$	1	mode of sampling distribution

sampling distribution is considered in more detail in Chapter 5, but briefly think about hypothetically obtaining a large number of data samples of size n from some unknown distribution. The ith order statistics from these samples will differ somewhat from each other, but each will correspond to some cumulative probability in the distribution from which the data were drawn. In aggregate over the large number of hypothetical samples there will be a distribution—the sampling distribution—of cumulative probabilities corresponding to the ith order statistic. One way to imagine this sampling distribution is as a histogram of cumulative probabilities for, say, the smallest (or any of the other order statistics) of the n values in each of the batches. This notion of the sampling distribution for cumulative probabilities is expanded upon more fully in a climatological context by Folland and Anderson (2002).

The mathematical form of the sampling distribution of cumulative probabilities corresponding to the ith order statistic is known to be a Beta distribution (see Section 4.4.6), with parameters $\alpha = i$ and $\beta = n - i + 1$, regardless of the distribution from which the x's have been independently drawn (Gumbel, 1958). Thus the Weibull ($a = 0$) plotting position estimator is the mean of the cumulative probabilities corresponding to a particular $x_{(i)}$, averaged over many hypothetical samples of size n. Similarly, the Benard & Bos-Levenbach ($a = 0.3$) and Tukey ($a = 1/3$) estimators approximate the medians of these distributions. The Gumbel ($a = 1$) plotting position locates the modal (single most frequent) cumulative probability, although it ascribes zero and unit cumulative probability to $x_{(1)}$ and $x_{(n)}$, respectively, leading to the generally unwarranted implication that the probabilities of observing data more extreme than these are zero. It is possible also to derive plotting position formulas using the reverse perspective, thinking about the sampling distributions of data quantiles x_i corresponding to particular, fixed cumulative probabilities (e.g., Cunnane, 1978; Stedinger et al., 1993). Plotting positions resulting from this approach depend on the distribution from which the data have been drawn, although the Cunnane ($a = 2/5$) plotting position is a compromise approximation to many of them.

In practice most of the various plotting position formulas produce quite similar results, especially when judged in relation to the intrinsic variability (Equation 4.59b) of the sampling distribution of the cumulative probabilities, which is much larger than the differences among the various plotting positions in Table 3.2. Generally very reasonable results are obtained using moderate (in terms of the parameter a) plotting positions such as Tukey or Cunnane. Reviewing and comparing properties of various plotting positions, Hyndman and Fan (1996) conclude with an overall preference for the Tukey estimator.

Figure 3.11a shows the cumulative frequency distribution for the January 1987 Ithaca maximum temperature data, using the Tukey ($a = 1/3$) plotting position to estimate the cumulative probabilities. Figure 3.11b shows the Ithaca precipitation data displayed in the same way. For example, the coldest of the 31 temperatures in Figure 3.11a is $x_{(1)} = 9°\text{F}$, and $p(x_{(1)})$ is plotted at $(1 - 0.333)/(31 + 0.333) = 0.0213$. The steepness in the center of the plot reflects the concentration of data values in the center of the distribution, and the flatness at high and low temperatures results from data being more rare there. The S-shaped character of this plot is indicative of a reasonably symmetrical data distribution, with comparable numbers of observations on either side of the median at a given distance from the median. The cumulative distribution function for the precipitation data (Figure 3.11b) rises quickly on the left because of the high concentration of data values there, and then rises more slowly in the center and right of the figure because of the relatively fewer large observations. The concave downward character of this cumulative distribution function is thus indicative of positively skewed data. A plot of cumulative probability for a batch of negatively skewed data would show just the reverse characteristics: a very shallow slope in

the left and center of the diagram, rising steeply toward the right, yielding a function that would be concave upward.

3.4. REEXPRESSION

It is possible that the original scale of measurement may obscure important features in a set of data. If so, an analysis can be facilitated, or may yield more revealing results, if the data are first subjected to a mathematical transformation. Such transformations can also be very useful for helping data conform to the assumptions of regression analysis (see Section 7.2), or allowing application of multivariate statistical methods that may assume Gaussian distributions (see Chapter 12). In the terminology of exploratory data analysis, such data transformations are known as *reexpression* of the data.

3.4.1. Power Transformations

Often data transformations are undertaken in order to make the distribution of values more nearly symmetrical, and the resulting symmetry may allow use of more familiar and traditional statistical techniques. Sometimes a symmetry-producing transformation can make exploratory analyses, such as those described in this chapter, more revealing. These transformations can also aid in comparing different batches of data, for example, by rendering the relationship between two variables more nearly linear. Another important use of transformations is to make the variations or dispersion (i.e., the spread) of one variable less dependent on the value of another variable, in which case the transformation is called *variance stabilizing*.

Undoubtedly the most commonly used (although not the only possible—see, e.g., Equation 12.12) symmetry-producing transformations are the *power transformations*, defined by one or the other of the two closely related functions

$$T_1(x) = \begin{cases} x^{\lambda} & , \lambda > 0 \\ \ln(x) & , \lambda = 0, \\ -(x^{\lambda}) & , \lambda < 0 \end{cases} \tag{3.20a}$$

and

$$T_2(x) = \begin{cases} \dfrac{x^{\lambda} - 1}{\lambda} & , \lambda \neq 0 \\ \ln(x) & , \lambda = 0 \end{cases}. \tag{3.20b}$$

These transformations are useful when dealing with unimodal (single-humped) distributions of strictly positive data variables. Each of these functions defines a family of transformations indexed by the single parameter λ. The name power transformation derives from the fact that the important work of these transformations—changing the shape of the data distribution—is accomplished by the exponentiation, or raising the data values to the power λ. Thus the sets of transformations in Equations 3.20a and 3.20b are actually quite comparable, and a particular value of λ produces the same effect on the overall shape of the data in either case. The transformations in Equation 3.20a are of a slightly simpler form and are often employed because of the greater ease. The transformations in Equation 3.20a are sometimes known as "*Tukey's ladder*". The transformations in Equation 3.20b, also known as the *Box-Cox transformations*, are simply shifted and scaled versions of Equation 3.20a, and are sometimes

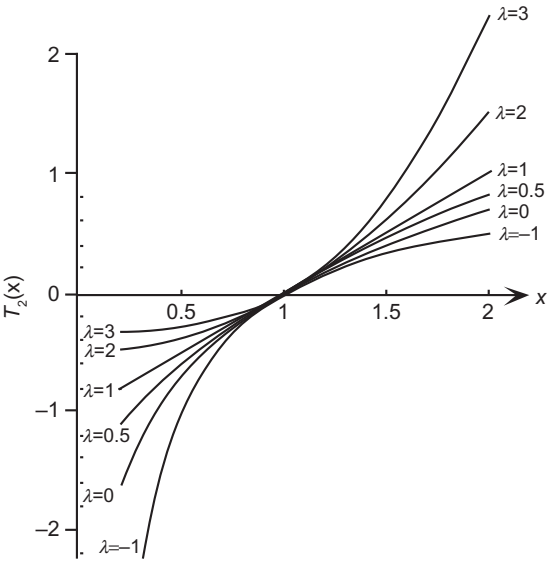


FIGURE 3.12 Graphs of the power transformations in Equation 3.20b for selected values of the transformation parameter λ. For $\lambda = 1$ the transformation is linear and produces no change in the shape of the data distribution. For $\lambda < 1$ the transformation reduces all data values, with larger values more strongly affected. The reverse effect is produced by transformations with $\lambda > 1$.

more useful when comparing among different transformations. Also, Equation 3.20b is mathematically "nicer" since the limit of the transformation in the upper equality as $\lambda \to 0$ is actually the function $\ln(x)$.

In both Equations 3.20a and 3.20b, adjusting the value of the parameter λ yields specific members of an essentially continuously varying set of smooth transformations. These transformations are sometimes referred to as the *ladder of powers*. A few of these transformation functions are plotted in Figure 3.12. The curves in this figure are functions specified by Equation 3.20b, although the corresponding curves from Equation 3.20a have the same shapes. Figure 3.12 makes it clear that use of the logarithmic transformation for $\lambda = 0$ fits neatly into the spectrum of the power transformations. This figure also illustrates another property of the power transformations, which is that they are all increasing functions of the original variable, x. This property is achieved in Equation 3.20a by the negative sign in the transformations with $\lambda < 1$. For the transformations in Equation 3.20b this sign reversal is achieved by dividing by λ. This strictly increasing property of the power transformations implies that they are order preserving, so that the smallest value in the original data set will correspond to the smallest value in the transformed data set, and likewise for the largest values. In fact, there will be a one-to-one correspondence between all the order statistics of the original and transformed distributions. Thus the median, quartiles, and so on, of the original data will be transformed to the corresponding quantiles of the transformed data.

Clearly for $\lambda = 1$ the shape of the data distribution remains unchanged. For $\lambda > 1$ the data values are increased (except for the subtraction of $1/\lambda$ and division by λ, if Equation 3.20b is used), with the larger values being increased more than the smaller ones. Therefore power transformations with $\lambda > 1$ will help produce symmetry when applied to negatively skewed data. The reverse is true for $\lambda < 1$, where larger data values are decreased more than smaller values. Power transformations with $\lambda < 1$ are therefore applied to data that are originally positively skewed, in order to produce more nearly symmetric distributions. Figure 3.13 illustrates the mechanics of this process for an originally positively

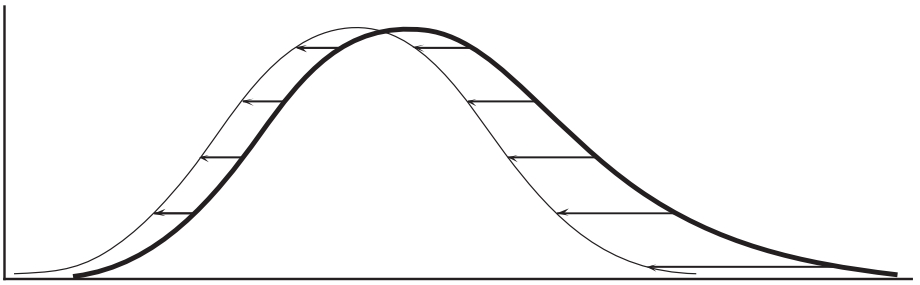


FIGURE 3.13 Effect of a power transformation with $\lambda < 1$ on a data distribution with positive skew (heavy curve). Arrows indicate that the transformation moves all the points to the left, with the larger values being moved much more. The resulting distribution (light curve) is reasonably symmetric.

skewed distribution (heavy curve). Applying a power transformation with $\lambda < 1$ reduces all the data values, but affects the larger values more strongly. An appropriate choice of λ can often produce at least approximate symmetry through this process (light curve). Choosing an excessively small or negative value for λ would yield an overcorrection, resulting in the transformed distribution being negatively skewed.

Initial inspection of an exploratory graphic such as a schematic plot can indicate quickly the direction and approximate magnitude of the skew in a batch of data. It is thus usually clear whether a power transformation with $\lambda > 1$ or $\lambda < 1$ is appropriate, but a specific value for the exponent will not be so obvious. A number of approaches to choosing an appropriate transformation parameter have been suggested. The simplest of these is the d_λ statistic (Hinkley, 1977),

$$d_\lambda = \frac{|\,\text{mean}(\lambda) - \text{median}(\lambda)\,|}{\text{spread}(\lambda)}. \tag{3.21}$$

Here, spread is some resistant measure of dispersion, such as the IQR or MAD. Each value of λ will produce a different mean, median, and spread in a particular set of data, and these functional dependencies on λ are indicated in the equation. The Hinkley d_λ is used to decide among power transformations essentially by trial and error, by computing its value for each of a number of different choices for λ. Usually these trial values of λ are spaced at intervals of 1/2 or 1/4. That choice of λ producing the smallest d_λ is then adopted to transform the data. One very easy way to do the computations is with a spreadsheet program on a desk computer.

The basis of the d_λ statistic is that the mean and median will be very close for symmetrically distributed data. Therefore as successively stronger power transformations (values of λ increasingly far from 1) move the data toward symmetry, the numerator in Equation 3.21 will move toward zero. As the transformations become too strong, the numerator will begin to increase relative to the spread measure, resulting in the d_λ increasing again.

Equation 3.21 is a simple and direct approach to finding a power transformation that produces symmetry or near symmetry in the transformed data. A more sophisticated approach was suggested in the original Box and Cox (1964) paper, which is particularly appropriate when the transformed data should have a distribution as close as possible to the bell-shaped Gaussian, for example, when the results of multiple transformations will be summarized simultaneously through the multivariate Gaussian or multivariate normal distribution (see Chapter 12). In particular, Box and Cox suggested choosing the power

transformation exponent to maximize the log-likelihood function (see Section 4.6) for the Gaussian distribution

$$L(\lambda) = -\frac{n}{2}\ln\left[s^2(\lambda)\right] + (\lambda - 1)\sum_{i=1}^{n}\ln[x_i]. \tag{3.22}$$

Here n is the sample size, and $s^2(\lambda)$ is the sample variance (computed with a divisor of n rather than $n-1$, see Equation 4.84b) of the data after transformation with the exponent λ. It is important to realize that the sum of the logarithms in the second term of Equation 3.22 pertains to the *un*transformed data. As was the case for using the Hinkley statistic (Equation 3.21), different values of λ may be tried, and the one yielding the largest value of $L(\lambda)$ is chosen as most appropriate. It is possible that the two criteria will yield different choices for λ since Equation 3.21 addresses only symmetry of the transformed data, whereas Equation 3.22 tries to accommodate all aspects of the Gaussian distribution, including but not limited to its symmetry. Note, however, that choosing λ by maximizing Equation 3.22 does not necessarily produce transformed data that are close to Gaussian if the original data are not well suited to the transformations in Equation 3.20.

Equations 3.20 and 3.22 are valid only if zero or negative values of the variable x cannot be realized. For transformation of data that include some zero or negative values, the original recommendation by Box and Cox (1964) was to modify the transformation by adding a positive constant to each data value, with the magnitude of the constant being large enough for all the data to be shifted onto the positive half of the real line. This easy approach is often adequate, but it is somewhat arbitrary and fails entirely if a future negative value of x is larger in absolute value than this constant. Yeo and Johnson (2000) have proposed a unified extension of the Box-Cox transformations that accommodate data anywhere on the real line:

$$T_3(x) = \begin{cases} \left[(x+1)^{\lambda} - 1\right]/\lambda & ,x \geq 0 \text{ and } \lambda \neq 0 \\ \ln(x+1) & ,x \geq 0 \text{ and } \lambda = 0 \\ -\left[(-x+1)^{2-\lambda} - 1\right]/(2-\lambda) & ,x < 0 \text{ and } \lambda \neq 2 \\ -\ln(-x+1) & ,x < 0 \text{ and } \lambda = 2 \end{cases}. \tag{3.23}$$

For $x > 0$, Equation 3.23 achieves the same effect as Equation 3.20b, although with the curves shifted to the left by one unit. The graphs of $T_3(x)$ resemble those in Figure 3.12, except that they pass through the origin. The simplest approach to choosing the transformation parameter λ for Equation 3.23 is again the Hinkley statistic (Equation 3.21), although Yeo and Johnson (2000) also provide a maximum likelihood estimation procedure.

Example 3.4. Choosing an Appropriate Power Transformation

Table 3.3 presents the 1933–1982 January Ithaca precipitation data from Table A.2 in Appendix A, sorted in ascending order and subjected to the power transformations $T_2(x)$ in Equation 3.20b, with $\lambda = 1$, $\lambda = 0.5$, $\lambda = 0$, and $\lambda = -0.5$. For $\lambda = 1$ this transformation amounts only to subtracting 1 from each data value. Note that even for the negative exponent $\lambda = -0.5$ the ordering of the original data is preserved in all the transformations, so that it is easy to determine the medians and the quartiles of the original and transformed data.

Figure 3.14 shows schematic plots for the data in Table 3.3. The untransformed data (leftmost plot) are clearly positively skewed, which is usual for distributions of precipitation amounts. All three of the

TABLE 3.3 Ithaca January Precipitation 1933–1982, From Table A.2 ($\lambda=1$)

Year	$\lambda=1$	$\lambda=0.5$	$\lambda=0$	$\lambda=-0.5$	Year	$\lambda=1$	$\lambda=0.5$	$\lambda=0$	$\lambda=-0.5$
1933	−0.56	−0.67	−0.82	−1.02	1948	0.72	0.62	0.54	0.48
1980	−0.48	−0.56	−0.65	−0.77	1960	0.75	0.65	0.56	0.49
1944	−0.46	−0.53	−0.62	−0.72	1964	0.76	0.65	0.57	0.49
1940	−0.28	−0.30	−0.33	−0.36	1974	0.84	0.71	0.61	0.53
1981	−0.13	−0.13	−0.14	−0.14	1962	0.88	0.74	0.63	0.54
1970	0.03	0.03	0.03	0.03	1951	0.98	0.81	0.68	0.58
1971	0.11	0.11	0.10	0.10	1954	1.00	0.83	0.69	0.59
1955	0.12	0.12	0.11	0.11	1936	1.08	0.88	0.73	0.61
1946	0.13	0.13	0.12	0.12	1956	1.13	0.92	0.76	0.63
1967	0.16	0.15	0.15	0.14	1965	1.17	0.95	0.77	0.64
1934	0.18	0.17	0.17	0.16	1949	1.27	1.01	0.82	0.67
1942	0.30	0.28	0.26	0.25	1966	1.38	1.09	0.87	0.70
1963	0.31	0.29	0.27	0.25	1952	1.44	1.12	0.89	0.72
1943	0.35	0.32	0.30	0.28	1947	1.50	1.16	0.92	0.74
1972	0.35	0.32	0.30	0.28	1953	1.53	1.18	0.93	0.74
1957	0.36	0.33	0.31	0.29	1935	1.69	1.28	0.99	0.78
1969	0.36	0.33	0.31	0.29	1945	1.74	1.31	1.01	0.79
1977	0.36	0.33	0.31	0.29	1939	1.82	1.36	1.04	0.81
1968	0.39	0.36	0.33	0.30	1950	1.82	1.36	1.04	0.81
1973	0.44	0.40	0.36	0.33	1959	1.94	1.43	1.08	0.83
1941	0.46	0.42	0.38	0.34	1976	2.00	1.46	1.10	0.85
1982	0.51	0.46	0.41	0.37	1937	2.66	1.83	1.30	0.95
1961	0.69	0.60	0.52	0.46	1979	3.55	2.27	1.52	1.06
1975	0.69	0.60	0.52	0.46	1958	3.90	2.43	1.59	1.10
1938	0.72	0.62	0.54	0.48	1978	5.37	3.05	1.85	1.21

The data have been sorted, with the power transformations $T_2(x)$ in Equation 3.20b applied for $\lambda=1$, $\lambda=0.5$, $\lambda=0$, and $\lambda=-0.5$. For $\lambda=1$ the transformation subtracts 1 from each data value. Schematic plots of these data are shown in Figure 3.14.

values outside the fences are large amounts, with the largest being far out. The three other schematic plots show the results of progressively stronger power transformations with $\lambda<1$. The logarithmic transformation ($\lambda=0$) both minimizes the Hinkley d_λ statistic (Equation 3.21) with IQR as the measure of spread and maximizes the Gaussian log-likelihood (Equation 3.22). The near symmetry exhibited by the schematic plot for the logarithmically transformed data supports the conclusion that it is best among the possibilities considered according to both criteria. The more extreme inverse square-root transformation ($\lambda=-0.5$) has evidently overcorrected for the positive skewness, as the three smallest amounts are now outside the lower fence. ◇

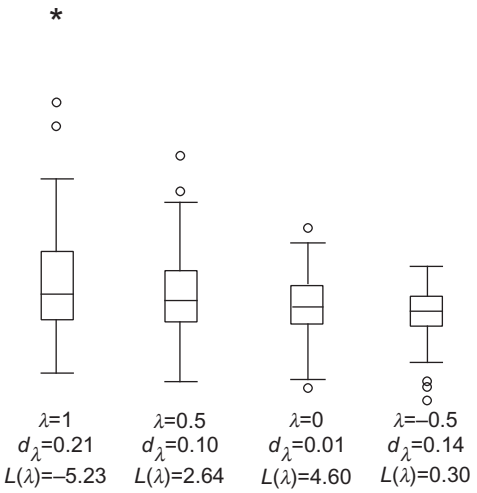


FIGURE 3.14 The effect of the power transformations $T_2(x)$ in Equation 3.20b on the January total precipitation data for Ithaca, 1933–1982 (Table A.2). The original data ($\lambda=1$) are skewed to the right, with the largest value being far out. The square-root transformation ($\lambda=0.5$) improves the symmetry somewhat. The logarithmic transformation ($\lambda=0$) produces a reasonably symmetric distribution. When subjected to the more extreme inverse square-root transformation ($\lambda=-0.5$) the data begin to exhibit negative skewness. The logarithmic transformation would be chosen as best by both the Hinkley d_λ statistic (Equation 3.21), and the Gaussian log-likelihood $L(\lambda)$ (Equation 3.22).

3.4.2. Some Other Nonlinear Transformations

In order for the shape of a transformed data distribution to be different from its untransformed counterpart, the transformation must be nonlinear. Equations 3.20 and 3.23 are by far the most frequently used such transformations, but they are not the only possibilities. This section presents a few others that can be useful in particular circumstances.

When the original data consist of proportions, probabilities, or other quantities p on the unit interval, $0<p<1$, it can be useful to transform them using

$$x = \ln\left(\frac{p}{1-p}\right), \tag{3.24}$$

which is known as the *log-odds, logit,* or *logistic* transformation. The bracketed quantity is called the *odds ratio*. The log-odds transformation yields transformed data on the full real line, $-\infty<x<\infty$.

Data that consist of correlations, so that $-1<r<1$, can be transformed to the full real line using the *Fisher Z*, or inverse hyperbolic tangent, transform,

$$Z = \frac{1}{2}\ln\left[\frac{1+r}{1-r}\right]. \tag{3.25}$$

Other nonlinear transformations can be designed to achieve particular goals. For example, Wang et al. (2012) propose the log-hyperbolic sine transformation for positively skewed data y,

$$z = \ln\left(\frac{e^y - e^{-y}}{2}\right). \tag{3.26}$$

This transformation was designed specifically for use with hydrological data which simultaneously exhibit positive skewness, and the tendency for variance to initially increase with y but stabilize for larger y. In contrast, Box-Cox transformations (Equations 3.20 or 3.23) with exponent $\lambda<1$ will be effective at improving symmetry for positively skewed data, but will increasingly suppress variance

as the data values become larger. The transformation in Equation 3.26 can be further elaborated by adjusting the tapering of the variance suppression by working with the linearly transformed variable $y = a + bx$, where now x is the original untransformed quantity, and a and b represent tuning constants.

3.4.3. Standardized Anomalies

Linear transformations, which do not change the shape of the data distribution, can nevertheless be useful when we are interested in working simultaneously with batches of data that are related but not strictly comparable because of differences in location and/or scale. One instance of this situation occurs when the data are subject to seasonal variations. Direct comparison of raw monthly temperatures, for example, will usually show little more than the dominating influence of the seasonal cycle: at most northern midlatitude locations a record warm January will still be much colder than a record cool July. In situations of this sort, reexpression of the data in terms of *standardized anomalies* can be very helpful.

The standardized anomaly, z, is computed simply by subtracting the sample mean of the raw data x and dividing by the corresponding sample standard deviation:

$$z = \frac{x - \overline{x}}{s_x} = \frac{x'}{s_x}. \tag{3.27}$$

In the jargon of the atmospheric sciences, an *anomaly* x' is understood to mean the subtraction from a data value of a relevant average, as in the numerator of Equation 3.27. The term anomaly does not connote a data value or event that is abnormal or necessarily even unusual. The standardized anomaly in Equation 3.27 is produced by dividing the anomaly in the numerator by the corresponding standard deviation. This transformation is sometimes also referred to as a *normalization*. It would also be possible to construct standardized anomalies using resistant measures of location and spread, for example, subtracting the median and dividing by IQR, but this is rarely done. Use of standardized anomalies is motivated by ideas related to the bell-shaped Gaussian distribution, which are explained in Section 4.4.2. However, it is not necessary to assume that a batch of data follows any particular distribution in order to reexpress them in terms of standardized anomalies, and transforming non-Gaussian data according to Equation 3.27 will not make their distribution shape be any more Gaussian, because linear transformations do not change the shape of a data distribution.

The idea behind the standardized anomaly is to try to remove the influences of location and spread from a data sample. The physical units of the original data cancel, so standardized anomalies are always dimensionless quantities. Subtracting the mean produces a series of anomalies, x', located somewhere near zero. Division by the standard deviation puts excursions from the mean in different batches of data on equal footings. Collectively, a data sample that has been transformed to a set of standardized anomalies will exhibit a mean of zero and a standard deviation of 1.

To illustrate, it is often the case that summer temperatures are less variable than winter temperatures. We might find that the standard deviation for average January temperature at some location is around 3°C, but that the standard deviation for average July temperature at the same location is close to 1°C. An average July temperature 3°C colder than the long-term mean for July would then be quite unusual, corresponding to a standardized anomaly of −3. An average January temperature 3°C warmer than the long-term mean January temperature at the same location would be a fairly ordinary occurrence, corresponding to a standardized anomaly of only +1. Another way to look at the standardized anomaly is as a measure of distance, in standard deviation units, between a data value and its mean.

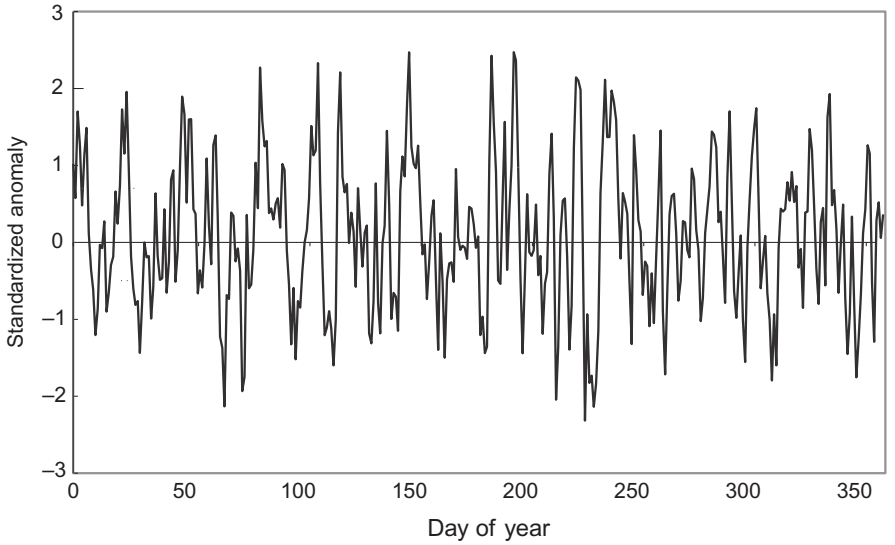


FIGURE 3.15 Standardized anomalies for the annual cycle of average daily temperatures at Boston, Massachusetts (USA), 1920–1999. *Adapted from Godfrey et al. (2002). © American Meteorological Society. Used with permission.*

Example 3.5. Standardizing a Nonstationary Time Series

Figure 3.15 demonstrates the idea of data standardization, comparing average daily temperatures throughout the year, at Boston, USA. Of course the raw temperatures exhibit warmer means in summer and colder means in winter, but also exhibit greater variability in winter than summer. The 365 daily standardized anomalies plotted in Figure 3.15 have been constructed by subtracting means for each day over the 80 years 1920–1999, and then dividing by the corresponding day-by-day standard deviations computed over the same period. That is, $z(t) = [x(t) - \overline{x}(t)]/s(t)$, where $\overline{x}(t)$ and $s(t)$ have been computed as smooth functions of the date t, as described in Section 11.4. The procedure achieves comparability of the values across seasons even though the means and standard deviations vary strongly throughout the year. Overall, the resulting standardized anomalies are centered near zero, and nearly all are smaller than 2 in absolute value. It is sometimes suggested that the peak around day 22 (the "January thaw") is somehow unusual, but Figure 3.15 indicates that this interpretation is simply an artifact of temperatures in winter being more variable than in other seasons (Godfrey et al., 2002).

Figure 3.15 illustrates the use of standardized anomalies to allow comparability across time, but the same idea can be applied for spatial comparisons by computing the standardized anomalies using location-specific means and standard deviations (e.g., Dabernig et al., 2017). ◇

Example 3.6. Expressing Climatic Data in Terms of Standardized Anomalies

Figure 3.16 illustrates the use of standardized anomalies in an operational context, where Equation 3.27 is applied twice. The plotted points are values of the *Southern Oscillation Index*, which is an index of the atmospheric component of El-Niño-Southern Oscillation (ENSO) phenomenon, that is used by the Climate Prediction Center of the U.S. National Centers for Environmental Prediction (Ropelewski and Jones, 1987). The values of this index in the figure are derived from month-by-month differences in the standardized anomalies of sea-level pressure at two tropical locations: Tahiti, in the central Pacific Ocean; and Darwin, in northern Australia. In terms of Equation 3.27 the first step toward generating

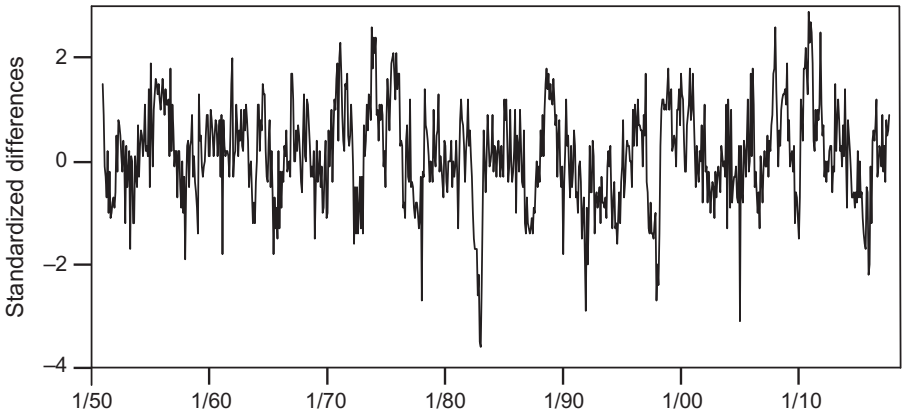


FIGURE 3.16 Standardized differences between the standardized monthly sea level pressure anomalies at Tahiti and Darwin (Southern Oscillation Index), January 1951–October 2017.

Figure 3.16 is to calculate the difference $\Delta z = z_{\text{Tahiti}} - z_{\text{Darwin}}$ for each month during the years plotted. The standardized anomaly z_{Tahiti} for January 1951, for example, is computed by subtracting the average pressure for all Januaries at Tahiti from the observed monthly pressure for January 1951. This difference is then divided by the standard deviation characterizing the year-to-year variations of January atmospheric pressure at Tahiti.

The curve in Figure 3.16 is based on monthly values that are themselves standardized anomalies of this difference of standardized anomalies Δz, so that Equation 3.27 has been applied three times to the original data. The first two of the standardizations are undertaken to minimize the influences of seasonal changes in the average monthly pressures and the year-to-year variability of the monthly pressures, separately at the two locations. The third standardization, calculating the standardized anomalies of the differences Δz, ensures that the resulting index will have unit standard deviation. For reasons that will be made clear in the discussion of the Gaussian distribution in Section 4.4.2, this attribute aids qualitative judgments about the unusualness of a particular index value.

Physically, during El Niño events the center of tropical Pacific precipitation activity shifts eastward from the western Pacific (near Darwin) to the central Pacific (near Tahiti). This shift is associated with higher than average surface pressures at Darwin and lower than average surface pressures at Tahiti, which together produce a negative value for the index plotted in Figure 3.16. The exceptionally strong El-Niño event of 1982–1983 is especially prominent in this figure. ◇

3.5. EXPLORATORY TECHNIQUES FOR PAIRED DATA

The ideas presented so far in this chapter have pertained mainly to the manipulation and investigation of single batches of data. Some comparisons have been made, such as the side-by-side schematic plots in Figure 3.5. There, several distributions of data from Appendix A were plotted, but potentially important aspects of the structure of those data were not shown. In particular, the relationships between variables observed on a given day were masked when the data from each batch were separately ranked prior to construction of the schematic plots. However, for each observation in one batch there is a corresponding observation from the same date in any one of the others. In this sense, these data are *paired*. Elucidating relationships among sets of data pairs often yields important insights.

3.5.1. Scatterplots

The nearly universal format for graphically displaying paired data is the familiar *scatterplot* or *x–y plot*. Geometrically, a scatterplot is simply a collection of points in the plane whose two Cartesian coordinates are the values of each member of the data pair. Scatterplots allow easy examination of such features in the data as trends, curvature in the relationship, clustering of one or both variables, changes of spread of one variable as a function of the other, and extraordinary points or outliers.

Figure 3.17 is a scatterplot of the maximum and minimum temperatures for Ithaca during January 1987. It is immediately apparent that very cold maxima are associated with very cold minima, and there is a tendency for the warmer maxima to be associated with the warmer minima. This scatterplot also shows that the central range of maximum temperatures is not strongly associated with minimum temperature, since maxima near 30°F occur with minima anywhere in the range of −5° to 20°F, or warmer.

Two optional but often useful embellishments on the scatterplot are also illustrated in Figure 3.17. The first of these is to sketch the individual data distributions, called the *marginal distributions*, using the lines on the upper and right-hand boundaries of the frame. These lines give the visual impression of fringe on a floor rug, and so Figure 3.17 is an example of a *rug plot*. The second embellishment is

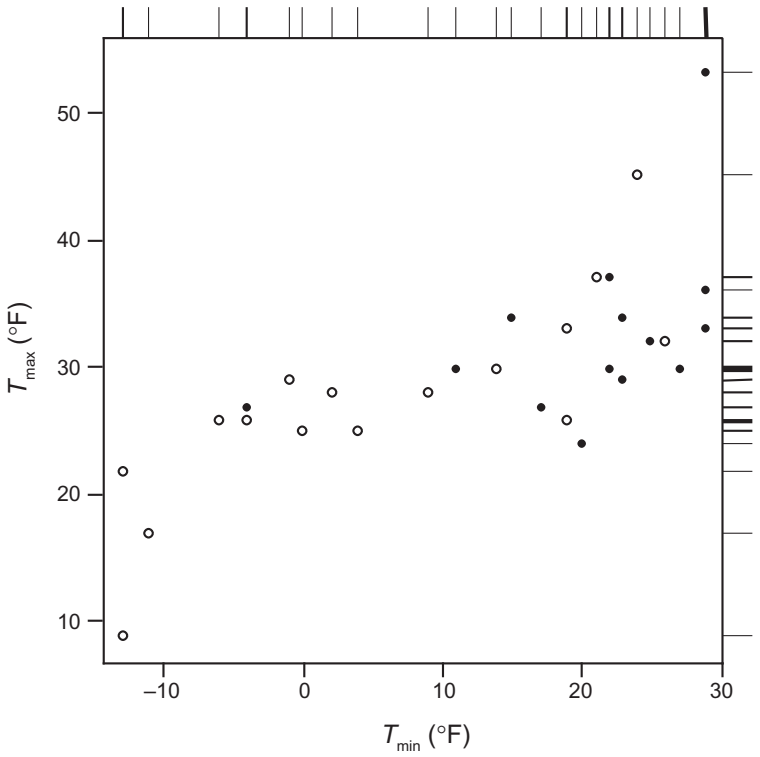


FIGURE 3.17 Scatterplot for daily maximum and minimum temperatures during January 1987 at Ithaca, New York. "Fringes" along the margins separately indicate the individual data distributions, with repeated data represented by heavier lines. Closed circles represent days with at least 0.01 in. of precipitation (liquid equivalent).

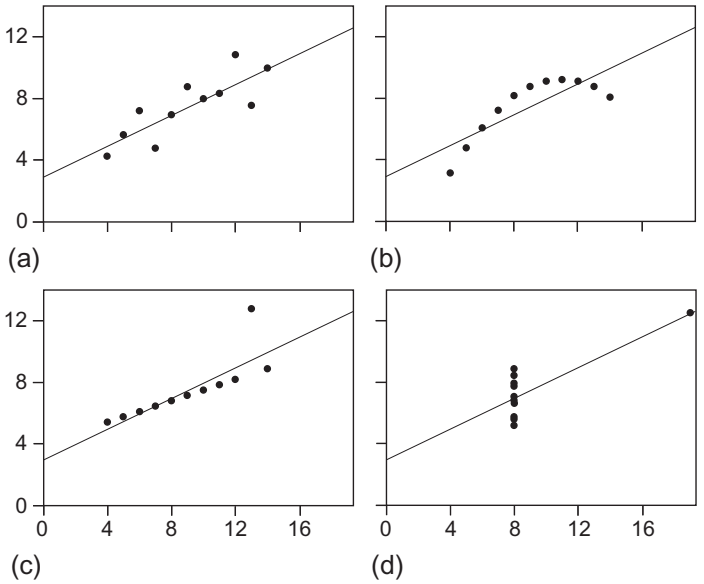


FIGURE 3.18 "Anscombe's quartet," illustrating the ability of graphical EDA to discern data features more powerfully than can a few numerical summaries. Each horizontal ("x") variable in panels (a) – (d) has the same mean (9.0) and standard deviation (11.0), as does each of the vertical ("y") variables (mean 7.5, standard deviation 4.12). Both the ordinary (Pearson) correlation coefficient ($r_{x,y}=0.816$) and the conventional least-squares regression relationship ($y=3+x/2$) are the same for all four of the panels.

the use of more than one type of plotting symbol. Here points representing days on which at least 0.01 in. (liquid equivalent) of precipitation were recorded are plotted using the filled circles. As was evident in Example 2.1 concerning conditional probability, precipitation days tend to be associated with warmer minimum temperatures. The scatterplot indicates that the maximum temperatures tend to be warmer as well on wet days, but that the effect is not as pronounced.

The scatterplots in Figure 3.18, known as *Anscombe's quartet* (Anscombe, 1973), illustrate the power of graphical EDA relative to computation of a few simple numerical summaries. The four sets of x–y pairs have been designed to have the same means and standard deviations in each panel, and the same ordinary (Pearson) correlation coefficient (Section 3.5.2) and the same least-squares linear regression relationship (Section 7.2.1). However, it is clear only from the graphical expositions that the relationships between the pairs of variables are very different in each case.

3.5.2. Pearson (Ordinary) Correlation

Often an abbreviated, single-valued measure of association between two variables, say x and y, is needed. In such situations, data analysts almost automatically (and sometimes somewhat uncritically) calculate a correlation coefficient. Usually, the term correlation coefficient is used to mean the "Pearson product-moment coefficient of linear correlation" between two variables x and y.

One way to view the *Pearson correlation* is as the ratio of the sample covariance between the two variables to the product of the two standard deviations,

$$r_{x,y} = \frac{\mathrm{Cov}(x,y)}{s_x s_y} = \frac{\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}[(x_i-\bar{x})(y_i-\bar{y})]}{\left[\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^{1/2}\left[\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(y_i-\bar{y})^2\right]^{1/2}} = \frac{\displaystyle\sum_{i=1}^{n}(x_i' y_i')}{\left[\displaystyle\sum_{i=1}^{n}(x_i')^2\right]^{1/2}\left[\displaystyle\sum_{i=1}^{n}(y_i')^2\right]^{1/2}}, \tag{3.28}$$

where the primes denote anomalies, or subtraction of mean values, as before. Note that the sample variance is a special case of the covariance (numerator in Equation 3.28), with $x = y$. One application of the covariance is in the mathematics used to describe turbulence, where the average product of, for example, the horizontal velocity anomalies u' and v' is called the *eddy covariance*, and is used in the framework of Reynolds averaging (e.g., Stull, 1988).

The Pearson product-moment correlation coefficient is neither robust nor resistant. It is not robust because strong but nonlinear relationships between the two variables x and y may not be recognized. It is not resistant since it can be extremely sensitive to one or a few outlying point pairs. Nevertheless it is often used, both because its form is well suited to mathematical manipulation, and because it is closely associated with regression analysis (see Section 7.2), and the bivariate (Equation 4.31) and multivariate (see Chapter 12) Gaussian distributions.

The Pearson correlation has two important properties. First, it is bounded by -1 and 1; that is, $-1 \le r_{x,y} \le 1$. If $r_{x,y} = -1$ there is a perfect, negative linear association between x and y. That is, the scatterplot of y versus x consists of points all falling along one line, and that line has negative slope. Similarly if $r_{x,y} = 1$ there is a perfect positive linear association. Note, however, that $|r_{x,y}| = 1$ says nothing about the slope of the perfect linear relationship between x and y except that it is not zero, and it says nothing about any vertical offset in their relationship that may be evident in a scatterplot. The second important property is that the square of the Pearson correlation, $r_{x,y}^2$, specifies the proportion of the variability of one of either x or y that is linearly accounted for, or described, by the other. It is sometimes said that $r_{x,y}^2$ is the proportion of the variance of one variable "explained" by the other, but this interpretation is imprecise at best and is sometimes misleading. The correlation coefficient provides no explanation at all about the relationship between the variables x and y, at least not in any physical or causative sense. It may be that x physically causes y or vice versa, but often both result physically from some other or many other quantities or processes.

The heart of the Pearson correlation coefficient is the covariance between x and y in the numerator of Equation 3.28. The denominator is in effect just a scaling constant and is always positive. Thus the Pearson correlation is essentially a nondimensionalized covariance. Consider the hypothetical cloud of (x, y) data points in Figure 3.19, recognizable immediately as exhibiting positive correlation. The two perpendicular lines passing through the two sample means define four quadrants, labeled conventionally using Roman numerals. For points in quadrant I, both the x and y values are larger than their respective means ($x' > 0$ and $y' > 0$), so that both factors being multiplied will be positive. Therefore points in quadrant I contribute positive terms to the sum in the numerator of Equation 3.28. Similarly, for points in quadrant III, both x and y are smaller than their respective means ($x' < 0$ and $y' < 0$), and again the product of their anomalies will be positive. Thus points in quadrant III will also contribute positive terms

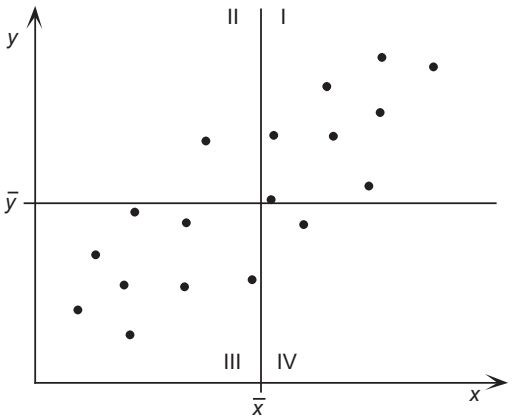


FIGURE 3.19 A hypothetical cloud of points in two dimensions, illustrating the mechanics of the Pearson correlation coefficient (Equation 3.28). The two sample means divide the plane into four quadrants, numbered I–IV.

to the sum in the numerator. For points in quadrants II and IV one of the two variables x and y is above its mean and the other is below. Therefore the product in the numerator of Equation 3.28 will be negative for points in quadrants II and IV, and these points will contribute negative terms to the sum.

Most of the points in Figure 3.19 are in either quadrants I or III, and therefore most of the terms in the numerator of Equation 3.28 are positive. Only the two points in quadrants II and IV contribute negative terms, and these are small in absolute value since the x and y values are relatively close to their respective means. The result is a positive sum in the numerator and therefore a positive covariance. The two standard deviations in the denominator of Equation 3.28 must always be positive, which yields a positive correlation coefficient overall for the points in Figure 3.19. If most of the points had been in quadrants II and IV, the point cloud would slope downward rather than upward, and the correlation coefficient would be negative. If the point cloud were more or less evenly distributed among the four quadrants the correlation coefficient would be near zero, since the positive and negative terms in the sum in the numerator of Equation 3.28 would tend to cancel.

Another way of looking at the Pearson correlation coefficient is produced by moving the scaling constants in the denominator (the standard deviations), inside the summation of the numerator. This operation yields

$$r_{x,y} = \frac{1}{n-1}\sum_{i=1}^{n}\left[\frac{(x_i-\bar{x})}{s_x}\frac{(y_i-\bar{y})}{s_y}\right] = \frac{1}{n-1}\sum_{i=1}^{n} z_{x_i} z_{y_i}, \tag{3.29}$$

showing that the Pearson correlation is (nearly) the average product of the variables after conversion to standardized anomalies.

From the standpoint of computational economy, the formulas presented so far for the Pearson correlation are awkward. This is true whether or not the computation is to be done by hand or by a computer program. In particular, they all require two passes through a data set before the result is achieved: the first to compute the sample means, and the second to accumulate the terms involving deviations of the data values from their sample means (the anomalies). Passing twice through a data set requires twice the effort and provides double the opportunity for keying errors when using a hand calculator, and can amount to substantial increases in computer time, especially when using small computing systems such as in data

loggers (Farrugia and Micallef, 2006). Therefore it is often useful to know the *computational form* of the Pearson correlation, which allows it to be calculated with only one pass through a data set.

The computational form arises through an easy algebraic manipulation of the summations in the correlation coefficient. Consider the numerator in Equation 3.28. Carrying out the indicated multiplication yields

$$\begin{aligned}\sum_{i=1}^{n}[(x_i-\bar{x})(y_i-\bar{y})] &= \sum_{i=1}^{n}[x_iy_i - x_i\bar{y} - y_i\bar{x} - \bar{x}\bar{y})] \\ &= \sum_{i=1}^{n}(x_iy_i) - \bar{y}\sum_{i=1}^{n}x_i - \bar{x}\sum_{i=1}^{n}y_i + \bar{x}\bar{y}\sum_{i=1}^{n}(1) \\ &= \sum_{i=1}^{n}(x_iy_i) - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y} \\ &= \sum_{i=1}^{n}(x_iy_i) - \frac{1}{n}\left[\sum_{i=1}^{n}x_i\right]\left[\sum_{i=1}^{n}y_i\right]\end{aligned} . \tag{3.30}$$

The second line in Equation 3.30 is arrived at through the realization that the sample means are constant, once the individual data values are determined, and therefore can be moved (factored) outside the summations. In the last term on this line there is nothing left inside the summation but the number 1, and the sum of n of these is simply n. The third step recognizes that the sample size multiplied by the sample mean yields the sum of the data values, which follows directly from the definition of the sample mean (Equation 3.2). The fourth step simply substitutes again the definition of the sample mean, to emphasize that all the quantities necessary for computing the numerator of the Pearson correlation can be known after one pass through the data. These are the sum of the x's, the sum of the y's, and the sum of their products.

It should be apparent from the similarity in form between Equation 3.30 and the summations in the denominator of the Pearson correlation that analogous formulas can be derived for them or, equivalently, for the sample standard deviation. The mechanics of the derivation are exactly as followed in Equation 3.30, with the result being

$$s_x = \left[\frac{\Sigma x_i^2 - n\bar{x}^2}{n-1}\right]^{1/2} = \left[\frac{\Sigma x_i^2 - \frac{1}{n}(\Sigma x_i)^2}{n-1}\right]^{1/2} . \tag{3.31}$$

A similar result, of course, is obtained for y. Mathematically, Equation 3.31 is exactly equivalent to the formula for the sample standard deviation in Equation 3.6. Thus Equations 3.30 and 3.31 can be substituted into the form of the Pearson correlation given in Equations 3.28 or 3.29, to yield the computational form for the correlation coefficient

$$r_{x,y} = \frac{\sum_{i=1}^{n}x_iy_i - \frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)}{\left[\sum_{i=1}^{n}x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2\right]^{1/2}\left[\sum_{i=1}^{n}y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}y_i\right)^2\right]^{1/2}} . \tag{3.32}$$

Analogously, a computational form for the sample skewness coefficient (Equation 3.9) is

$$\gamma = \frac{\frac{1}{n-1}\left[\Sigma x_i^3 - \frac{3}{n}(\Sigma x_i)(\Sigma x_i^2) + \frac{2}{n^2}(\Sigma x_i)^3\right]}{s^3}. \tag{3.33}$$

It is important to mention a cautionary note regarding the computational forms just derived. There is a potential problem inherent in their use, which stems from the fact that they are very sensitive to roundoff errors. The problem arises because these formulas involve differences of two numbers that may be of comparable magnitude. To illustrate, suppose that the two terms on the last line of Equation 3.30 have each been saved to five significant digits. If the first three of these digits are the same, their difference will then be known only to two significant digits rather than five. The remedy to this potential problem is to retain as many as possible (preferably all) of the significant digits in each calculation, for example, by using the double-precision representation when programming floating-point calculations on a computer.

Example 3.7. Some Limitations of Linear Correlation

Consider the two artificial data sets in Table 3.4. The data values are few and small enough that the computational form of the Pearson correlation can be used without discarding any significant digits. For Set I, the Pearson correlation is $r_{x,y}=+0.88$, and for Set II the Pearson correlation is $r_{x,y}=+0.61$. Thus moderately strong linear relationships appear to be indicated for both sets of paired data.

The Pearson correlation is neither robust nor resistant, and these two small data sets have been constructed to illustrate these deficiencies. Figure 3.20 shows scatterplots for the two data sets, with Set I in panel (a) and Set II in panel (b). For Set I the relationship between x and y is actually stronger than indicated by the linear correlation of 0.88. The data points all fall very nearly on a smooth curve, but since that curve is not a straight line the Pearson coefficient underestimates the strength of the relationship. It is not robust to deviations from linearity in a relationship.

TABLE 3.4 Artificial Paired Data Sets for Correlation Examples

Set I		Set II	
x	y	x	y
0	0	2	8
1	3	3	4
2	6	4	9
3	8	5	2
5	11	6	5
7	13	7	6
9	14	8	3
12	15	9	1
16	16	10	7
20	16	20	17

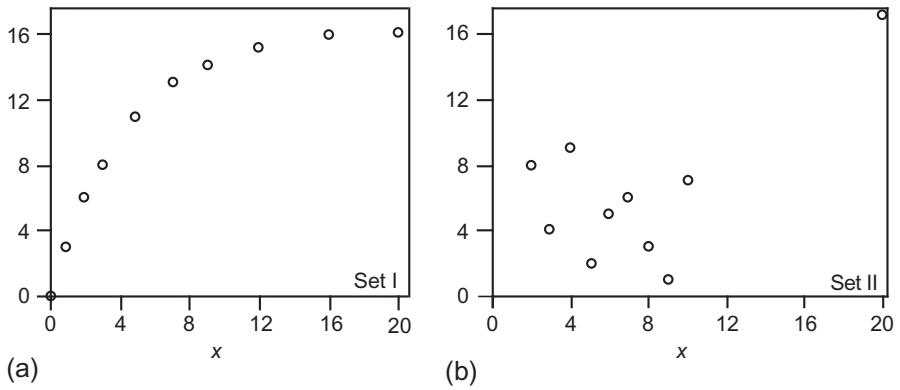


FIGURE 3.20 Scatterplots of the two artificial sets of paired data in Table 3.4. The Pearson correlation for the data in panel (a) (Set I in Table 3.4) of only 0.88 underrepresents the strength of the relationship, illustrating that this measure of correlation is not robust to nonlinearities. The Pearson correlation for the data in panel (b) (Set II) is 0.61, reflecting the overwhelming influence of the single outlying point, and illustrating lack of resistance.

Figure 3.20b illustrates that the Pearson correlation coefficient is not resistant to outlying data. Except for the single outlying point, the data in Set II exhibit very little structure. If anything these remaining nine points are weakly negatively correlated. However, the data pair $x=20$ and $y=17$ are so far from their respective sample means that the product of the resulting two large positive differences in the numerator of Equation 3.28 or Equation 3.29 dominate the entire sum, which erroneously indicates a moderately strong positive relationship among the ten data pairs overall. ◇

3.5.3. Spearman Rank Correlation and Kendall's τ

Robust and resistant alternatives to the Pearson product-moment correlation coefficient are available. The first of these is known as the *Spearman rank correlation* coefficient. The Spearman correlation is simply the Pearson correlation coefficient computed using the ranks of the data. Conceptually, either Equation 3.28 or Equation 3.29 is applied, but to the ranks of the data rather than to the data values themselves. For example, consider the first data pair, (2, 8), in Set II of Table 3.4. Here $x=2$ is the smallest of the 10 values of x and therefore has rank 1. Being the eighth smallest of the 10, $y=8$ has rank 8. Thus this first data pair would be transformed to (1, 8) before computation of the correlation. Similarly, both x and y values in the outlying pair (20,17) are the largest of their respective batches of 10 and would be transformed to (10, 10).

In practice it is not necessary to use Equation 3.28, 3.29, or 3.32 to compute the Spearman rank correlation. Rather, the computations are simplified because we know in advance what the transformed values will be. Because the data are ranks, they consist simply of all the integers from 1 through the sample size n. For example, the average of the ranks of any of the four data batches in Table 3.4 is $(1+2+3+4+5+6+7+8+9+10)/10=5.5$. Similarly, the standard deviation (Equation 3.31) of these first ten positive integers is about 3.028. More generally, the average of the integers from 1 to n is $(n+1)/2$, and their variance is $n(n+1)/12$. Taking advantage of this information, computation of the Spearman rank correlation can be simplified to

$$r_{rank} = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n\left(n^2-1\right)}, \tag{3.34}$$

where D_i is the difference in ranks between the ith pair of data values. In cases of ties, where a particular data value appears more than once, all of these equal values are assigned their average rank before computing the D_i's.

Kendall's τ is a second robust and resistant alternative to the conventional Pearson correlation. Kendall's τ is calculated by considering the relationships among all possible matchings of the data pairs (x_i, y_i), of which there are $n(n-1)/2$ in a sample of size n. Any such matching in which both members of one pair are larger than their counterparts in the other pair are called *concordant*. For example, the pairs (3, 8) and (7, 83) are concordant because both numbers in the latter pair are larger than their counterparts in the former. Matchups in which each pair has one of the larger values, for example (3, 83) and (7, 8), are called *discordant*. The slope of the line segment connecting a concordant pair will be positive, and the slope of the segment connecting a discordant pair will be negative. Kendall's τ is calculated by subtracting the number of discordant pairs, N_D, from the number of concordant pairs, N_C, and dividing by the number of possible matchups among the n observations,

$$\tau = \frac{N_C - N_D}{n(n-1)/2}. \tag{3.35}$$

Pairs for which one or both elements are identical contribute 1/2 to both N_C and N_D.

Example 3.8. Comparison of Spearman and Kendall Correlations for the Table 3.4 Data
In Set I of Table 3.4, there is a monotonic relationship between x and y, so that each of the two batches of data is already arranged in ascending order. Therefore both members of each of the n pairs have the same rank within its own batch, and the differences D_i are all zero. Actually, because of rounding the two largest y values are equal, so that each would be assigned the rank 9.5. Other than this tie, the sum in the numerator of the second term in Equation 3.34 is zero, and the Spearman rank correlation is essentially 1. This result better characterizes the strength of the relationship between x and y than does the Pearson correlation of 0.88. The Pearson correlation coefficient reflects the strength of linear relationships, but the Spearman rank correlation reflects the strength of monotone relationships.

Because the data in Set I exhibit an essentially perfect positive monotone relationship, all of the 10 $(10-1)/2 = 45$ possible matchups between data pairs yield concordant relationships. For data sets with perfect negative monotone relationships (one of the variables is strictly decreasing as a function of the other), all comparisons among data pairs yield discordant relationships. Except for the one tie, all comparisons for Set I are concordant relationships. $N_C = 44.5$, so that Equation 3.31 would produce $\tau = (44.5 - 0.5)/45 = 0.978$.

For the data in Set II, the x values are presented in ascending order, but the y values with which they are paired are jumbled. The difference of ranks for the first record is $D_1 = 1 - 8 = -7$. There are only three data pairs in Set II for which the ranks match (the fifth, sixth, and the outliers of the tenth pair). The remaining seven pairs will contribute nonzero terms to the sum in Equation 3.34, yielding $r_{rank} = 0.018$ for Set II. This result reflects much better the very weak overall relationship between x and y in Set II than does the Pearson correlation of 0.61.

Calculation of Kendall's τ for Set II is facilitated by their being sorted according to increasing values of the x variable. Given this arrangement, the number of concordant combinations can be determined by

counting the number of subsequent y variables that are larger than each of the first through $(n-1)$st listings in the table. Specifically, there are two y variables larger than 8 in (2, 8) among the nine values below it, five y variables larger than 4 in (3, 4) among the eight values below it, one y variable larger than 9 in (4, 9) among the seven values below it, …, and one y variable larger than 7 in (10, 7) in the single value below it. Together there are $2+5+1+5+3+2+2+2+1=23$ concordant combinations, and $45-23=22$ discordant combinations, yielding $\tau=(23-22)/45=0.022$. ◇

3.5.4. Serial Correlation

In Chapter 2 meteorological persistence, or the tendency for weather in successive time periods to be similar, was illustrated in terms of conditional probabilities for the two discrete events "precipitation" and "no precipitation." For continuous variables (e.g., temperature), persistence typically is characterized in terms of *serial correlation* or *temporal autocorrelation*. The prefix "auto" in autocorrelation denotes the correlation of a variable with itself, so that temporal autocorrelation indicates the correlation of a variable with its own future and past values. Sometimes such correlations are referred to as *lagged correlations*. Almost always, autocorrelations are computed as Pearson product-moment correlation coefficients, although there is no reason why other forms of lagged correlation cannot be computed as well.

The process of computing autocorrelations can be visualized by imagining two copies of a sequence of data values being written, with one of the series shifted by one unit of time. This shifting is illustrated in Figure 3.21, using the January 1987 Ithaca maximum temperature data from Table A.1. This data series has been rewritten, with the middle part of the month represented by ellipses, on the first line. The same record has been recopied on the second line, but shifted to the right by one day. This process results in 30 pairs of temperatures within the box, which are available for the computation of a correlation coefficient.

Autocorrelations are computed by substituting the lagged data pairs into the formula for the Pearson correlation (Equation 3.28). For the lag-1 autocorrelation there are $n-1$ such pairs. The only real confusion arises because the mean values for the two series will in general be slightly different. In Figure 3.21, for example, the mean of the 30 boxed values in the upper series is 29.77°F, and the mean for the boxed values in the lower series is 29.73°F. This difference arises because the upper series does not include the temperature for 1 January, and the lower series does not include the temperature for 31 January. Denoting the sample mean of the first $n-1$ values with the subscript "−" and that of the last $n-1$ values with the subscript "+," the lag-1 autocorrelation is

$$r_1 = \frac{\sum_{i=1}^{n-1}[(x_i-\bar{x}_-)(x_{i+1}-\bar{x}_+)]}{\left[\sum_{i=1}^{n-1}(x_i-\bar{x}_-)^2\right]^{1/2}\left[\sum_{i=2}^{n}(x_i-\bar{x}_+)^2\right]^{1/2}} \tag{3.36}$$

For the January 1987 Ithaca maximum temperature data, for example, $r_1=0.52$.

33	32	30	29	25	30	53	· · ·	17	26	27	30	34	
	33	32	30	29	25	30	53	· · ·	17	26	27	30	34

FIGURE 3.21 Construction of a shifted time series of January 1987 Ithaca maximum temperature data. Shifting the data by one day leaves 30 data pairs (enclosed in the box) with which to calculate the lag-1 autocorrelation coefficient.

The lag-1 autocorrelation is the most commonly computed measure of persistence, but it is also sometimes of interest to compute autocorrelations at longer lags. Conceptually, this is no more difficult than the procedure for the lag-1 autocorrelation, and computationally the only difference is that the two series are shifted by more than one time unit. Of course, as a time series is shifted increasingly relative to itself there is progressively less overlapping data to work with. Equation 3.36 can be generalized to the lag-k autocorrelation coefficient using

$$r_k = \frac{\sum_{i=1}^{n-k}\left[(x_i-\overline{x}_-)(x_{i+k}-\overline{x}_+)\right]}{\left[\sum_{i=1}^{n-k}(x_i-\overline{x}_-)^2\right]^{1/2}\left[\sum_{i=k+1}^{n}(x_i-\overline{x}_+)^2\right]^{1/2}}. \tag{3.37}$$

Here the subscripts "−" and "+" indicate sample means over the first and last $n-k$ data values, respectively. Equation 3.37 is valid for $0 \le k < n-2$, although it is usually only the lowest few values of k that will be of interest. So much data is lost at large lags that lagged correlations for roughly $k > n/2$ or $k > n/3$ rarely are computed.

In situations where a long data record is available it is sometimes acceptable to use an approximation to Equation 3.37, which simplifies the calculations and allows use of a computational form. In particular, if the data series is sufficiently long, the overall sample mean will be very close to the subset averages of the first and last $n-k$ values. The overall sample standard deviation will be close to the two subset standard deviations for the first and last $n-k$ values as well. Invoking these assumptions leads to the commonly used approximation

$$r_k \approx \frac{\sum_{i=1}^{n-k}\left[(x_i-\overline{x})(x_{i+k}-\overline{x})\right]}{\sum_{i=1}^{n}(x_i-\overline{x})^2} = \frac{\sum_{i=1}^{n-k}(x_i x_{i+k}) - \frac{n-k}{n^2}\left(\sum_{i=1}^{n}x_i\right)^2}{\sum_{i=1}^{n}x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2}. \tag{3.38}$$

3.5.5. Autocorrelation Function

Together, the collection of autocorrelations computed for various lags is called the *autocorrelation function*. Often autocorrelation functions are displayed graphically, with the autocorrelations plotted as a function of lag. Figure 3.22 shows the first seven values of the sample autocorrelation function for the January 1987 Ithaca maximum temperature data. An autocorrelation function always begins with $r_0 = 1$, since any unshifted series of data will exhibit perfect correlation with itself. It is typical for an autocorrelation function to exhibit a more or less gradual decay toward zero as the lag k increases, reflecting the generally weaker statistical relationships between pairs of data points further removed from each other in time. It is instructive to relate this observation to the context of weather forecasting. If the autocorrelation function did not decay toward zero after a few days, making reasonably accurate forecasts at that range would be very easy: simply forecasting today's observation (the persistence forecast) or some modification of today's observation would give good results.

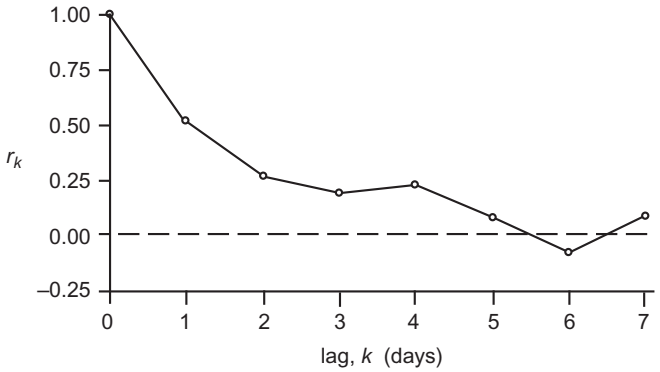


FIGURE 3.22 Sample autocorrelation function for the January 1987 Ithaca maximum temperature data. The correlation is 1 for $k=0$, since the unlagged data are perfectly correlated with themselves. The autocorrelation function decays to essentially zero for $k \geq 5$.

Sometimes it is useful to rescale the autocorrelation function, by multiplying all the autocorrelations by the variance of the data. The result, which is proportional to the numerators of Equations 3.37 and 3.38, is called the *autocovariance function*,

$$\gamma_k = \sigma^2 r_k, \; k = 0, 1, 2, \cdots. \tag{3.39}$$

The existence of autocorrelation in meteorological and climatological time series has important implications regarding the applicability of some standard statistical methods to atmospheric data. In particular, uncritical application of classical methods requiring independence of data within a sample will often give badly misleading results when applied to strongly persistent series. In some cases it is possible to successfully modify these techniques, by accounting for the temporal dependence using sample autocorrelations. This topic will be discussed in Chapter 5.

3.6. VISUALIZATION FOR HIGHER-DIMENSIONAL DATA

Graphical methods are essential when exploration, analysis, or comparison of matched data consisting of more than two variables is required. The methods presented so far can be applied only to pairwise subsets of the variables. Simultaneous display of three or more variables is more difficult due to a combination of geometric and cognitive problems. The geometric problem is that most available display media (e.g., paper and computer screens) are two-dimensional, so that directly plotting higher-dimensional data requires a geometric projection onto the plane, during which process information is inevitably lost. The cognitive problem derives from the fact that our brains have evolved to deal with life in a three-dimensional world, and visualizing four or more dimensions simultaneously is difficult or impossible.

Systematic study and articulation of principles for data visualization are relatively recent. In addition to Tukey's (1977) pathbreaking book, notable contributions include Tufte (1983, 1990), Cleveland (1994), and Wilkinson (2005). Some broad principles emerge from the work of these authors. Notably, Tufte (1983) calls for maximization of the ratio of "data-ink" to "total ink," and minimization of *chartjunk*, or gratuitous decoration. Cleveland (1994) expresses these two ideas as making the data stand out and avoiding superfluity. These books are also full of detailed suggestions for better

graphic design, such as (from Cleveland, 1994) drawing tick-marks on frame exteriors in order not to obscure the data, adjusting plot aspect ratios to best allow assessment of rates of change, and using logarithmic scales when understanding proportional changes or multiplicative factors is important, to name a few.

A variety of clever graphical tools have been devised for multivariate (three or more variables simultaneously) EDA, and there is plenty of room for the design of new graphical tools suited to particular purposes. In addition to the ideas presented in this section, some multivariate graphical EDA devices designed particularly for ensemble forecasts are described in Section 8.5, and a high-dimensional EDA approach based on principal component analysis is described in Section 13.7.3.

3.6.1. The Star Plot

If the number of variables, K, is not too large, each of a set of n K-dimensional observations can be displayed graphically as a *star plot*. The star plot is based on K coordinate axes sharing the same origin, spaced $360°/K$ apart on the plane. For each of the n observations, the value of the kth of the K variables is proportional (with perhaps some minimum value subtracted) to the radial plotting distance on the corresponding axis. The "star" consists of line segments connecting these points to their counterparts on adjacent radial axes.

For example, Figure 3.23 shows star plots for the last 5 (of $n=31$) days of the January 1987 data in Table A.1. Since there are $K=6$ variables, the six axes are separated by $360°/6=60°$, and each is identified with one of the variables as indicated in the plot for 27 January. In general the scales of proportionality on star plots are different for different variables and are designed so the smallest value (or some value near but below it) corresponds to the origin and the largest value (or some value near and above it) corresponds to the full length of the axis. Because the variables in Figure 3.23 are matched in type, the scales for the three types of variables have been chosen identically in order to better compare them. For example, the origin for both the Ithaca and Canandaigua maximum temperature axes corresponds to 10°F, and the ends of these axes correspond to 40°F. The precipitation axes have zero at the origin and 0.15 in. at the ends, so that the double-triangle shapes for 27 and 28 January indicate zero precipitation at both locations for those days. The near symmetry of the stars suggests strong correlations for the pairs of like variables (since their axes have been plotted 180 degree apart), and the tendency for the stars to get larger through time indicates warmer and wetter days at the end of the month.

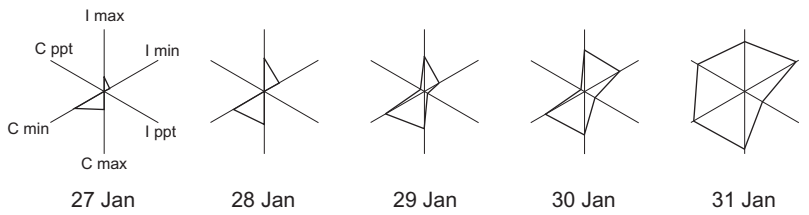


FIGURE 3.23 Star plots for the last five days in the January 1987 data in Table A.1, with axes labeled for the 27 January star only. Approximate radial symmetry in these plots reflects correlation between like variables at the two locations, and expansion of the stars through the time period indicates warmer and wetter days at the end of the month.

3.6.2. The Glyph Scatterplot

The *glyph scatterplot* is an extension of the ordinary scatterplot, in which the simple dots locating points on the two-dimensional plane defined by two variables are replaced by "glyphs," or more elaborate symbols that encode the values of additional variables in their sizes, shapes, and/or colors. Figure 3.15 is a primitive glyph scatterplot, with the filled/open circular glyphs indicating the binary precipitation/no-precipitation variable.

Figure 3.24 is a simple glyph scatterplot displaying three variables relating to evaluation of a small set of winter maximum temperature forecasts. The two scatterplot axes are the forecast and observed temperatures, rounded to 5°F bins, and the circular glyphs are drawn so that their areas are proportional to the numbers of forecast-observation pairs in a given 5×5°F square bin. Choosing area to be proportional to the third variable (here, counts in each bin) is preferable to radius or diameter because the glyph areas correspond better to the visual impression of size.

Figure 3.24 is essentially a two-dimensional histogram for this bivariate set of temperature data, but is more effective than a direct generalization to three dimensions of a conventional two-dimensional histogram for a single variable. Figure 3.25 shows such a perspective-view bivariate histogram for the same data, which is usually ineffective because projection of the three dimensions onto the two-dimensional page has introduced ambiguities about the locations of individual points. This is so, even though each point in Figure 3.25 is tied to its location on the forecast-observed plane at the apparent base of the plot through the vertical tails, and the points falling exactly on the diagonal are indicated by open plotting symbols. Figure 3.24 speaks more clearly than Figure 3.25 about the data, for example, showing immediately that there is an overforecasting bias (forecast temperatures systematically warmer than the corresponding observed temperatures, on average), particularly for the colder forecasts.

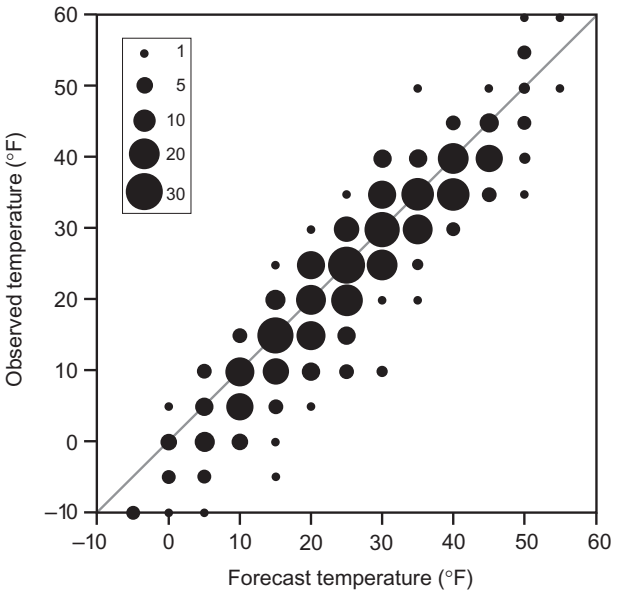


FIGURE 3.24 Glyph scatterplot of the bivariate frequency distribution of forecast and observed winter daily maximum temperatures for Minneapolis, 1980–1981 through 1985–1986. Temperatures have been rounded to 5°F intervals, and the circular glyphs have been scaled to have areas proportional to the counts (inset).

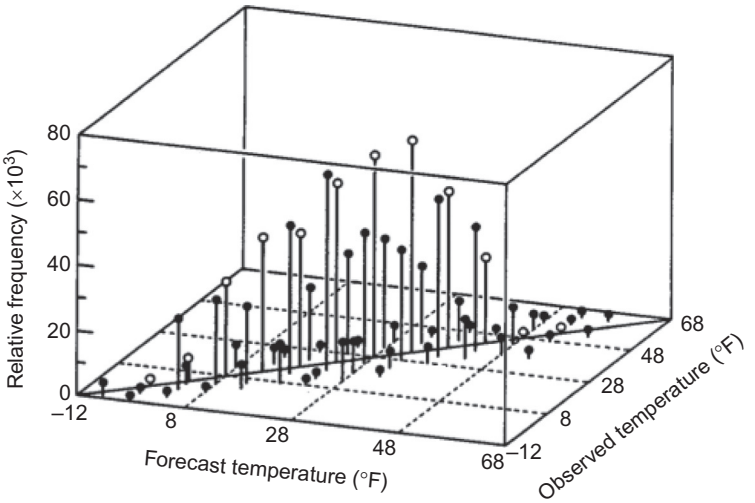
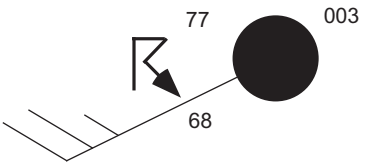


FIGURE 3.25 Bivariate histogram rendered in perspective view, of the same data plotted as a glyph scatterplot in Figure 3.24. Even though data points are located on the forecast-observation plane by the vertical tails and points on the 1:1 diagonal are further distinguished by open circles, the projection from three dimensions to two makes the figure difficult to interpret. *From Murphy et al. (1989). © American Meteorological Society. Used with permission.*

Effective alternatives to the glyph scatterplot in Figure 3.24 for displaying the bivariate frequency distribution might be a contour plot of the bivariate kernel density estimate (see Section 3.3.6) for these data; or a *hexbin plot*, which divides the Cartesian plane into a tessellation of hexagons and indicates the third dimension (in the present case, numbers of forecast-observation pairs) using either shading or color.

More elaborate glyphs than the circles in Figure 3.24 can be used to simultaneously display multivariate data with more than three variables. For example, star glyphs as described in Section 3.6.1 could be used as the plotting symbols in a glyph scatter plot. Virtually any shape that might be suggested by the data or the scientific context can be used in this way as a glyph. For example, Figure 3.26 shows a glyph that simultaneously displays seven meteorological quantities: wind direction, wind speed, sky cover, temperature, dew point temperature, pressure, and current weather condition. When these glyphs are plotted as a scatterplot defined by longitude (horizontal axis) and latitude (vertical axis), the result is a raw weather map, which is, in effect, a graphical EDA depiction of a nine-dimensional data set describing the spatial distribution of weather at a particular time. Similarly, Figure 3.27 shows a glyph map of observed linear temperature trends for the period 1950–2010 over much of North America.

FIGURE 3.26 An elaborate glyph, known as a meteorological station model, simultaneously depicting seven quantities. When plotted on a map, two location variables (latitude and longitude) are added as well, increasing the dimensionality of the depiction to nine, in what amounts to a glyph scatterplot of the weather data.



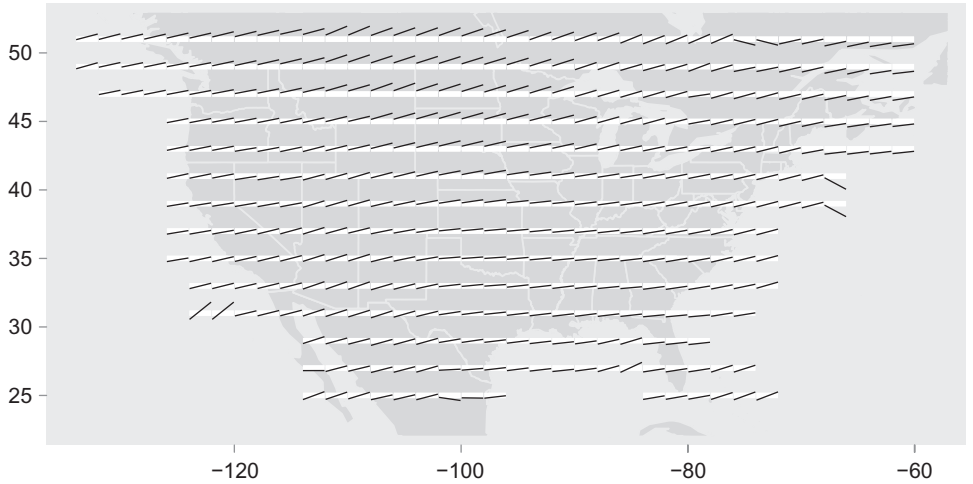


FIGURE 3.27 Glyph map of observed linear temperature trends for the period 1950–2010 over much of North America. *From Wickham et al. (2012).*

3.6.3. The Rotating Scatterplot

Figure 3.25 illustrates that it is generally unsatisfactory to attempt to extend the two-dimensional scatterplot to three dimensions by rendering it as a perspective view. The problem occurs because the three-dimensional counterpart of the scatterplot consists of a point cloud located in a volume rather than on the plane, and geometrically projecting this volume onto any one plane results in ambiguities about distances perpendicular to that plane. One solution to this problem is to draw larger and smaller symbols that are respectively closer to and further from the front in the direction of the projection, in a way that mimics the change in apparent size of physical objects with distance.

More effective, however, is to view the three-dimensional data in a computer animation known as a *rotating scatterplot*. At any instant the rotating scatterplot is a projection of the three-dimensional point cloud, together with its three coordinate axes for reference, onto the two-dimensional surface of the computer screen. But the plane onto which the data are projected can be changed smoothly in time, typically using the computer mouse, in a way that produces the illusion that we are viewing the points and their axes rotating around the three-dimensional coordinate origin, "inside" the computer monitor. The apparent motion can be rendered quite smoothly, and it is this continuity in time that allows a subjective sense of the shape of the data in three dimensions to be developed as we watch the changing display. In effect, the animation substitutes time for the missing third dimension.

It is not really possible to convey the power of this approach in the static form of a book page. However, an idea of how this works can be had from Figure 3.28, which shows four snapshots from a rotating scatterplot sequence, using the June Guayaquil data for temperature, pressure, and precipitation in Table A.3, with the 5 El Niño years indicated by open circles. Initially (Figure 3.28a) the temperature axis is oriented out of the plane of the page, so what appears is a simple two-dimensional scatterplot of precipitation versus pressure. In Figure 3.28b–d, the temperature axis is rotated into the plane of the page, which allows a gradually changing perspective on the arrangement of the points relative to each other and relative to the projections of the coordinate axes. Figure 3.28 shows only about 90

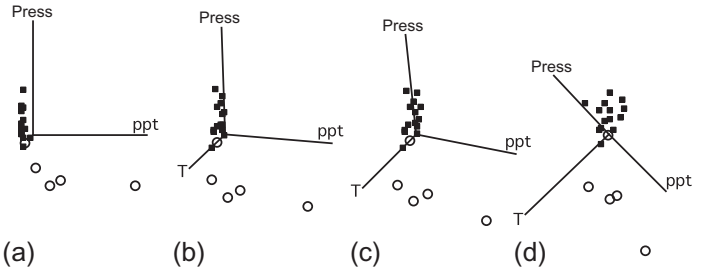


FIGURE 3.28 Four snapshots of the evolution of a three-dimensional rotating plot of the June Guayaquil data in Table A.3, in which the 5 El Niño years are shown as circles. The temperature axis is perpendicular to, and extends out of, the page in panel (a), and the three subsequent panels (b) – (d) show the changing perspectives as the temperature axis is rotated into the plane of the page, in a direction down and to the left. The visual illusion of a point cloud suspended in a three-dimensional space is much greater in a live rendition with continuous motion.

degree of rotation. A "live" examination of these data with a rotating plot usually would consist of choosing an initial direction of rotation (here, down, and to the left), allowing several full rotations in that direction, and then possibly repeating the process for other directions of rotation until an appreciation of the three-dimensional shape of the point cloud had developed.

3.6.4. The Correlation Matrix

The *correlation matrix* is a very useful device for simultaneously displaying correlations among more than two batches of matched data. For example, the data set in Table A.1 contains matched data for six variables. Correlation coefficients can be computed for each of the 15 distinct pairings of these six variables. In general, for K variables, there are $(K)(K-1)/2$ distinct pairings, and the correlations between them can be arranged systematically in a square array, with as many rows and columns as there are matched data variables whose relationships are to be summarized. Each entry in the array, $r_{i,j}$, is indexed by the two subscripts, i and j, that point to the identity of the two variables whose correlation is represented. For example, $r_{2,3}$ would denote the correlation between the second and third variables in a list. The rows and columns in the correlation matrix are numbered correspondingly, so that the individual correlations are arranged as shown in Figure 3.29.

FIGURE 3.29 The layout of a correlation matrix, $[R]$. Correlations $r_{i,j}$ between all possible pairs of variables are arranged so that the first subscript, i, indexes the row number, and the second subscript, j, indexes the column number.

$$[R] = \begin{bmatrix} r_{1,1} & r_{1,2} & r_{1,3} & r_{1,4} & \cdots & r_{1,J} \\ r_{2,1} & r_{2,2} & r_{2,3} & r_{2,4} & \cdots & r_{2,J} \\ r_{3,1} & r_{3,2} & r_{3,3} & r_{3,4} & \cdots & r_{3,J} \\ r_{4,1} & r_{4,2} & r_{4,3} & r_{4,4} & \cdots & r_{4,J} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{I,1} & r_{I,2} & r_{I,3} & r_{I,4} & \cdots & r_{I,J} \end{bmatrix}$$

Row number, i

Column number, j

The correlation matrix was not designed for exploratory data analysis, but rather as a notational shorthand that allows mathematical manipulation of the correlations in the framework of linear algebra (see Chapter 11). As a format for an organized exploratory arrangement of correlations, parts of the correlation matrix are redundant, and some are simply uninformative. Consider first the diagonal elements of the matrix, arranged from the upper left to the lower right corners, that is, $r_{1,1}, r_{2,2}, r_{3,3}, \ldots, r_{K,K}$. These are the correlations of each of the variables with themselves and are always equal to 1. Realize also that the correlation matrix is symmetric. That is, the correlation $r_{i,j}$ between variables i and j is exactly the same number as the correlation $r_{j,i}$, between the same pair of variables, so that the correlation values above and below the diagonal of 1's are mirror images of each other. Therefore as noted earlier, only $(K)(K-1)/2$ of the K^2 entries in the correlation matrix provide distinct information.

Table 3.5 presents correlation matrices for the data in Table A.1. The matrix on the left contains Pearson product-moment correlation coefficients, and the matrix on the right contains Spearman rank correlation coefficients. As is consistent with usual practice when using correlation matrices for display rather than computational purposes, only one of the upper and lower triangles of each matrix actually is printed. Omitted are the uninformative diagonal elements and the redundant upper triangular elements. Only the $(6)(5)/2 = 15$ distinct correlation values are presented.

Important features in the underlying data can be discerned by studying and comparing these two correlation matrices. First, notice that the six correlations involving only temperature variables have comparable values in both matrices. The strongest Spearman correlations are between like temperature variables at the two locations. Correlations between maximum and minimum temperatures at the same location are moderately large, but weaker. The correlations involving one or both of the precipitation variables differ substantially between the two correlation matrices. There are only a few very large precipitation amounts for each of the two locations, and these tend to dominate the Pearson correlations, as explained previously. On the basis of this comparison between the correlation matrices, we therefore would suspect that the precipitation data contained some outliers, even without the benefit of knowing the type of data, or of having seen the individual numbers. The rank correlations would be expected to better reflect the degree of association for data pairs involving one or both of the precipitation variables. Subjecting the precipitation variables to a monotonic transformation appropriate to reducing the skewness would produce no changes in the matrix of Spearman correlations, but would be expected to improve the agreement between the Pearson and Spearman correlations.

TABLE 3.5 Correlation Matrices for the Data in Table A.1

	Ith. Ppt	Ith. Max	Ith. Min	Can. Ppt	Can. Max	Ith. Ppt	Ith. Max	Ith. Min	Can. Ppt	Can. Max
Ith. Max	−0.024					0.319				
Ith. Min	0.287	0.718				0.597	0.761			
Can. Ppt	0.965	0.018	0.267			0.750	0.281	0.546		
Can. Max	−0.039	0.957	0.762	−0.015		0.267	0.944	0.749	0.187	
Can. Min	0.218	0.761	0.924	0.188	0.810	0.514	0.790	0.916	0.352	0.776

Only the lower triangle of the matrices is shown, to omit redundancies and the uninformative diagonal values. The left matrix contains Pearson product-moment correlations, and the right matrix contains Spearman rank correlations.

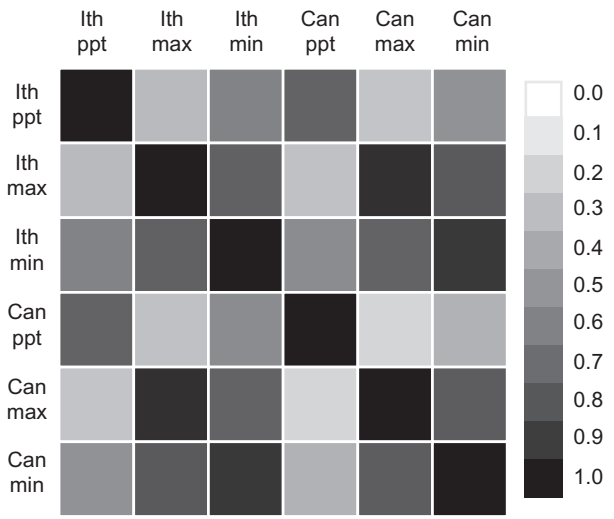


FIGURE 3.30 Heatmap of the Spearman correlations in Table 3.5, using a grayscale to indicate the magnitudes.

3.6.5. Heatmaps

When there are a large number of variables being related through their correlations, the very large number of pairwise comparisons can be overwhelming, in which case this arrangement of the numerical values in a correlation matrix is not particularly effective as an EDA device. However, different colors or shading levels can be assigned to particular ranges of correlation, and then plotted in the same two-dimensional arrangement as the numerical correlations on which they are based, in order to more directly gain a visual appreciation of the patterns of relationship. Figure 3.30 shows such a plot for the Spearman correlations presented in Table 3.5. This is an example of what is known as a *heatmap*, which can also be plotted using color instead of a grayscale. Here the full correlation matrix has been plotted, including the unit correlations along the main diagonal. The shading shows clearly that the strongest correlations are between like temperature variables and that the weakest correlations are between precipitation variables and temperature variables.

3.6.6. The Scatterplot Matrix

The *scatterplot matrix* is a graphical extension of the correlation matrix. The physical arrangement of the correlation coefficients in a correlation matrix is convenient for quick comparisons of relationships between pairs of variables, but distilling these relationships to a single number such as a correlation coefficient inevitably hides important details. A scatterplot matrix is an arrangement of individual scatterplots according to the same logic governing the placement of individual correlation coefficients in a correlation matrix.

Figure 3.31 is a scatterplot matrix for the January 1987 data in Table A.1, with the scatterplots arranged in the same pattern as the correlation matrices in Table 3.5. The complexity of a scatterplot matrix can be bewildering at first, but a large amount of information about the joint behavior of the data is displayed very compactly. For example, quickly evident from a scan of the precipitation rows and columns in Figure 3.31 is the fact that there are just a few large precipitation amounts at each of the

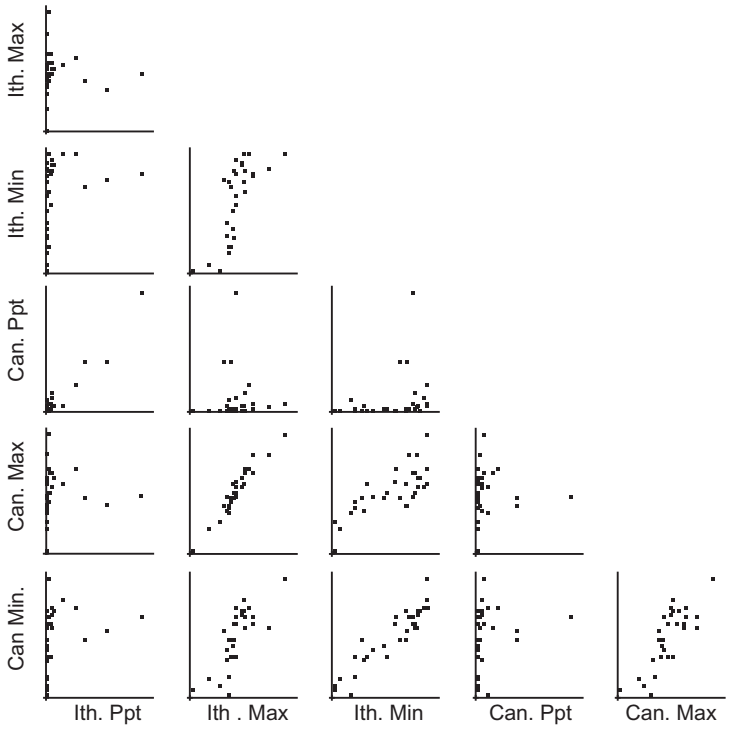


FIGURE 3.31 Scatterplot matrix for the January 1987 data in Table A.1 of Appendix A.

two locations. Looking vertically along the column for Ithaca precipitation, or horizontally along the row for Canandaigua precipitation, the eye is drawn to the largest few data values, which appear to line up. Most of the precipitation points correspond to small amounts and therefore cluster along the opposite axes. Focusing on the plot of Canandaigua versus Ithaca precipitation, it is apparent that the two locations received most of their precipitation for the month on the same few days. Also evident is the association of precipitation with milder minimum temperatures that was seen in previous looks at these same data. The closer relationships between maximum and maximum, or minimum and minimum temperature variables at the two locations—as compared to the maximum versus minimum temperature relationships at one location—can also be seen clearly.

The scatterplot matrix in Figure 3.31 has been drawn without the diagonal elements in the positions that correspond to the unit correlation of a variable with itself in a correlation matrix. A scatterplot of any variable with itself would be equally uninteresting, consisting only of a straight-line collection of points at a 45-degree angle. However, it is possible to use the diagonal positions in a scatterplot matrix to portray useful univariate information about the variable corresponding to that matrix position. One simple choice would be schematic plots of each of the variables in the diagonal positions. Another potentially useful choice is the *Q-Q* plot (Section 4.5.2) for each variable, which graphically compares the data with a reference distribution; for example, the bell-shaped Gaussian distribution. Sometimes the diagonal positions are used merely to contain labels for the respective variables.

The scatterplot matrix can be even more revealing if constructed using software allowing *brushing* of data points in related plots. When brushing, the analyst can select a point or set of points in one plot, and

the corresponding points in the same data record then also light up or are otherwise differentiated in all the other plots then visible. For example, when preparing Figure 3.17, the differentiation of Ithaca temperatures occurring on days with measurable precipitation was achieved by brushing another plot (that plot was not reproduced in Figure 3.17) involving the Ithaca precipitation values. The solid circles in Figure 3.17 thus constitute a temperature scatterplot conditional on nonzero precipitation. Brushing can also sometimes reveal surprising relationships in the data by keeping the brushing action of the computer mouse in motion. The resulting "movie" of brushed points in the other simultaneously visible plots essentially allows the additional dimension of time to be used for differentiating relationships in the data.

3.6.7. Correlation Maps

Correlation matrices such as those in Table 3.5 are understandable and informative, so long as the number of quantities represented (six, in the case of Table 3.5) remains reasonably small. When the number of variables becomes large it may not be possible to easily make sense of the individual values, or even to fit their correlation matrix on a single page. A frequent cause of atmospheric data being excessively numerous for effective display in a correlation or scatterplot matrix is the necessity of working simultaneously with data from a large number of locations. In this case the geographical arrangement of the locations can be used to organize the correlation information in map form.

Consider, for example, summarization of the correlations among surface pressure at perhaps 200 locations around the world. This would be only a modestly large set of data by modern standards. However, this many batches of pressure data would lead to $(200)(199)/2 = 19{,}100$ distinct station pairs, and as many correlation coefficients. A technique that has been used successfully in such situations is construction of a series of *one-point correlation maps*.

Figure 3.32, taken from the classic paper by Bjerknes (1969), is a one-point correlation map for annual surface pressure data. Displayed on this map are contours of Pearson correlations between the pressure data at roughly 200 locations with those at Djakarta, Indonesia. Djakarta is thus the "one point" in this one-point correlation map. Essentially, the quantities being contoured are the values in the row or

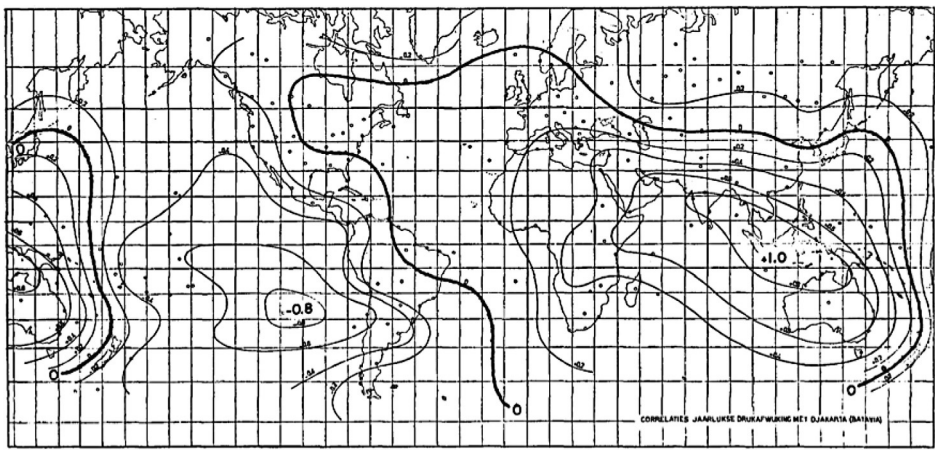


FIGURE 3.32 One-point correlation map of annual surface pressures at locations around the globe with those at Djakarta, Indonesia. The strong negative correlation of −0.8 at Easter Island reflects the atmospheric component of the El Niño-Southern Oscillation phenomenon. *From Bjerknes (1969). © American Meteorological Society. Used with permission.*

column corresponding to Djakarta in the very large correlation matrix containing all the 19,100 or so correlation values. A complete representation of that large correlation matrix in terms of one-point correlation maps would require as many maps as stations, or in this case about 200. However, not all the maps would be as interesting as Figure 3.32, although the maps for nearby stations (e.g., Darwin, Australia) would look very similar.

Clearly Djakarta is located under the +1.0 on the map, since the pressure data there are perfectly correlated with themselves. Not surprisingly, pressure correlations for locations near Djakarta are quite high, with gradual declines toward zero at locations somewhat further away. This pattern is the spatial analog of the tailing off of the (temporal) autocorrelation function indicated in Figure 3.22. The surprising feature in Figure 3.32 is the region in the eastern tropical Pacific, centered on Easter Island, for which the correlations with Djakarta pressure are strongly negative. These negative correlations imply that in years when average pressures at Djakarta (and nearby locations, such as Darwin) are high, pressures in the eastern Pacific are low, and vice versa. This correlation pattern is an expression in the surface pressure data of the ENSO phenomenon, sketched earlier in this chapter, and is an example of what has come to be known as a *teleconnection* pattern. In the ENSO warm phase, the center of tropical Pacific convection moves eastward, producing lower than average pressures near Easter Island and higher than average pressures at Djakarta. When the precipitation shifts westward during the cold phase, pressures are low at Djakarta and high at Easter Island.

Not all geographically distributed correlation data exhibit teleconnection patterns such as the one shown in Figure 3.32. However, many large-scale fields, especially pressure (or geopotential height) fields, show one or more teleconnection patterns. A device used to simultaneously display these aspects of the large underlying correlation matrix is the *teleconnectivity* map. To construct a teleconnectivity map, the row (or column) for each station or gridpoint in the correlation matrix is searched for the largest negative value. The teleconnectivity value for location i, T_i, is the absolute value of that most negative correlation,

$$T_i = \left| \min_j \left(r_{i,j} \right) \right|. \tag{3.40}$$

Here the minimization over j (the column index for $[R]$) implies that all correlations $r_{i,j}$ in the ith row of $[R]$ are searched for the smallest (most negative) value. For example, in Figure 3.32 the largest negative correlation with Djakarta pressures is with Easter Island, is -0.80. The teleconnectivity for Djakarta surface pressure would therefore be 0.80, and this value would be plotted on a teleconnectivity map at the location of Djakarta. To construct the full teleconnectivity map for surface pressure, the other 199 or so rows of the correlation matrix, each corresponding to another station, would be examined for the largest negative correlation (or, if none were negative, then the smallest positive one), and its absolute value would be plotted at the map position of that station.

Figure 3.33 shows the teleconnectivity map for northern hemisphere winter 500 mb heights. The density of the shading indicates the magnitude of the individual gridpoint teleconnectivity values. The locations of local maxima of teleconnectivity are indicated by the positions of the numbers, $\times 100$. The arrows in Figure 3.33 point from the teleconnection centers (i.e., the local maxima in T_i) to the location with which each maximum negative correlation is exhibited. The unshaded regions indicate gridpoints for which the teleconnectivity is relatively low. The one-point correlation maps for locations in these unshaded regions would tend to show gradual declines toward zero at increasing distances, analogously to the time correlations in Figure 3.22, but without declining much further to large negative values.

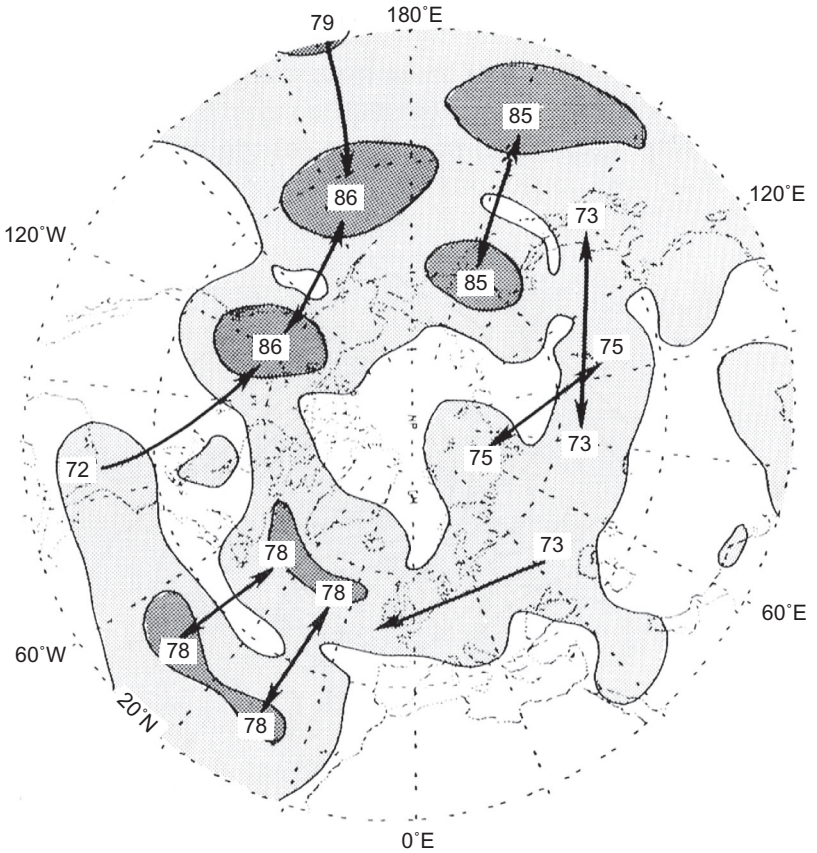


FIGURE 3.33 Teleconnectivity, or absolute value of the strongest negative correlation from each of many one-point correlation maps plotted at the base grid point, for winter 500 mb heights. *From Wallace and Blackmon (1983).*

It has become apparent that a fairly large number of these teleconnection patterns exist in the atmosphere, and the many double-headed arrows in Figure 3.33 indicate that these group naturally. Especially impressive is the four-center pattern arcing from the central Pacific to the southeastern United States, known as the *Pacific-North America*, or *PNA pattern*. Notice, however, that these patterns emerged here from a statistical, exploratory analysis of a large quantity of atmospheric data. This type of work actually had its roots in the early part of the 20th century (see Brown and Katz, 1991) and is a good example of exploratory data analysis in the atmospheric sciences turning up interesting features in very large data sets.

3.6.8. On the Use of Color

Modern computing interfaces allow easy manipulation of aspects of color and pattern that can be used to enhance quantitative communication. Color can be very effective at graphically conveying information if used intelligently. Unfortunately color can be distracting and ineffective if used poorly.

Both principles and tools for choosing effective color schemes in graphical displays are available. Cleveland (1994) distinguishes two distinct uses of color for statistical graphics. The first is to aid in the discrimination among groups or categories of graphed elements. An example would be use of colored dots as glyphs in a glyph scatterplot. For this purpose, choosing a small number of distinct hues (basic colors) is most effective. However, be aware that many readers will be unable to distinguish red and green hues (e.g., Light and Bartlein, 2004).

In contrast, visual estimation of continuously varying quantitative information is better achieved using a range of saturation (lightness) together with one or a small number of hues, whereas a more traditional "rainbow" palette is often less effective for this purpose. Further exposition and elaboration on these ideas are available in Stauffer et al. (2015) and Retchless and Brewer (2016). Online tools to aid in the selection of appropriate and effective color schemes are available at www.colorbrewer2.org, and www.hclwizard.org.

3.7. EXERCISES

3.1. Compare the median, trimean, and the mean of the precipitation data in Table A.3.
3.2. Compute the MAD, the IQR, and the standard deviation of the pressure data in Table A.3.
3.3. Draw a stem-and-leaf display for the temperature data in Table A.3.
3.4. Compute the Yule-Kendall Index and the skewness coefficient using the temperature data in Table A.3.
3.5. Draw the empirical cumulative frequency distribution for the pressure data in Table A.3. Compare it with a histogram of the same data.
3.6. Compare the boxplot and the schematic plot representing the precipitation data in Table A.3.
3.7. Use Hinkley's d_λ to find an appropriate power transformation for the precipitation data in Table A.2 using Equation 3.19a, rather than Equation 3.19b as was done in Example 3.4. Use IQR in the denominator of Equation 3.20.
3.8. Construct side-by-side schematic plots for the candidate, and final, transformed distributions derived in Exercise 3.7. Compare the result to Figure 3.13.
3.9. Express the June 1951 temperature in Table A.3 as a standardized anomaly.
3.10. Plot the autocorrelation function up to lag 3, for the Ithaca minimum temperature data in Table A.1.
3.11. Construct a scatterplot of the temperature and pressure data in Table A.3.
3.12. Construct correlation matrices for the data in Table A.3 using
   a. The Pearson correlation.
   b. The Spearman rank correlation.
3.13. Draw and compare star plots of the data in Table A.3 for each of the years 1965 through 1969.