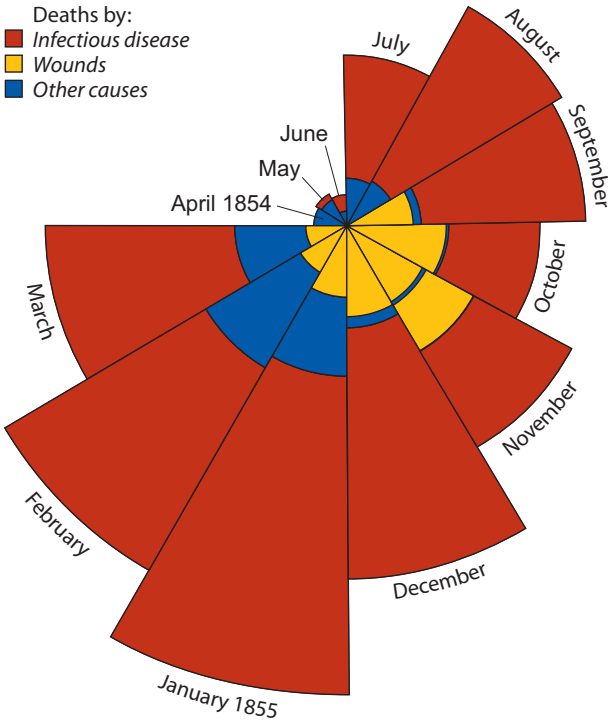


2 Displaying data

The human eye is a natural pattern detector, adept at spotting trends and exceptions in visual displays. For this reason, biologists spend hours creating and examining visual summaries of their data—graphs and, to a lesser extent, tables. Effective graphs enable visual comparisons of measurements between groups, and they expose relationships between different variables. They are also the principal means of communicating results to a wider audience.

Florence Nightingale (1858) was one of the first persons to put graphs to good use. In her famous wedge diagrams, redrawn in the figure above, she visualized the causes of death of British troops during the Crimean War. The number of cases is indicated by the area of a wedge, and the cause of death by color. The diagrams showed convincingly that disease was the main cause of soldier deaths during the wars, not wounds or other causes. With these vivid graphs, she successfully campaigned for military and public health measures that saved many lives.

Effective graphs are a prerequisite for good data analysis, revealing general patterns in the data that bare numbers cannot. Therefore, the first step in any data analysis or statistical procedure is to graph the data and look at it. Humans are a visual species, with brains evolved to process visual information. Take advantage of millions of years of evolution, and look at visual representations of your data before doing anything else. We'll follow this prescription throughout the book.

In this chapter, we explain how to produce effective graphical displays of data and how to avoid common pitfalls. We'll then review which types of graphs best show the data. The choice will depend on the type of data, numerical or categorical, and whether the goal is to show measurements of one variable or the association between two variables. There is often more than one way to show the same pattern in data, and we will compare and evaluate successful and unsuccessful approaches. We will also mention a few tips for constructing tables, which should also be laid out to show patterns in data.

2.1 Guidelines for effective graphs

Graphs are vital tools for analyzing data. They are also used to communicate patterns in data to a wider audience in the form of reports, slide shows, and web content. The two purposes, analysis and presentation, are largely coincident because the most revealing displays will be the best both for identifying patterns in the data and for communicating these patterns to others. Both purposes require displays that are clear, honest, and efficient.

To motivate principles of effective graphs, let's highlight some common ways in which researchers might get it wrong.

How to draw a bad graph

Figure 2.1-1 shows the results of an experiment in which maize plants were grown in pots under three nitrogen regimes and two soil water contents. Height of bars represents the average maize yield (dry weight per plant) at the end of the experiment under the six combinations of water and nitrogen. The data are real (Quaye et al. 2009), but we made the bad graph intentionally to highlight four common defects. Examine the graph before reading further and try to recognize some of them. Many graphics packages on the computer make it easy to produce flawed graphs like this one, which is probably why we still encounter them so often.

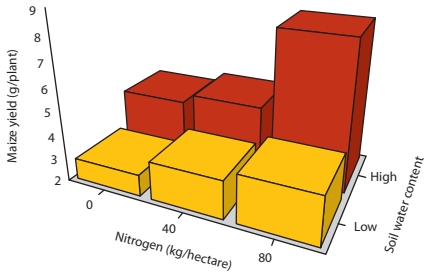


FIGURE 2.1-1
An example of a defective graph showing mean plant height of maize grown in pots under different nitrogen and water treatments.

Mistake #1: Where are the data? Each bar in Figure 2.1-1 represents average yield of four plant pots assigned to that nitrogen and water treatment. The data points—yields of all the experimental units (pots)—are nowhere to be seen. This means we are unable to see the variation in yield between pots and compare it with the magnitude of differences between treatments. It means that any unusual observations that might distort the calculation of average yield remain hidden. It would be a challenge to add the data points to this particular graph because the bars are in the way. We'll say more later about when bars are appropriate and when they are not.

Mistake #2: Patterns in the data are difficult to see. The three dimensions and angled perspective make it difficult to judge bar height by eye, which means that average plant growth is difficult to compare between treatments. In his classic book on information graphics, Tufte (1983) referred to 3-D and other visual embellishments as "chartjunk". Chartjunk adds clutter that dilutes information and interferes with the ability of the eye and brain to "see" patterns in data.

Mistake #3: Magnitudes are distorted. The vertical axis on the graph, plant yield, ranges from 2 to 9 g/plant, rather than 0 to 9, which means that bar height is out of proportion to actual magnitudes.

Mistake #4: Graphical elements are unclear. Text and other figure elements are too small to read easily.

How to draw a good graph

A few straightforward principles will help to make sure that your graphs do not end up with the kind of problems illustrated in Figure 2.1-1. We follow these four rules ourselves in the remainder of the book.

- Show the data.
- Make patterns in the data easy to see.
- Represent magnitudes honestly.
- Draw graphical elements clearly.

Show the data, first and foremost (Tufte 1983). A good graph allows you to visualize the measurements and helps the eye detect patterns in the data. Showing the

data makes it possible to evaluate the shape of the distribution of data points and to compare measurements between groups. It helps you to spot potential problems, such as extreme observations, which will be useful as you decide the next step of your data analysis.

Figure 2.1-2 gives an example of what it means to show data. The study examined the role of the neurotransmitter serotonin[1] in bringing about a transition in social behavior, from solitary to gregarious, in a desert locust (Anstey et al. 2009). This behavior change is a critical point in the production of huge locust swarms that blacken skies and ravage crops in many parts of the world. Each data point is the serotonin level of one of 30 locusts experimentally caged at high density for 0, 1, or 2 hours, with 0 representing the control. The panel on the left of Figure 2.1-2 *shows* the data (this type of graph is called a strip chart or dot plot). The panel on the right of Figure 2.1-2 *hides* the data, using bars to show only treatment averages.

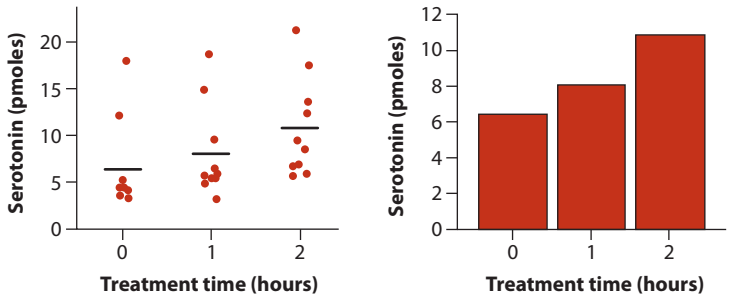


FIGURE 2.1-2 A graph that shows the data (left) and a graph that hides the data (right). Data points are serotonin levels in the central nervous system of desert locusts, *Schistocerca gregaria,* that were experimentally crowded for 0 (the control group), 1, and 2 hours. The data points in the left panel were perturbed a small amount to the left or right to minimize overlap and make each point easier to see. The horizontal bars in the left panel indicate the mean (average) serotonin level in each group. The graph on the right shows only the mean serotonin level in each treatment (indicated by bar height). Note that the vertical axis does not have the same scale in the two graphs.

In the left panel of Figure 2.1-2, we can see lots of scatter in the data in each treatment group and plenty of overlap between groups. We see that most points fall below the treatment average, and that each group has a few extreme observations. Nevertheless, we can see a clear shift in serotonin levels of locusts between treatments. All this information is missing from the right panel of Figure 2.1-2, which uses more ink yet shows only the averages of each treatment group.

Make patterns easy to see. Try displaying your data in different ways, possibly with different types of graphs, to find the best way to communicate the findings. Is the main pattern in the data recognizable right away? If not, try again with a different method. Stay away from 3-D effects and elaborate chartjunk that obscures the

1. Serotonin is a neurotransmitter in most animals, including humans. Some antidepressant drugs improve feelings of well-being by manipulating serotonin levels.

patterns in the data. In the rest of the chapter we'll compare alternative ways of graphing the same data sets and discuss their effectiveness.

Avoid putting too much information into one graph. Remember the purpose of a graph: to communicate essential patterns to eyes and brains. The purpose is not to cram as much data as possible into each graph. Think about getting the main point across with one or two key graphs in the main body of your presentation. Put the remainder into an appendix or online supplement if it is important to show them to a subset of your audience.

Represent magnitudes honestly. This sounds easy, but misleading graphics are common in the scientific literature. One of the most important decisions concerns the smallest value on the vertical axis of a graph (the "baseline"). A bar graph must always have a baseline at zero, because the eye instinctively reads bar height and area as proportional to magnitude. The upper bar graph in Figure 2.1-3 shows an example, depicting government spending on education each year since 1998 in British Columbia. The area of each bar is not proportional to the magnitude of the value displayed. As a result, the graph exaggerates the differences. The figure falsely suggests that spending increased twenty-fold over time, but the real increase is less than 20%. It is more honest to plot the bars with a baseline of zero, as in the lower graph in

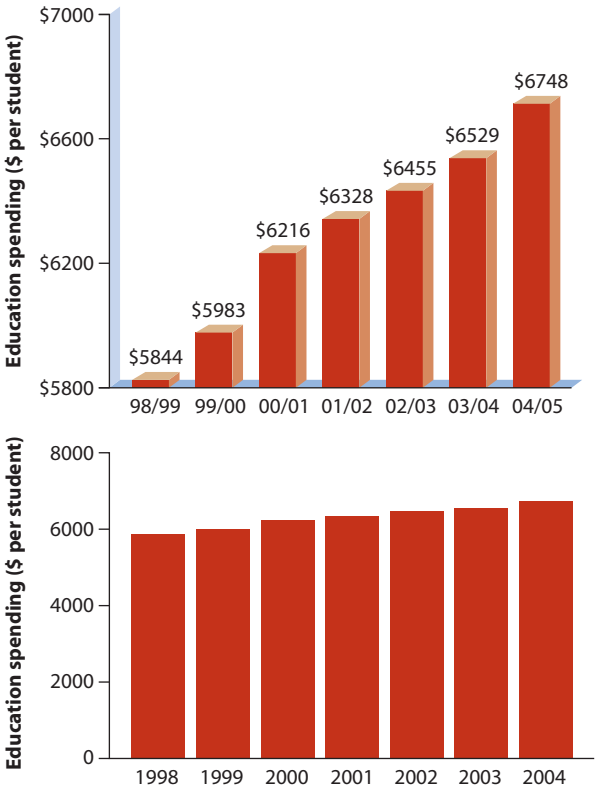


FIGURE 2.1-3
Upper graph. A bar graph, taken from a British Columbia government brochure, indicating education spending per student in different years. *Lower graph:* A revised presentation of the same data, in which the magnitude of the spending is proportional to the height and area of bars. This revision also removed the 3-D effects and the numbers above bars to make the pattern easier to see. The upper graph is modified from British Columbia Ministry of Education (2004).

Figure 2.1-3 (the revised graph also removed the 3-D effects and the numbers above bars to make the pattern easier to see).

Other types of graphs, such as strip charts, don't always need a zero baseline if the main goal is to show differences between treatments rather than proportional magnitudes.

Draw graphical elements clearly. Clearly label the axes and choose unadorned, simple typefaces and colors. Text should be legible even after the graph is shrunk to fit the final document. Always provide the units of measurement in the axis label. Use clearly distinguishable graphical symbols if you plot with more than one kind. Don't always accept the default output of statistical or spreadsheet programs.

Up to a tenth of your male audience is red-green color-blind, so choose colors that differ in intensity and apply redundant coding to distinguish groups (for example, use distinctive shapes or patterns as well as different colors).[2]

A good graph is like a good paragraph. It conveys information clearly, concisely, and without distortion. A good graph requires careful editing. Just as in writing, the first draft is rarely as good as the final product.

2.2 Showing data for one variable

To examine data for single variable, we show its **frequency distribution**. Recall from Chapter 1 that the frequency of occurrence of a specific measurement in a sample is the number of observations having that particular measurement. The frequency distribution of a variable is the number of occurrences of all values in the data.

Relative frequency is the proportion of observations having a given measurement, calculated as the frequency divided by the total number of observations. The **relative frequency distribution** is the proportion of occurrences of each value in the data set.

The *relative frequency distribution* describes the fraction of occurrences of each value of a variable.

Showing categorical data: frequency table and bar graph

Let's start with displays for a categorical variable. A **frequency table** is a text display of the number of occurrences of each category in the data set. A **bar graph** uses the height of rectangular bars to visualize the frequency (or relative frequency) of occurrence of each category.

2. If you are in doubt, load your graphic file into a colorblindness simulator such as Vischeck (vischeck.com).

A *bar graph* uses the height of rectangular bars to display the frequency distribution (or relative frequency distribution) of a categorical variable.

Example 2.2A illustrates both kinds of displays.

EXAMPLE 2.2A **Crouching tiger**

Conflict between humans and tigers threatens tiger populations, kills people, and reduces public support for conservation. Gurung et al. (2008) investigated causes of human deaths by tigers near the protected area of Chitwan National Park, Nepal. Eighty-eight people were killed by 36 individual tigers between 1979 and 2006, mainly within 1 km of the park edge. Table 2.2-1 lists the main activities of people at the time they were killed. Such information may be helpful to identify activities that increase vulnerability to attack.

Table 2.1-1 is a frequency table showing the number of deaths associated with each activity. Here, alternative values of the variable "activity" are listed in a single column, and frequencies of occurrence are listed next to them in a second column. The categories have no intrinsic order, but comparing the frequencies of each activity is made easier by *arranging the categories in order of their importance,* from the most frequent at the top to the least frequent at the bottom.

Table 2.2-1 Frequency table showing the activities of 88 people at the time they were attacked and killed by tigers near Chitwan National Park, Nepal, from 1979 to 2006.

Activity	Frequency (number of people)
Collecting grass or fodder for livestock	44
Collecting non-timber forest products	11
Fishing	8
Herding livestock	7
Disturbing tiger at its kill	5
Collecting fuel wood or timber	5
Sleeping in a house	5
Walking in forest	3
Using an outside toilet	2
Total	88

The table shows that more people were killed while collecting grass and fodder for their livestock than when doing any other activity. The number of deaths under this activity was four times that of the next category of activity (collecting non-timber forest products) and is related to the amount of time people spend carrying out these activities.

The differences in frequency stand out even more vividly in the bar graph shown in Figure 2.2-1. In a bar graph, frequency is depicted by the height of rectangular bars. Unlike a frequency table, a bar graph does not usually present the actual numbers. Instead, the graph gives a clear picture of how steeply the numbers drop between categories. Some activities are much more common than others, and we don't need the actual numbers to see this.

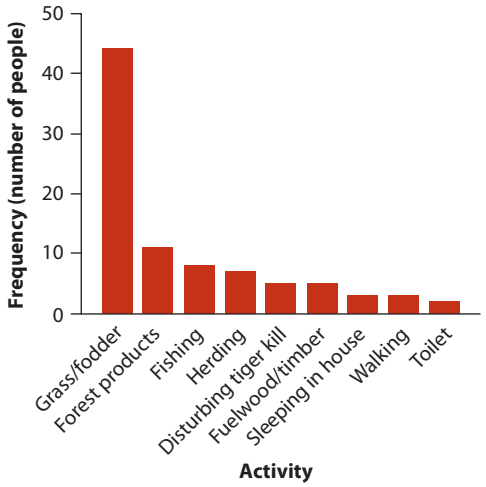


FIGURE 2.2-1
Bar graph showing the activities of people at the time they were attacked and killed by tigers near Chitwan National Park, Nepal, between 1979 and 2006. Total number of deaths: $n = 88$. The frequencies are taken from Table 2.2-1, which also gives more detailed labels of activities.

Making a good bar graph

The top edge of each bar conveys all the information about frequency, but the eye also compares the areas of the bars, which must therefore be of equal width. It is crucial that the baseline of the *y*-axis is at zero—otherwise, the area and height of bars are out of proportion with actual magnitudes and so are misleading.

When the categorical variable is nominal, as in Figure 2.2-1 and Table 2.2-1, the best way to arrange categories is by frequency of occurrence. The most frequent category goes first, the next most frequent category goes second, and so on. This aids in the visual presentation of the information. For an ordinal categorical variable, such as snakebite severity score, the values should be in the natural order (e.g., minimally severe, moderately severe, and very severe). Bars should stand apart, not be fused together. It is a good habit to provide the total number of observations (n) in the figure legend.

A bar graph is usually better than a pie chart

The pie chart is another type of graph often used to display frequencies of a categorical variable. This method uses colored wedges around the circumference of a circle to represent frequency or relative frequency. Figure 2.2-2 shows the tiger data again, this time in a pie chart. This graphical method is reminiscent of Florence Nightingale's wedge diagram shown at the beginning of this chapter.

FIGURE 2.2-2
Pie chart of the activities of people at the time they were attacked and killed by tigers near Chitwan National Park, Nepal. The frequencies are taken from Table 2.2-1.Total number of deaths: $n = 88$.

The pie chart has received a lot of criticism from experts in information graphics. One reason is that while it is straightforward to visualize the frequency of deaths in the first and most frequent category (Collecting grass/fodder), it is more difficult to compare frequencies in the remaining categories by eye. This problem worsens as the number of categories increases. Another reason is that it is very difficult to compare frequencies between two or more pie charts side by side, especially when there are many categories. To compensate, pie charts are often drawn with the frequencies added as text around the circle perimeter. The result is not better than a table. The shape of a frequency distribution is more readily perceived in a bar graph than a pie chart, and it is easier to compare frequencies between two or more bar graphs than between pie charts. Use the bar graph instead of the pie chart for showing frequencies in categorical data.

Showing numerical data: frequency table and histogram

A frequency distribution for a numerical variable can be displayed either in a frequency table or in a **histogram**. A histogram uses area of rectangular bars to display frequency. The data values are split into consecutive intervals, or "bins," usually of equal width, and the frequency of observations falling into each bin is displayed.

A *histogram* uses the area of rectangular bars to display the frequency distribution (or relative frequency distribution) of a numerical variable.

We discuss how histograms are made in greater detail using the data in Example 2.2B.

EXAMPLE 2.2B Abundance of desert bird species

How many species are common in nature and how many are rare? One way to address this question is to construct a frequency distribution of species abundance. The data in Table 2.2-2 are from a survey of the breeding birds of Organ Pipe Cactus National Monument in southern Arizona, USA. The measurements were extracted from the North American Breeding Bird Survey, a continent-wide data set of estimated bird numbers (Sauer et al. 2003).

Table 2.2-2 Data on the abundance of each species of bird encountered during four surveys in Organ Pipe Cactus National Monument.

Species	Abundance	Species	Abundance
Greater roadrunner	1	Turkey vulture	23
Black-chinned hummingbird	1	Violet-green swallow	23
Western kingbird	1	Lesser nighthawk	25
Great-tailed grackle	1	Scott's oriole	28
Bronzed cowbird	1	Purple martin	33
Great horned owl	2	Black-throated sparrow	33
Costa's hummingbird	2	Brown-headed cowbird	59
Canyon wren	2	Black vulture	64
Canyon towhee	2	Lucy's warbler	67
Harris's hawk	3	Gilded flicker	77
Loggerhead shrike	3	Brown-crested flycatcher	128
Hooded oriole	4	Mourning dove	135
Northern mockingbird	5	Gambel's quail	148
American kestrel	7	Black-tailed gnatcatcher	152
Rock dove	7	Ash-throated flycatcher	173
Bell's vireo	10	Curve-billed thrasher	173
Common raven	12	Cactus wren	230
Northern cardinal	13	Verdin	282
House sparrow	14	House finch	297
Ladder-backed woodpecker	15	Gila woodpecker	300
Red-tailed hawk	16	White-winged dove	625
Phainopepla	18		

We treated each bird species in the survey as the unit of interest and the abundance of a species in the survey as its measurement. The range of abundance values was divided into 13 intervals of equal width (0–50, 50–100, and so on), and the number of species falling into each abundance interval was counted and presented in a frequency table to help see patterns (Table 2.2-3).

Table 2.2-3 Frequency distribution of bird species abundance at Organ Pipe Cactus National Monument.

Abundance	Frequency (Number of species)
0–50	28
50–100	4
100–150	3
150–200	3
200–250	1
250–300	2
300–350	1
350–400	0
400–450	0
450–500	0
500–550	0
550–600	0
600–650	1
Total	43

Source: Data are from Table 2.2-2.

Although the table shows the numbers, the shape of the frequency distribution is more obvious in a histogram of these same data (Figure 2.2-3). Here, frequency (number of species) in each abundance interval is perceived as bar area.

The frequency table and histogram of the bird abundance data reveal that the majority of bird species have low abundance. Frequency falls steeply with increasing abundance.[3] The white-winged dove (pictured in Example 2.2B) is exceptionally

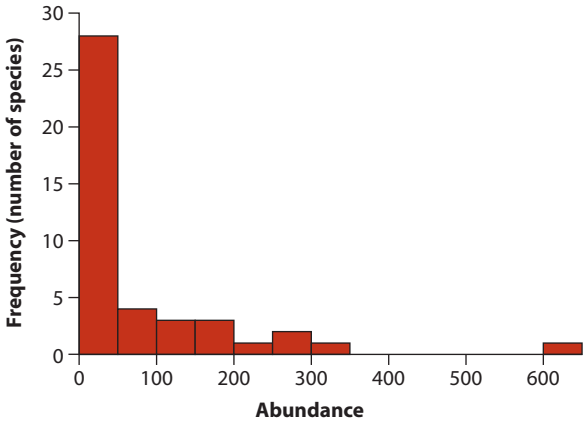


FIGURE 2.2-3
Histogram illustrating the frequency distribution of bird species abundance at Organ Pipe Cactus National Monument. Total number of bird species: $n = 43$.

3. This pattern is a remarkably general one in nature, found in many types of organisms. Typically, only a few species are common, whereas most species are rare.

abundant at Organ Pipe Cactus National Monument, accounting for a large fraction of all individual birds encountered in the survey.

Describing the shape of a histogram

The histogram reveals the shape of a frequency distribution. Some of the most common shapes are displayed in Figure 2.2-4. Any interval of the frequency distribution that is noticeably more frequent than surrounding intervals is called a peak. The **mode** is the interval corresponding to the highest peak. For example, a bell-shaped frequency distribution has a single peak (the mode) in the center of the range of observations. A frequency distribution having two distinct peaks is said to be **bimodal**.

The *mode* is the interval corresponding to the highest peak in the frequency distribution.

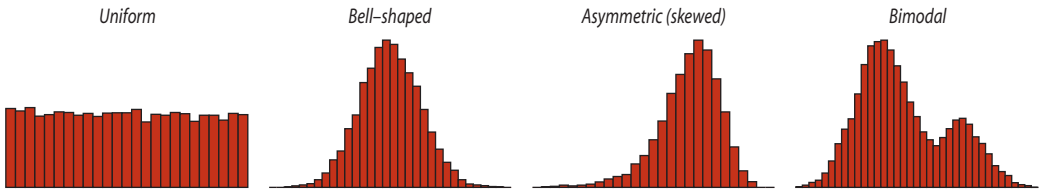


FIGURE 2.2-4 Some possible shapes of frequency distributions.

A frequency distribution is **symmetric** if the pattern of frequencies on the left half of the histogram is the mirror image of the pattern on the right half. The uniform distribution and the bell-shaped distribution in Figure 2.2-4 are symmetric. If a frequency distribution is not symmetric, we say that it is **skewed**. The distribution in Figure 2.2-4 labeled "Asymmetric" has left or negative skew: it has a long tail extending to the left. The distribution in Figure 2.2-4 labeled "Bimodal" is also asymmetric but is positively skewed: its long tail is to the right.[4] The abundance data for desert bird species also have positive skew (Figure 2.2-3), which means they have a long tail extending to the right.

Skew refers to asymmetry in the shape of a frequency distribution for a numerical variable.

Extreme data points lying well away from the rest of the data are called **outliers**. The histogram of desert bird abundance (Figure 2.2-3) includes one extreme observa-

4. The nomenclature of skew seems backward to many people. Focus on the sharp tail of the distribution extending to the left in the third distribution in Figure 2.2-4. We say it is skewed left (or has a negative skew) because it seems to have a "skewer" sticking out to the left toward negative numbers, like the skewer through a shish kebab.

tion (the white-winged dove) that falls well outside the range of abundance of other bird species. The white-winged dove, therefore, is an outlier. Outliers are common in biological data. They can result from mistakes in recording the data or, as in the case of the white-winged dove, they may represent real features of nature. Outliers should always be investigated. They should be removed from the data only if they are found to be errors.

An *outlier* is an observation well outside the range of values of other observations in the data set.

How to draw a good histogram

When drawing a histogram, the choice of interval width must be made carefully because it can affect the conclusions. For example, Figure 2.2-5 shows three different histograms that depict the body mass of 228 female sockeye salmon (*Oncorhynchus nerka*) from Pick Creek, Alaska, in 1996 (Hendry et al. 1999). The leftmost histogram of Figure 2.2-5 was drawn using a narrow interval width. The result is a somewhat bumpy frequency distribution that suggests the existence of two or even more peaks. The rightmost histogram uses a wide interval. The result is a smoother frequency distribution that masks the second of the two dominant peaks. The middle histogram uses an intermediate interval that shows two distinct body-size groups. The fluctuations from interval to interval within size groups are less noticeable.

To choose the ideal interval width we must decide whether the two distinct body-size groups are likely to be "real," in which case the histogram should show both, or whether a bimodal shape is an artifact produced by too few observations.[5]

When you draw a histogram, each bar must rise from a baseline of zero, so that the area of each bar is proportional to frequency. Unlike bar graphs, adjacent histogram bars are contiguous, with no spaces between them. This reinforces the perception of

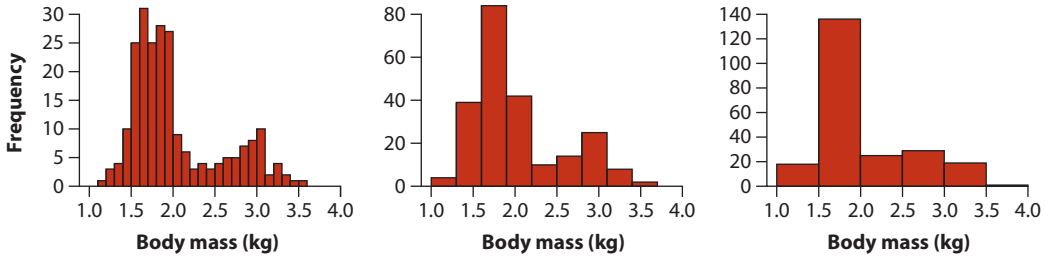


FIGURE 2.2-5 Body mass of 228 female sockeye salmon sampled from Pick Creek in Alaska (Hendry et al. 1999). The same data are shown in each case, but the interval widths are different: 0.1 kg (left), 0.3 kg (middle), and 0.5 kg (right).

5. The two distinct body-size classes in this salmon population correspond to two age groups.

a numerical scale, with bars grading from one into the next. In this book we follow convention by placing an observation whose value is exactly at the boundary of two successive intervals into the higher interval. For example, the Gila woodpecker, with a total of 300 individuals observed (Table 2.2-2), is recorded in the interval 300–350, not in the interval 250–300.

There are no strict rules about the number of intervals that should be used in frequency tables and histograms. Some computer programs use Sturges's rule of thumb, in which the number of intervals is $1 + \ln(n)/\ln(2)$, where n is the number of observations and ln is the natural logarithm. The resulting number is then rounded up to the higher integer (Venables and Ripley 2002). Many regard this rule as overly conservative, and in this book we tend to use a few more intervals than Sturges. The number of intervals should be chosen to best show patterns and exceptions in the data, and this requires good judgment rather than strict rules. Computers allow you to try several alternatives to help you determine the best option.

When breaking the data into intervals for the histogram, use readable numbers for breakpoints—for example, break at 0.5 rather than 0.483. Finally, it is a good idea to provide the total number of individuals in the accompanying legend.

Other graphs for numerical data

The histogram is recommended for showing the frequency distribution of a single numerical variable. The *box plot* and the *strip chart* are alternatives, but most often these are used to show differences when there are data from two or more groups. We describe these graphs in the next section. Another type of graph, the *cumulative frequency distribution,* is explained in Chapter 3.

2.3 Showing association between two variables

Here we illustrate how to show data for two variables simultaneously, rather than one at a time. The goal is to create an image that visualizes association or correlation between two variables and differences between groups. The most suitable type of graph depends on whether both variables are categorical, both are numerical, or one is of each data type.

Showing association between categorical variables

If two categorical variables are associated, the relative frequencies for one variable will differ among categories of the other variable. To reveal such association, show the frequencies using a contingency table, a mosaic plot, or a grouped bar graph. Here's an example.

EXAMPLE 2.3A Reproductive effort and avian malaria

Is reproduction hazardous to health? If not, then it is difficult to explain why adults in many organisms seem to hold back on the number of offspring they raise. Oppliger et al. (1996) investigated the impact of reproductive effort on the susceptibility to malaria[6] in wild great tits (*Parus major*) breeding in nest boxes. They divided 65 nesting females into two treatment groups. In one group of 30 females, each bird had two eggs stolen from her nest, causing the female to lay an additional egg. The extra effort required might increase stress on these females. The remaining 35 females were left alone, establishing the control group. A blood sample was taken from each female 14 days after her eggs hatched to test for infection by avian malaria.

The association between experimental treatment and the incidence of malaria is displayed in Table 2.3-1. This table is known as a **contingency table**, a frequency table for two (or more) categorical variables. It is called a contingency table because it shows how the frequencies of the categories in a response variable (the incidence of malaria, in this case) are contingent upon the value of an explanatory variable (the experimental treatment group).

Table 2.3-1 Contingency table showing the incidence of malaria in female great tits in relation to experimental treatment.

	Experimental treatment group		
	Control group	Egg-removal group	Row total
Malaria	7	15	22
No malaria	28	15	43
Column total	35	30	65

Each experimental unit (bird) is counted exactly once in the four "cells" of Table 2.3-1, and so the total count (65) is the number of birds in the study. A cell is one combination of categories of the row and column variables in the table. The explanatory variable (experimental treatment) is displayed in the columns, whereas the response variable, the variable being predicted (incidence of malaria), is displayed in the rows. The frequency of subjects in each treatment group is given in the column totals, and the frequency of subjects with and without malaria is given in the row totals.

According to Table 2.3-1, malaria was detected in 15 of the 30 birds subjected to egg removal, but in only seven of the 35 control birds. This difference between treat-

6. Malaria is a common cause of death in humans, but avian forms of the disease are even more prevalent in many bird species. For example, many native bird species in Hawaii are threatened with extinction after the inadvertent introduction of mosquitoes and avian malaria by humans in the 19th century.

ments suggests that the stress of egg removal, or the effort involved in producing one extra egg, increases female susceptibility to avian malaria.

A *contingency table* gives the frequency of occurrence of all combinations of two (or more) categorical variables.

Table 2.3-1 is an example of a 2 × 2 ("two-by-two") contingency table, because it displays the frequency of occurrence of all combinations of two variables, each having exactly two categories. Larger contingency tables are possible if the variables have more than two categories.

Two types of graph work best for displaying the relationship between a pair of categorical variables. The **grouped bar graph** uses heights of rectangles to graph the frequency of occurrence of all combinations of two (or more) categorical variables. Figure 2.3-1 shows the grouped bar graph for the avian malaria experiments. Grouped bar graphs are like bar graphs for single variables, except that different categories of the response variable (e.g., malaria and no malaria) are indicated by different colors or shades. Bars are grouped by the categories of the explanatory variable treatment (control and egg removal), so make sure that the spaces between bars from different groups are wider than the spaces between bars separating categories of the response variable. We can see from the grouped bar graph in Figure 2.3-1 that incidence of malaria is associated with treatment, because the relative heights of the bars for malaria and no malaria differ between treatments. Most birds in the control group had no malaria (the yellow bar is much taller than the red bar), whereas in the experimental group, the frequency of subjects with and without malaria was equal.

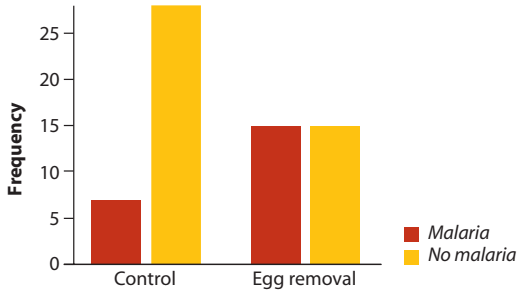


FIGURE 2.3-1
Grouped bar graph for reproductive effort and avian malaria in great tits. The data are from Table 2.3-1, where $n = 65$ birds.

A *grouped bar graph* uses the height of rectangular bars to display the frequency distributions (or relative frequency distributions) of two or more categorical variables.

A **mosaic plot** is similar to a grouped bar plot except that bars within treatment groups are stacked on top of one another (Figure 2.3-2). Within a stack, bar area and height indicate the relative frequencies (i.e., the proportion) of the responses. This makes it easy to see the association between treatment and response variables:

if an association is present in the data, then the vertical position at which the colors meet will differ between stacks. If no association is present, then the meeting point between the colors will be at the same vertical position between stacks. In Figure 2.3-2, for example, few individuals in the control group were infected with malaria, so the red bar (malaria) meets the yellow bar (no malaria) at a higher vertical position than in the egg removal stack, where the incidence of malaria was greater.

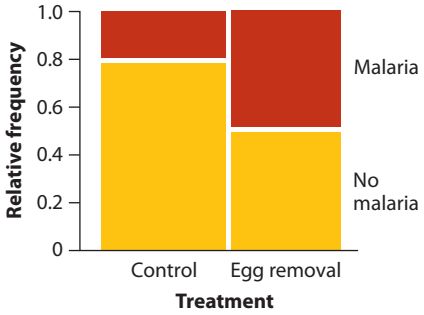


FIGURE 2.3-2
Mosaic plot for reproductive effort and avian malaria in great tits. Red indicates birds with malaria, whereas yellow indicates birds free of malaria. The data are from Table 2.3-1, where $n = 65$ birds.

Another feature of the mosaic plot is that the width of each vertical stack is proportional to the number of observations in that group. In Figure 2.3-2, the wider stack for the control group reflects the greater total number of individuals in this treatment (35) compared with the number in the egg-removal treatment (30). As a result, the total area of each box is proportional to the relative frequency of that combination of variables in the whole data set.

A mosaic plot provides only relative frequencies, not the absolute frequency of occurrence in each combination of variables. This might be considered a drawback, but keep in mind that the most important goal of graphs is to depict the *pattern* in the data rather than exact figures. Here, the pattern is the association between treatment and response variables: the difference in the relative frequencies of diseased birds in the two treatments.

The *mosaic plot* uses the area of rectangles to display the relative frequency of occurrence of all combinations of two categorical variables.

Of the three methods for presenting the same data—the contingency table, the mosaic plot, and the grouped bar graph—which is best? The answer depends on the circumstances, and it is a good idea to try all three to evaluate their effectiveness in any data set. It is usually easier to see differences in relative frequency between groups when the data are visualized in a grouped bar plot or mosaic plot than in a contingency table. On the other hand, a contingency table might work best if one of the response categories is vastly more frequent than the other, making it difficult to see the bars corresponding to rare categories in a graph, or if the explanatory and response variables have many categories, thus increasing the complexity of the graph.

We find that association, or lack of association, is easier to see in a mosaic plot than in a grouped bar graph, but this will not always be the case. Deciding which type of display is most effective for a given circumstance is best done by trying several methods and choosing among them on the basis of information, clarity, and simplicity.

Showing association between numerical variables: scatter plot

Use a scatter plot to show the association between two numerical variables. Position along the horizontal axis (the *x*-axis) indicates the measurement of the explanatory variable. The position along the vertical axis (the *y*-axis) indicates the measurement of the response variable. The pattern in the resulting cloud of points indicates whether an association between the two variables is positive (in which case the points tend to run from the lower left to the upper right of the graph), negative (the points run from the upper left to the lower right), or absent (no discernible pattern). Example 2.3B shows an example.

EXAMPLE 2.3B Sins of the father

The bright colors and elaborate courtship displays of the males of many species posed a problem for Charles Darwin: how can such elaborate traits evolve? His answer was that they evolved because females are attracted to them when choosing a mate. But why would females choose those kinds of males? One possible answer: females that choose fancy males have attractive sons as well. A recent laboratory study examined how attractive traits in guppies are inherited from father to son (Brooks 2000). The attractiveness of sons (a score representing the rate of visits by females to corralled males, relative to a standard) was compared with their fathers' ornamentation (a composite index of several aspects of male color and brightness). The father's ornamentation is the explanatory variable in the resulting scatter plot of these data (Figure 2.3-3).

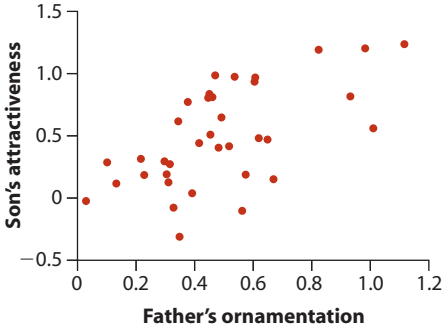


FIGURE 2.3-3
Scatter plot showing the relationship between the ornamentation of male guppies and the average attractiveness of their sons. Total number of families: $n = 36$.

Each dot in the scatter plot is a father-son pair. The father's ornamentation is the explanatory variable and the son's attractiveness is the response variable. The plot shows a *positive* association between these variables (note how the points tend to run

from the lower left to the upper right of the graph). Thus, the sexiest sons come from the most gloriously ornamented fathers, whereas unadorned fathers produce less attractive sons on average.

A *scatter plot* is a graphical display of two numerical variables in which each observation is represented as a point on a graph with two axes.

Showing association between a numerical and a categorical variable

There are several good methods to show an association between a numerical variable and a categorical variable. Three that we recommend are the *strip chart* (which we first saw in Figure 2.1-2), the *box plot,* and the *multiple histograms* method. Here we compare these methods with an example. We recommend against the common practice of using a bar graph because the bars make it difficult to show the data (bar graphs are ideal for frequency data). Showing an association between a numerical and a categorical variable is the same as showing a difference in the numerical variable between groups.

EXAMPLE 2.3C Blood responses to high elevation

The amount of oxygen obtained in each breath at high altitude can be as low as one-third that obtained at sea level. Studies have begun to examine whether indigenous people living at high elevations have physiological attributes that compensate for the reduced availability of oxygen. A reasonable expectation is that they should have more hemoglobin, the molecule that binds and transports oxygen in the blood. To test this, researchers sampled blood from males in three high-altitude human populations: the high Andes, high-elevation Ethiopia, and Tibet, along with a sea-level population from the USA (Beall et al. 2002). Results are shown in Figures 2.3-4 and 2.3-5.

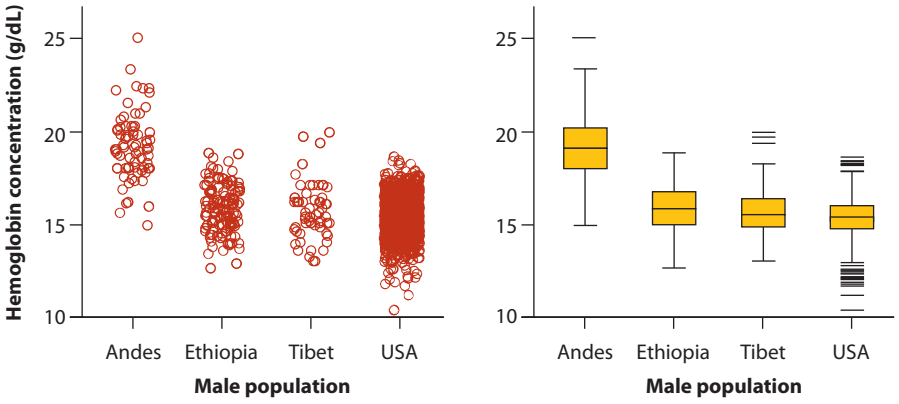


FIGURE 2.3-4 Strip chart (left) and box plot (right) showing hemoglobin concentration in males living at high altitude in three different parts of the world: the Andes (71), Ethiopia (128), and Tibet (59). A fourth population of 1704 males living at sea level (USA) is included as a control.

The left panel of Figure 2.3-4 shows the hemoglobin data with a **strip chart** (sometimes also called a dot plot). In a strip chart, each observation is represented as a dot on a graph showing its numerical measurement on one axis (here, the vertical or *y*-axis) and the category (group) to which it belongs on the other (here, horizontal or *x*-axis). A strip chart is like a scatter plot except the explanatory variable is categorical rather than numerical. It is usually necessary to spread, or "jitter," the points along the horizontal axis to reduce overlap of points, so that they can be more easily seen. The strip chart method worked well in Figure 2.1-2, where there were few data points in each group. However, several of the male populations in Example 2.3C have so many observations that the points overlap too much in the strip chart, making it difficult to see the individual dots and their distribution (left panel of Figure 2.3-4).

The *strip chart* is a graphical display of a numerical variable and a categorical variable in which each observation is represented as a dot.

An alternative method to "show the data" is the **box plot**, which uses lines and rectangles to display a compact summary of the frequency distribution of the data (right panel of Figure 2.3-4). A box plot doesn't show most of the data points, but instead uses lines and boxes to indicate where the bulk of the observations lie. The scale of the vertical axis is the same in both panels of Figure 2.3-4 so that you can see the correspondence between boxes and data points.

The line inside each box (right panel of Figure 2.3-4) is the *median,* the middle measurement of the group of observations. Half the observations lie above the median and half lie below. The lower and upper edge of each box are first and third quartiles. One-fourth of the observations lie below the *first quartile* (three-fourths lie above). Conversely, three-fourths of the observations lie below the *third quartile* (one-quarter lie above). Two lines, called whiskers, extend outward from a box at each end. The whiskers stop at the smallest and largest "non-extreme" values in the data. Extreme values are plotted as isolated dots past the ends of the whiskers. We explain these quantities in more detail, and how to calculate them, in Chapter 3.

A *box plot* is a graph that uses lines and a rectangular box to display the median, quartiles, range, and extreme measurements of the data.

The box plot in Figure 2.3-4 shows key features of the four frequency distributions using just a few graphical elements. We can clearly see in this graph how only men from the high Andes had elevated hemoglobin concentrations, whereas men from high-elevation Ethiopia and Tibet were not noticeably different in hemoglobin concentration from the sea-level group.[7] We can see that the shapes of distributions are

7. Mysteriously, oxygen levels in the blood of highland Ethiopian men are just as high as those in men living at sea level (data not shown), despite their similar concentrations of hemoglobin. The physiological mechanism behind this feat is not yet known.

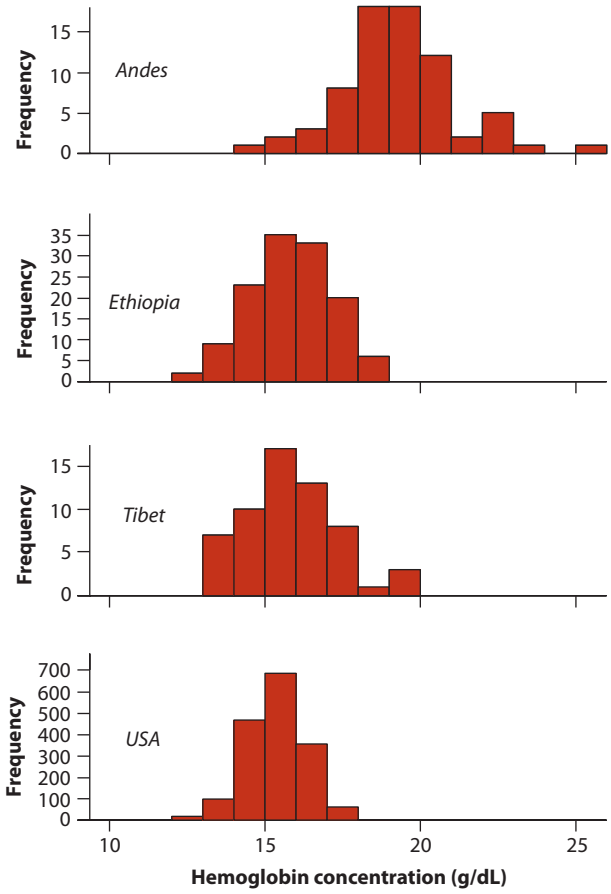


FIGURE 2.3-5
Multiple histograms showing the hemoglobin concentration in males of the four populations. The number of measurements in each group is given in Figure 2.3-4.

relatively similar in the four groups of males—the span of the boxes is similar in the four groups, although greatest in the Andean males and least in the USA males. The lengths of the whiskers are also fairly similar. USA males have the most extreme observations, but the shape of the distribution, as indicated by the box and whiskers, is similar to that in the other groups.

The third method uses multiple histograms, one for each category, to show the data, as shown in Figure 2.3-5. It is important that the histograms be stacked above one another as shown so that the position and spread of the data are most easily compared. Side-by-side histograms lose most of the advantages of the multiple histogram method for visualizing association, because differences in the position of bars between groups are difficult to see. Use the same scale along the horizontal axis to allow comparison.

Of the three methods for showing association between a numeric and a categorical variable (difference between groups), which is the best? The strip chart shows all the data points, which is ideal when there are only a few observations in each category. The box plot picks out a few of the most important features of the frequency distri-

bution and is more suitable when the number of observations is large. The multiple histogram plot shows more features of the frequency distribution but takes up more space than the other two options. It works best when there are only a few categories. As usual, the best strategy is to try all three methods on your data and judge for that situation which method shows the association most clearly.

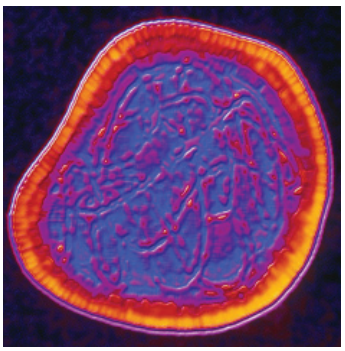
2.4 Showing trends in time and space

Often a variable of interest represents a summary measurement taken at consecutive points in time or space. In this case, *line graphs* and *maps* are excellent visual tools.

Line graph

A **line graph** is a powerful tool for displaying trends in time or other ordered series. Typically, one y-measurement is displayed for each point in time or space, which is displayed along the x-axis. Adjacent points along the x-axis are connected by a line segment. Example 2.4A illustrates the line graph.

EXAMPLE 2.4A **Bad science can be deadly**



Since the introduction of a measles vaccine, the number of cases in the U.K. dropped dramatically. A disease that once killed hundreds of people per year in the U.K. became a negligible risk as most of the population was immunized. However, recent declines in the fraction of people vaccinated, in part from unfounded fears concerning the safety of the vaccine,[8] has caused a resurgence in the number of cases of measles (Jansen et al. 2003). The number of cases quarterly between 1995 and 2011 is shown in a line plot in Figure 2.4-1 (data from Health Protection Agency 2012).

8. Vaccination rates dropped below 85% in the years after a 1998 publication that appeared to link the measles, mumps, and rubella vaccine to an increased risk of autism (e.g., see Gilmour et al. 2011). This controversial conclusion was refuted in subsequent reviews (e.g., Canadian Paediatric Society 2007). The original article has since been retracted by most of its authors and by the medical journal that published it. In 2010, the General Medical Council of the U.K. found the senior author of the research guilty of ethical breaches, dishonesty, and conflict of interest and banned him from practicing medicine.

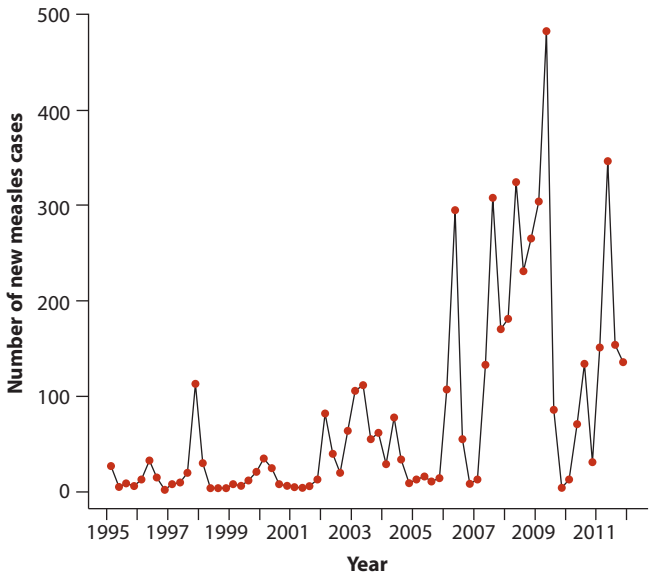


FIGURE 2.4-1 Confirmed cases of measles in England and Wales from 1995 to 2011. The four numbers in each year refer to new cases in each quarter.

The trends in the number of measles cases over time are made more visible by the lines connecting the points in Figure 2.4-1. The steepness of the line segments reflects the speed of change in the number of cases from one quarter-year to the next. Notice, for example, how steeply the number of cases rises when an outbreak begins, and then how cases decline just as quickly afterward, as immunity spreads. When the baseline for the vertical axis is zero, as in the present example, the area under the curve between two time points is proportional to the total number of new cases in that period.

A *line graph* uses dots connected by line segments to display trends in a measurement over time or other ordered series.

Maps

A **map** is the spatial equivalent of the line graph, using a color gradient to display a numerical response variable at multiple locations on a surface. The explanatory variable is location in space. One measurement is displayed for each point or interval of the surface, as shown in Example 2.4B.

EXAMPLE 2.4B Biodiversity hotspots

South America is renowned for its extraordinary numbers of species. We tend to think of the vast expanse of lowland Amazon rainforest as the seat of this abundance. The data shown in Figure 2.4-2 are the numbers of plant species recorded at many points on a fine grid covering the northern part of South America. Points are colored such that "hotter" colors represent more plant species at each point. The image shows that peak diversities actually occur at the northwest coast, the nearby inland where the Andes mountains meet the lowland rainforest, and along the southeast coast of Brazil.

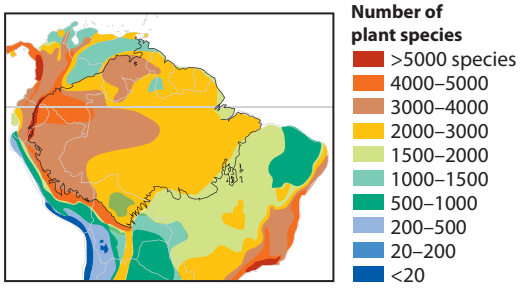


FIGURE 2.4-2
Map displaying numbers of plant species in northern South America. Colors reflect the numbers of species estimated at many points in a fine grid, with each point consisting of an area 100 km x 100 km. The color scale is on the right, with hotter colors reflecting more species. The horizontal gray line is the equator. Modified from Bass et al. (2010).

The map in Figure 2.4-2 contains an enormous amount of data, yet the pattern is easy to see. The regions of peak diversity, and those of relatively low diversity, are clearly evident.

Maps can be used for measurements at points on any two-dimensional surface, including a spatial grid (such as in the plant species richness example) or at political or geological boundaries on a map. They can also be used to indicate measurements at locations on the surface of two- or three-dimensional objects, such as the brain or the body. For example, a visual representation of an MRI scan is also a map.

2.5 How to make good tables

Tables have two functions: to communicate patterns and to record quantitative data summaries for further use. When the main function of a table is to display the patterns in the data—a "display table"—numerical detail is less important than the effective communication of results. This is the kind of table that would appear in the main body of a report or publication. Compact frequency tables are examples of display tables; for example, look at Table 2.2-3, which shows the number of bird species in a sequence of abundance categories. In this section we summarize strategies for making good display tables.

The purpose of the second kind of table is to store raw data and numerical summaries for reference purposes. Such "data tables" are often large and are not ideal for

recognizing patterns in data. They are inappropriate for communicating general findings to a wider audience. They are nevertheless often invaluable. Table 2.2-2, which lists the abundances of every bird species in a survey, is an example. Data tables aren't usually included in the main body of a report. When published, they usually appear as appendices or online supplements, so specialized readers interested in more details can find them.

Follow similar principles for display tables

Producing clear, honest, and efficient display tables should follow many of the same principles discussed already for graphs. In particular,

- Make patterns in the data easy to see.
- Represent magnitudes honestly.
- Draw table elements clearly.

Make patterns easy to see. Make the table compact and present as few significant digits as are necessary to communicate the pattern. Avoid putting too much data into one table. Arrange the rows and columns of numbers to facilitate pattern detection. For example, a series of numbers listed above one another in a single column are easier to compare with one another than the same numbers listed side by side in different columns. Our earlier recommendations for frequency tables apply here (Section 2.2). For example, list unordered categorical (nominal) data in order of importance (frequency) rather than alphabetically or otherwise. If the categorical variables have a natural order (such as life stages: zygote, fetus, newborn, adolescent, adult), they should be listed in that order.

Represent magnitudes honestly. For example, when combining numbers into bins in frequency tables, use intervals of equal width so that the numbers can be more accurately compared.

Draw table elements clearly. Clearly label row and column headers, and always provide units of measurement. Choose unadorned, simple fonts.

Let's look at an example of a display table and then consider how it might be improved. The data in Table 2.5-1 were put together by Alvarez et al. (2009) to investigate the idea that a strong preference for consanguineous marriages (inbreeding) within the line of Spanish Habsburg kings, which ruled Spain from 1516 to 1700, contributed to its downfall. The quantity F is a measure of inbreeding in the offspring. F is zero if king and queen were unrelated, and F is 0.25 if they were brother and sister whose own parents were unrelated. Values may be lower or higher if there was inbreeding further generations back.

There is a tendency for less related kings and queens to produce a higher proportion of surviving offspring, but it is not so easy to see this in Table 2.5-1. Before reading any further, examine the table and make a note of any deficiencies. How might these deficiencies be overcome by modifying the table?

Table 2.5-1 Inbreeding coefficient (*F*) of Spanish Habsburg kings and queens and survival of their progeny.

King/Queen	*F*	Pregnancies	Miscarriages & stillbirths	Neonatal deaths	Later deaths	Survivors to age 10	Survival (total)	Survival (postnatal)
Ferdinand of Aragon								
Elizabeth of Castile	0.039	7	2	0	0	5	0.714	1.000
Philip I								
Joanna I	0.037	6	0	0	0	6	1.000	1.000
Charles I								
Isabella of Portugal	0.123	7	1	1	2	3	0.429	0.600
Philip II								
Elizabeth of Valois	0.008	4	1	1	0	2	0.500	1.000
Anna of Austria	0.218	6	1	0	4	1	0.167	0.200
Philip III								
Margaret of Austria	0.115	8	0	0	3	5	0.625	0.625
Philip IV								
Elizabeth of Bourbon	0.050	7	0	3	2	2	0.286	0.500
Mariana of Austria	0.254	6	0	1	3	2	0.333	0.400

Source: Data are from Alvarez et al. (2009).

Let's apply the principles of effective display to improve this table. Consider that the main goal of producing the table should be to show a pattern, in this case a possible association between *F* and offspring survival. Here is a list of features of Table 2.5-1 that we felt made it difficult to see this pattern.

- King/queen pairs are not ordered in such a way as to make it easy for the eye to see any association.
- The main variables of interest, *F* and survival, are separated by intervening columns.
- Blank lines are inserted for every new king listed, fragmenting any pattern.
- The number of decimal places is overly large, making it difficult to read the numbers.

To overcome these problems, we have extracted the most crucial columns and reorganized them in Table 2.5-2. In this revised table, king and queen pairs are ordered by *F* value of the offspring, and survival has been placed in the adjacent column. Blank lines have been eliminated along with unnecessary columns, and decimals have been rounded to two places.

The revised Table 2.5-2 suggests that survival of more inbred progeny tends to be lower, at least when measured as postnatal survival. The trend appears weaker for total survival, which includes prenatal and neonatal survival.

Just as for a graph, a good table must convey information clearly, concisely, and without distortion. A good table requires careful editing. See Ehrenberg (1977) for further insights into how to draw tables.

Table 2.5-2 Inbreeding coefficient (*F*) of Spanish kings and queens and survival of their progeny. These data are extracted and reorganized from Table 2.5-1.

King/Queen	*F*	Survival (postnatal)	Survival (total)	Number of pregnancies
Philip II/Elizabeth of Valois	0.01	1.00	0.50	4
Philip I/Joanna I	0.04	1.00	1.00	6
Ferdinand/Elizabeth of Castile	0.04	1.00	0.71	7
Philip IV/Elizabeth of Bourbon	0.05	0.50	0.29	7
Philip III/Margaret of Austria	0.12	0.63	0.63	8
Charles I/Isabella of Portugal	0.12	0.60	0.43	7
Philip II/Anna of Austria	0.22	0.20	0.17	6
Philip IV/Mariana of Austria	0.25	0.40	0.33	6

2.6 Summary

- Graphical displays must be clear, honest, and efficient.
- Strive to show the data, to make patterns in the data easy to see, to represent magnitudes honestly, and to draw graphical elements clearly.
- Follow the same rules when constructing tables to reveal patterns in the data.
- A frequency table is used to display a frequency distribution for categorical or numerical data.
- Bar graphs and histograms are recommended graphical methods for displaying frequency distributions of categorical and numerical variables:

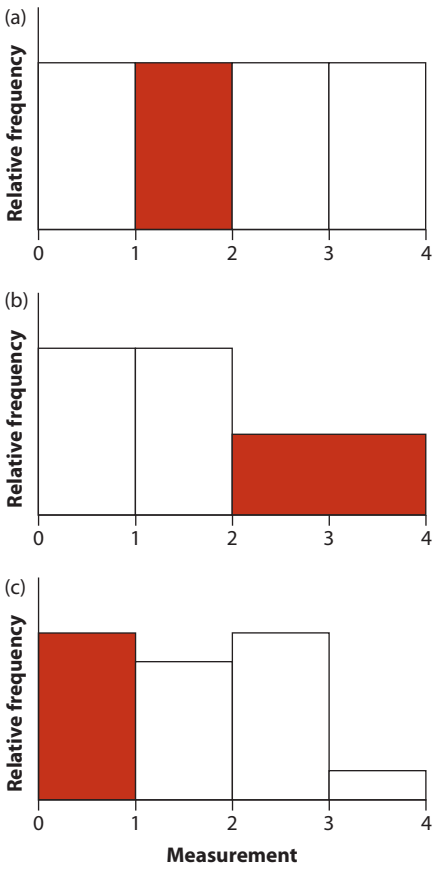
Type of data	Graphical method
Categorical data	Bar graph
Numerical data	Histogram

- Contingency tables describe the association between two (or more) categorical variables by displaying frequencies of all combinations of categories.
- Recommended graphical methods for displaying associations between variables and differences between groups include the following:

Types of data	Graphical method
Two numerical variables	Scatter plot Line plot (space or time) Map (space)
Two categorical variables	Grouped bar graph Mosaic plot
One numerical variable and one categorical variable	Strip chart Box plot Multiple histograms Cumulative frequency distributions (Chapter 3)

PRACTICE PROBLEMS

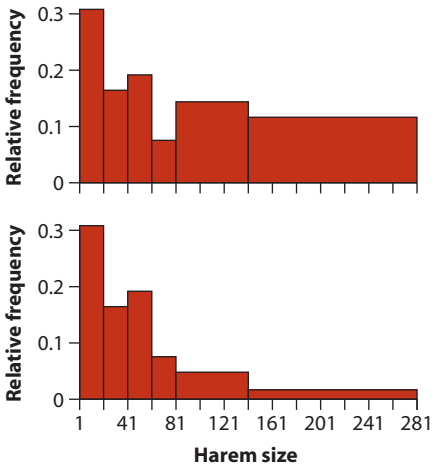
1. Estimate by eye the relative frequency of the shaded areas in each of the following histograms.



2. Using a graphical method from this chapter, draw three frequency distributions: one that is symmetric, one that is skewed, and one that is bimodal.

- **a.** Identify the mode in each of your frequency distributions.
- **b.** Does your skewed distribution have negative or positive skew?
- **c.** Is your bimodal distribution skewed or symmetric?

3. In the southern elephant seal, males defend harems that may contain hundreds of reproductively active females. Modig (1996) recorded the numbers of females in harems in a population on South Georgia Island. The histograms of the data (below, drawn from data in Modig 1996) are unusual because the rarer, larger harems have been divided into wider intervals. In the upper histogram, bar *height* indicates the relative frequency of harems in the interval. In the lower histogram, bar height is adjusted such that bar *area* indicates relative frequency. Which histogram is correct? Why?



4. Draw scatter plots for invented data that illustrate the following patterns:

- **a.** Two numerical variables that are positively associated
- **b.** Two numerical variables that are negatively associated
- **c.** Two numerical variables whose relationship is nonlinear

5. A study by Miller et al. (2004) compared the survival of two kinds of Lake Superior rainbow trout fry (babies). Four thousand fry were from a government hatchery on the lake, whereas 4000 more fry came from wild trout. All 8000 fry were released into a stream flowing into the

lake, where they remained for one year. After one year, the researchers found 78 survivors. Of these, 27 were hatchery fish and 51 were wild. Display these results in the most appropriate table. Identify the type of table you used.

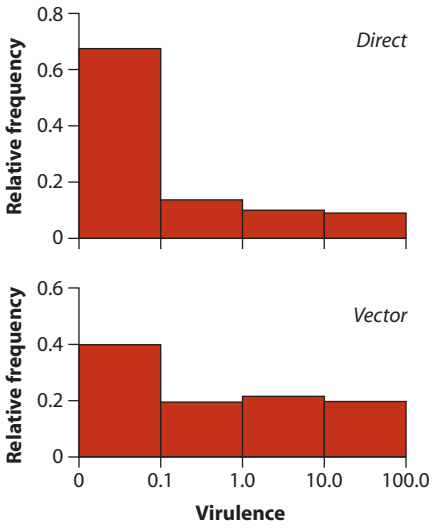
6. The following data are the occurrences in 2012 of the different taxa in the list of endangered and threatened species under the U.S. Endangered Species Act (U.S. Fish and Wildlife Service 2012). The taxa are listed in no particular order in the table.

Taxon	Number of species
Birds	93
Clams	83
Reptiles	36
Fish	152
Crustaceans	22
Mammals	85
Snails	40
Flowering plants	782
Amphibians	26
Insects	66
Arachnids	12

a. Rewrite the table, but list the taxa in a more revealing order. Explain your reasons behind the ordering you choose.
b. What kind of table did you construct in part (a)?
c. Choosing the most appropriate graphical method, display the number of species in each taxon. What kind of graph did you choose? Why?
d. Should the baseline for the number of species in your graph in part (c) be 0 or 12, the smallest number in the data set? Why?
e. Create a version of this table that shows the relative frequency of endangered species by taxon.

7. Can environmental factors influence the incidence of schizophrenia? A recent project measured the incidence of the disease among children born in a region of eastern China: 192 of 13,748 babies born in the midst of a severe famine in the region in 1960 later developed schizophrenia. This compared with 483 schizophrenics out of 59,088 births in 1956, before the famine, and 695 out of 83,536 births in 1965, after the famine (St. Clair et al. 2005).
a. What two variables are compared in this example?
b. Are the variables numerical or categorical? If numerical, are they continuous or discrete; if categorical, are they nominal or ordinal?
c. Effectively display the findings in a table. What kind of table did you use?
d. In each of the three years, calculate the relative frequency (proportion) of children born who later developed schizophrenia. Plot these proportions in a line graph. What pattern is revealed?

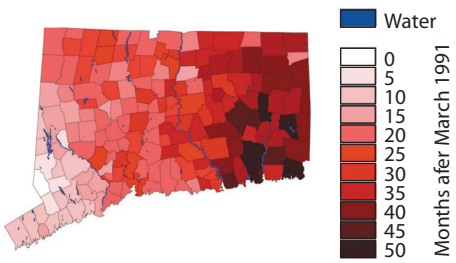
8. Human diseases differ in their virulence, which is defined as their ability to cause harm. Scientists are interested in determining what features of different diseases make some more dangerous to their hosts than others. The graph below depicts the frequency distribution of virulence measurements, on a log base 10 scale, of a sample of human diseases (modified from Ewald 1993). Diseases that spread from one victim to



another by direct contact between people are shown in the upper graph. Those transmitted from person to person by insect vectors are shown in the lower graph.

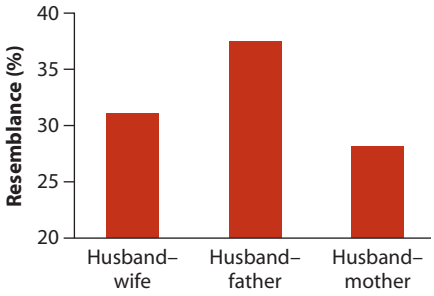
a. Identify the type of graph displayed.
b. What are the two groups being compared in this graph?
c. What variable is being compared between the two groups? Is it numerical or categorical?
d. Explain the units on the vertical (y) axis.
e. What is the main result depicted by this graph?

9. Examine the figure below, which indicates the date of first occurrence of rabies in raccoons in the townships of Connecticut, measured by the number of months following March 1, 1991 (modified from Smith et al. 2002).



a. Identify the type of graph shown.
b. What is the response variable?
c. What is the explanatory variable?
d. What was the direction of spread of the disease (from where, to where, approximately)?

10. The following graph is taken from a study of married women who had been raised by adoptive parents (Bereczkei et al. 2004). It shows the facial resemblance between the women and their husbands (first bar), between their husbands and the women's adoptive fathers (second bar), and between their husbands and the women's adoptive mothers (third bar). Facial resemblance of a given woman to her husband was scored by 242 "judges," each of whom was given a photograph of the woman, photos of three other women, and a photo of her husband. The judges were asked to decide from the photos which of the four women was the wife of the husband based on

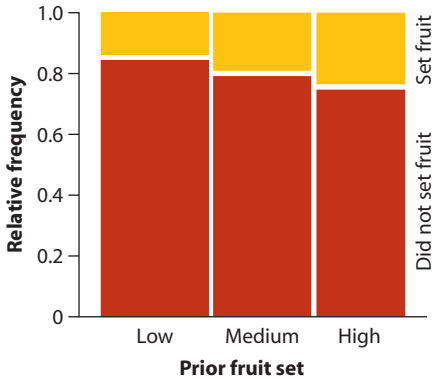


facial similarity. Her resemblance was scored as the percentage of judges who chose correctly. If there was no resemblance between a given woman and her husband, then by chance only one in four judges (25%) should have chosen correctly. Resemblance of the woman's husband to the wife's adoptive father and to the wife's adoptive mother was measured in the same way.

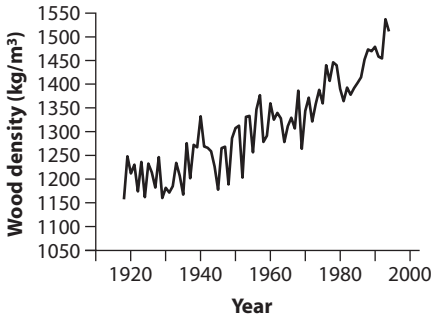
a. Describe the essential findings displayed in the figure.
b. Which two principles of good graph design are violated in this figure?

11. Each of the following graphs illustrates an association between two variables. For each graph identify (1) the type of graph, (2) the explanatory and response variables, and (3) the type of data (whether numerical or categorical) for each variable.

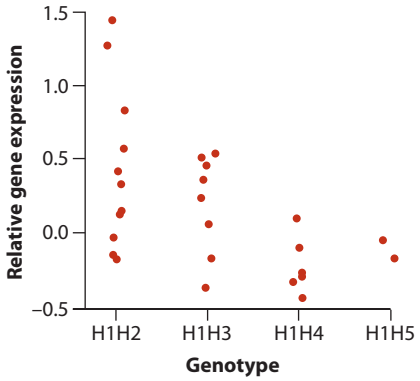
a. Observed fruiting of individual plants in a population of *Campanula americana* according to the number of fruits produced previously (Richardson and Stephenson 1991):



b. The maximum density of wood produced at the end of the growing season in white spruce trees in Alaska in different years (modified from Barber et al. 2000):



c. Relative expression levels of *Neuropeptide Y* (*NPY*), a gene whose activity correlates with anxiety and is induced by stress, in the brains of people differing in their genotypes at the locus (Zhou et al. 2008).



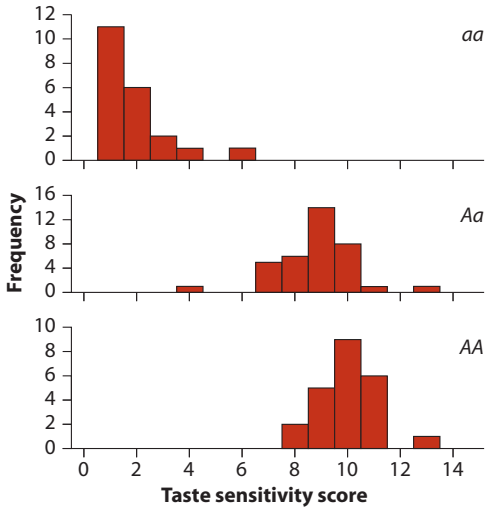
12. The following data are from the Cambridge Study in Delinquent Development (see Problem 22). They examine the relationship between the occurrence of convictions by the end of the study and the family income of each boy when growing up. Three categories described income level: inadequate, adequate, and comfortable.

	Income level		
	Inadequate	Adequate	Comfortable
No convictions	47	128	90
Convicted	43	57	30

a. What type of table is this?
b. Display these same data in a mosaic plot.
c. What type of variable is "income level"? How should this affect the arrangement of groups in your mosaic plot in part (b)?
d. By viewing the table above and the graph in part (b), describe any apparent association between family income and later convictions.
e. In answering part (d), which method (the table or the graph) better revealed the association between conviction status and income level? Explain.

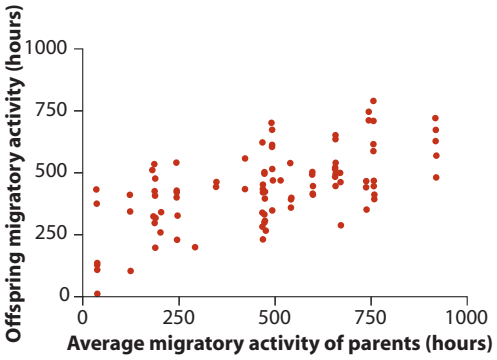
13. Each of the following graphs illustrates an association between two variables. For each graph, identify (1) the type of graph, (2) the explanatory and response variables, and (3) the type of data (whether numerical or categorical) for each variable.

a. Taste sensitivity to phenylthiocarbamide (PTC) in a sample of human subjects grouped according to their genotype at the PTC gene—namely, *AA*, *Aa*, or *aa* (Kim et al. 2003):

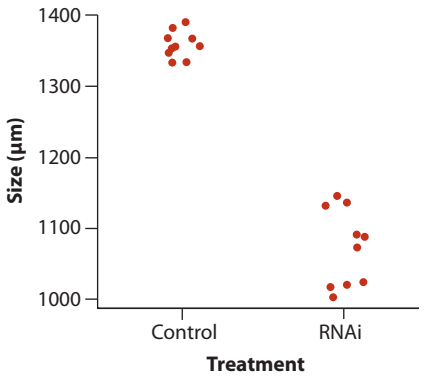


b. Migratory activity (hours of nighttime restlessness) of young captive blackcaps (*Sylvia atricapilla*) compared with the migratory

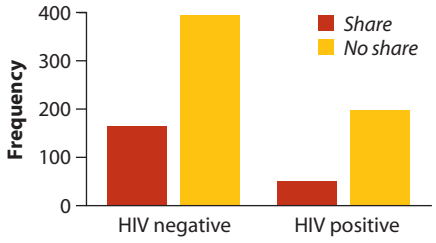
activity of their parents (Berthold and Pulido 1994):



c. Sizes of the second appendage (middle leg) of embryos of water striders, in 10 control embryos and 10 embryos dosed with RNAi for the developmental gene *Ultrabithorax* (Khila et al. 2009).



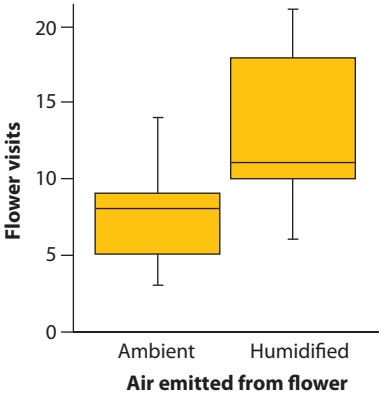
d. The frequency of injection-heroin users that share or do not share needles according to their known HIV infection status (Wood et al. 2001):



14. *Spot the flaw.* In an experimental study of gender and wages, Moss-Racusin et al. (2012) presented professors from research-intensive universities each with a job application for a laboratory manager position. The application was randomly assigned a male or female name, and the professors were asked to state the starting salary they would offer the candidate if hired. The average starting salary reported is compared in the following figure between applications with male names and female names. (The vertical lines at the top edge of each bar are "standard error bars"—we'll learn about them in Chapter 4).

a. Identify at least two of the four principles of good graph design that are violated.
b. What alternative graph type is ideal for these data?
c. Identify the main pattern in the data (interestingly, this pattern was similar when male professors and female professors were examined separately).

15. How do the insects that pollinate flowers distinguish individual flowers with nectar from empty flowers? One possibility is that they can detect the slightly higher humidity of the air—produced by evaporation—in flowers that contain nectar. von Arx et al. (2012) recently tested this idea by manipulating the humidity of air emitted from artificial flowers that were otherwise identical. The following graph summarizes the number of visits to the two types of flowers by hawk moths (*Hyles lineata*).

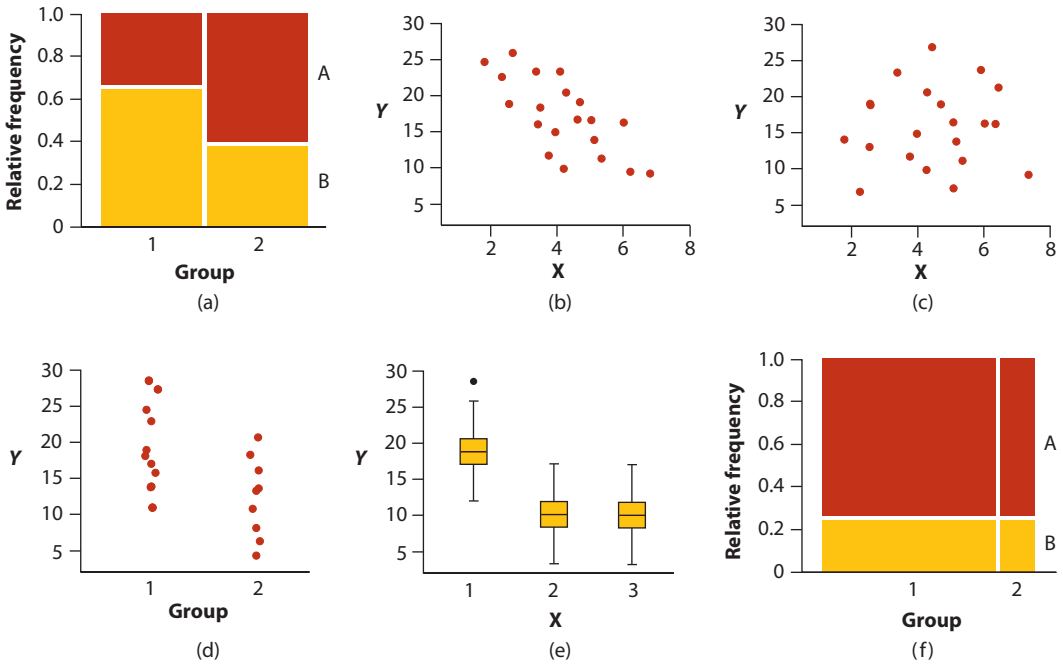


a. What type of graph is this?
b. What does the horizontal line in the center of each rectangle represent?
c. What do the top and bottom edges of each rectangle represent?

d. What are the vertical lines extending above and below each rectangle?
e. Is an association apparent between the variables plotted? Explain.

16. For each of the graphs shown below, based on hypothetical data, identify the type of graph and say whether or not the two variables exhibit an association. Explain your answer in each case.

FIGURE FOR PROBLEM 16



17. "Animal personality" has been defined as the presence of consistent differences between individuals in behaviors that persist over time. Do sea anemones have it? To investigate, Briffa and Greenaway (2011) measured the consistency of the startle response of individuals of wild beadlet anemones, *Actinia equina,* in tide pools in the U.K. When disturbed, such as with a mild jet of water (the method used in this study), the anemones retract their feeding tentacles to cover the oral disc, opening them again some time later. The accompanying table records the duration of the startle response (time to reopen, in seconds) of 12 individual anemones. Each anemone was measured twice, 14 days apart.

a. Choose the best method, and make a graph to show the association between the first and second measurements of startle response.

b. Is a strong association present? In other words, does the beadlet anemone have animal personality?

18. Refer to the previous question.

a. Draw a frequency distribution of startle durations measured on the first occasion.

b. Describe the shape of the frequency distribution: is it skewed or symmetric? If skewed, say whether skew is positive or negative.

TABLE FOR PROBLEM 17

Anemone:	1	2	3	4	5	6	7	8	9	10	11	12
Occasion one	1065	248	436	350	378	410	232	201	267	687	688	980
Occasion two	939	268	460	261	368	467	303	188	401	690	711	571

ASSIGNMENT PROBLEMS

19. Male fireflies of the species *Photinus ignitus* attract females with pulses of light. Flashes of longer duration seem to attract the most females. During mating, the male transfers a spermatophore to the female. Besides containing sperm, the spermatophore is rich in protein that is distributed by the female to her fertilized eggs. The data below are measurements of spermatophore mass (in mg) of 35 males (Cratsley and Lewis 2003).

0.047, 0.037, 0.041, 0.045, 0.039, 0.064, 0.064, 0.065, 0.079, 0.070, 0.066, 0.059, 0.075, 0.079, 0.090, 0.069, 0.066, 0.078, 0.066, 0.066, 0.055, 0.046, 0.056, 0.067, 0.075, 0.048, 0.077, 0.081, 0.066, 0.172, 0.080, 0.078, 0.048, 0.096, 0.097

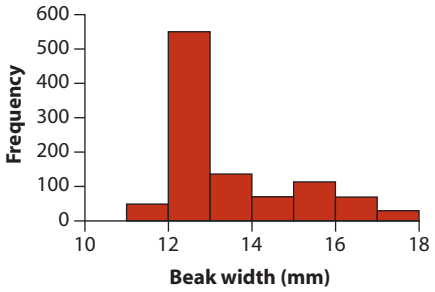
a. Create a graph depicting the frequency distribution of the 35 mass measurements.

b. What type of graph did you choose in part (a)? Why?

c. Describe the shape of the frequency distribution. What are its main features?

d. What term would be used to describe the largest measurement in the frequency distribution?

20. The accompanying graph depicts a frequency distribution of beak widths of 1017 black-bellied seedcrackers, *Pyrenestes ostrinus,* a finch from West Africa (Smith 1993).

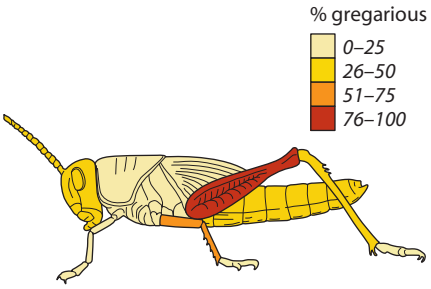


a. What is the mode of the frequency distribution?

b. Estimate by eye the fraction of birds whose measurements are in the interval representing the mode.

c. There is a hint of a second peak in the frequency distribution between 15 and 16 mm. What strategy would you recommend be used to explore more fully the possibility of a second peak?
d. What name is given to a frequency distribution having two distinct peaks?

21. When its numbers increase following favorable environmental conditions, the desert locust, *Schistocerca gregaria,* undergoes a dramatic transformation from a solitary, cryptic form into a gregarious form that swarms by the billions. The transition is triggered by mechanical stimulation—locusts bumping into one another. The accompanying figure shows the results of a laboratory study investigating the degree of gregariousness resulting from mechanical stimulation of different parts of the body (modified from Simpson et al. 2001).



a. Identify the type of graph displayed.
b. Identify the explanatory and response variables.

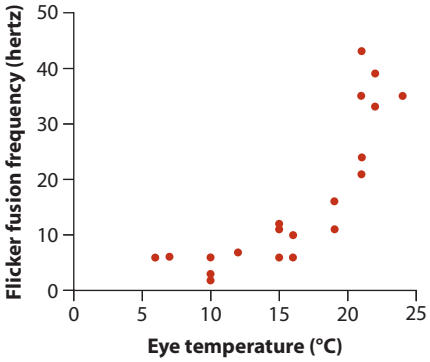
22. The *Cambridge Study in Delinquent Development* was undertaken in north London (U.K.) to investigate the links between criminal behavior in young men and the socioeconomic factors of their upbringing (Farrington 1994). A cohort of 395 boys was followed for about 20 years, starting at the age of 8 or 9. All of the boys attended six schools located near the research office. The following table shows the total number of criminal convictions by the boys between the start and end of the study.

Number of convictions	Frequency
0	265
1	49
2	21
3	19
4	10
5	10
6	2
7	2
8	4
9	2
10	1
11	4
12	3
13	1
14	2
	Total: 395

a. What type of table is this?
b. How many variables are presented in this table?
c. How many boys had exactly two convictions by the end of the study?
d. What fraction of boys had no convictions?
e. Display the frequency distribution in a graph. Which type of graph is most appropriate? Why?
f. Describe the shape of the frequency distribution. Is it skewed or is it symmetric? Is it unimodal or bimodal? Where is the mode in number of criminal convictions? Are there outliers in the number of convictions?
g. Does the sample of boys used in this study represent a random sample of British boys? Why or why not?

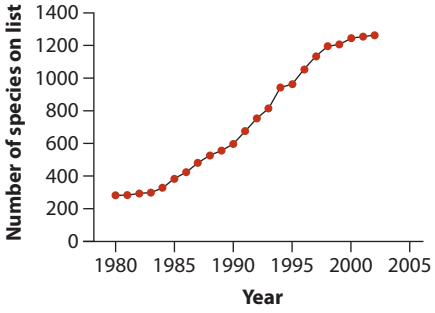
23. Swordfish have a unique "heater organ" that maintains elevated eye and brain temperatures when hunting in deep, cold water. The following graph illustrates the results of a study by Fritsches et al. (2005) that measured how

the ability of swordfish retinas to detect rapid motion, measured by the flicker fusion frequency, changes with eye temperature.



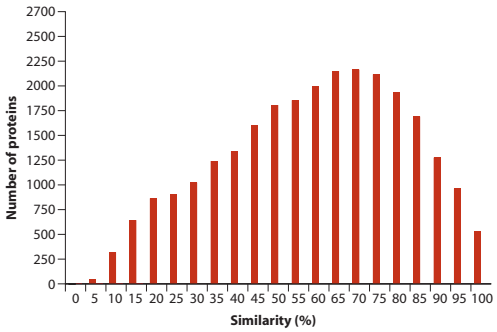
a. What types of variables are displayed?
b. What type of graph is this?
c. Describe the association between the two variables. Is the relationship between flicker fusion frequency and temperature positive or negative? Is the relationship linear or nonlinear?
d. The 20 points in the graph were obtained from measurements of six swordfish. Can we treat the 20 measurements as a random sample? Why or why not?

24. The following graph displays the net number of species listed under the U.S. Endangered Species Act between 1980 and 2002 (U.S. Fish and Wildlife Service 2001):



a. What type of graph is this?
b. What does the steepness of each line segment indicate?
c. Explain what the graph tells us about the relationship between the number of species listed and time.

25. *Spot the flaw.* Examine the following figure, which displays the frequency distribution of similarity values (the percentage of amino acids that are the same) between equivalent (homologous) proteins in humans and pufferfish of the genus *Fugu* (modified from Aparicio et al. 2002).



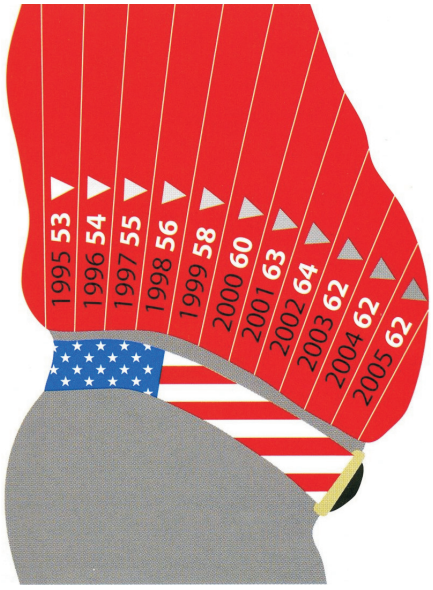
a. What type of graph is this?
b. Identify the main flaw in the construction of this figure.
c. What are the main results displayed in the figure?
d. Describe the shape of the frequency distribution shown.
e. What is the mode in the frequency distribution?

26. The following data give the photosynthetic capacity of nine individual females of the neotropical tree *Ocotea tenera*, according to the number of fruits produced in the previous reproductive season (Wheelwright and Logan 2004). The goal of the study was to investigate how reproductive effort in females of these trees impacts subsequent growth and photosynthesis.

Number of fruits produced previously	Photosynthetic capacity (μmol $O_2/m^2/s$)
10	13.0
14	11.9
5	11.5
24	10.6
50	11.1
37	9.4
89	9.3
162	9.1
149	7.3

a. Graph the association between these two variables using the most appropriate method. Identify the type of graph you used.
b. Which variable is the explanatory variable in your graph? Why?
c. Describe the association between the two variables in words, as revealed by your graph.

27. Examine the accompanying figure, which displays the percentage of adults over 18 with a "body mass index" greater than 25 in different years (modified from *The Economist* 2005, with permission). Body mass index is a measure of weight relative to height.

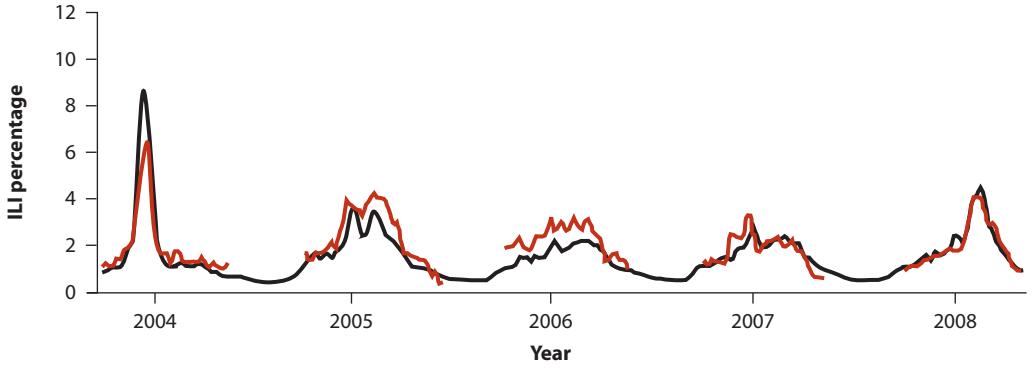


a. What is the main result displayed in this figure?
b. Which of the four principles for drawing good graphs are violated here? How are they violated?
c. Redraw the figure using the most appropriate method discussed in this chapter. What type of graph did you use?

28. When a courting male of the small Indonesian fish *Telmatherina sarasinorum* spawns with a female, other males sometimes sneak in and release sperm, too. The result is that not all of the female's eggs are fertilized by the courting male. Gray et al. (2007) noticed that courting males occasionally cannibalize fertilized eggs immediately after spawning. Egg eating took place by 61 of 450 courting males who fathered the entire batch; the remaining 389 males did not cannibalize eggs. In contrast, 18 of 35 courting males ate eggs when a single sneaking male also participated in the spawning event. Finally, 16 of 20 males ate eggs when two or more sneaking males were present.
a. Display these results in a table that best shows the association between cannibalism and the number of sneaking males. Identify the type of table you used.
b. Illustrate the same results using a graphical technique instead. Identify the type of graph you used.

29. The graph at the top of page 62, in red, shows the number of new cases of influenza in New York, Pennsylvania, and New Jersey, according to data from the Centers for Disease Control (Ginsberg et al. 2009). The black line shows predictions based on the number of Google searches of words like "flu" or "influenza."
a. What type of graph is this?
b. Describe some of the scientific conclusions you might draw from looking at this graph.

FIGURE FOR PROBLEM 29



Source: Jeremy Ginsberg et al., "Detecting Influenza Epidemics Using Search Engine Query Data," *Nature* 457 (2009): 1012–1014.

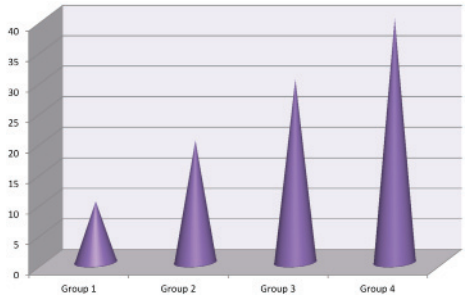
c. Can you suggest an improvement to the axes labels?

30. The following graph was drawn using a very popular spreadsheet program in an attempt to show the frequencies of observations in four hypothetical groups. Before reading further, estimate by eye the frequencies in each of the four groups.

a. Identify two features of this graph that cause it to violate the principle, "Make patterns in the data easy to see."

b. Identify at least two other features of the graph that make it difficult to interpret.

c. The actual frequencies are 10, 20, 30, and 40. Draw a graph that overcomes the problems identified above.



31. In Poland, students are required to achieve a score of 21 or higher on the high-school Polish language "maturity exam" to be eligible for university. The following graph shows the frequency distribution of scores (Freakonomics 2011).

a. Examine the graph and identify the most conspicuous pattern in these data.

b. Generate a hypothesis to explain the pattern.

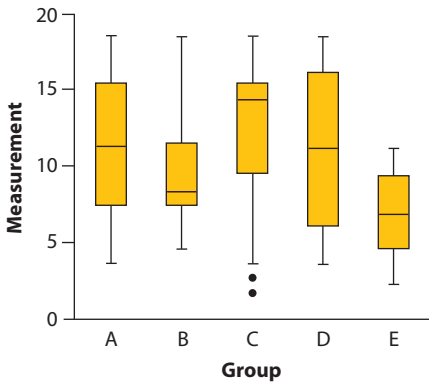
32. More than 10% of people carry the parasite *Toxoplasma gondii*. The following table gives data from Prague on 15- to 29-year-old drivers who had been involved in an accident. The table gives the number of drivers who were infected with *Toxoplasma gondii* and who were uninfected. These numbers are compared with a control sample of 249 drivers of the same age living in the same area who had not been in an accident.

	Infected	Uninfected
Drivers with accidents	21	38
Controls	38	211

a. What type of table is this?
b. What are the two variables being compared? Which is the explanatory variable and which is the response?
c. Depict the data in a graph. Use the results to answer the question: are the two variables associated in this data set?

33. The cutoff birth date for school entry in British Columbia, Canada, is December 31. As a result, children born in December tend to be the youngest in their grade, whereas those born in January tend to be the oldest. Morrow et al. (2012) examined how this relative age difference influenced diagnosis and treatment of attention deficit/hyperactivity disorder (ADHD). A total of 39,136 boys aged 6 to 12 years and registered in school in 1997–1998 had January birth dates. Of these, 2219 were diagnosed with ADHD in that year. A total of 38,977 boys had December birth dates, of which 2870 were diagnosed with ADHD in that year. Display the association between birth month and ADHD diagnosis using a table or graphical method from this chapter. Is there an association?

34. Examine the following figure, which displays hypothetical measurements of a sample of individuals from several groups.



a. What type of graph is this?
b. In which of the groups is the frequency distribution of measurements approximately symmetric?
c. Which of the frequency distributions show positive skew?
d. Which of the frequency distributions show negative skew?
e. Which group has the largest value for the upper quartile?
f. Which group has the smallest value for the median?
g. Which group has the most extreme observation?

35. The following data are from Mattison et al. (2012), who carried out an experiment with rhesus monkeys to test whether a reduction in food intake extends life span (as measured in years). The data are the life spans of 19 male and 15 female monkeys who were randomly assigned a normal nutritious diet or a similar diet reduced in amount by 30%. All monkeys were adults at the start of the study.

Females—reduced: 16.5, 18.9, 22.6, 27.8, 30.2, 30.7, 35.9

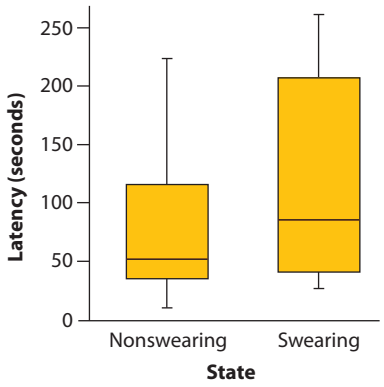
Females—control: 23.7, 24.5, 24.7, 26.1, 28.1, 33.4, 33.7, 35.2

Males—reduced: 23.7, 28.1, 29.8, 31.1, 36.3, 37.7, 39.9, 39.9, 40.2, 40.2

Males—control: 24.9, 25.2, 29.6, 33.2, 34.1, 35.4, 38.1, 38.8, 40.7

a. Graph the results, using the most appropriate method and following the four principles of good graph design.
b. According to your graph, which difference in life span is greater: that between the sexes, or that between diet groups?

36. The accompanying graph indicates the amount of time (latency) that female subjects were willing to leave their hand in icy water while they were swearing ("words you might use after hitting yourself on the thumb with a hammer") or while not swearing, using other words instead ("words to describe a table"). The data are from Stephens et al. (2009).



a. Identify the type of graph.
b. Is any association apparent between the variables? Explain.
c. What do the "whiskers" indicate in this graph?
d. List two other types of graphs that would also be appropriate for showing these results.

37. Following is a list of all the named hurricanes in the Atlantic between 2001 and 2010, along with their category on the Saffir-Simpson Hurricane Scale,[9] which categorizes each hurricane by a label from 1 to 5 depending on its power.

2001: Erin, 3; Feliz, 3; Gabrielle, 1; Humberto, 2; Iris, 4; Karen, 1; Michelle, 4; Noel, 1; Olga, 1.

2002: Gustav, 2; Isidore, 3; Kyle, 1; Lili, 4;

2003: Claudette, 1; Danny, 1; Erika, 1; Fabian, 4; Isabel, 5; Juan, 2; Kate, 3.

2004: Alex, 3; Charley, 4; Danielle, 2; Frances, 4; Gaston, 1; Ivan, 5; Jeanne, 3; Karl, 4; Lisa, 1.

2005: Cindy, 1; Dennis, 4; Emily, 5; Irene, 2; Katrina, 5; Maria, 3; Nate, 1; Ophelia, 1; Philippe 1; Rita, 5; Stan, 1; Vince, 1; Wilma, 5; Beta, 3; Epsilon, 1.

2006: Ernesto, 1; Florence, 1; Gordon, 3; Helene, 3; Isaac, 1.

2007: Dean, 5; Felix, 5; Humberto, 1; Karen, 1; Lorenzo, 1; Noel, 1.

2008: Bertha, 3; Dolly, 2; Gustav, 4; Hanna, 1; Ike, 4; Kyle, 1; Omar, 4; Paloma, 4.

2009: Bill, 4; Fred, 3; Ida, 2.

2010: Alex, 2; Danielle, 4; Earl, 4; Igor, 4, Julia, 4; Karl, 3; Lisa, 1; Otto, 1; Paula, 2; Richard, 2; Shary, 1; Toas, 2.

a. Make a frequency table showing the frequency of hurricanes in each severity categories during the decade.
b. Make a frequency table that shows the frequency of hurricanes in each year.
c. Explain how you chose to order the categories in your tables.

9. Data taken from http://en.wikipedia.org/wiki/2010_Atlantic_hurricane_season and similar articles for other years.